

Große KI-Modelle

FÜR DEUTSCHLAND

Version mit vollständigem Anhang

Gefördert durch:



Bundesministerium
für Wirtschaft
und Klimaschutz

aufgrund eines Beschlusses
des Deutschen Bundestages

LEAM:AI



KI BUNDESVERBAND

MACHBARKEITSSTUDIE



IMPRESSUM

Die Machbarkeitsstudie zu LEAM - Large European AI Models wurde erstellt von:

**Akademie für Künstliche Intelligenz AKI gGmbH
im KI Bundesverband**

Schiffbauerdamm 40, 10117 Berlin
+ 49 (0) 30 30 340 388
info@ki-verband.de • www.ki-verband.de

Auftraggeber

**Bundesministerium
für Wirtschaft und Klimaschutz**

11019 Berlin
+ 49 (0) 30 18 615 - 0
info@bmwk.bund.de • www.bmwk.de

Layout & Satz

Dr. Sandra Bütow
KI Bundesverband

Grafik & Gestaltung

Victoria Patricia
contact@victoriapatricia.com • www.victoriapatricia.com

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Wirtschaft und Klimaschutz unter dem Förderkennzeichen 01MT22006 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autor:innen.

© 2023



Grußwort Jörg Bienert



Jörg Bienert, Präsident KI Bundesverband

Im Jahr 2002 erschien in der New York Times ein Artikel mit der Überschrift „Google's toughest search is for a Business Model“ (Hansell, 2002). Der Autor des Artikels war überzeugt, dass Google sich gegen die damaligen Konkurrenten im Online-Advertising Geschäft nicht behaupten würde und dass das damalige Kerngeschäft, die Lizenzierung der Suchmaschine an andere Web-Portale, kein hinreichendes Wachstum bringen wird.

Seitdem hat sich viel getan. Innerhalb weniger Jahre beherrschte die Google-Suchmaschine das Internet, den Online-Werbemarkt und spielte eine dominierende Rolle in vielen weiteren Bereichen. Ohne massive Investments in Google-Ads lässt sich heute kein Online-Geschäftsmodell mehr starten, in vielen Autos ist die Navigation von Google-Maps integriert, der meistgenutzte Browser Chrome sammelt vielfältige Daten über unser Surf-Verhalten und mit Online-Diensten wie Gmail, Docs oder Drive vertrauen weltweit Millionen Menschen ihre Daten dem Alphabet Konzern an. Europa hat erfolglos versucht, mit einer eigenen Suchmaschine eine Alternative aufzubauen.

Das alles sehen wir nur bedingt als problematisch an, weil Google es geschafft hat, einen enormen Vertrauensvorsprung aufzubauen. Was aber, wenn sich dies ändert? Denken wir doch nur einmal an das hypothetische Szenario, Elon Musk würde Alphabet übernehmen. Wären wir in der Lage, das Internet ohne die Google Services sinnvoll zu nutzen? Oder hat Alphabet es durch die Ansammlung von Daten und darauf aufbauenden Diensten bereits geschafft, eine Lock-in-Situation zu erzeugen, aus der wir uns kaum noch befreien können?

Künstliche Intelligenz hat das Potential, die Welt ähnlich stark zu verändern, wie es das Internet getan hat. Es wird viele neue Funktionalitäten, Produkte und Geschäftsmodelle geben. Prozessverbesserungen werden zu Effizienzsteigerungen und zu disruptiven Veränderungen führen, die wir heute noch gar nicht absehen können. Im Jahr 2000 hätten wir bei der Vorhersage, dass das Internet zur Insolvenz der größten deutschen Versand- und Warenhäuser führen würde, wahrscheinlich nur den Kopf geschüttelt.

Wir erleben gerade den Beginn der zweiten Welle der KI-Revolution, die mit der Veröffentlichung von GPT-3 durch OpenAI im Sommer 2020 begonnen hat. Auf Basis riesiger Datenmengen und mit enormem Einsatz von Entwicklerressourcen, Geld und Rechenkapazität hat OpenAI ein Sprachmodell geschaffen, das eine bis dahin unerreichte Performance aufweisen konnte. In dieser Studie beleuchten wir die Entwicklung und den Status Quo genauer.

Was also wird sich in den nächsten 2-3 Jahren verändern? Wir werden eine Vielzahl von neuen Modellen und Anwendungen sehen. Der Umfang und die Auswirkungen sind derzeit kaum abzuschätzen, aber es werden eine Menge Fragen auftauchen wie z.B.

- Wann und wie werden große Sprachmodelle eine nahezu hundertprozentige inhaltliche Korrektheit in den generierten Texten erreichen?
- Welche Auswirkungen hat dies für Unternehmen aus allen Branchen und mögliche Anwendungen in den unterschiedlichen Bereichen der Wertschöpfungskette?
- Wie gut werden neben Sprachmodellen Bild- und Video-Generatoren sowie die Kombination aus diesen? Was bedeutet dies für die Kreativbranche, den Journalismus und die Ausbildung?
- Wie wird sich das Internet mit einer Flut von automatisch generierten Inhalten verändern? Werden die großen Sprachmodelle die Google-Suche ablösen und damit den Online-Werbemarkt auch für andere Akteure öffnen?
- Können wir durch die Demokratisierung von KI den Arbeitsmarkt so umgestalten, dass die demographisch immer kleineren Jahrgänge die Arbeit gesund bewältigen können und gleichzeitig den Fachkräftemangel insbesondere bei "menschlichen" Arbeitsplätzen ausgleichen?
- Inwiefern können die drängendsten Probleme der Menschheit, etwa in Bezug auf die Energiewende, Klimawandel, Gesundheit, mit Hilfe großer Modelle gelöst werden?

Eine übergeordnete Fragestellung ist dabei von entscheidender Bedeutung. Wer besitzt die Technologie, die Daten und die Ressourcen, um große Modelle zu erstellen und die Entwicklung und revolutionären Durchbrüche zu steuern? Bleibt dies in der Hand weniger großer Konzerne, und werden diese die Nutzung, die Regulierung und auch die Behandlung von ethischen Fragestellungen bestimmen, so wie dies bereits im Internet, bei Suchmaschinen und Sozialen Netzwerken der Fall ist?

Genau hier liegt aktuell eine große Herausforderung für Deutschland und Europa. Nur wer die Technologie beherrscht, wird in der Lage sein, deren Nutzung nach eigenem Ermessen zu gestalten und wirtschaftliche und gesellschaftliche Entwicklungen positiv zu beeinflussen. Hier geht es bei weitem nicht nur um die Künstliche Intelligenz als Plattform-Technologie. Große KI-Modelle werden komplett neue Anwendungen ermöglichen und bestehende Geschäftsmodelle und Wertschöpfung disruptiv verändern – in allen Industrien.

Wenn wir nicht in der Lage sind, diese Basis-Technologie eigenständig zu entwickeln und bereitzustellen, wird die deutsche Industrie auf ausländische Dienste ausweichen müssen, Teile der Wertschöpfungskette verlieren und an Wettbewerbsfähigkeit einbüßen. Wir werden schwierige Diskussionen um Datenschutz, Datensicherheit und die ethische Nutzung von KI-Modellen haben. Die aktuellen Kontroversen um Google und Facebook erscheinen dagegen trivial.

Um dies zu verhindern und nicht auch in der KI die digitale Souveränität zu verlieren, müssen wir uns in Deutschland in die Lage versetzen, auf internationalem Niveau zu forschen, Daten zu sammeln und zu veredeln, große Modelle zu trainieren und diese offen für die Anwendung durch die Wirtschaft, Konzerne, Mittelstand und Start-ups bereitzustellen.

Dies ist das Ziel unserer Initiative Large European AI Models, kurz LEAM. Ein Team von 40 Vertretern aus Wissenschaft, Wirtschaft und Gesellschaft hat im vergangenen Jahr zunächst erste Ideen für eine Infrastruktur zur Schaffung von großen Modellen entwickelt. Im Auftrag des Bundesministeriums für Wirtschaft und Klimaschutz wurde nun diese Machbarkeitsstudie erstellt, die wir Ihnen hier vorstellen.

Kernpunkt des Konzeptes ist der Aufbau einer dedizierten KI-Supercomputing-Infrastruktur. Ein Team von Spezialist:innen betreibt eine dedizierte Hardware-Infrastruktur, die auf große KI-Modelle spezialisiert ist. Es entwickelt diese KI-Modelle weiter und stellt diese anderen zur Verfügung. Darüber hinaus sammelt und veredelt das Team die zum Betrieb und den Anwendungen notwendigen Daten und implementiert Software und Services rund um diese KI-Modelle, die das Training und Tuning von großen Modellen vereinfachen und diese für unterschiedliche Zielgruppen einfach nutzbar machen.

Mit LEAM planen wir ein zentrales KI-Leuchtturmprojekt, um das sich ein leistungsfähiges Ökosystem aus Wissenschaft, Wirtschaft und Start-ups bilden wird – in enger Zusammenarbeit auch mit bestehenden Aktivitäten wie Open GPT-X, Aleph Alpha oder Bloom und als wichtiger Player im europäischen Kontext. Ein Schwerpunkt liegt dabei auf der Berücksichtigung europäischer Werte und kommender Standards und Regulierungen.

Wir sind sehr froh über den breiten Zuspruch aus Wissenschaft, Wirtschaft und Politik. Denn nur gemeinsam mit allen Beteiligten können wir die Herausforderung meistern, die Möglichkeiten der Künstlichen Intelligenz zum Wohle der Menschen einzusetzen, durch leistungsfähige Forschung und Produkte international wettbewerbsfähig zu bleiben und den Wohlstand in Deutschland zu sichern.

KERNERGEBNISSE

KI-Foundation-Modelle sind die Zukunft.

KI-Foundation-Modelle werden zu disruptiven Anwendungen auf Basis Künstlicher Intelligenz führen. Diese Entwicklung wird in naher Zukunft viele neue Anwendungen, Plattformen und Geschäftsmodelle in beinahe allen Bereichen von Wirtschaft und Gesellschaft ermöglichen. LEAM zeigt einen Fahrplan auf, wie Deutschland an dieser Zukunftstechnologie teilhaben kann.

Die Entwicklung bei KI-Foundation-Modellen findet hauptsächlich in den USA statt.

US-amerikanische Modelle dominieren den Markt. Seit 2017 stammen 73 % der KI-Foundation-Modelle aus den USA und 15 % aus China. In den USA wird die Entwicklung vor allem von den großen Technologieunternehmen mit Investitionen im Milliardenbereich vorangetrieben.

73 % USA

Deutschland steht vor erheblichen Herausforderungen.

Es besteht die Gefahr, dass Deutschland im Bereich der KI-Foundation-Modelle den Paradigmenwechsel verpasst und zunehmend in die Abhängigkeit amerikanischer Modelle gerät. Die digitale Souveränität im Bereich KI und nachgelagerten Anwendungen ist akut bedroht und damit die Wettbewerbsfähigkeit der gesamten Wirtschaft in Deutschland.

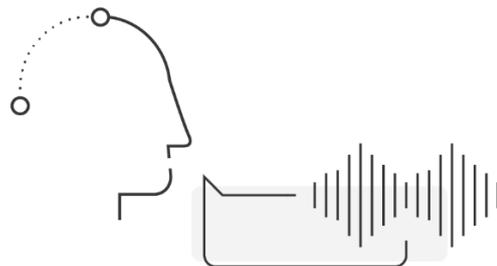
Die Relevanz für die Wirtschaft ist hoch.

Zwei Drittel der befragten KI-Unternehmen arbeiten bereits mit dieser Art von Modellen. Mehr als 80% sehen den Aufbau eines deutschen Ökosystems sowie die Entwicklung von Foundation-Modellen auf Basis europäischer Werte als relevant an. Die befragten Großunternehmen experimentieren zum Teil schon mit verschiedenen Anwen-

dungen und evaluieren den Einsatz für interne Prozesse und ihre Produkte. Die Bedeutung der Modelle wird in Zukunft weiter steigen.

Der Bedarf für Sprachmodelle ist besonders groß.

Von den KI-Foundation-Modellen besitzen Sprachmodelle aktuell die größte Relevanz in Nutzung und Entwicklung aller Modellarten. 71 % der KI-Unternehmen setzen hier den Fokus. Dahinter folgen multimodale Modelle (38 %) und Modelle basierend auf Geschäfts- und Fertigungsprozessen (34 %).



Die Nachfrage nach Services ist signifikant.

Die Wirtschaft benötigt Anwendungen, die auf Basis frei verfügbarer Foundation-Modellen auf ihre Bedürfnisse und Anwendungsfälle zugeschnitten sind. Voraussetzung dafür sind die Bereitstellung erforderlicher Rechenkapazität, Datenpools und Unterstützung beim Training und Beratungsleistungen.

Deutschland braucht eine leistungsfähige KI-Recheninfrastruktur. Eine Voraussetzung für die Entwicklung von KI-Foundation-Modellen ist der Zugang zu einer leistungsfähigen KI-Supercomputinginfrastruktur, die derzeit in Deutschland in dieser Form nicht vorhanden ist. Der Aufbau eines geeigneten KI-Rechenzentrums in Deutschland ist hard- und softwaretechnisch machbar und wird rund 350 - 400 Millionen Euro kosten.

bringt LEAM relevante Akteure aus dem Bereich KI-Foundation-Modelle zusammen. Die Initiative steht in engem Austausch mit Leuchtturmprojekten und Unternehmen wie OpenGPT-X, Aleph Alpha und Huggingface, um Synergien zu schöpfen.

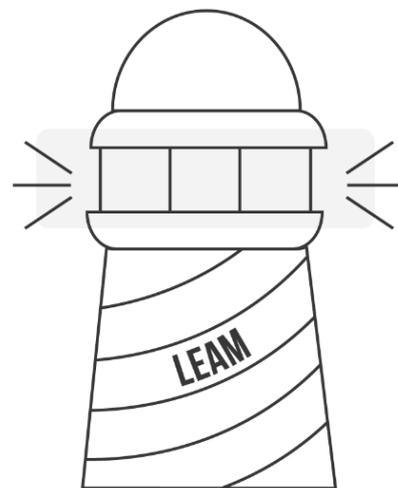
LEAM ermöglicht den Aufbau von vertrauenswürdigen Foundation-Modellen in Deutschland.

Die LEAM-Initiative (Large European AI Models) hat im Rahmen dieser Machbarkeitsstudie ein Konzept entwickelt, mit dem die notwendige Infrastruktur und die zugehörigen Services bereitgestellt werden können, um vertrauenswürdige Open Source Foundation-Modelle nach europäischen Ethikstandards zu entwickeln. Hiermit wird die Wirtschaft in die Lage versetzt, die dort entwickelten Foundation-Modelle für eigene Anwendungen zu nutzen.

LEAM kann die Wettbewerbsfähigkeit Deutschlands durch gemeinsame Investitionen sichern. Ein LEAM KI-Servicezentrum kann durch eine gemeinsame Initiative von Wirtschaft, Wissenschaft und Politik in Deutschland erfolgreich aufgebaut und betrieben werden. Die Kernfinanzierung durch die öffentliche Hand wird durch privatwirtschaftliche Investitionen ergänzt.

LEAM kann als KI-Leuchtturm Fachkräfte in Deutschland halten. Die Abwanderung von gut ausgebildeten Talenten in die USA aufgrund fehlender fachlicher Entwicklungsmöglichkeiten in Deutschland ist zunehmend ein Problem. Ein KI-Leuchtturm bietet die Chance durch State-of-the-Art Projekte, Talente in Deutschland zu halten und neue Fachkräfte auszubilden.

LEAM verbindet das bestehende KI-Ökosystem und baut es aus. Als einzigartige Initiative aus Unternehmen, KI-Start-ups und -Forschung



Inhaltsverzeichnis

Grußwort Jörg Bienert.....	3
Kernergebnisse.....	6
Inhaltsverzeichnis.....	8
Einleitung.....	11
Ziele der Machbarkeitsstudie	13
Autor:innen der Machbarkeitsstudie	13
Leseanleitung und Dokumentenstruktur	16
1. Das Paradigma der KI-Foundation-Modelle	19
2. Technologische Grundlagen.....	23
2.1 Die Bedeutung der Größe von KI-Foundation-Modellen.....	31
2.2 Anwendungsgebiete von KI-Foundation-Modellen.....	33
2.3 Vertrauenswürdige KI-Foundation-Modelle	45
2.4 Offene Forschungsfragen, neueste Entwicklungen und Erwartungen.....	51
3. KI-Foundation-Modelle im internationalen Vergleich	55
4. Bedarf der Wirtschaft an KI-Foundation-Modellen	67
5. Unterstützung bei der Entwicklung durch Forschung und Wissenschaft.....	76
6. Chancen und Pläne bei der Entwicklung europäischer KI-Foundation-Modelle....	85
6.1 Erste europäische multilinguale Foundation-Sprachmodelle.....	88
6.2 Vermeidung von Falschaussagen, Bias und Toxizität.....	97
6.3 Verbindung von Foundation-Modellen mit großen Wissensbeständen	99
6.4 Kombination von Sprache mit anderen Modi und Medien.....	100
6.5 Fragestellungen und Weiterentwicklungen.....	105
6.6 Foundation-Modelle in anderen Datendomänen	106
6.7 Zusammenfassung	107
7. Voraussetzungen bei Software und Personal	109
7.1 Applikations-Layer: Trainings- & Inference-Technologien.....	112
7.2 Data-Storage & -Loading-Layer.....	118
7.3 System-Layer	120
7.4 Framework- & Service-Layer	121
7.5 LEAM als Leuchtturmprojekt für die Zukunft des KI-Ökosystems.....	123
7.6 Zusammenfassung	125
8. Aufbau eines KI-Hochleistungsrechenzentrums.....	127
8.1 Definition Rechenzentrum	127
8.2 Anforderungen an ein KI-Hochleistungsrechenzentrum	130
8.3 Nachhaltigkeitsaspekte	136
8.4 Infrastrukturanforderungen im Detail.....	137
8.5 Standortauswahl.....	139
8.6 Betrieb eines KI-Rechenzentrums	156
8.7 Zusammenfassung und Empfehlung.....	164
9. Die organisatorische Struktur von LEAM.....	167
9.1 Zielgruppen des LEAM KI-Servicezentrums	167
9.2 Organisationseinheiten des LEAM KI-Servicezentrums.....	168
9.3 Das LEAM-Board	181
9.4 Zusammenfassung	182

10.	Betriebswirtschaftliche Aspekte	184
10.1	Kosten	184
10.2	Einnahmen	189
11.	Finanzierungsmodelle von LEAM	193
11.1	Öffentliche Finanzierung	194
11.2	Private Finanzierung	195
11.3	Public-Private-Partnership	197
11.4	Rechtliche Rahmenbedingungen.....	199
11.5	Auswirkungen der Rechtsmaterien auf die Finanzierungsmodelle.....	205
11.6	Abschließende Übersicht	218
12.	Gesellschaftsstruktur von LEAM.....	220
12.1	Öffentliche Finanzierung	220
12.2	Private Finanzierung	221
12.3	Public-Private-Partnership	221
13.	Szenario für ein LEAM KI-Servicezentrum	223
14.	Fazit.....	227
14.1	Beurteilung der Machbarkeit.....	227
14.2	Ausblick.....	229
I.	Quellenverzeichnis	231
II.	Abbildungsverzeichnis	240
III.	Tabellenverzeichnis	242
IV.	Abkürzungsverzeichnis	243
V.	Methodik der Machbarkeitsstudie	245
Anhang	251

MACHBARKEITSSTUDIE

zum Aufbau und Betrieb eines dedizierten
KI-Hochleistungsrechenzentrums für das
Trainieren großer KI-Modelle in
Deutschland

Einleitung

Ein Interview mit dem von Open AI trainierten Modell: ChatGPT

Autor:innen

Was sind große KI-Modelle?

Chatbot ChatGPT

Große KI-Modelle sind KI-Modelle, die auf sehr großen Datenmengen trainiert wurden und in der Lage sind, sehr komplexe Aufgaben zu lösen. Sie werden oft in der Sprachverarbeitung verwendet, um Aufgaben wie das Verstehen von menschlicher Sprache und das Erkennen von Objekten in Bildern und Videos zu erledigen. Ein bekanntes Beispiel für ein großes KI-Modell ist das GPT-3-Modell von OpenAI. Dieses Modell wurde auf einer riesigen Datenmenge trainiert und kann viele verschiedene Aufgaben im Bereich der Natur- und Sprachverarbeitung ausführen.

Warum sind große KI-Modelle wichtig?

Große KI-Modelle sind wichtig, weil sie in der Lage sind, sehr komplexe Aufgaben zu lösen, die für kleinere KI-Modelle zu schwierig wären. Durch das Training auf sehr großen Datenmengen sind sie in der Lage, sehr genaue Vorhersagen und Klassifizierungen durchzuführen und können daher in vielen Bereichen eingesetzt werden, in denen präzise Ergebnisse erforderlich sind, zum Beispiel in der Medizin, in der Finanzindustrie und in der Sprachverarbeitung.

Wo werden große KI-Modelle hauptsächlich entwickelt?

Große KI-Modelle werden hauptsächlich von Technologieunternehmen und forschungsorientierten Institutionen entwickelt. Diese Unternehmen und Institutionen verfügen über die Ressourcen und das Fachwissen, um die enormen Datenmengen zu beschaffen und zu verarbeiten, die für das Training von großen KI-Modellen erforderlich sind. Zu den bekannten Unternehmen und Institutionen, die sich mit der Entwicklung von KI-Modellen beschäftigen, gehören Google, OpenAI und das Massachusetts Institute of Technology (MIT).

Warum braucht Europa eine Initiative, um in der Entwicklung von großen KI-Modellen konkurrenzfähig zu bleiben?

Die Entwicklung von großen KI-Modellen ist ein wichtiger Bereich in der Welt der Technologie und viele Unternehmen und Länder investieren viel Zeit und Geld, um Fortschritte in diesem Bereich zu erzielen. Europa möchte in diesem Bereich konkurrenzfähig bleiben, um auch weiterhin eine führende Rolle in der Entwicklung von Technologien und Innovationen zu spielen. Eine Initiative in diesem Bereich könnte dazu beitragen, dass Europa weiterhin im Bereich der KI-Forschung und -Entwicklung führend bleibt und die Vorteile, die KI bieten kann, für seine Bürger und Unternehmen nutzen kann.

Wie leistungsstark große KI-Modelle bzw. KI-Foundation-Modelle¹ bereits sind, zeigen diese einleitenden Worte. Sie wurden nicht von einem der vielen Forscher:innen und Expert:innen geschrieben, die an dieser Studie mitgewirkt haben, sondern von dem auf GPT-3 basierenden Chatbot ChatGPT.

Die Antworten von ChatGPT zeigen eindrucksvoll, wozu KI-Foundation-Modelle bereits heute fähig sind. Dabei steht die Entwicklung noch ganz Anfang und findet bisher vor allem in den USA und China statt. Um eine mittelfristige Abhängigkeit ausländischer Technologiekonzerne zu verhindern und dem Wettbewerb standzuhalten, müssen Deutschland und Europa in die Lage versetzt werden, diese nächste Generation innovativer KI-Technologien mitzugestalten.

Zu diesem Zweck hat der KI Bundesverband 2021 die Initiative LEAM - Large European Language Models - ins Leben gerufen, die von über 40 namhaften Institutionen aus Forschung und Wirtschaft sowie weiteren europäischen KI-Verbänden unterstützt wird. Eine zentrale Forderung der Initiative ist die Förderung einer europäischen KI-Recheninfrastruktur, die von Wissenschaft, Industrie und Start-ups gleichermaßen genutzt werden soll, sowie der Aufbau eines dedizierten KI-Hochleistungsrechenzentrums in Deutschland.

Um die Umsetzbarkeit dieser Ziele zu überprüfen, hat das Bundesministerium für Wirtschaft und Klimaschutz (BMWK) den KI Bundesverband beauftragt, eine Machbarkeitsstudie zu LEAM durchzuführen. In dieser Studie werden die Herausforderungen und Potentiale von KI-Foundation-Modellen für die deutsche KI-Entwicklung kritisch untersucht, Strategien und Instrumente für die Umsetzung der LEAM-Ziele benannt und inhaltlich konkretisiert sowie Handlungsoptionen ausgearbeitet und miteinander verglichen. Die Erkenntnisse der Machbarkeitsstudie sind wegweisend für die Entwicklung der Künstlichen Intelligenz in Deutschland und entscheidend für die Innovationskraft des europäischen KI-Ökosystems und der digitalen Souveränität von Deutschland in der Zukunft.

¹ Die Begriffe "Große KI-Modelle" und "KI-Foundation-Modelle" sind synonym. Für diese Studie wird der Begriff "KI-Foundation-Modelle" benutzt, der international für diese Art der KI anerkannt ist.

Ziele der Machbarkeitsstudie

In kürzester Zeit hat Künstliche Intelligenz gigantische Entwicklungssprünge gezeigt und damit selbst Technologieexpert:innen ins Staunen versetzt. Zurückzuführen ist dies insbesondere auf die großen Fortschritte im Bereich der großen KI-Modelle. Seit OpenAI GPT-3 eingeführt hat, wurden auf Basis der großen KI-Modelle, die auch Foundation-Modelle genannt werden, viele Anwendungen entwickelt.

Neben den enormen Chancen, die sich daraus für die Arbeit und das Leben eröffnen, ergeben sich daraus auch einige Herausforderungen für die deutsche Wirtschaft, Wissenschaft und Gesellschaft, denn Europa ist im Wettbewerb um KI-Foundation-Modelle abgeschlagen. Für Deutschland heißt das konkret, Datenschutz und Datensicherheit verfolgen geringere Standards, Verzerrungen und mangelnde Datenqualität können aufgrund fehlender Transparenz nicht identifiziert und entgegengewirkt werden, und deutsche Unternehmen werden lediglich Nutznießer und keine Gestalter von Foundation-Modelle. Technologisch rutscht Deutschland damit in die Abhängigkeit.

Eine grundlegende Herausforderung dabei: Für die Erstellung von KI-Foundation-Modellen sind enorme Rechenkapazitäten und Ressourcen nötig, die im Vergleich zu den USA nicht ausreichend in Deutschland bereitstehen.

Ziel dieser Studie ist es ein Konzept zu erarbeiten, wie in Deutschland eine dezidierte KI-Recheninfrastruktur aufgebaut werden kann, die es ermöglicht KI-Foundation-Modelle zu trainieren und der Wirtschaft bereitzustellen. Dabei sollen vor allem die Bedürfnisse deutscher Unternehmen berücksichtigt werden.

Indem die Studie Empfehlungen für die Ausgestaltung eines KI-Rechenzentrums und einer entsprechenden Servicegesellschaft gibt, möchte sie der privaten und öffentlichen Hand als eine Entscheidungsgrundlage für die Umsetzung von LEAM in Deutschland dienen.

Autor:innen der Machbarkeitsstudie

Die LEAM Machbarkeitsstudie wurde in Zusammenarbeit mit Alexander Thamm GmbH, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), eco – Verband der Internetwirtschaft e. V., Fieldfisher LLP, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS), Merantix Momentum GmbH, Simmons & Simmons und Ubermetrics Technologies GmbH durchgeführt. Die Gesamtprojektleitung oblag dem KI Bundesverband e.V..

Autor:innen:



Jörg Bienert
Alexander Thamm GmbH
KI Bundesverband e.V.



Roland Broch
eco – Verband der
Internetwirtschaft e.V.



Patrick Bunk
Ubermetrics Technologies
GmbH



Dr. Aljoscha Burchardt
Deutsches
Forschungszentrum für
Künstliche Intelligenz (DFKI)



Vanessa Cann
KI Bundesverband e.V.



Ben Deetjen
Deutsches
Forschungszentrum für
Künstliche Intelligenz (DFKI)



Alex Dickmann
KI Bundesverband e.V.



Dr. Nicolas Flores-Herr
Fraunhofer-Institut für
Intelligente Analyse- und
Informationssysteme (IAIS)



Laszlo Friedmann
Fraunhofer-Institut für
Intelligente Analyse- und
Informationssysteme (IAIS)



Christopher Götz
Simmons & Simmons LLP



Fabio Haenel
Merantix Momentum GmbH



Dennis Hillemann
Fieldfisher LLP



Kim Lambers
eco – Verband der
Internetwirtschaft e.V.



Lukas Lux
Alexander Thamm GmbH



Dr. Johannes Otterbach
Merantix Momentum GmbH



Dr. Gerhard Paaß
Fraunhofer-Institut für
Intelligente Analyse- und
Informationssysteme (IAIS)



Charlene Röhl
Deutsches
Forschungszentrum für
Künstliche Intelligenz (DFKI)



Dr. Sven Schmeier
Deutsches
Forschungszentrum für
Künstliche Intelligenz (DFKI)



Dr. Philipp Slusallek
Deutsches
Forschungszentrum für
Künstliche Intelligenz (DFKI)



Boris Strauch-Rötting
Simmons & Simmons LLP



Oliver Süme
Fieldfisher LLP



Dr. Christian Temath
Fraunhofer-Institut für
Intelligente Analyse- und
Informationssysteme (IAIS)



Jakob Tesch
Ubermetrics Technologies
GmbH



Hauke Timmermann
eco – Verband der
Internetwirtschaft e.V.



Dr. Hans Uszkoreit
Deutsches
Forschungszentrum für
Künstliche Intelligenz (DFKI)

Unser Dank gilt auch dem gesamten Team des KI Bundesverbandes für die Umsetzung der LEAM-Konferenz und ihr Mitwirken im Rahmen der Studie: Daniel Abbou, Stefanie Baade, Dr. Sandra Bütow, Katharina Fischer, Franziska Fink, Phillip Handy, Benjamin Rodatz, Valentin Roth, Julia Sartison und Esther Schragmann.

Leseanleitung und Dokumentenstruktur

Kapitel 1 bis 5 sind der Bestandsaufnahme gewidmet. Kapitel 1 beginnt mit einem Abschnitt über die wichtigsten Eigenschaften des neuen Paradigmas, gefolgt von einem Abriss des Standes der Technologie in der internationalen Forschung und Entwicklung (Kapitel 2). Dieser Abriss schildert auch vielfältige Anwendungen, die bereits realisiert und evaluiert wurden (Kapitel 2.2). Ein spezieller Abschnitt widmet sich den Technologien, die die Vertrauenswürdigkeit der KI-Foundation-Modelle sichern und kritische Fälle inadäquater Performanz verhindern sollen (Kapitel 2.3). Dem schließt sich eine Analyse des internationalen Wettbewerbs an, die auf einer Zusammenstellung aller bisher veröffentlichten KI-Foundation-Modelle beruht (Kapitel 3). In dieser Analyse werden auch die Ursachen des festgestellten Ungleichgewichts zwischen den USA, China und Deutschland bzw. Europa diskutiert.

Nach dem Überblick zum aktuellen Stand der internationalen KI-Entwicklung konzentrieren wir uns in den Kapiteln 4 und 5 auf die Situation in Deutschland. Dazu beleuchten wir insbesondere die Bedarfe der deutschen Wirtschaft und legen dar, wie die Forschung die Entwicklung der benötigten europäischen KI-Foundation-Modelle unterstützen kann. Zu diesem Zweck wurden zahlreiche Expert:innen in Wissenschaft und Wirtschaft befragt.

Das sechste Kapitel zieht die Konsequenzen aus den Ergebnissen der vorangegangenen Kapitel. Die Untersuchung hat das große Potenzial der Technologie, gleichzeitig aber auch einen immensen Bedarf an Foundation-Modellen ermittelt. Dieses Kapitel beschreibt, welche Modelle vordringlich entwickelt werden sollten und könnten, ob und wie man die Daten für diese Modelle bekommen kann und welche Optionen es für künftige Modellentwicklungen gibt.

Die folgenden zwei Kapitel untersuchen die Anforderungen, die an High-Performance-Computing-Systeme (HPC) und speziell an ein benötigtes KI-Hochleistungsrechenzentrum gestellt werden. In Kapitel 7 wird erklärt, wie die dafür notwendige Software-Architektur aussehen kann und wie daraus Anforderungen an Hardware und Personal abgeleitet werden können. Dabei zeigen wir auf, wie sich diese Architektur mittels Open-Source-Software (OSS) aufbauen lässt, um die Wichtigkeit von OSS für ein wettbewerbsfähiges KI-Ökosystem zu betonen. Dabei wird die technische Machbarkeit unmittelbar in Zusammenhang mit den wissenschaftlichen und infrastrukturellen Kapiteln dieser Machbarkeitsstudie gestellt, um so den gesellschaftlichen Mehrwert des Betriebs eines kompetitiven KI-Hochleistungsrechenzentrums zu erläutern.

Kapitel 8 betrachtet insbesondere die technischen Voraussetzungen, die von KI-Foundation-Modellen an die Infrastruktur eines entsprechenden Rechenzentrums gestellt werden. Außerdem wird eine Übersicht über die HPC-Kapazitäten im Bereich der Forschung und der kommerziellen Anbieter gegeben, die bereits für KI-Anwendungen geeignete HPC-Cloudlösungen anbieten. Es werden außerdem die nötigen Schritte für die Standortermittlung und den Aufbau eines KI-Hochleistungsrechenzentrums mit passender Gebäudeinfrastruktur beleuchtet. Weitere Möglichkeiten wie Collocation-Betrieb oder der mögliche Bezug von HPC-Rechenleistungen aus der Cloud werden anhand von Beispielen untersucht. Beim Betrieb eines HPC-Rechenzentrums sind mit Blick auf die Entwicklung der Strompreise und Anforderungen des Klimaschutzes insbesondere die Energieeffizienz der Systeme sowie der Nachhaltigkeitsaspekt beim Betrieb eines KI-Rechenzentrums wichtige zu untersuchende Faktoren.

Kapitel 9 bis 13 beleuchten die organisatorische, wirtschaftliche und juristische Machbarkeit eines KI-Hochleistungsrechenzentrums. Wir schlagen die Einrichtung eines LEAM-KI-Servicezentrums vor, das seine Dienste der Wirtschaft und Wissenschaft anbietet. Dafür werden verschiedene Organisationsmodelle, Gesellschaftsstrukturen und Finanzierungsmöglichkeiten näher beleuchtet und diskutiert. Darüber hinaus findet eine rechtliche Bewertung der Optionen statt. Das Kapitel schließt mit einem Szenario zur Realisierung eines LEAM-KI-Servicezentrums.

Die Ergebnisse werden in Kapitel 14 zusammengefasst. Das Kapitel gibt außerdem die Empfehlung ab, eine Projektentwicklungsgesellschaft zu gründen, die das Thema weiter vorantreibt.



Das Paradigma der KI-Foundation-Modelle

.....●

1. Das Paradigma der KI-Foundation-Modelle

KI-Foundation-Modelle schreiben das neueste Kapitel in der Erfolgsgeschichte der Künstlichen Intelligenz.

KI-Foundation-Modelle sind große neuronale KI-Modelle, die auf gigantischen Mengen generischer Daten vortrainiert wurden. Das Besondere an diesen Modellen ist, dass sich das im sogenannten Vortraining (pre-training) erworbene implizite Wissen als Grundlage für viele verschiedene Anwendungen eignet. Für manche Anwendungen ist sogar kein oder nur minimales zusätzliches Training erforderlich. Andere Anwendungen erfordern zwar weiterhin ein Training auf anwendungsspezifischen Lerndaten, wobei aber oft viel weniger Lerndaten benötigt werden oder aber eine Leistungsfähigkeit erreicht wird, die ohne das generische Vortrainieren nicht möglich war.

Der breiten Öffentlichkeit bekannt geworden ist das neue Paradigma spätestens seit dem großen Medienecho auf GPT-3, einem großen Sprachmodell des kalifornischen KI-Labors OpenAI. GPT-3 wurde 2020 veröffentlicht und gilt als spektakulärer Durchbruch auf dem Gebiet der intelligenten Sprachtechnologie. Das KI-Modell kann anspruchsvolle Texte verfassen, Fragen beantworten, Sätze ergänzen und nebenbei auch noch ganz passabel übersetzen. Derzeit gilt die mediale Aufmerksamkeit dem neuen KI-System ChatGPT, einer Variante des GPT-3 Modells, das Journalist:innen und Testbenutzer:innen gleichermaßen fasziniert aber manchmal auch verschreckt, weil es so eloquent parliert und meist informative, bedachte und ausgewogene Antworten auch auf schwierige Fragen gibt und nach Benutzervorgaben mitunter sogar druckreife Texte für die verschiedensten Zwecke formuliert. Und all das in mehreren Sprachen.

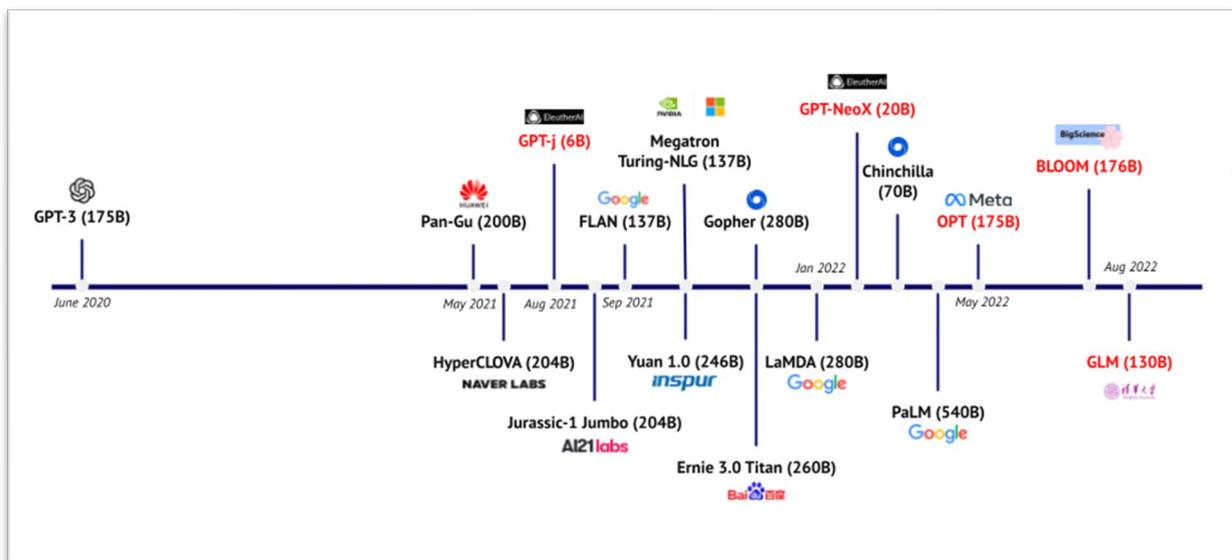


Abb. 1: Anzahl der Parameter großer KI-Sprachmodelle seit GPT-3 (Open Source Modelle rot markiert). Quelle: state of ai Report 2022 (Benaic & Hogarth, 2022)

Möglich wurde diese erfolgreiche Innovation durch eine neue Architektur für neuronale Netze, Transformer genannt, die 2017 in der Google-Forschung entwickelt und zuerst erfolgreich in der automatischen Textübersetzung getestet wurde. 2018 folgten zwei KI-Foundation-Modelle, die jeweils nur Teile der Transformer-Architektur realisieren: GPT, der Vorläufer von GPT-3 von OpenAI und BERT, ein einflussreiches bidirektionales Sprachmodell aus der Google-Forschung.

Die Grundidee des Transferlernens ist die Nutzung von bereits trainierten Netzen für neue Aufgabenstellungen. Anstatt für eine Anwendung ein neues Netzwerk zu trainieren, verwendet man ein Netzwerk, das bereits für eine andere Aufgabenstellung vortrainiert wurde. So kann durch den Einsatz des vortrainierten Netzes der Bedarf an annotierten (labeled) anwendungsspezifischen Lerndaten reduziert werden. Gelingt es nun, Aufgaben für das Vortraining so zu wählen, dass entweder hinreichend große Mengen an bereits annotierten Lerndaten vorhanden sind oder aber die annotierten Daten vollautomatisch hergestellt werden können, dann verringert sich der Aufwand für die Datenannotierung immens.

Die ersten KI-Foundation-Modelle waren große Sprachmodelle. Sie unterscheiden sich von früheren KI-Modellen durch ihre vielseitige Verwendbarkeit. Möglich geworden ist diese Vielseitigkeit durch eine zentrale inhärente Eigenschaft der neuen KI-Kerntechnologie, die als Emergenz bezeichnet wird. Damit ist eine neue Stufe in der Evolution der Künstlichen Intelligenz erreicht. Bisherige neuronale KI-Modelle waren immer auf eine bestimmte Anwendung ausgerichtet. Sie beruhten in der Regel auf Training durch überwachtes (supervised) oder semi-überwachtes (semi-supervised) Lernen. Aber im Gegensatz zur menschlichen Intelligenz konnten die erworbenen Fähigkeiten nicht für weitere Aufgabenstellungen genutzt werden.

Deshalb wurde diese Künstliche Intelligenz auch treffend als schmale KI (Narrow AI) bezeichnet und als solche der sogenannten Künstlichen Allgemeinen Intelligenz (Artificial General Intelligence - AGI) gegenübergestellt.

Nun gibt es aber viele Anwendungen, welche die Kombination von Aufgaben erfordern. So kann ein Voice-Chatbot aus der Kombination von Spracherkennung, semantischer Analyse, Fragenbeantwortung und akustischer Sprachausgabe bestehen, die jeweils durch eigene Modelle realisiert sind. Hier kann es zu Inkonsistenzen kommen, indem zum Beispiel eins dieser Modelle Wörter oder Begriffe kennt, die ein anderes nicht gelernt hat. Die Homogenisierung der Modelle hat diese Inkonsistenzen reduziert. Die KI-Foundation-Sprachmodelle können das vortrainierte implizite Sprachwissen bereits für mehrere Anwendungen einsetzen. Es kommt nicht mehr vor, dass eine Anwendung oder Komponente, Begriffe oder Satzstrukturen nicht kennt, die von einer anderen beherrscht werden. Das Paradigma der versatilen Foundation-Modelle stellt so den nächsten Schritt in Richtung einer breiteren KI dar.

Die Bezeichnung *Foundation-Modelle* wurde erst im August 2021 durch das neue Center for Research on Foundation Models (CRFM) an der Stanford Universität vorgeschlagen (Bommasani et al., 2021), das sich als Teil des Stanford Institute for Human-Centered Artificial Intelligence's (HAI) ganz dediziert dem neuen KI-Paradigma widmet. Der Terminus ist treffender als der Begriff *große Sprachmodelle*, denn Foundation-Modelle müssen nicht immer Sprachmodelle sein, sondern können z.B. auch auf Bildern, Videos oder DNA-Sequenzen trainiert und dann jeweils für eine Vielzahl von Anwendungen eingesetzt werden. Zudem gibt es auch große Sprachmodelle, die mit entsprechend ausgewählten und annotierten Daten nur für eine spezielle Anwendung, zum Beispiel maschinelle Übersetzung, trainiert wurden und sich somit nicht als Grundlage (Foundation) für viele verschiedene Anwendungen eignen.

Weil die vortrainierten, vielseitigen Foundation-Modelle gegenwärtig die Diskussion um die nächsten Durchbrüche der KI dominieren, wurde der intuitive neue Begriff sehr schnell von der internationalen Forschungsgemeinschaft aufgegriffen. In ihrem initialen Positionspapier zu dem Forschungsthema schildern die Stanford-Wissenschaftler nicht nur das Anwendungspotenzial der bereits existierenden Modelle, sondern sie argumentieren auch überzeugend, dass das Paradigma der Foundation-Modelle die nächste Entwicklungsstufe der KI bestimmen wird, in der die Modelle Fähigkeiten aufweisen werden, die bis vor Kurzem noch undenkbar schienen und die den Menschen bei vielen Aufgaben übertreffen.

Im gleichen Artikel schildern sie aber auch die Risiken, die entstehen können, wenn diese mächtige Technologie die Konzentration von technologischer und wirtschaftlicher Macht in der Hand einiger weniger IT-Konzerne verschärft. Die bloße Verfügbarmachung von fertigen Modellen genügt nicht, um die wirtschaftlichen und sozialen Interessen der Gesellschaft zu sichern und den Missbrauch der Technologie wirksam zu verhindern.



Technologische Grundlagen



2. Technologische Grundlagen

Sprachmodelle gehören zu den Grundwerkzeugen der maschinellen Sprachverarbeitung (Natural Language Processing). Das sind mathematische Modelle, die bestimmen können, ob gewisse Sätze oder Äußerungen zur Sprache gehören oder nicht, beziehungsweise mit welcher Wahrscheinlichkeit sie das tun. Bereits die ersten Versionen von Siri, Alexa oder Google Translate nutzten stochastische Sprachmodelle, die für jede Abfolge von drei, vier oder fünf Wörtern die Wahrscheinlichkeit gelernt hatten, in genau dieser Reihenfolge in Texten oder gesprochenen Äußerungen vorzukommen. Mit solchen Modellen, die damals noch nicht als neuronale Netze realisiert wurden, konnte die Korrektheit oder Natürlichkeit von generierten Sätzen verbessert werden. Man hat die Wahrscheinlichkeiten aber auch verwendet, um bei der Analyse gesprochener Eingaben Unsicherheiten in der akustischen Erkennung von Wörtern aufzulösen. Diese Wahrscheinlichkeiten waren durch syntaktische und semantische Faktoren bestimmt, insbesondere durch grammatische Regularitäten und Wortbedeutungen.

Schon früh wurde in der Sprachverarbeitung deshalb die Idee entwickelt, die Bedeutung von Wörtern durch die Wörter in der Nachbarschaft zu erklären. Leider gibt die Schreibweise von Wörtern nur wenig Aufschluss über ihre Bedeutung. Daher entstand schon vor längerer Zeit die Idee, die Bedeutung eines jeden Wortes durch einen langen Vektor, eine Einbettung, zu repräsentieren. Allerdings stellte sich heraus, dass viele Wörter je nach Kontext unterschiedliche Bedeutungen haben. Beispielsweise kann „Bank“ ein Sitzmöbel oder ein Finanzinstitut sein. Vor fünf Jahren wurde von Google-Wissenschaftler:innen der Transformer entwickelt, welcher diese Bedeutungsunterschiede mit kontextsensitiven Einbettungen erfassen kann (Vaswani et al., 2017). Transformer sind eine Variante der tiefen Neuronalen Netze, die seit 2012 weite Teile der KI revolutioniert haben.

Das wirklich Neue an diesen Transformern ist, dass sie die Einbettung des Kontextes in vergleichbar effizienter Weise berechnen können. Dies war zuvor nicht möglich, die Kontexte beschränkten sich vorher nur auf wenige Worte. Ähnlich wie wir Menschen, erkennen die Transformer dabei die relevanten Worte in weiteren Kontexten, auch über Satzgrenzen hinweg, und können dadurch den semantischen Bezug herstellen.

Da die Anzahl der unterschiedlichen Wörter der Sprache durch Verbindung von einfachen Wörtern zu zusammengesetzten Wörtern nahezu unbeschränkt ist, verwendet neuere Sprachmodelle statt der Wörter ein beschränktes Vokabular von Token (Teilwörtern und häufigen Wörtern), aus denen man jedes Wort zusammensetzen kann. Die Tokenisierung wird, genauso wie das KI-Modell, auf Trainingsdaten trainiert und ist somit Bestandteil des Modells.

Herzstück des Transformers ist der Self-Attention Block, der in sehr flexibler Weise die Bedeutungsrelationen zwischen verschiedenen Token durch Korrelationen der jeweiligen Einbettungen ermittelt und damit neue Einbettungen konstruiert. Mit ihm wurden Sprachmodelle mit Milliarden von Parametern trainiert. Paradebeispiel ist GPT-3 (Brown et al., 2020), das syntaktisch und inhaltlich stimmige Texte von bisher unerreichter Qualität produzieren kann.

Die Ableitung von kontextsensitiven Einbettungen lässt sich am besten mit dem BERT-Modell (Devlin et al., 2019) erläutern: Jedem Token des Eingabetextes wird ein Einbettungsvektor zugeordnet, der die semantische Bedeutung des Tokens repräsentiert und ein weiterer Vektor, der die Position des Tokens im Text markiert. Diese Einbettungsvektoren sind Parameter und werden im Laufe des Trainings angepasst.

Das Verfahren der *Self-Attention* kann nun in dem Satz „Die Bank verleiht Geld“ die Einbettung von „Bank“ durch die Berücksichtigung der Einbettung von „Geld“ so abändern, dass die Bedeutung „Finanzinstitut“ betont wird. Dazu berechnet es die „Korrelation“ (Skalarprodukt) der Einbettung von Bank mit denen sämtlicher anderen Token (Abb. 2). Dies geschieht für komplementäre „Aspekte“ der Einbettungen, die durch Parameter ausgedrückt werden (*Attention-Heads*). Schließlich werden die mit den Korrelationen gewichteten Einbettungen aller Token addiert, um eine neue Einbettung für das Token „Bank“ zu erzeugen, die der Bedeutung von „Geld“ Rechnung trägt. Jeweils eine Self-Attention Schicht mit mehreren Attention-Heads sowie eine anschließende nichtlineare voll verbundene Schicht von Neuronen bilden einen Encoderblock, welcher grundlegender Bestandteil fast aller Foundation-Modelle ist.

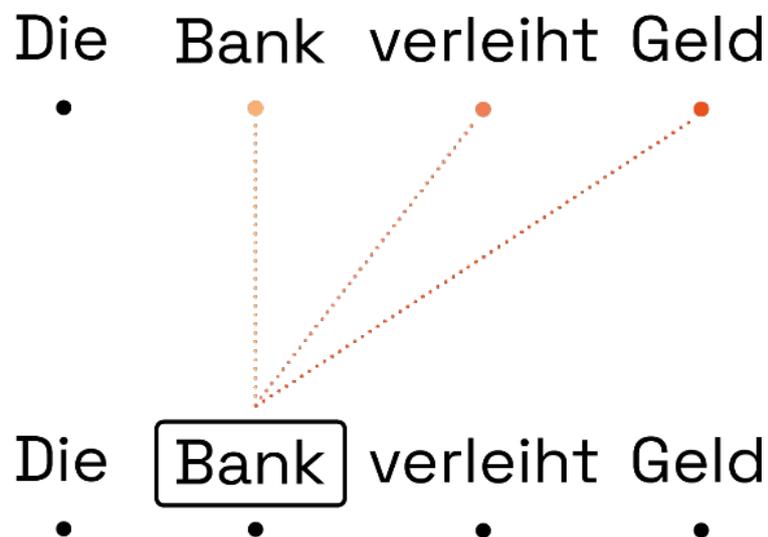


Abb. 2: Verdeutlichung der Self-Attention am Satz "Die Bank verleiht Geld". Das Token Bank (unten) hat eine hohe Korrelation mit dem Token Geld (oben), wobei die Korrelation zu den anderen Token geringer ausfällt.

Primäres Ziel des *BERT*-Modells (Devlin et al., 2019) ist es, kontextsensitive Einbettungsvektoren zu bestimmen, die möglichst viele Informationen über die jeweiligen Token eines Textes enthalten. Zunächst werden die Parameter mit Zufallszahlen initialisiert. Weiter werden im Eingabetext ein Teil der Eingabetoken durch „MASK“ ersetzt. Für jeden Eingabetoken berechnet BERT als Autoencoder mit Hilfe einer Reihe von Encoderblöcken die kontextsensitiven Einbettungen, welche die semantische Bedeutung des Tokens erfassen und mit jeder Schicht besser werden. Trainingsziel ist es, aus den besonders aussagekräftigen Einbettungen in der obersten Schicht an der Position eines maskierten Tokens eine möglichst hohe Wahrscheinlichkeit für das maskierte Token zu prognostizieren. Dies geschieht durch Optimierung der Parameter. Da Teile der beobachteten Daten prognostiziert werden müssen und keine menschlichen Annotationen benötigt werden, spricht man hier von selbstüberwachtem Lernen.

Das erworbene Wissen über Sprache kann man in einem zweiten Schritt auswerten, indem man BERT für eine weitere Klassifikationsaufgabe trainiert. Im einfachsten Fall wird für einen zusätzlichen „CLS“-Token eine kontextsensitive Einbettung berechnet, aus der mit Hilfe eines logistischen Regressionsmodells die Wahrscheinlichkeit der beobachteten Klasse prognostiziert werden kann. Ein Beispiel ist die Sentimentanalyse, bei der das Modell entscheiden muss, ob der Eingabetext eine negative oder positive Bewertung enthält. Diese zweite Trainingsaufgabe, auch Finetuning genannt, benötigt einen von Menschen annotierten Trainingsdatensatz. Obwohl das Finetuning alle Parameter des Modells anpasst, benötigt es nur einen kleinen Bruchteil des Trainingsaufwandes für das Basismodell, sodass die annotierten Trainingsdaten für das Finetuning meist relativ wenige Beispiele umfassen muss. Zur Unterscheidung wird die erste Trainingsaufgabe mit einem großen allgemeinen Textkorpus ohne Annotationen auch als Vortraining bezeichnet. BERT kann für viele semantische Klassifikationsaufgaben die Genauigkeit bisheriger Modelle wesentlich verbessern. Die Übertragung von erworbenem Wissen von einem Lernproblem auf ein anderes, aber verwandtes Problem nennt man Transferlernen.

Sprachmodelle verwenden ebenfalls Schichten von Self-Attention Modulen. Sie werden aber nicht darauf trainiert, maskierte Token innerhalb eines Textes zu prognostizieren, sondern sollen für einen bestehenden Anfangstext das nächste Token vorhersagen. Hierbei werden mehrere Schichten von Encoderblöcken auf die bisher bekannten Worte des Textes angewendet. Die kontextsensitive Einbettung des letzten bekannten Wortes in der obersten Schicht bildet dann die Eingabe für ein logistisches Regressionsmodell, das die Wahrscheinlichkeit der unterschiedlichen Token an der nächsten Position prognostiziert. Während des Trainings wird das Modell so angepasst, dass diese letzte Einbettung möglichst viel Informationen über das nächste Token enthält und die Token der Trainingsmenge eine hohe Wahrscheinlichkeit erhalten. Sprachmodelle wie GPT-3 (Brown et al., 2020) und PaLM (Chowdhery et al., 2022) sind in der Lage, Anfangstexte syntaktisch fehlerfrei und inhaltlich überwiegend stimmig fortzusetzen, indem sie einen Token nach dem anderen generieren. Die Auswahl des nächsten Tokens findet dabei zufällig gemäß den abgeschätzten Wahrscheinlichkeiten statt, so dass bei einer Wiederholung immer ein neuer Text entsteht.

Darüber hinaus kann ein Sprachmodell auch Anweisungen ausführen. Beispielsweise antwortet GPT-3 auf den Starttext "Create an outline for an essay about Walt Disney: I: Introduction" mit einem detaillierten Text über Walt Disney. Man kann also ohne Zusatztraining erreichen, dass GPT-3 eine bisher unbekannte Aufgabe löst. Häufig kann man die Qualität der Antworten noch durch zusätzliche Beispiele verbessern, z.B. durch die Instruktion „English: I do not speak French. French: Je ne parle pas français. English: Where is the restroom? French:“. GPT-3 erkennt die Eingabe als Aufforderung zur Übersetzung und liefert die französische Übersetzung. Dieses „k-shot Learning“ eröffnet völlig neue Möglichkeiten zur Nutzung von Sprachmodellen, ohne zusätzliches Finetuning. Allerdings ist die Genauigkeit oft höher, wenn das Sprachmodell durch Finetuning für die neue Aufgabe trainiert wird.

Sequence-to-Sequence-Modelle (seq2seq) übersetzen eine Sequenz von Token in eine andere Sequenz. Wichtigstes Anwendungsgebiet ist die Übersetzung eines Textes in eine andere Sprache. Der Prototyp dieser Architektur wurde von (Vaswani et al., 2017) vorgestellt:

- Der **Encoder** ist ein BERT-Modell, welches kontextsensitive Einbettungen der Eingabetoken berechnet.
- Der **Decoder** arbeitet wie ein Sprachmodell und wird auf die bisher erzeugten Token der Übersetzung angewendet. Jeder Decoderblock enthält mehrere Self-Attentions, die die Korrelation mit den schon generierten Token der Übersetzung auswerten. Zum anderen werden über sogenannte Cross-Attentions die Informationen in den Einbettungsvektoren der Eingabetoken berücksichtigt. Aus den Einbettungen des obersten Decoderblocks wird die Wahrscheinlichkeit der Token für die nächste Wortposition der Übersetzung berechnet.

Die Trainingsmenge enthält Paare von Eingabetexten und deren Übersetzungen. Beim Training werden die Parameter von Encoder und Decoder gleichzeitig so angepasst, dass die Wahrscheinlichkeiten der korrekten Ausgabewörter möglichst hoch werden.

Es ist bemerkenswert, dass der ursprüngliche Encoderblock mit Multi-Head-Self-Attention immer noch von fast allen Foundation-Modellen verwendet wird. Abbildung 3 zeigt die Gemeinsamkeiten in der Struktur von BERT-Encoder, Sprachmodell und Transformer Encoder-Decoder.

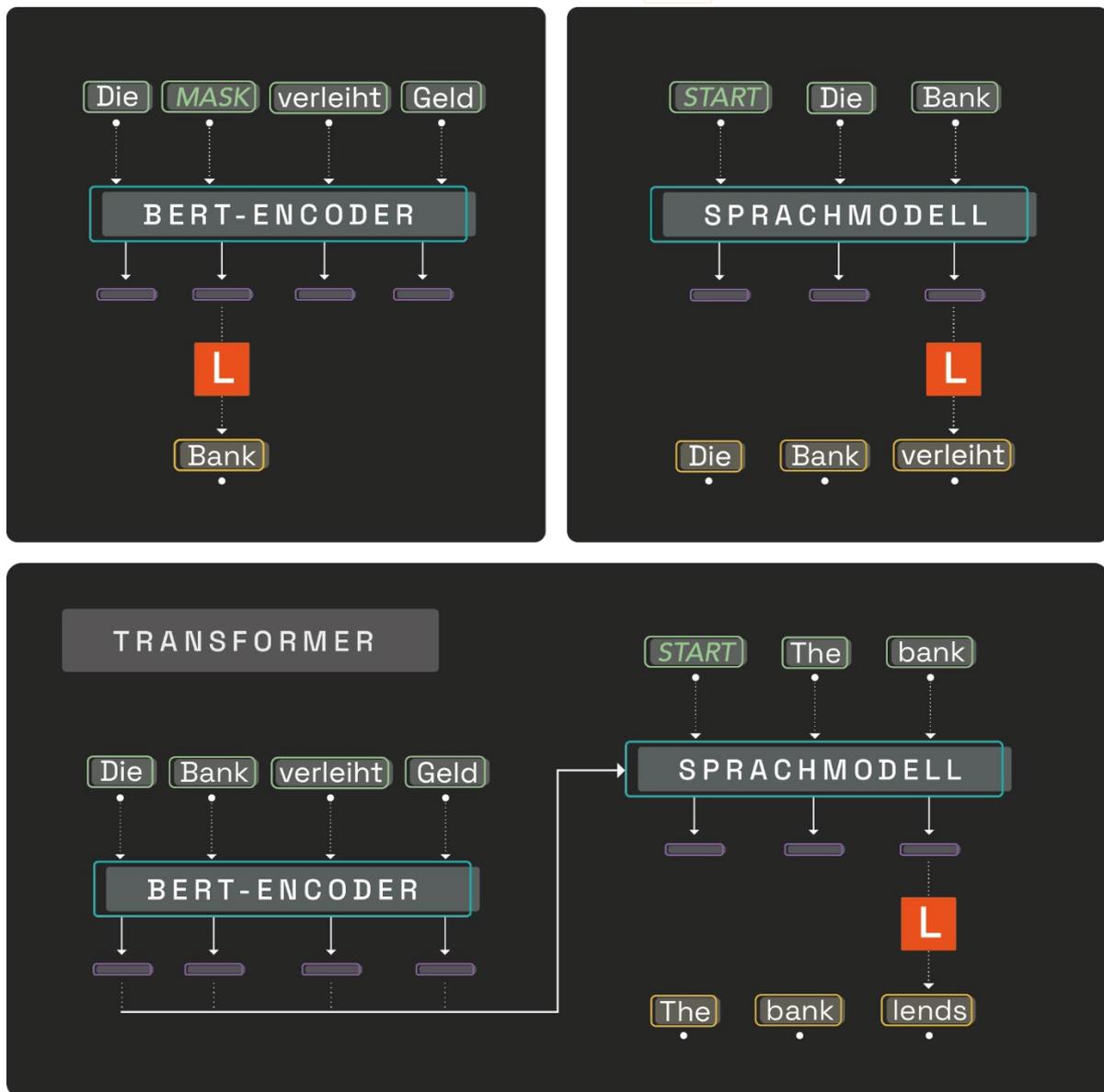


Abb. 3: Zentraler Bestandteil der Foundation-Modelle sind Schichten mit Self-Attention Blöcken (blau), die kontextsensitive Einbettungsvektoren (violett) von Eingabetoken (grün) berechnen. Die logistische Schicht L prognostiziert die Wahrscheinlichkeit der Ausgabedaten. Beim Training werden die Parameter so optimiert, dass die Wahrscheinlichkeiten der korrekten fehlenden bzw. nächsten Token (gelb) möglichst hoch sind.

In den letzten Jahren wurde eine Reihe von Verbesserungen für die ursprüngliche Architektur gesucht. Bei der Self-Attention wächst der Rechenaufwand und der Speicherbedarf quadratisch mit der Länge der Eingabesequenz. Daher wurden Varianten entwickelt, bei denen der Aufwand nur noch linear ansteigt. Dies ist eine signifikante Verbesserung, die es ermöglicht, sehr viel längere Texte als Eingabesequenz zu geben. Dadurch können Sprachmodelle auf Textstellen weiter vorne im Text Bezug nehmen.

Foundation-Sprachmodelle können gleichzeitig auf Dokumenten unterschiedlicher Sprachen trainiert werden, wenn ein gemeinsames Tokenvokabular verfügbar ist. Dies ermöglicht z.B. Frage-Antwort Systeme, die Informationen aus unterschiedlichen Sprachen in eine Antwort integrieren können. Dabei stellte sich heraus, dass die Modelle selbst dann die Bedeutung der Wörter in unterschiedlichen Sprachen in Beziehung setzen können, wenn jedes Trainingsdokument nur in einer Sprache verfasst ist (Liu et al., 2020).

Auch große Foundation-Modelle können nur begrenzt Informationen in ihren Parametern speichern. Daher wurde mit unterschiedlichen Ansätzen versucht, zusätzliche Informationen in Foundation-Modelle einzubeziehen. Dabei werden hauptsächlich folgende Ansätze verfolgt:

- Die zusätzlichen Informationen werden sprachlich formuliert. Dies funktioniert sowohl für Tabelleninhalte (Yin et al., 2020) als auch für Wissensbasen. Beispielsweise verbalisiert TekGen (Agarwal et al., 2021) die komplette Wikidata Wissensbasis mit Hilfe des T5 Seq2seq-Modells. Dieser Datenbestand kann dann zum Beispiel als zusätzliche Trainingsdaten für ein Sprachmodell verwendet werden.
- Der vielversprechendste Ansatz ist die Verwendung von Retrievaltechniken. Viele aktuelle Sprachmodelle nutzen ein Retriever-Reader-Modul, um die gefundenen Dokumente zu berücksichtigen (Izacard and Grave, 2021).

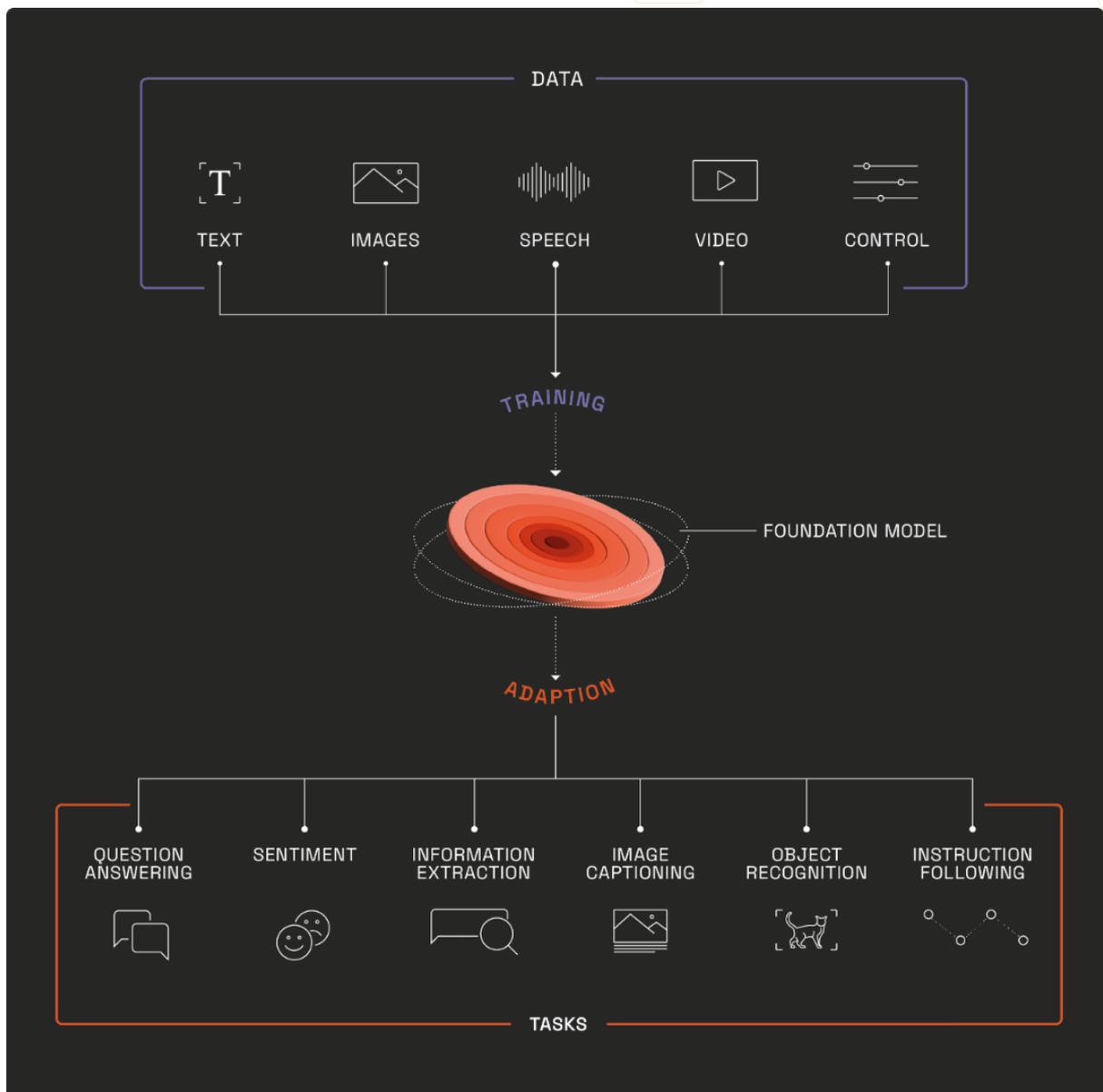


Abb. 4: Ein Foundation-Modell kann Informationen aus verschiedenen Datenquellen verschiedener Modalitäten berücksichtigen. Dieses eine Modell kann dann eine Vielzahl von nachgelagerten Aufgaben lösen (Bommasani et al., 2021).

Es ist nun möglich, auch andere Medieninhalte durch Token zu repräsentieren. Ein Bild kann zum Beispiel in kleine quadratische Pixelbereiche unterteilt und so durch Bildtoken dargestellt werden. Die Sprachmodelle lassen sich in nahezu unveränderter Form auf diese alternativen Tokensequenzen anwenden und sind in der Lage, die Leistung bestehender

Modelle zu verbessern. Besonders beeindruckend ist die Mischung verschiedener Modalitäten mit denen beispielsweise aus einem Text ein Bild erzeugt werden kann. Abbildung 4 verdeutlicht das Vortraining dieser Modelle mit gleicher Architektur auf verschiedenen Medien und die Anwendung auf unterschiedlichste Aufgabenbereiche. Diese Modelle haben für eine extrem große Anzahl von Fragestellungen der Künstlichen Intelligenz den Stand der Kunst verbessert und werden daher als Foundation-Modelle bezeichnet (Bommasani et al., 2021). Sie bilden die Grundlage für eine Vielzahl weiterer KI-Anwendungen. Einen aktuellen Überblick über Foundation-Modelle gibt die Monographie von Paaß et al. (2023).

Die hervorstechenden Merkmale des neuen Paradigmas der Foundation-Modelle sind Emergenz und Homogenisierung:

- **Emergenz** bezeichnet den Umstand, dass ein Foundation-Modell Fähigkeiten aufweist, die nicht explizit konstruiert, sondern implizit gelernt werden. Ein Beispiel ist das GPT-3-Modell, welches eine neue Aufgabe lösen kann, indem man es durch eine natürlichsprachliche Beschreibung der Aufgabe, den so genannten Prompt, instruiert. Das Modell kann also eine Aufgabe lösen, obwohl es nie dafür trainiert wurde.
- **Homogenisierung** rührt daher, dass nahezu alle aktuellen Modelle die Architektur von wenigen Foundation-Modellen (insb. BERT, RoBERTa, T5, GPT-3) implementieren. Dies bewirkt eine Vereinheitlichung von der Modellierung für ein breites Spektrum von Anwendungen. Folglich kann ein Modell durch Anpassungen für viele verschiedene Aufgaben verwendet werden. Ähnliche Ansätze zur Sequenzmodellierung können für Texte, Bilder, gesprochene Sprache, Video, DNA-Sequenzen u.ä. verwendet werden. Dies hat aber auch den möglichen Nachteil, dass diese KI-Systeme die gleichen problematischen Vorurteile oder Fehler einiger weniger Grundmodelle erben können.

2.1 Die Bedeutung der Größe von KI-Foundation-Modellen

Die Größe eines Modells, insbesondere die Anzahl der Parameter, hat einen entscheidenden Einfluss auf die Leistung des Modells, seinen Speicherbedarf und die Rechenressourcen für das Training. Kaplan et al. (2020) untersuchten empirisch die Abhängigkeit zwischen der Anzahl der Modellparameter, dem Umfang der Trainingsdaten und dem Rechenaufwand für das Training. Sie bewerteten eine große Anzahl von Modellen und zogen die folgenden Schlussfolgerungen:

- Die Leistung der Modelle hängt weitgehend von diesen drei Größen ab. Andere Architekturmerkmale wie Breite oder Tiefe haben nur einen schwachen Einfluss.
- Werden Modellgröße und Trainingsdaten in gleichem Maße erhöht, wächst die Modellgenauigkeit zuverlässig über einen großen Bereich des Rechenaufwandes. Wenn einer dieser Faktoren konstant gehalten wird, fällt die Verbesserung geringer aus und nähert sich einer Schranke.

Dies erklärt auch den Erfolg von großen Foundation-Modellen wie T5, GPT-3 oder PaLM. Allerdings erfordert das Training großer Modelle eine extrem leistungsfähige Infrastruktur.

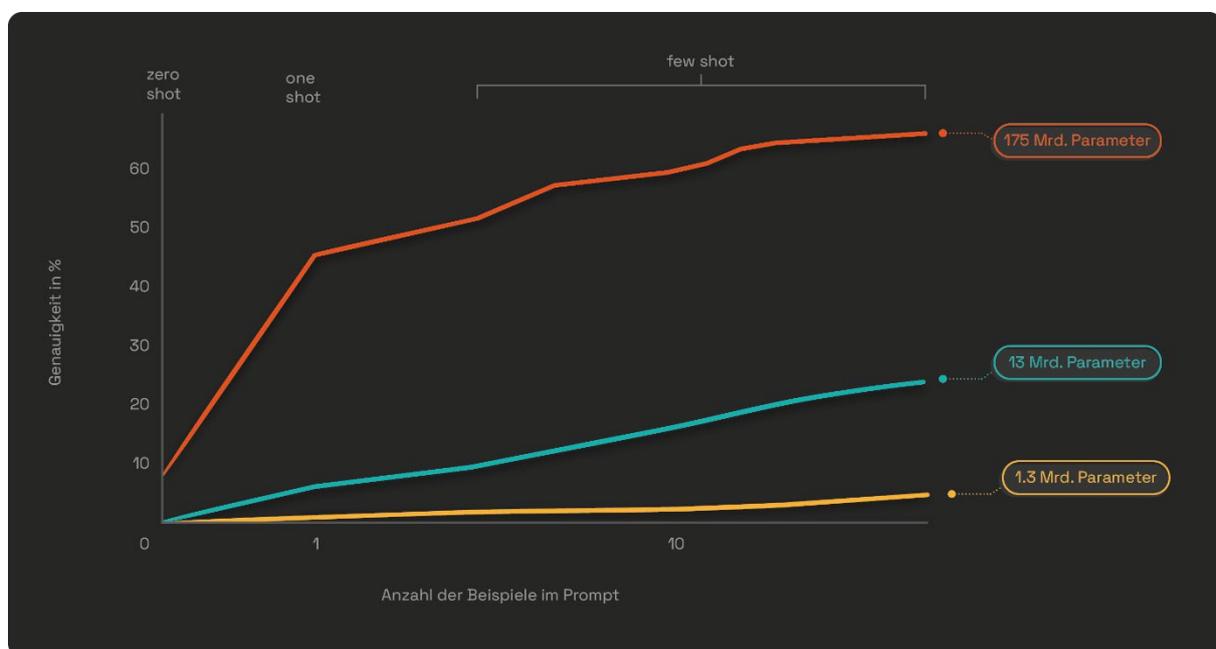


Abb. 5: Die Genauigkeit des „few-shot“-Lernens von GPT-3 wird durch die Erweiterung der Modellgröße und der Anzahl der präsentierten Beispiele erhöht (Brown et al., 2020).

Abbildung 5 zeigt, dass wichtige Eigenschaften von Foundation-Modellen erst für große Modelle beobachtet werden können (Emergenz). Während das GPT-3-Modell mit 13 Mrd. Parametern natürlichsprachige Anweisungen mit einer Genauigkeit von etwa 20 % beantworten kann, steigt dieser Anteil bei 175 Mrd. Parametern auf über 60 % (Brown et al., 2020). Offenbar benötigen die Modelle ein großes Geflecht von Beziehungen zwischen Begriffen, um korrekt auf natürlichsprachige Prompts zu reagieren. Eine mögliche Folge der Emergenz ist, dass es eine Reihe von Aufgaben gibt, die für die derzeitigen Foundation-Modelle unerreichbar sind, die aber bald erfolgreich bewältigt werden könnten.

Eine einfache Möglichkeit, die Anzahl der Parameter ohne höheren Trainingsaufwand zu erhöhen, ist eine Mixture-of-Experts-Architektur. Sie besteht aus einem einzigen Gating-Modul und einer Reihe von Expertenmodulen mit identischer Architektur, aber unterschiedlichen Parametern. Jedes Expertenmodul ist nur auf eine Teilmenge der Daten spezialisiert, und das Gating-Modul ordnet jede Eingabe den wenigen (z.B. 2) Expertenmodulen zu. Diese Zuordnung wird automatisch optimiert, so dass das gesamte Modell die optimale Leistung erbringt. Eine Verringerung des Rechenaufwands kann erreicht werden, da nur wenige Expertenmodule für eine Eingabe tatsächlich verwendet werden. Die Architektur ermöglicht massive Modelle und ist besonders effizient für verteilte Systeme, bei denen die Experten auf verschiedene Recheneinheiten verteilt sind. Beispiele sind Switch, GLaM und WuDao-2.0.

Da die ersten Versionen erfolgreicher Modelle oft extrem groß sind, wurden verschiedene Techniken zur Komprimierung und Beschleunigung der Modelle entwickelt. Wissensdestillation (Hinton et al., 2015) überträgt das Wissen von einem großen Lehrermodell auf ein kleineres Schülermodell. Der Vorteil dieses Ansatzes ist, dass das Schülermodell beim Training die internen Aktivierungen des Lehrermodells nutzen kann. Für eine Reihe von Modellen ergab sich eine deutliche Reduzierung des Speicher- und Rechenaufwands bei nahezu identischer Leistung.

Foundation-Modelle benötigen enorm viele Daten für das Training. Beispielsweise verwendet das PaLM-Modell 780 Milliarden Token (25-mal mehr als alle Texte in Wikipedia), die einen großen Bereich natürlicher Sprache abdecken (Chowdhery et al., 2022). Es ist wichtig, dass die Texte eine hohe syntaktische und inhaltliche Qualität besitzen, da sonst grammatikalische und fachliche Fehler sowie Vorurteile von dem Modell reproduziert werden. Texte mit niedriger Qualität findet man oft in sozialen Medien oder Nutzer:innenkommentaren im Internet. Sie sind nicht zum Training geeignet. Die Kuratierung der Datenqualität ist daher einer der aufwändigsten und kostspieligsten Aspekte des Trainings von Foundation-Modellen. Diese Aspekte werden im Folgenden noch ausführlich diskutiert.

Die Leistung der ersten Foundation-Modelle wurde mit wenigen Benchmarks (z.B. GLUE) überprüft. Mittlerweile sind KI-Foundation-Modelle so leistungsfähig, dass sie Weltwissen aus vielen Bereichen abdecken und sogar Schlussfolgerungen durch die Kombination von Fakten ziehen können (Zhang et al., 2022). Es hat sich herausgestellt, dass viele Benchmarks nach einer gewissen Zeit saturiert sind und die Fortschritte aktueller Modelle nicht mehr erfassen können. Daher werden nun Batterien von mehreren hundert Benchmarks verwendet (z.B. BIG-Bench), die eine Vielzahl von Aufgaben abdecken, darunter logisches Denken, Übersetzung, Beantwortung von Fragen, Mathematik und andere. Mittlerweile übertrifft PaLM mit 5-shot Prompts die Leistung durchschnittlicher menschlicher Bearbeiter (Chowdhery, et al., 2022).

2.2 Anwendungsgebiete von KI-Foundation-Modellen

In diesem Abschnitt stellen wir die vielfältigen Anwendungsmöglichkeiten der Foundation-Modelle vor.

Dabei gehen wir erst auf Anwendungen im Sprachbereich ein, und beschreiben dann Anwendungen im Bereich Multimedia, Biowissenschaften und Robotersteuerung. Die hier aufgeführten Anwendungsgebiete stellen lediglich eine Übersicht dar und sind nicht vollständig. In vielen Bereichen lassen sie die möglichen Anwendungen aktuell noch gar nicht abschätzen.

Die folgenden Abschnitte diskutieren exemplarisch relevante Anwendungsmöglichkeiten aus ganz unterschiedlichen Bereichen.

2.2.1 Anwendungen im Sprachbereich

Anwendung	Beschreibung
Informationsextraktion	extrahiert Konzepte und Namen aus dem Text und gegebenenfalls deren Relationen
Dokumentensuche	findet passende Texte zu einer Anfrage. Dabei werden auch sinnverwandte Formulierungen berücksichtigt.
Beantwortung von Fragen	erzeugt eine Antwort zu einer Frage. Grundlage sind die Informationen aus dem Vortraining und die Ergebnisse einer Dokumentensuche. Die Antwort kann ggf. erklärt werden.
Maschinelle Übersetzung	übersetzt einen Text aus einer Sprache in eine andere. Dies ist mit dem gleichen Modell für mehr als 100 Sprachen möglich.
Erzeugung von Computercode	erzeugt aus einer natürlichsprachlichen Beschreibung lauffähigen Source-Code in ganz unterschiedlichen Programmiersprachen
Zusammenfassung und Vereinfachung von Dokumenten	erfasst die wichtigsten Aussagen eines oder mehrerer Dokumente und verfasst eine vereinfachte Version
Generierung neuer Texte	erzeugt eine inhaltliche kohärente Fortsetzung eines Textes. Dabei können inhaltliche Vorgaben berücksichtigt werden.
Textkorrektur	verbessert und kontrolliert Texte in Bezug auf Rechtschreibung, Grammatik, Stil, Formatierung, Wirksamkeit oder Terminologie
Dialogsysteme	führen ein längeres Gespräch mit einem menschlichen Dialogpartner. Dabei werden Informationen über den Dialogverlauf gespeichert und wiederverwendet.
Meinungs- und Sentimentanalyse	erkennt und klassifiziert Meinungen und emotionale Einstellungen zu Produkten, Personen, Organisationen, Ereignissen usw.
Entdeckung von Fake News und Bot Texten	erkennt Falschaussagen und automatische hergestellte Nachrichten

Tabelle 1: Eine Auswahl möglicher Anwendungen auf Basis von Sprachmodellen

[Sprachmodell GPT-3]

Dem privatwirtschaftlichen Forschungsunternehmen OpenAI gelang mit dem Sprachmodell GPT-3 ein wissenschaftlicher Durchbruch: Das Modell ist in der Lage, das nächste Wort in einem Satz vorherzusagen. Schnell bildete sich eine Community aus Entwickler:innen und Nutzer:innen um das Modell und Applikationen wurden entwickelt, die zeigen, dass GPT-3 überzeugende Aufsätze schreiben, Diagramme und Websites aus Textbeschreibungen erstellen, Computercode generieren und vieles mehr kann (Tamkin & Ganguli, 2021).

Seit der Veröffentlichung von GPT-3 im Juni 2020 werden immer mehr konkurrierende, teilweise sogar als Open Source Lösung angebotene Modelle publiziert. Dabei bietet OpenAI GPT-3 als API an und ermöglicht Nutzer:innen so, das Modell über die OpenAI-Plattform zu nutzen oder GPT-3 in eigenen Anwendungen einzubauen. Daraus hat sich ein wachsender Markt an Tools und Anwendungen entwickelt, der heute viele Industrien und Geschäftsbereiche als auch Kreative beeinflusst.

Das KI-Foundation-Modell macht es möglich, Text in Sekunden zusammenzufassen oder zu ergänzen. Mittels Chatbots können Kundeninteraktionen einfacher gesteuert werden. Weiterführend ermöglicht GPT-3 die Generierung von Programmcode durch die Eingabe von Sprachbefehlen. Realisiert werden Anwendungen dieser Art durch Start-ups, welche die API von OpenAI gegen ein Entgelt nutzen. Gerade im Bereich Gesundheitswesen, eCommerce sowie im Medien- und Communications-Bereich profitieren Unternehmen, die Chatbots sowie Anwendungen für Natural Language Understanding (NLU) entwickeln, massiv von der Weiterentwicklung großer Sprachmodelle.

INFOBOX

Informationsextraktion

Informationsextraktion ist die Aufgabe, automatisch für die Nutzer:innen relevante strukturierte Informationen aus unstrukturierten und/oder halbstrukturierten maschinenlesbaren Dokumenten zu extrahieren. Diese Funktionen sind von großer Bedeutung für die automatische Erfassung von Textdokumenten und Verwaltungsvorgängen. Hierbei sind insbesondere die Klassifikation von Dokumenten nach inhaltlichen Kriterien, die Eigennamenerkennung und die Relationsextraktion zu nennen, wobei letztere Relationen zwischen Eigennamen und Begriffen aus einem Text extrahiert und in einer Datenbank speichert. Foundation-Modelle haben bei diesen Aufgaben die Genauigkeit stark verbessert und übertreffen oft die Performanz von Menschen.

Dokumentensuche

Eine extrem wichtige Anwendung ist die Dokumentensuche. Dabei werden sowohl die Dokumente einer Textsammlung als auch die Anfrage durch eine Einbettung kodiert und nach Berechnung die zur Anfrage ähnlichsten Dokumente zurückgegeben. Vorteil ist, dass auch Synonyme und alternative Formulierungen des gleichen Sachverhaltes berücksichtigt werden. Diese einbettungsbasierten Retrievalverfahren übertreffen die klassische Stichwortsuche und werden mittlerweile bei allen Internetsuchmaschinen genutzt.

Frage: when are hops added to the brewing process?

Einer der gefundenen Texte: After mashing , the beer wort is boiled with hops (and other flavourings if used) in a large tank known as a " copper " or brew kettle – though historically the mash vessel was used and is still in some small breweries . The **boiling process** is where chemical reactions take place , including sterilization of the wort to remove unwanted bacteria , releasing of hop flavours , bitterness and aroma compounds through isomerization , stopping of enzymatic processes , precipitation of proteins , and concentration of the wort . Finally , the vapours produced during the boil volatilise off - flavours , including dimethyl sulfide precursors . The boil is conducted so that it is even and intense – a continuous " rolling boil " . The boil on average lasts between 45 and 90 minutes , depending on its intensity , the hop addition schedule , and volume of water the brewer expects to evaporate . At the end of the boil , solid particles in the hopped wort are separated out , usually in a vessel called a " whirlpool " .

Antwort: The boiling process



Von einem Foundation-Modell mit Hilfe von Retrieval gefundene Antwort auf eine Frage im Natural Questions Benchmark. Aktuelle Modelle erreichen eine Genauigkeit (F1) von 80% (Zhang et al. 2021).

Abb. 6: Von einem Foundation-Modell mit Hilfe von Retrieval gefundene Antwort auf eine Frage im Natural Question Benchmark. Aktuelle Modelle erreichen eine Genauigkeit (F1) von 80% (Zhanag et al., 2021).

Beantwortung von Fragen

Bei der Beantwortung von Fragen (Question Answering, QA) erhält ein System eine natürlichsprachliche Anfrage und generiert automatisch eine Antwort in natürlicher Sprache. Fortschrittliche Systeme arbeiten in der Regel in zwei Stufen (Abb. 6): Für eine Frage findet ein einbettungs-basiertes Retriever-Modul eine Reihe von passenden Dokumenten aus einer Textsammlung. Hierbei werden auch Dokumente mit ähnlichen Inhalten gefunden, die unterschiedlich ausgedrückt wurden. Anschließend prozessiert ein Reader die Frage und die gefundenen Dokumente und generiert eine natürlichsprachliche Antwort. Retriever-Reader Module werden von vielen fortgeschrittenen Sprach- und Dialogmodellen eingesetzt und produzieren wesentlich bessere Antworten als sehr große Sprachmodelle ohne diese Erweiterung. Die Antwort kann durch die gefundenen Dokumente erklärt und begründet werden.

Maschinelle Übersetzung

Zur maschinellen Übersetzung gibt es mittlerweile Modelle, z.B. M2M von Facebook AI (Fan et al., 2022), die die Übersetzung zwischen beliebigen Paaren von über 100 Sprachen gestatten. Durch das gleichzeitige Training des Encoder-Decoders mit vielen Sprachen wird die Übersetzungsqualität für fast alle Sprachpaare wesentlich verbessert, insbesondere für regionale Sprachen mit wenigen Trainingsdaten.

Erzeugung von Computercode

Die Erzeugung von Computercode aus einer textuellen Beschreibung ist eine spezielle Übersetzungsaufgabe (M. Chen et al., 2021), die mittlerweile relativ zuverlässig laufenden Code erzeugen kann und ein großes Anwendungspotential besitzt.

Dokumentenzusammenfassung

Die automatische Zusammenfassung von Dokumenten kann helfen, die wichtigsten Informationen in Dokumenten zu erfassen. Modelle zur Zusammenfassung nutzen meist ein Seq2seq-Modell, welche als Eingabe ein Dokument erhalten und die Zusammenfassung ausgeben. Dabei werden insbesondere Modelle mit einer langen Eingabesequenz benötigt. Die Qualität der Zusammenfassung ließ sich mit Foundation-Sprachmodellen sehr stark erhöhen.

Referenz-Zusammenfassung: alejandro valverde won ahead of julian alaphilippe and michael albasini. chris froome finished 123rd after a crash during the final 12 kilometres. team sky's sports director gabriel rasch praised froome for finishing. rasch said froome was 'banged up' but expects to ride tour de romandie.

Zusammenfassung durch BRIO: alejandro valverde defended his fleche wallonne title in belgium on wednesday. movistar rider finished ahead of julian alaphilippe and michael albasini. team sky's chris froome fell in the final 12km of the race but finished in 123rd. froome was involved in a crash but finished the race despite being 'banged up'

Zusammenfassung eines Textes von 800 Wörtern durch das Modell BRIO im Vergleich zu der von Experten erstellten Zusammenfassung. (Liu et al., 2022)

Abb. 7: Zusammenfassung eines Textes von 800 Wörtern durch das Modell BRIO im Vergleich zu der von Experten erstellten Zusammenfassung (Liu et al., 2022).

Textgenerierung

Die Generierung neuer Texte ist die zentrale Aufgabe von Sprachmodellen. Hierbei wird ein vorgegebener Starttext syntaktisch fehlerfrei und inhaltlich stimmig fortgesetzt. GPT-3 ist hier das bekannteste Modell. Durch einen Starttext mit Anweisungen (Prompt) kann man GPT-3 beauftragen, eine Reihe von Punkten in den erzeugten Text aufzunehmen. Diese Anweisungen können auch Lösungsbeispiele enthalten, wodurch das Modell bei der Konstruktion der Ausgabe unterstützt wird (k-shot prompts). Da die Worte des neuen Textes entsprechend ihrer Wahrscheinlichkeit generiert werden, entsteht bei einer Wiederholung immer ein anderer Text.

Mittlerweile gibt es auch interaktive Verfahren zur Generierung von Texten (A. Yuan et al., 2022), in denen die Nutzer:innen die Ausgestaltung des Textes steuern können. PaLM ist ein mehrsprachiges, fortgeschrittenes Sprachmodell mit 540 Milliarden Parametern, welches auf mehr als 150 Benchmarks die Leistung durchschnittlicher menschlicher Bearbeiter:innen übertraf. Zudem konnte PaLM nach entsprechender Anleitung komplexe Aufgaben in einfache Schritte zerlegen und Erklärungen für seine Schlußfolgerungen liefern (Chowdhery, et al., 2022). Starttexte können Sprachmodelle im Prinzip dazu bringen, beleidigende Äußerungen und Fake News zu produzieren. Durch nachträgliche Filtertechniken, Finetuning und Retrieval können allerdings verletzende Äußerungen und Falschinformationen weitgehend vermieden werden.

Dialogsysteme

Dialogsysteme (Chatbots) generieren automatisch adäquate Antworten auf die Äußerungen menschlicher Gesprächspartner:innen im Laufe eines längeren Gesprächs. Sie kombinieren Techniken zur Beantwortung von Fragen mit der Generierung von Geschichten und dem Retrieval zusätzlicher Informationen. Dabei wird auch der Stand der Diskussion und Informationen über die „Persönlichkeit“ des Chatbots in einer Datenbank gespeichert und durch Retrieval weiterverwendet. Eine Bewertung durch menschliche Prüfer:innen zeigt, dass z.B. das LaMDA System (Thoppilan et al., 2022) in Bezug auf Sensibilität, Sicherheit und Faktentreue der menschlichen Leistung nahekommt. Im Bereich des Kundenkontakts gibt es sehr viele Anwendungsfälle für Chatbots. Das Modell ChatGPT (OpenAI, 2022b) hat eine ähnliche Architektur, ist aber frei im Internet nutzbar. Es kann Fragen beantworten, Code generieren, Texte zusammenfassen, aber auch einen kohärenten Dialog führen.

[ChatGPT]

ChatGPT ist ein am 30. November 2022 innerhalb einer Open-Beta veröffentlichtes Sprachmodell, welches in der Lage ist, sinnvolle Konversationen mittels Text zu führen. Trainiert mit großen Datensätzen von Konversationen, ist es in der Lage, realistische Gespräche zu führen, Sachverhalte zu erklären oder Programmiercode zu erstellen. Die wichtigste Eigenschaft von ChatGPT ist dabei, dass es in der Lage ist, kontextbezogen zu antworten. Dabei können weiterführende Fragen innerhalb einer Konversation gedeutet und interpretiert werden.

ChatGPT hat eine Vielzahl potenzieller Anwendungen, da es in der Lage ist, angemessene Antworten auf eine breite Palette von Aufforderungen zu generieren und Gespräche in verschiedenen Kontexten zu führen. Einige Beispiele für diese Anwendungen sind die Entwicklung von Chatbots für den Kundendienst oder die Bereitstellung von Informationen, die Entwicklung virtueller Assistenten, die den Nutzern bei Aufgaben und der Beantwortung von Fragen helfen, und die Entwicklung interaktiver Tutorensysteme. Dies sind nur einige der möglichen Einsatzgebiete von ChatGPT zur Entwicklung intelligenter und interaktiver Systeme, die mit Menschen auf natürliche und intuitive Weise kommunizieren können.

Da mittels konversationeller KI-Modelle Fragestellungen direkt beantwortet werden können, stellt ein Modell wie ChatGPT sowie eventuelle zukünftige Entwicklungen dieser Art eine ernstzunehmende Herausforderung für konventionelle Suchanbieter wie Google dar. Mittels konversationeller KI-Modelle lassen sich Fragestellungen von Nutzern direkt beantworten sowie weiter spezifizieren, statt nur auf weiterführende Webseiten zu verweisen.

Obwohl sich ChatGPT noch in der Beta-Version befindet, existieren bereits Suchmaschinen auf Basis des KI-Modells. Damit wird dem Nutzer ein viel spezifischeres Suchergebnis geboten. Gerade miteinander verknüpfte Inhalte lassen sich deutlich leichter finden und in Zusammenhang bringen. Auch als unternehmensinternes Tool, im Sinne einer "Single Source of Truth" sind KI-Modelle wie ChatGPT denkbar, da sie dem/der Nutzer:in ein sehr authentisches und simples Suchergebnis bieten - ähnlich einem Gespräch mit einem/einer Mitarbeitenden.

INFOBOX

2.2.2 Anwendungen von multimodalen Foundation-Modellen

KI-Foundation-Modelle sind in der Lage, nicht nur Token der natürlichen Sprache zu modellieren, sondern auch Token-Elemente beliebiger anderer Sequenzen. Im Folgenden werden Anwendungen des Paradigmas auf unterschiedlichen Datendomänen wie gesprochene Sprache, Bilder, Videos, DNA und Proteine vorgestellt, die teilweise völlig neue Problemlösungen ermöglichen.

Dabei erschließt insbesondere die Verknüpfung von Daten aus zwei oder mehr Domänen zusätzliche anspruchsvolle Anwendungsfelder. Hierbei gibt es Modelle, die nur auf eine Anwendung spezialisiert sind, aber auch Foundation-Modelle wie NÜWA, OFA und GATO, die viele Aufgaben gleichzeitig lösen können. Da bei dieser Art der Modelle verschiedene Datendomänen kombiniert werden, bezeichnet man sie als multimodale Modelle.

Anwendung	Beschreibung
Sprache zu Text	übersetzt gesprochene Sprache in Text für viele unterschiedliche Sprecher:innen
Text zu Sprache	transformiert Text in gesprochene Sprache unter Berücksichtigung von Sprachmelodie und Sprecher:innenstimme
Erkennung von Objekten	identifiziert die wichtigsten Objekte eines Bildes und deren Positionen
Erkennung von Aktionen	identifiziert Aktionen in Videos und deren Positionen
Bild nach Text	generiert zu einem Bild eine Bildbeschreibung
Text in Bild	erzeugt zu einer Bildbeschreibung ein passendes Bild
Video nach Text	beschreibt die Objekte und Abläufe in einem Video durch Text
Text nach Video	erzeugt zu einer Bildbeschreibung ein kompatibles Video
Genomik	DNA-Sequenzen werden analysiert und die daraus erzeugten Proteine prognostiziert
Proteomik	Die 2D- und 3D-Struktur der Proteine wird vorhergesagt
Verstärkungslernen	Sequenzen von Aktionen und Zuständen für Steuerungsprobleme werden prognostiziert
Livesynchronisation	übersetzen der Sprache eines Videos

Tabelle 2: Eine Auswahl möglicher Anwendungen auf Basis von multimodalen Modellen

Die Anwendungsbereiche multimedialer Foundation-Modelle stellen wir in diesem Kapitel genauer vor.

Gesprochene Sprache

Ein Anwendungsbereich ist gesprochene Sprache. Das Audiosignal wird dabei oft durch sein Frequenzspektrum (MFCC) für jedes 10 msec Zeitintervall repräsentiert. Wav2vec 2.0 (Baevski et al., 2020) führt unüberwachtes Lernen auf Sprachdaten ohne Transkription mit Convolution- und Self-Attention-Schichten durch. Derartige Convolution-Schichten wurden ursprünglich für die Bilderkennung mit Convolutional-Neural-Networks (CNNs) entwickelt. Ähnlich wie das BERT-Modell für Text lernt es, maskierte „Sound-Tokens“ vorherzusagen. Die Transkription von Sprache in Text kann dann durch eine Kombination von Convolution- und Self-Attention-Schichten erfolgen (Zhang et al., 2020). Zur Transkription von Text in gesprochene Sprache erzeugt beispielsweise FastSpeech 2 (Ren et al., 2022) aus den eingegebenen Phonemen mit einem Seq2seq-Modell das Frequenzspektrum für kleine Zeitintervalle, aus dem direkt die Sprachausgabe produziert werden kann. Dabei werden verschiedene Informationen wie Dauer, Tonhöhe und Energie berücksichtigt. FastSpeech 2 wird von menschlichen Juroren besser beurteilt als konkurrierende Systeme.

Bilder

Bilder können in eine Sequenz von Pixelbereichen zerlegt werden, die als Bild-Token genutzt werden können. Der Vision-Transformer (Dosovitskiy et al., 2020) verwendet Pixelbereiche der Größe 14x14 als Token und führt ein Vortraining mit einem sehr großen Datenbestand von 300 Mio. nicht-annotierten Bildern durch. Dabei sind maskierte Bild-Token zu prognostizieren. Anschließend erfolgt ein Finetuning des Modells auf den ImageNet-Daten zur Klassifikation von Bildern in 1000 Klassen. Der Vision-Transformer erzielte eine höhere Genauigkeit als alle CNNs bei gleichzeitig wesentlich geringerem Trainingsaufwand. Eines der ersten Modelle zur Kombination von Bildern und deren textueller Beschreibung ist das CLIP-Modell (Radford et al., 2021). Mit separaten Encodern erzeugt es aus einem Bild und dem zugehörigen Text je eine Einbettung. Die Differenz zwischen den beiden Einbettungen wird dann durch Training minimiert. Damit lässt sich zu einer Bildunterschrift das am besten passende Bild finden und umgekehrt.

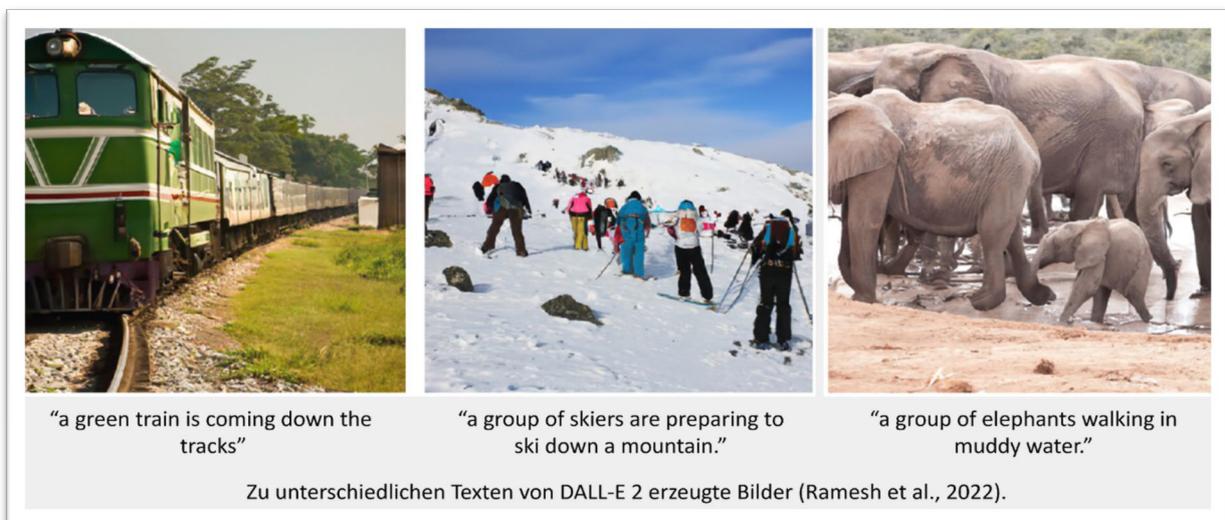


Abb. 8: Zu unterschiedlichen Texten von DALL-E 2 erzeugte Bilder (Ramesh et al., 2022)

Die Erzeugung von Bildern aus Text verwendet oft CLIP, um zu einem Text passende Bild-Einbettung zu finden. Ein Diffusionsmodell kann den Prozess der Degradierung eines Bildes durch Zufallsänderungen modellieren. Hierbei wird sukzessive die Farbe einzelner Pixel zufällig geändert, so dass sich mit der Zeit eine graue Fläche ergibt. Dieser Prozess kann umgekehrt und zur Rekonstruktion von Bildern in hoher Auflösung aus den Einbettungen genutzt werden. Bekanntestes Modell ist DALL-E 2 (Ramesh et al., 2022), welches zusätzlich noch Bereiche eines Bildes modifizieren kann. Das OFA-Modell (Wang et al., 2022) kann gleichzeitig viele Aufgabe erledigen: die Position von Objekten in Bildern bestimmen, Bildunterschriften erzeugen, Fragen zu einem Bild beantworten, Objekte in einem Bild erkennen, fehlende Bereiche in einem Bild ausfüllen und Bilder zu einem Text erzeugen. Darunter lassen sich viele Anwendungen heute schon produktiv einsetzen, beispielsweise in der Bearbeitung von Fotos. Mittels des Modells Stable Diffusion (Rombach et al., 2022) ist es etwa möglich, unerwünschte Objekte aus Bildern zu entfernen, mehrere Bilder zu verbinden oder auch ausgehend von einer textuellen Bildbeschreibung ganz neue Bilder zu Werbezwecken bzw. als Ersatz für sogenannte „Stock Photos“ zu generieren. Bei den so erzeugten Bildern handelt es sich um Unikate. Zwar spiegeln sie den Datensatz wider, mit dem das Modell trainiert wurde, sind aber keine Kopien. Dadurch, dass es sich bei den generierten Bildern nicht um Kopien handelt, können diese potentiell kommerziell genutzt werden. Allerdings sind die Copyright-Fragen noch nicht abschließend geklärt.

Video

Um Videos mit Foundation-Modellen verarbeiten zu können, verwendet man meist Video-Token, die einen Pixelbereich in mehreren hintereinander folgenden Videobildern beschreiben. Flamingo (Alayrac et al., 2022) ist ein visuelles Sprachmodell, das Sequenzen von beliebig hintereinander folgenden Bildern, Videos und Texten verarbeiten kann. Es nutzt im Hintergrund ein großes Sprachmodell für Text. Das Modell kann einerseits Fragen zu Bildern beantworten oder Bilder beschreiben. Zum anderen kann es Aktionen in Videos beschreiben oder klassifizieren. Schließlich kann es durch gemischte Few-shot-Prompts aus Texten und Bildern/Videos instruiert werden, eine neue Beschreibungsaufgabe zu erledigen.



Abb. 9: Zu unterschiedlichen Texten von CogVideo erzeugte Videos (Hong et al., 2022)

NÜWA (Wu et al., 2021) ist ein Encoder-Decoder-Modell, das ein Video zu einem Text erzeugen kann. Es verwendet einen speziellen Attention-Mechanismus, um die Relation der Token sowohl für räumliche als auch für zeitliche Achsen zu erfassen. Das Modell kann einerseits eine Reihe von Bild-Aufgaben lösen, etwa Erzeugung eines Bildes zu einem gegebenen Text. Weiter kann es ein Video zu einem Text generieren, die Fortsetzung eines Videos prognostizieren oder Videos manipulieren.

Imagen-Video (Ho et al., 2022) wurde mit 60 Millionen Bild-Video-Paaren und 14 Millionen Text-Video-Paaren jeweils aus dem öffentlich zugänglichen LAION 400M-Datensatz (Schuhmann, 2021) trainiert. Mit Hilfe der Bilder können auch bestimmte Kunststile imitiert werden und z.B. ein Video im Stil von Monet erzeugt werden. Das Modell Make-a-Video (Singer et al., 2021) bietet die zusätzliche Möglichkeit zum unüberwachten Lernen auf Videodaten ohne textuelle Beschreibung, um realistische Bewegungen von Objekten und Szenen zu lernen. Zudem kann es zwischen einem Paar von Bildern einen dynamischen Übergang in Form eines Videos erzeugen. Beide Modelle sind nicht nur in der Lage, Videos mit hoher Wiedergabetreue zu generieren, sondern die Modelle umfassen ein hohes Maß an Weltwissen und sind im Detail kontrollierbar mit der Fähigkeit, verschiedene Videos und Textanimationen in verschiedenen künstlerischen Stilen und mit 3D-Objekt-verständnis zu erzeugen. Leider sind Make-a-Video und Imagen Video proprietär und der Modell-Code ist nicht frei verfügbar. Die beschränkte Länge von hochauflösenden Videos von derzeit fünf Sekunden ist offenbar der limitierende Faktor. Insgesamt sind die erzeugten Videos noch nicht perfekt.

Genomik und Proteomik

Die Entschlüsselung der Sprache der DNA ist eines der wichtigsten Ziele der biologischen Forschung. Der genetische Code ist universell und erklärt, wie die DNA in Proteine übersetzt wird. Im Gegensatz dazu variiert der regulatorische Code, der bestimmt, wann und wie die Gene exprimiert werden, zwischen verschiedenen Zelltypen und Organismen. Dies ist ähnlich zur Polysemie und entfernten semantischen Beziehungen bei natürlichsprachigen Texten. DNABERT wurde auf einer großen Menge von DNA-Sequenzen vortrainiert und kann durch Finetuning den Stand der Technik für viele spezifische Prognoseaufgaben verbessern. Darunter sind die Analyse von Sequenzmotiven (DNA-Abschnitten mit biologischer Relevanz) und die Prognose der Promotor-Regionen (Nukleotid-Sequenz, die die regulierte Expression eines Gens ermöglicht). MoDNA (An et al., 2022) und GeneBERT (Mo et al., 2021) haben eine ähnliche Funktionalität.

Proteine sind lineare Ketten von Aminosäuren, die durch kovalente Bindungen verbunden sind. Aminosäuren lassen sich durch ein Alphabet mit 25 Zeichen repräsentieren. Die Zeichenketten eignen sich hervorragend für viele Natural-Language-Processing (NLP) Methoden (Ofer et al., 2021). AminoBERT ist ein Sprachmodell (Chowdhury et al., 2022), welches aus einer Proteinsequenz als Eingabe die 3D-Proteinstruktur prognostiziert. Dabei wird auch eine natürliche Methode zur Beschreibung der Polypeptidgeometrie verwendet, die auf der Ebene des Polypeptids als Ganzes rotations- und translationsinvariant ist. Im Durchschnitt übertrifft das Modell AlphaFold2 (Jumper et al., 2021) und RoseTTAFold (Baek et al., 2021) bei verwaisten Proteinen und Klassen von konstruierten Proteinen und erreicht dabei eine bis zu 10^6 -fache Reduzierung der Rechenzeit. Es gibt eine Reihe weiterer Modelle mit ähnlichen Ergebnissen, z.B. das

Protein-Sprachmodell ESMFold (Lin et al., 2022). Es erzeugt Einbettungen, die in nachgelagerten Aufgaben eingesetzt werden können, zum Beispiel zur Erfassung der strukturellen Eigenschaften von Proteinen.

[AlphaFold]

Mit der Software AlphaFold gelang dem privaten Forschungsunternehmen DeepMind 2021 ein großer Durchbruch in der Forschung an der Faltung von Proteinen. Mithilfe eines KI-Modells erreichte das Team, die Faltung von Proteinen vorherzusagen und dadurch deren Form und Funktion innerhalb des Organismus genauer zu erforschen. Über 200 Millionen Strukturen von Proteinen konnten so gefunden und in einer Protein-Datenbank gespeichert werden - im Gegensatz zu den vorher nur circa 1 Millionen verfügbaren Proteinstrukturen.

AlphaFold hat bereits heute einen bedeutenden, direkten Einfluss auf die menschliche Gesundheit. Bei einem Treffen mit Forscher:innen der European Society of Human Genetics wurde deutlich, wie wichtig die AlphaFold-Strukturen für Biolog:innen und Kliniker:innen sind, die versuchen, die Ursachen seltener genetischer Krankheiten zu entschlüsseln. Darüber hinaus beschleunigt AlphaFold die Entdeckung von Medikamenten, indem es ein besseres Verständnis neu identifizierter Proteine ermöglicht, die als Angriffspunkte für Medikamente in Frage kommen. Es hilft Wissenschaftler:innen, schneller potenzielle Medikamente zu finden, die sich an diese Proteine binden.

Ende Infobox

Für das Wirkstoffdesign ist die Vorhersage der Interaktion zwischen einem Arzneimittel und dem Zielorgan wichtig. Sie ist für die Entdeckung neuer Medikamente und die Umwidmung bestehender Medikamente von entscheidender Bedeutung. Yazdani-Jahromi et al. (2021) beschreiben ein Sprachmodell für derartige Anwendungen.

INFOBOX

Steuerung dynamischer Systeme

Verstärkungslernen (auch als Reinforcement Learning bekannt) betrachtet ein System mit Zuständen, Aktionen und Belohnungen zu bestimmten Zeitpunkten. Der Agent wählt zu den gegebenen Zuständen eine neue Aktion, während die nächsten Zustände und die Belohnung durch die Umgebung bestimmt wird. Ziel ist, eine Strategie zu erlernen, die jedem Zustand eine Aktion zuordnet und die Summe der Belohnungen maximiert. Mit derartigen Systemen lassen sich Brett- und Videospiele beschreiben, aber auch Robotersteuerungen und selbstfahrende Autos. Der Decision-Transformer (L. Chen et al., 2021) ist ein Sprachmodell, welches die Aktionen nacheinander prognostiziert. Dabei schätzt er die optimale Summe aller zukünftigen Belohnungen. Das Modell wird auf einer großen Menge von beobachteten Zeitreihen trainiert. Anschließend kann der Agent zu einem gegebenen Zustand die Aktion auswählen, welche zur höchsten prognostizierten Belohnungssumme führt. GATO (Reed et al., 2022) ist ein multimodales Modell, welches Text, Bilder und Sequenzen von Werten verarbeiten und daraus Steuerungsstrategien ableiten kann. Es erzielte auf mehr als 600 Benchmarks gute Resultate.

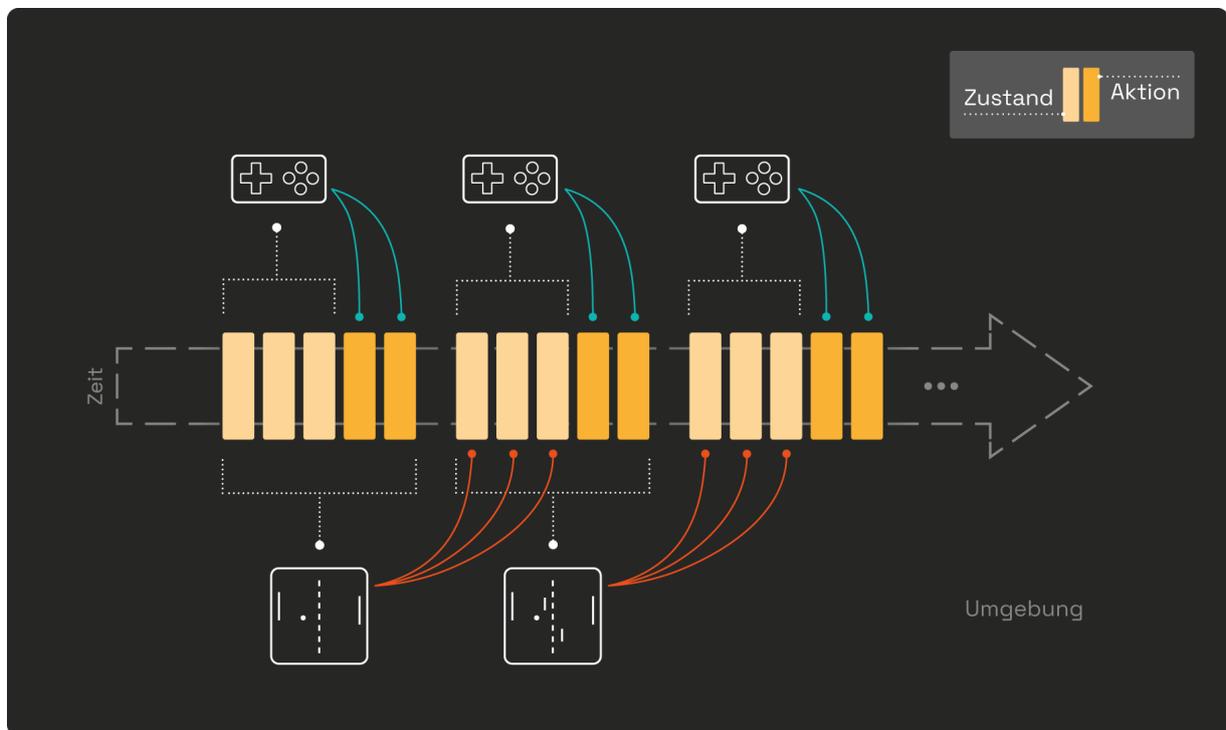


Abb. 10: Das Gato-Modell generiert aus den aktuellen Zuständen (hellgelb) neue Aktionen (dunkelgelb). Die Umgebung produziert daraus neue Zustände, usw. Das Modell kann Texte, Messwerte, Bilder, usw. verarbeiten (Reed et al. 2022).

2.3 Vertrauenswürdige KI-Foundation-Modelle

Der großflächige Einsatz von KI-Systemen und das damit verbundene technische Innovationspotential wird erwartungsgemäß Auswirkungen von breiter wirtschaftlicher, aber auch gesellschaftlicher Relevanz haben. Bei ChatGPT haben sich innerhalb von fünf Tagen über eine Millionen Nutzer:innen angemeldet, um das System selbst auszuprobieren. Dieses große Interesse der breiten Öffentlichkeit zeigt, wie schnell und stark große KI-Systeme die öffentliche Diskussion prägen können. Aus dem Black-Box ähnlichen Charakter der eingesetzten trainierten Modelle ergeben sich Risiken, die besondere Maßnahmen im Hinblick auf den vertrauenswürdigen Einsatz von KI notwendig machen. Die vielfältigen Anwendungsmöglichkeiten von KI-Foundation-Modellen erfordern einen systematischen Ansatz zur Bewertung bzw. Abschwächung der entstehenden Risiken. Sie müssen parallel zur Modellkonstruktion im LEAM-Projekt angegangen werden.

Die folgende Darstellung orientiert sich an dem im Kontext des KI.NRW-Flagship-Projekts „Zertifizierte KI“ (Poretschkin, 2022) erarbeiteten risikobasierten Ansatz zur vertrauenswürdigen KI und den vorhergehenden Arbeiten (Cremers et al., 2019; Poretschkin et al., 2021). Dieser Ansatz zielt darauf ab, KI-spezifische Risiken systematisch zu erfassen, messbar zu machen und unter Berücksichtigung von verschiedenen Dimensionen der Vertrauenswürdigkeit hinreichend zu reduzieren.

Wir stellen im Folgenden sechs Dimensionen der Vertrauenswürdigkeit dar und skizzieren aktuelle Maßnahmen zu deren Identifikation und Mitigation.

Fairness

Die Dimension Fairness soll sicherstellen, dass die KI-Anwendung nicht zu ungerechtfertigter Diskriminierung führt. Typische Ursachen hierfür sind unausgewogene (mit Bias behaftete) Trainingsdaten oder auch die statistische Unterrepräsentation von Personengruppen, welche zu einer verringerten Qualität der KI-Anwendung in Bezug auf diese Gruppen führen kann. Da die Foundation-Modelle mit Texten trainiert werden, die von Menschen verfasst wurden, spiegeln diese Texte häufig die in der Gesellschaft vorhandenen Stereotype wider. Hierbei kann es insbesondere zu einer Bevorzugung oder Benachteiligung von geschlechtsspezifischen oder ethnischen Gruppen kommen.

Die Entwickler:innen großer KI-Foundation-Modelle sind sich vieler der immanenten Risiken bewusst und implementieren korrektive Maßnahmen, die darauf ausgerichtet sind, unerwünschte Effekte zu vermeiden. So berichten die Entwickler:innen von DALL-E (OpenAI, 2022a), dass spezielle Techniken eingesetzt werden, um einem in genutzten Trainingsdaten tatsächlich vorhandenen Bias (z.B. „heroic firefighter“ wird zumeist als männliche Person dargestellt) entgegenzuwirken. Große Sprachmodelle wie PaLM und LaMDA verwenden mit gutem Erfolg filter-basierte Techniken, um unerwünschte Ausgaben („toxic language“) zu vermeiden. Diese Methoden sind aber sehr spezifischer Natur und mit großer Wahrscheinlichkeit im breiten Einsatz nicht ausreichend.

Autonomie und Kontrolle

Diese Dimension zielt auf zwei Dinge ab: zum einen die Autonomie der KI-Anwendung und zum anderen die Autonomie des Menschen. Einerseits ist hier zu beurteilen, welcher Grad an Autonomie (Gehman et al., 2020) für die Anwendung angemessen ist. Andererseits wird untersucht, ob der Mensch durch die KI-Anwendung angemessen unterstützt wird und ausreichend Handlungsspielraum in der Interaktion mit der KI-Anwendung erhält. Die spezifische Herausforderung liegt darin, dass die Möglichkeiten der Interaktion mit dem Menschen meist erst im Design der konkreten nachgelagerten KI-Anwendung festgelegt werden und nicht durch das Foundation-Modell an sich kontrolliert werden kann, das in dem Prozess aber eine zentrale Rolle spielt. Hier ist noch substantielle Forschungsarbeit notwendig, um die Querbeziehungen zwischen den Anwendungen und den Foundation-Modellen kontrolliert abdecken zu können.

Ähnlich zur zwischenmenschlichen Kommunikation, können „toxische“ Modellausgaben (Gehman et al., 2020) (etwa Beleidigungen oder Mobbing) zu psychologisch-emotionalen Beeinträchtigungen der Nutzer:innen führen. Jenseits von verletzender Sprache können (manipulative) Kommunikationsstrategien oder die Vorspiegelung falscher Tatsachen emotionale Abhängigkeiten schaffen und damit die menschliche Autonomie potentiell einschränken. Einen frühen Ansatz, diesem Risiko zu begegnen, schlagen Glaese et al. (2022) vor. Sie optimieren ihre Conversational AI dergestalt, dass jene ihre maschinelle Natur, wenn immer nötig, offen kommuniziert. Auch hier stehen wir erst ganz am Anfang und die deutsche KI-Forschung könnte wesentliche Beiträge liefern.

Transparenz

Unter diesem Oberbegriff sind Aspekte der Nachvollziehbarkeit, Reproduzierbarkeit und Erklärbarkeit subsumiert. Die Dimension Transparenz untersucht insbesondere, ob die grundlegende Funktionsweise der KI-Anwendung für Nutzer:innen und Experten:innen angemessen nachvollziehbar ist und ob Ergebnisse der KI-Anwendung reproduziert und ggf. begründet werden können. Die Transparenz-Dimension wird in Foundation-Modellen zwar bereits auf der Ebene der Dokumentation und Beschreibung der Daten/Modelle (z.B. durch Modelcards (Gehman et al., 2020) adressiert, bedarf aber noch einer systematischen Herangehensweise in Bezug auf den tatsächlichen Einsatz der KI-Foundation-Modelle in konkreten Anwendungen.

Wie andere soziale Medien, kann ein Chatbot durch Fine-tuning oder Prompts dazu gebracht werden, dem/der Nutzer:in nur bestimmte Aspekte zu kommunizieren. Sie befinden sich dann in einer „Filterblase“, in der Nachrichten, die nicht der geäußerten Meinung entsprechen, ausgeblendet werden. Für diese Problematik gibt es mittlerweile Audit-Verfahren (Cen & Shah, 2021), mit denen überprüft werden kann, ob die Plattform unerwünschte inhaltliche Filter verwendet, wobei nur ein Black-Box-Zugriff auf den Filteralgorithmus erforderlich ist. In allen diesen Bereichen haben wir bisher wenig Kontrolle speziell bei Foundation-Modellen und es sind weitere Anstrengungen im Bereich der Forschung und dann wahrscheinlich auch der Regulierung erforderlich.

Verlässlichkeit

Diese Dimension bezieht sich vornehmlich auf die Qualität der KI-Komponente und beurteilt u.a. deren Robustheit, das heißt die Konsistenz ihrer Ausgaben. Ein erschwerender Umstand dabei ist, dass die Erzeugung von Text meist zufallsgesteuert ist. Ein wichtiger Aspekt ist auch die Faktentreue der Ausgaben, da Foundation-Modelle nicht zwischen plausiblen und korrekten Zusammenhängen unterscheiden können (Lin et al., 2021). Beispielsweise hat sich der Wahrheitsgehalt von „Angela Merkel ist Bundeskanzlerin“ im Laufe der Zeit geändert. Im Bereich von Foundation-Modellen wird hier bereits mit verschiedenen Ansätzen (z.B. „safety fine-tuning“) untersucht, wie unerwünschte („unsafe“) Ausgaben möglichst vermieden werden können.

Die „factfulness“ von Sprachmodellen zu erhöhen, ist zurzeit eine stark bearbeitete Forschungsrichtung im Bereich NLU (Natural Language Understanding, siehe z.B. Glaese et al. (2022), Nakano et al. (2021)). Retrieval-basierte Foundation-Modelle wie WebGPT, Retro, und LaMDA können auf eine große Sammlung von aktuellen Textdokumenten zugreifen, um den zu erstellenden Text durch relevante abgerufenen Informationen zu verbessern. Shuster et al. (2021) haben gezeigt, dass der Einsatz von Retrieval die Rate der „Halluzinationen“ reduziert. Insgesamt erlaubt Retrieval die gezielte Verwendung aktuellen Wissens, um die Antwortqualität zu verbessern.

Ein weiteres Problem sind (mutwillige) Zweckentfremdungen, etwa zur Generierung von „Fake News“ oder zur Erstellung gewaltverherrlichender Texte, für die Brundage et al. (2022) einen Maßnahmenkatalog vorschlagen. Allgemeiner ist es schwierig, die Fähigkeiten und damit das Missbrauchspotential von großen Modellen einzuschätzen, da deren „abilities“ (Fähigkeiten) oftmals hochgradig sensitiv auf die Details des jeweiligen Modellinputs (Prompting) sind - z.B. Chain-of-Thought Prompting (Suzgun et al., 2022) und Prompt Injection Attacks (Branch et al., 2022) - und sich ex-ante nicht vorhersagen lassen. Die HELM-Initiative (Liang et al., 2022) nähert sich diesem Problem, indem sie eine Vielzahl an Sprachmodellen standardisiert evaluiert und vergleicht und dabei neben Performanz auch Aspekte wie Kalibrierung, Robustheit und Fairness berücksichtigt.

Ein wichtiger Aspekt in diesem Kontext ist „Trusted-AI“: Die Möglichkeiten, Garantien über die Korrektheit der Ergebnisse der Foundation-Modelle geben zu können, wie sie insbesondere in sicherheitskritischen Kontexten wichtig sind. Hier gibt es vor allem zwei Ansätze: Deduktive Verifikation, die auf Basis von grundlegenden Theorien über die KI-Algorithmen formale Beweise über die Korrektheit bestimmter Aspekte führt, und induktive Validierung, die durch systematisches Testen Aussagen über die Eigenschaften von KI-Systemen ableiten. Beide Bereiche werden in der KI teilweise schon erfolgreich angewendet, müssen aber insbesondere für die hier betrachteten, sehr komplexen Modelle noch deutlich weiterentwickelt werden.

Sicherheit

Diese Dimension adressiert sowohl Eigenschaften der funktionalen Sicherheit als auch die Absicherung gegenüber Angriffen und Manipulationen der KI-Anwendung. Da sich die Maßnahmen in dieser Dimension primär auf die Einbettung der KI-Komponente beziehen, können Foundation-Modelle hier unter anderem durch klassische Methoden der IT-Sicherheit geschützt werden. Dennoch bleiben auch Fragen offen, die sich insbesondere aus dem Zielkonflikt ergeben, einerseits immer mehr auch öffentlich zugängliche Trainingsdaten zu nutzen, diese aber gleichzeitig vor Angriffen, die sich auf das Modellverhalten auswirken könnten, zu schützen.

Dialogsysteme wie BlenderBot 3 verwenden fine-tuning auf „korrekten“ Dialogen, um z.B. das System darin zu verbessern, auf kritische Anmerkungen auch angemessen zu reagieren (Ung et al., 2022). Dennoch erfordern viele der bekannten Herausforderungen der vertrauenswürdigen KI für Foundation-Modellen bessere Lösungen. Als ein Beispiel sei hier nur das Problem der „visuellen Synonyme“ genannt (Mishkin et al., 2022), mit denen sprachbasierte Filtertechniken umgangen werden können (z.B. „rote Flüssigkeit“ statt „Blut“).

Datenschutz

Diese Dimension bezieht sich auf den Schutz sensibler Daten im Kontext von Entwicklung und Betrieb einer KI-Anwendung. Dabei wird sowohl der Schutz personenbezogener Daten als auch von Geschäftsgeheimnissen adressiert. In Zusammenhang mit sind hier insbesondere Risiken zu beachten, die sich im Umfeld der Techniken um „Model-Inversion“ ergeben. So konnten bereits durch gezielte und systematische Abfragen von Modellen sensible Daten wie Sozialversicherungsnummern oder auch realistische Abbildungen (vorher unbekannter) Personen erzeugt werden.

Ferner sind für die oftmals generativ verwendeten Sprachmodelle sogenannte „training data extractions“ relevant, bei denen die erzeugten Modellausgaben Fragmente der Trainingsdaten enthalten. Sofern letztere nicht vollständig von personenbeziehbaren oder urheberrechtlich geschützten Informationen bereinigt wurden, können solche „Extraktionen“ Datenschutz- oder Eigentumsrechte von Dritten verletzen (Carlini et al., 2021). Nichol (2022) schlägt hierzu eine Ad-hoc-Mitigationsstrategie vor.

Es gibt mehrere Möglichkeiten, Datenschutzprobleme bei Foundation-Modellen zu entschärfen. Ein Ansatz wäre, Sequenzen aus den generierten Daten herauszufiltern, die bereits in den Trainingsdaten vorkommen, etwa durch einen Bloom Filter. Ein anderer Ansatz ist das Training mit differential privacy Ansätzen. Hier gibt es einen hohen Forschungsbedarf.

Forschungsrichtungen und -ansätze

Es bedarf weiterer intensiver Forschung und Entwicklung, um die Anforderungen aus allen beschriebenen sechs Dimensionen der vertrauenswürdigen KI für Foundation-Modelle systematisch abzudecken. Die bestehenden Ansätze zur Adressierung spezieller Risiken von Foundation-Modellen müssen erweitert und systematisch zusammengeführt werden. Dabei stellt neben der Komplexität und der schieren Größe der Foundation-Modelle auch die Vielfalt möglicher Anwendungen eine besondere Herausforderung dar. Ohne diese Ansätze ist es meist nicht möglich, die entwickelten Foundation-Modelle in der Praxis verantwortungsvoll zu nutzen.

Zusätzlich zu den technischen Maßnahmen während der Entwicklungs- und Testphase benötigt man einen kontinuierlichen Prozess und Regeln zur Governance, um den Einsatz eines Foundation-Modells zu begleiten. Diese sind bei den bisherigen Modellen aus den USA und China nicht gegeben bzw. wenig kontrollierbar. Während des Modellbetriebs ist weiterhin zu überwachen und sicherzustellen, dass die Prinzipien einer vertrauenswürdigen KI erfüllt bleiben. Grundsätzlich sind geeignete organisatorische Maßnahmen zu ergreifen, um in Situationen, in denen z.B. ein mögliches Fehlverhalten eines Modells auftritt, reagieren zu können. Hierbei ist auch das Wechselspiel zwischen möglicherweise verschiedenen beteiligten Organisationen, wie dem Entwickler des Foundation-Modells und dem Anbieter einer darauf aufbauenden Anwendung, zu berücksichtigen. Dabei ist insbesondere darauf zu achten, dass zur Behebung gefundene Fehler auch erneute technische Maßnahmen, wie z.B. Modellverbesserungen oder Einführung weiterer Filter, angestoßen werden können.

Mögliche und nötige Schritte umfassen:

- Tieferes Verständnis der Strukturen und Funktionsweise von Foundation-Modelle
- Verfahren zur Risikobewertung und Tests entlang spezifischer Proxy-Aufgaben
- Etablierung von Benchmarks zur Vertrauenswürdigkeit von Foundation-Modellen
- Untersuchung und Test von semantischen Eigenschaften des latenten Raums
- Validierung von Modellen zur Input- oder Output-Überwachung
- Definition geeigneter organisatorischer Maßnahmen zur Überwachung des laufenden Betriebs

SPOTLIGHT

2txt NLG GmbH

An verschiedenen Stellen in der Studie stellen wir ausgewählte deutsche Startups und Unternehmen vor, die KI-Foundation-Modelle bereits nutzen. Sie berichten über ihre Anwendungsfälle, Herausforderungen und die Vorteile europäischer Modelle.



Johannes Bubenzer,
Founder und CEO von 2txt

2txt ist Anbieter einer Software zur automatischen Textgenerierung auf Basis von KI-Sprachmodellen. Die Software ist für Anwendungsfälle im Travel-Bereich, in der Finanzbranche und im E-Commerce optimiert. 2txt zeichnet sich durch besonders einfaches und schnelles Setup, leichte Integrierbarkeit in Enterprise-Anwendungen und vor allem durch eine konstant zuverlässige und sehr hohe Textqualität aus.

Wo setzt ihr Foundation-Modelle ein? Was ist euer Use-Case?

Wir setzen Foundation-Modelle mit linguistischen Modellen ein, um beschreibende Texte für diverse

Themengebiete im B2B zu generieren. So erzeugen wir z.B. Produktbeschreibungen, Ortsbeschreibungen oder Finanzberichte komplett automatisiert. Wir kombinieren Foundation-Modelle mit klassischen Sprachmodellen, um das beste aus beiden Welten zu erhalten: Foundation-Modelle macht unser Produkt skalierbar und kreativ, während linguistischen Modelle unsere Produkte kontrollierbar und zuverlässig machen.

Welchen Einfluss haben KI-Foundation-Modelle auf euer Geschäftsmodell bzw. eure Projekte?

Foundation-Modelle ermöglichen es uns, unser Produkt in kürzester Zeit zu skalieren. Arbeitsschritte, die früher Monate gedauert haben, können mit Hilfe von Foundation-Modellen in Sekunden erledigt werden. Das spart Ressourcen, Kosten und Zeit und ermöglicht uns, einen wesentlich größeren Markt zu adressieren.

Welche Schwierigkeiten und Probleme siehst du, dass nur USA und China derzeit KI-Foundation-Modelle umfassend bereitstellen?

Heutzutage: Die Anbieter von KI-Modellen betreiben diese als digitale Services. Das bedeutet, dass sie Zugriff auf alle eingehenden Daten und ausgehende Texte haben. Das ist ein Albtraum für Datenschutz und Geschäftsgeheimnisse der nutzenden Unternehmen.

Für die nahe Zukunft: Es ist abzusehen, dass die Entwicklungen im Bereich der KI-Foundation-Modelle eine der zentralen technologischen Revolutionen der Menschheit auslösen wird. Es werden intelligente Maschinen entstehen und es wäre ein unermesslicher sozialer und wirtschaftlicher Fehler, den Wettlauf um diese Technologien privatwirtschaftlichen Firmen in den USA oder China zu überlassen.

Wie würden euch europäische Modelle - Open Source, alle europäischen Sprachen abdeckend, mit hohen Datenschutzstandards und minimalem Bias - helfen?

Unternehmen können nur dann nachhaltige KI-Geschäftsmodelle aufbauen, wenn sie einen technologischen Vorsprung aufbauen und verteidigen können. Deswegen ist es essentiell, dass wir Downstream Modelle auf freien und offenen Foundation-Modellen mit eigenen Daten trainieren können. Ansonsten wäre jeglicher technologische Fortschritt in diesem Feld zu leicht replizierbar.

Zusätzlich sind einige der interessantesten Anwendungen für KI-Foundation-Modelle solche, in denen persönliche Daten involviert sind, z.B., wenn es um Kundenkommunikation oder Business Intelligence geht. Solche Services sind völlig undenkbar, wenn die Daten unkontrollierbar in die USA oder nach China transferiert werden müssen.

2.4 Offene Forschungsfragen, neueste Entwicklungen und Erwartungen

Die Entwicklung von Foundation-Modellen befindet sich noch in ihren Anfängen und es bestehen viele offene Fragen und potenzielle Risiken. Gleichzeitig sind die Chancen insbesondere für die Wirtschaft sehr vielversprechend und Foundation-Modelle prägen bereits heute die Geschäftsmodelle und zugrundeliegenden KI-Lösungen von Unternehmen mit steigender Tendenz.

Die folgenden Entwicklungen sind abzusehen:

Ausweitung auf weitere Domänen

Die starke Entwicklung und Nutzung von Foundation-Modellen hat im NLP-Bereich seinen Anfang genommen und wird sich voraussichtlich auf alle anderen KI-Bereiche ausbreiten und weitere Datentypen abdecken: Bilder, Ton, Videos, Genom- und Proteinsequenzen, Sensordaten und so weiter.

Multimodalität und das Limit von Textdaten

Multimodalität ist ein wichtiger Faktor. Es ist bereits absehbar, dass multimodale Foundation-Modelle höhere Genauigkeiten als unimodale Modelle erreichen werden und eine noch größere Vielfalt von Anwendungsfällen abdecken können.

Darüber hinaus ist festzustellen, dass die ausschließliche Verwendung von Textdaten zum Training von Foundation-Modellen bereits heute an ihre Grenzen stößt, da sehr große Teile der verwendbaren digitalen Texte des Internets schon genutzt werden. Hier könnten multimodale Datensammlungen eine Lösung darstellen, bspw. die enorme Menge von Videos auf YouTube.

Hürde zum Training von Foundation-Modellen

Die existierende Kluft zwischen einer überschaubaren Anzahl von Technologiekonzernen mit den notwendigen Ressourcen, um Foundation-Modelle zu trainieren, und all jenen, denen es an den finanziellen Mitteln und entsprechenden Rechenressourcen mangelt, könnte in Zukunft weiter wachsen. Diesem Trend wirken Initiativen wie LEAM entgegen. Weiteren Einfluss nehmen Bestrebungen, die eine KI-Demokratisierung durch verteiltes Lernen vorantreiben oder Open Source Kollektive wie Hugging Face, die mit der Big Science Initiative das Modell BLOOM entwickelt und öffentlich gemacht haben.

Keine Monopolbildung, aber verzögerte Entwicklung

Trotz des eingeschränkten Zugangs zu Rechenressourcen ist eine Monopolbildung zur Entwicklung von Foundation-Modellen jedoch zum aktuellen Zeitpunkt eher unwahrscheinlich. Im Fall von GPT-3 wurden vergleichbare Modelle veröffentlicht: Jurassic-1-Modelle von A21 Labs, OPT von Meta, die Modelle GPT-Neo und GPT-J von Eleuther AI, bis hin zu nicht-englischen Modellen wie dem russischen ruGPT-3 von Sber, dem koreanischen HyperCLOVA von Naver, den chinesischen CPM-1/CPM-2-Modellen der Tsinghua-Universität, PanGu- α von Huawei und Wu Dao 2.0 von der Beijing Academy of Artificial Intelligence. Dennoch ist darauf hinzuweisen, dass die Entwicklung der Wissenschaft dem Stand der amerikanischen Wirtschaft um bis zu zwei Jahre hinterher ist. Ein solcher Umstand ist höchst ungewöhnlich.

Enge Zusammenarbeit von Entwicklung und Anwendung

Die Grenze zwischen KI-Entwickler:innen und Anwender:innen wird in den kommenden Jahren vermutlich unschärfer. Grund dafür ist, dass immer mehr Menschen ohne KI-Expertise in der Lage sein werden, Foundation-Modelle erfolgreich für ihre eigenen Fälle anzupassen. Gleichzeitig ist somit ein exponentieller Anstieg neuer KI-basierter Produkte zu erwarten.

Grounding

Foundation-Sprachmodelle lernen alleine Korrelationen zwischen Begriffen und sprachlichen Konzepten. Dabei ist z.B. ein Hund mit den Begriffen Leine, Ohren, Katze, Säugetier, Bein, Fell, Schwanz, Spielzeug, Bellen usw. verbunden. Was fehlt sind Aspekte wie z.B. die dreidimensionale Gestalt des Hundes, seine Art, sich zu bewegen, der Klang seines Bellens, seine dynamische Reaktion auf Katzen oder Menschen. Damit zusammen hängen Gesetzmäßigkeiten der Physik, wie z.B. die Permanenz und Verformbarkeit von Objekten, die Wirkung der Schwerkraft. Daher lässt sich das Konzept des Hundes am besten lernen, wenn es in mehreren Medien auftaucht, zum Beispiel als Bild, in Worten oder in einem Film, in dem er eine Katze jagt. Die Verwendung von multimodalen Foundation-Modellen bietet die Möglichkeit für ein solches integriertes Lernen von Konzepten in der Welt. Yann LeCun sagt: „Anstelle von Sprache oder Bildern wird die nächste KI-Generation jedoch direkt aus Videos lernen. Meta unternimmt derzeit große Anstrengungen, um Videodaten aus der Ich-Perspektive für diese neue KI-Generation zu sammeln, aber auch YouTube-Videos sind als Trainingsmaterial geeignet" (Schreiner, 2022; Jawahar, 2021). Das kürzlich vorgeschlagene Foundation-Modell PLATO ist ein erster Versuch, um intuitive Physik aus Videos zu lernen (Piloto et al., 2022).

Planung und Simulation

Daniel Kahneman hat auf der Grundlage langjähriger Studien des menschlichen Verhaltens eine Hypothese über zwei verschiedene Denksysteme entwickelt (Kahneman, 2011). *Fast Thinking* ist schnell, instinktiv und emotional. Beispiele hierfür sind das Verstehen eines einfachen gesprochenen Satzes oder das Erkennen eines Objekts auf einem Bild. *Fast Thinking* läuft ständig und erzeugt Eindrücke, Intuitionen und schnelle Urteile auf der Grundlage unserer unmittelbaren Wahrnehmungen. *Fast Thinking* kann zum großen Teil schon heute mit den existierenden Foundation-Modellen geleistet werden.

Slow Thinking ist langsamer, überlegter und logischer. Es ist z.B. dafür verantwortlich in einer engen Parklücke zu parken oder die Rechenaufgabe $16 \cdot 34$ zu lösen. *Slow Thinking* kommt nur dann zum Einsatz, wenn Probleme mit *Fast Thinking* auftreten, das heißt, wenn es die Wahrnehmungen nicht gut erklären kann.

Slow Thinking ist in der Lage, Probleme mental durchzuspielen und Ergebnisse für verschiedene Randbedingungen zu simulieren. Es entspricht daher weitgehend dem Konzept der Steuerung durch Prognosemodelle (model predictive control). Diese Fähigkeiten können ansatzweise von Foundation-Modellen wie GATO (Reed et al., 2022) realisiert werden. Notwendig ist eine Integration dieser Modelle mit den Perzeptionsmodellen für unterschiedliche Medien und die flexible Anwendbarkeit auf neue Planungsprobleme. Yann LeCun zufolge ist die „Fähigkeit, Modelle der Welt zu konstruieren, im Grunde das Wesen der Intelligenz“. Diese Modelle werden nicht nur benötigt, um physische Bewegungen vorherzusagen, sondern auch das Verhalten von Menschen, wirtschaftliche Aktivitäten usw. Die große Herausforderung der Künstlichen Intelligenz im nächsten Jahrzehnt besteht darin, prädiktive Modelle der Welt zu erlernen, um mit Unsicherheiten umzugehen (Fridman, 2022).



KI-Foundation-Modelle im internationalen Vergleich

.....●

3. KI-Foundation-Modelle im internationalen Vergleich

Dieses Kapitel beleuchtet die Entwicklung von KI-Foundation-Modellen im internationalen Vergleich. Es zeigt, dass die USA und China aktuell führend sind und erläutert die Gründe, warum Europa bislang keine Vorreiterrolle in der Entwicklung großer KI-Modelle einnimmt.

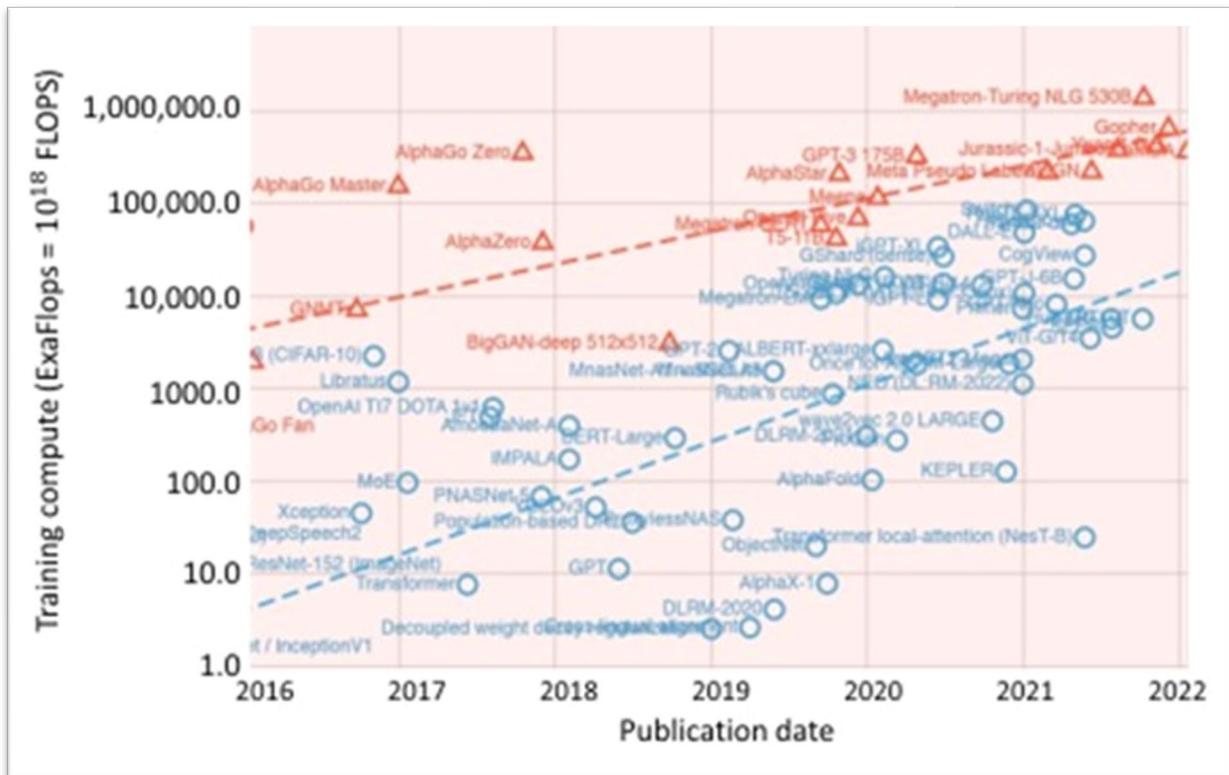


Abb. 11: Um 2016 tauchte ein neuer Trend zu sehr großen Modellen auf, die von großen Internetfirmen trainiert wurden (rot). Diese waren in der Lage, die notwendigen Investitionen zu finanzieren. Die untere blaue Linie veranschaulicht den Berechnungsaufwand der anderen Modelle, z.B. von Universitäten (Sevilla et al., 2022).

Im internationalen Vergleich zeichnen sich folgende Trends deutlich ab:²

- Von der Wissenschaft oder KI-Community entwickelte Modelle zur öffentlichen Verwendung im Text- und Bildbereich liegen etwa 14 bzw. 15 Monate nach der Erstentwicklung aus dem wirtschaftlichen Sektor (Benaic & Hogarth, 2022).
- Der Großteil der veröffentlichten Foundation-Modelle stammt aus den USA und China: Seit 2017 stammen 73 % aus den USA, 15 % aus China.
- Die meisten der entwickelten Modelle stammen aus dem wirtschaftlichen, nicht wissenschaftlichen Umfeld: 86 % der Modelle wurden seit 2017 aus dem wirtschaftlichen Sektor heraus entwickelt, 13 % im wissenschaftlichen Sektor.



Abb. 12: Trainingsleistung (1 ExaFLOPs = 10^{18} FLOPs)³ unterschiedlicher Foundation-Modelle gegenüber Veröffentlichungsjahr nach Ländern

² Für den internationalen Vergleich wurden 125 Foundation-Modelle ab der initialen Entwicklung des Transformers in 2017 betrachtet. Dazu wurde der Datensatz von Sevilla et al. (2022) nach Foundation-Modellen gefiltert und um aktuelle Veröffentlichungen erweitert.

³ Gleitkommaoperationen, Plural von FLOP (floating-point operation) - hier eine Maßeinheit zum Vergleich von Trainingsaufwänden gemessen in den kleinsten Operationen, nicht zu verwechseln mit FLOPS (floating-point operations per second).

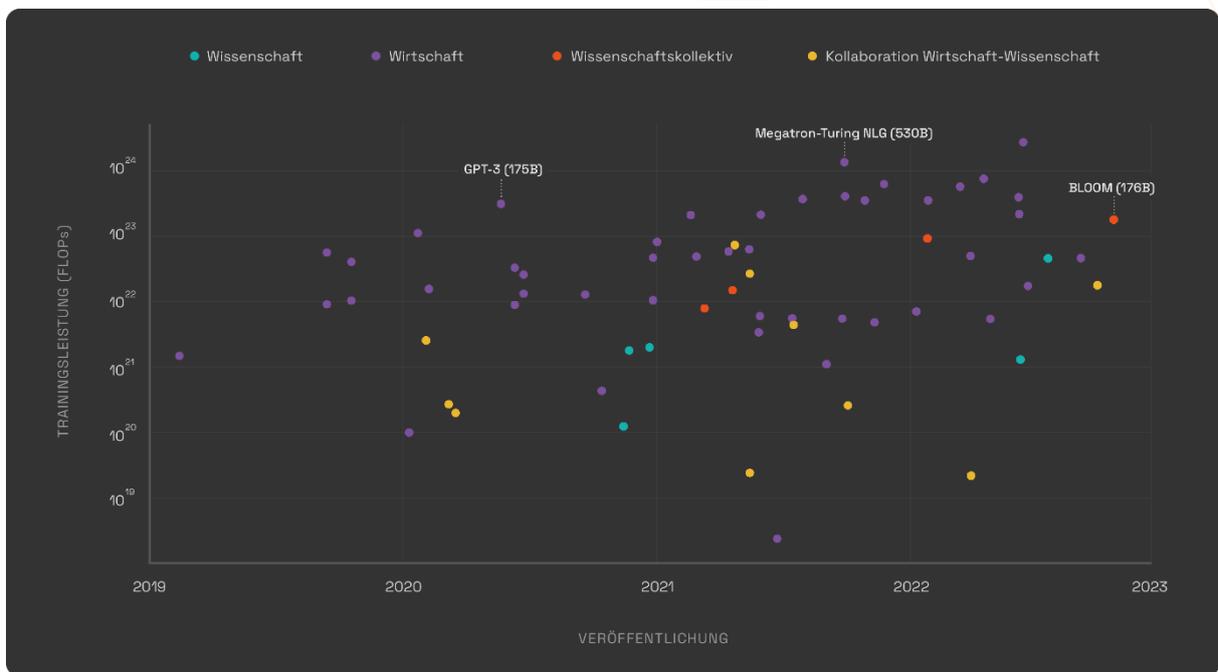


Abb. 13: Trainingsleistung (1 ExaFLOPs = 10^{18} FLOPs) unterschiedlicher Foundation-Modelle gegenüber Veröffentlichungsjahr nach vier Kategorien von Organisationen

Die Übersichtsgrafiken (Abb. 12 und 13) zeigen ganz eindeutig, dass das Technologiegebiet von zwei Ländern, den USA und China, dominiert wird, und dass Europa einschließlich Deutschland im Verhältnis zu seiner Größe und Rolle in der globalen Gesellschaft stark unterrepräsentiert ist. 73 % der Modelle stammen aus den USA und 15 % aus China.

Aus Deutschland und anderen EU-Ländern wurden tendenziell kleinere Foundation-Modelle mit niedrigerem Trainingsaufwand veröffentlicht. Das BLOOM Modell stellt eine Ausnahme dar und wurde von einem Wissenschaftskollektiv bestehend aus über 250 Institutionen auf dem Jean Zay Supercomputer in Frankreich trainiert. Als einziges Modell aus Deutschland taucht in den Abbildungen 12 und 13 das Modell Luminous des deutschen KI-Unternehmens Aleph Alpha auf.

[Aleph Alpha]

Aleph Alpha aus Heidelberg ist ein unabhängiges, deutsches KI-Forschungsunternehmen, gegründet von Jonas Andrulis und Samuel Weinbach. Jonas ist erfolgreicher KI-Serienunternehmer und ehemaliger Senior Manager aus Apples KI-Forschungsabteilung für geheime Innovationsprojekte in Kalifornien. Samuel hat 10 Jahren Erfahrung für KI-Innovation und ist aktuell einer der führenden Köpfe für das Engineering von großen Sprachmodellen.

Aleph Alpha hat auf europäischen Daten und in fünf Sprachen ein GPT3-Äquivalent entwickelt, das in der größten Ausbaustufe die doppelte Anzahl von Parametern verglichen zu OpenAIs bestem Angebot bietet. Zusätzlich entwickelte das Team um Jonas und Samuel eine multimodale Erweiterung, die nicht nur Text, sondern auch Bilder im Kontext versteht. Damit hat Aleph Alpha nach eigenen Aussagen Anfang 2023 das weltweit einzige multimodale Angebot für große Sprachmodelle. Diese und weitere Innovationen werden in zahlreichen akademischen Publikationen und Open-Source Veröffentlichungen mit der Community geteilt. Mit vielen der wissenschaftlichen Spitzenforschern besteht eine enge Zusammenarbeit.

Aleph Alpha gelang der Aufbau eines Teams aus 50 internationalen Experten von den besten Unternehmen aus den USA und dem Rest der Welt. In zwei Finanzierungsrunden konnte das Team mit der Unterstützung von einigen der Spitzeninvestoren aus Deutschland und Europa eine unabhängige Rekordfinanzierung in Höhe von 28 Mio. EUR sichern.

Zum souveränen Betrieb auch für sicherheitskritische Anwendungsszenarien hat Aleph Alpha ein spezialisiertes Rechenzentrum aufgebaut und betreibt damit aktuell das leistungsstärkste kommerzielle KI-Rechenzentrum Europas. Die Technologie von Aleph Alpha wird aktuell in Unternehmen verschiedenster Größe und Branche im Finanzsektor, bei Gesundheit, Recht und in Verwaltung und Sicherheit eingesetzt.

Durch die angedachte LEAM-Infrastruktur könnte Aleph Alpha in den Forschungs- und Open Source-Projekten auch für moderne Foundation-Modelle mit hohen Anforderungen Unterstützung anbieten. Auch für Aleph Alphas KMU-Kunden ohne eigene Rechenzentren sind diese Möglichkeiten entscheidende Zutaten für den Eintritt in ein neues Technologiezeitalter.

INFOBOX

Alle erheblichen Entwicklungssprünge im Bereich der Foundation-Modelle kommen aus den USA und zu einem geringeren Grad auch aus China. Das betrifft sowohl die Arten der Modelle und neue Funktionalitäten, als auch Sprünge in der Modellgröße und die Erschließung neuer Datendomänen.

In den USA sind die wichtigen Durchbrüche in wenigen Industrielaboren erzielt worden oder in Labs, die durch privatwirtschaftliche Investitionen über gigantische Infrastrukturen und großzügige Personalmittel für Spitzenforscher:innen verfügen. Dazu gehören die Labs von Google, Meta, Microsoft, Amazon sowie OpenAI und zu einem geringeren Anteil auch AllenAI.

In China wurde mit dem Modell Wu Dao 2 an der Beijing Academy of AI der erste große Durchbruch erreicht. Dieses Labor vereint Expert:innen aus mehreren Spitzenuniversitäten und arbeitet eng mit industriellen Partnern (u.a. Xiaomi, Meituan und Kuaishou) zusammen. Ein großer Teil der Investition stammt vom chinesischen Staat. Weitere große Modelle kommen aus den Forschungslaboren von Alibaba, Baidu, Tencent, Huawei, Inspur und anderen chinesischen Hightech-Konzernen.

Unter dem Aspekt der Innovationskraft können wir grob drei Klassen von Modellen unterscheiden. Als diese gelten:

- i. Die bekanntesten Modelle, die jeweils bahnbrechende Durchbrüche repräsentieren,
- ii. Modelle, die bestehende Modelltypen in der Größe, den Daten oder ihren Funktionalitäten verbessern oder erweitern sowie
- iii. Modelle, die bestehende Modelle mehr oder weniger genau kopieren und sich technologisch nur unwesentlich von ihren Vorbildern unterscheiden. Dazu gehören auch verkleinerte oder vereinfachte Nachahmungen.

Während Modelle der Klassen (i) und (ii) wesentlich zur Evolution der Technologie beitragen, sorgen Modelle der Klasse (iii) für eine Verfügbarkeit der Modelle in Organisationen und Geschäftsmodellen, in denen die Nutzung aus wirtschaftlichen, technischen oder regulatorischen Gründen nicht möglich war.

Die Entwicklung von Modellen der Klassen (ii) und (iii) hat einen wichtigen Seiteneffekt: Sie bewirkt den Aufbau oder die Verstärkung der technologischen Kompetenz für zukünftige Technologieentwicklungen

Alle Modelle der Klasse (i) und die meisten Modelle der Klasse (ii) wurden in den USA und China geschaffen. Die meisten Modelle aus anderen Ländern kopieren die GPT-Architektur, manchmal mit geringen Modifikationen, und erweitern die Lerndaten, um gewisse Sprachen (mitunter auch Anwendungen) besser abzudecken. Dazu gehören die Modelle aus Israel, Russland, Schweden, Frankreich und Deutschland. Es gibt derzeit einige Modelle der Klasse (ii), u.a. das Modell GPT-SW3 von Sweden AI, das Modell Luminous der deutschen Firma Aleph Alpha und das Modell BLOOM, das von der französischen Initiative OpenScience gemeinsam mit Hugging Face geschaffen wurde. Alle

diese Modelle wurden allerdings erst 15-24 Monate nach der Veröffentlichung von GPT-3 fertiggestellt.

Um Foundation-Sprachmodelle entwickeln zu können, müssen drei essentielle Voraussetzungen erfüllt sein: Kompetenz, Infrastruktur und Daten.

Verfügbarkeit von intellektueller Kompetenz: Gibt es genügend Expert:innen, die das notwendige Wissen und u.U. auch Erfahrung für die Modellentwicklung mitbringen, und sind diese Personen verfügbar?

Verfügbarkeit von Infrastruktur: Reicht die Recheninfrastruktur für die Modellentwicklung, das heißt minimal für Datenaufbereitung, Training und Evaluation?

Verfügbarkeit von Daten: Gibt es hinreichende Mengen von Daten in einer oder mehreren Sprachen, um damit die emergenten Fähigkeiten zu erzeugen? Das ist nicht alleine eine Frage der Menge an Daten, denn durch hohe Diversität und Qualität kann ein Mangel an Masse bis zu einem gewissen Grad kompensiert werden.

Die Auswertung der Expert:inneninterviews und Umfragen (s. Kapitel 4 und 5) hat gezeigt, dass sowohl Deutschland als auch Europa in allen drei Voraussetzungen vor erheblichen Herausforderungen stehen. Im Rahmen der Studie wird ein Fokus auf die verfügbare Infrastruktur gelegt. Dennoch muss betont werden, dass für erfolgreiche europäische KI-Foundation-Modelle alle drei Voraussetzungen erfüllt werden müssen.

Die Verfügbarkeit von hinreichenden Infrastrukturen ist derzeit zum Flaschenhals für die erfolgreiche Beteiligung bei der Technologieentwicklung und -kommerzialisierung geworden. Infrastrukturen, die neue Modelle der Klasse (i) ermöglichen, sind weitaus größer als die Mindestvoraussetzungen für die Entwicklung von Modellen der Klasse (ii). Die Forschung an der Spitze der Technologieentwicklung erfordert das kreative Experimentieren mit vielen Kandidaten für neue Architekturen und Lernmethoden und deren zahlreichen Varianten. Das erschwert die Teilnahme von Universitäten an dieser Forschung. 86 % der Modelle wurden aus dem wirtschaftlichen Sektor heraus entwickelt und nur 13 % im wissenschaftlichen Sektor. Von der Wissenschaft oder KI-Community entwickelte Modelle zur öffentlichen Verwendung im Text- und Bildbereich liegen etwa 14 bzw. 15 Monate nach der Erstentwicklung in den Industrielaboren.

Selbst die großen US-Universitäten können sich ohne Kooperationen mit den Forschungslaboren der Industrie nicht mehr am Wettbewerb beteiligen. Die US-Regierung hat daher die National Artificial Intelligence Research Resource Task Force (NAIRRTF) eingerichtet, um die Infrastrukturen für die KI-Forschung deutlich zu verbessern (*THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH RESOURCE TASK FORCE (NAIRRTF)*, o.D.).

Versuche, die durchaus beachtlichen Hochleistungscomputer der wissenschaftlichen Hochleistungsrechenzentren für das Trainieren von Foundation-Modellen zu nutzen, haben nur begrenzten Erfolg. Es lassen sich zwar mit guter Planung neue GPT-Modelle trainieren und das auch mit zusätzlichen Daten und kleinen Modifikationen. Für die systematische Optimierung und Weiterentwicklung der Modelle und für das Experimentieren mit neuen Modelltypen fehlt jedoch die kontinuierliche Verfügbarkeit von hinreichend großen Computerressourcen. Auch für die Entwicklung und Evaluation neuer Anwendungsklassen können Hochleistungscomputer der Wissenschaft nicht genutzt werden.

Um die Entwicklung von Modellen der Klasse (iii) muss man sich weniger Sorgen machen. Durch die kontinuierliche Zunahme von Rechenleistung in Industrie und Wissenschaft und durch die ständige Verbesserung der Lernverfahren wird das Kopieren von Modelltypen leichter werden.

Während die Anforderungen an Rechenkapazität für das Trainieren von Modellen, selbst mit Erweiterungen der Lerndaten, eher abnehmen, nimmt die erforderliche Rechenleistung für die Entwicklung von Modellen der nächsten Generationen noch stark zu. Der Grund dafür ist die bevorstehende Fusion von Modalitäten zum Erwerb von Weltwissen und zusätzlichen Funktionalitäten, die die Nutzung von großen Volumina an Filmdaten und analogen Daten aus der realen Welt erfordern. Das Kapitel 8 beleuchtet das Thema der Infrastruktur näher.

[OpenGPT-X]

OpenGPT-X ist ein Kooperationsprojekt mit Partnern aus Wissenschaft, Wirtschaft und Technologie. Das Ziel des Projektes ist die Schaffung großer KI-Sprachmodelle, um innovative Sprachanwendungen für Europa und nach europäischen Werten voranzutreiben. Mittels Gaia-X, der sicheren Dateninfrastruktur zur Förderung der Innovation in Europa, wird OpenGPT-X KI-Sprachmodelle und Sprachdienste europaweit offen und in mehreren Sprachen zur Verfügung stellen.

Als Anwendungsbeispiele für zukünftige Produktentwicklungen werden beispielsweise im Bereich Medien KI-Sprachmodelle zur Fragenbeantwortung bei interaktiven Medienformaten entwickelt. In der Domäne Finanzwesen werden die Modelle für eine effizientere Dokumentenverarbeitung bei der Schadensabwicklung durch Versicherungen eingesetzt und in der Mobilität sollen sie als persönliche Assistenten beim Autofahren zu mehr Sicherheit und Fahrkomfort beitragen.

Aktuell trainieren die Projektpartner:innen ein erstes Modell mit zwölf Milliarden Parametern. Dabei liegt der Fokus darauf, möglichst inhaltlich korrekte Antworten von dem Sprachmodell zu erhalten.

Das Vorhaben wird vom Bundesministerium für Wirtschaft und Klimaschutz von Januar 2022 bis Dezember 2024 im Rahmen des Förderprogramms Innovative und praxisnahe Anwendungen und Datenräume im digitalen Ökosystem Gaia-X gefördert. Beteiligt sind das Fraunhofer-Institut für intelligente Analyse- und Informationssysteme (IAIS), das Fraunhofer-Institut für Intelligente Schaltungen (IS), 1&1 IONOS SE, das Forschungszentrum Jülich, die Technische Universität Dresden, die Alexander Thamm GmbH, das Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Aleph Alpha, Control Expert, Westdeutscher Rundfunk (WDR) und der KI Bundesverband.

Zwischen OpenGPT-X und LEAM gibt es einen engen Austausch. Die Ergebnisse und Erfahrungen des Projektes sind in die Ausarbeitung der Machbarkeitsstudie eingeflossen und werden auch weiterhin in die Planung von LEAM einfließen.

INFOBOX

Bedeutung für die technologische Souveränität Deutschlands

Im internationalen Vergleich wird deutlich: Deutsche Unternehmen sind aus Mangel an leistungsfähigen europäischen Alternativen maßgeblich auf die Nutzung amerikanischer und chinesischer Foundation-Modelle angewiesen. Daraus ergeben sich eine Reihe an Herausforderungen, denn die bestehenden Modelle erfüllen in vielerlei Hinsicht nicht die europäischen Qualitätsstandards.

Diese Defizite fassen wir in fünf Punkten zusammen:

- (1) **Inhaltliche Funktionalität:** Deutsche und europäische Inhalte sind unterrepräsentiert. Das gilt insbesondere für wirtschaftliche, gesellschaftspolitische, rechtliche, regionale und kulturelle Themen. Die großen Modelle sind zudem eher auf den Massenmarkt hin ausgerichtet als auf die Anforderungen der Anwendungen in Wirtschaft, Politik, Bildung.
- (2) **Sprachliche Funktionalität:** Die bestehenden Modelle sind durch die verwendeten Lerndaten, aber auch durch die Evaluationsdaten und Benchmarks viel stärker auf das Englische, im Falle der chinesischen Modelle auch auf das Chinesische, hin ausgerichtet als auf andere Sprachen. Sprachen, die im Internet stark vertreten sind, wie das Deutsche, Französische und Spanische, sind zwar auch gut repräsentiert, aber in der messbaren Funktionalität längst nicht so wirksam abgedeckt. Die meisten europäischen Sprachen sind nur sehr unzureichend repräsentiert. Ein wichtiges Ergebnis der linguistischen Forschung ist, dass Sprache das Denken beeinflusst und sogar die grundlegenden Aspekte menschlicher Erfahrung verändert: Raum, Zeit, Kausalität und die Beziehung zu anderen (Boroditsky, 2012). Um die deutsche Kultur adäquat zu erfassen, sind daher Foundation-Modelle für die deutsche Sprache erforderlich.
- (3) **Verfügbarkeit:** Die großen internationalen Modelle sind wegen ihrer Eigentumsverhältnisse und Lizenzmodelle nur eingeschränkt für kommerzielle Anwendungen einsetzbar. Das gilt für die Anpassung durch modifizierte (erweiterte, korrigierte, gefilterte) Lerndaten, zusätzliches Pretraining durch andere Trainingsaufgaben, Inferenz (also praktischen Einsatz), few-shot prompting, large-scale fine-tuning und Integration in umfangreichere Anwendungen.
- (4) **Sicherheit, Verlässlichkeit:** Für viele europäische Anwendungen wären zusätzliche Maßnahmen zur Gewährleistung besserer Performanz in Hinblick auf Sicherheit und Verlässlichkeit erforderlich, insbesondere in Korrektheit, Konsistenz und Datenschutz. Zu den Sicherheitsanforderungen gehört aber auch die Vertraulichkeit der Eingabedaten im Test- und Inferenzgebrauch. Eine weitere Anforderung ist die Verlässlichkeit im Hinblick auf Persistenz, das heißt die langfristige ständige Verfügbarkeit der eingesetzten vortrainierten und insbesondere der durch aufwendiges Nachtraining angepassten Modelle. Dies gilt vor allem dann, wenn diese in Anwendungen eingesetzt werden, die aus wirtschaftlichen oder Sicherheitsinteressen ausfallsicher und unterbrechungsfrei betrieben werden müssen.

(5) **Ethische Akzeptabilität:** Die Ausgaben der Modelle verletzen mitunter durch Bias (Aussagen/Entscheidungen basierend auf falschen Vorurteilen) und Toxizität (Verwendung von ethisch oder stilistisch-ästhetisch nicht akzeptablen sprachlichen Ausdrucksweisen) die de-facto Standards für den Einsatz in Wirtschaft, Politik und Bildung. Daher muss es für die Anwendung möglich sein, Korrektur- und Filtermaßnahmen eigenständig zu definieren, in die Modelle zu integrieren und anzupassen.

Die europäische Forschung kann mit Recht stolz sein auf ihre ersten europäischen Foundation-Sprachmodelle wie Aleph Alpha, BLOOM oder GPT-SW3. Weitere Modelle wie Open GPT-X sind in der Vorbereitung. Diese europäischen Modelle sind ermutigende Beispiele dafür, dass Europa, wenn auch mit etwas Verspätung, Foundation-Sprachmodelle entwickeln kann. Sie sind aber noch kein Indiz für ein Aufschließen der europäischen Forschung in die Avant-Garde der internationalen Forschung zu diesem Thema. Zudem decken diese Modelle trotz großer Fortschritte bisher weder die deutsche Sprache noch die Bandbreite der anderen europäischen Sprachen in dem Maße ab, wie es heute bereits für das Englische erreicht ist.

Ein Grund dafür sind die materiellen Forschungsbedingungen. In der Künstlichen Intelligenz, wie auch in der gesamten Informatik gilt, dass nur Forschungsgruppen, die für eine neue Technologie spielerisch viele Möglichkeiten der Realisierung, Adaption und Optimierung ausprobieren können, längerfristig erfolgreich sein können. Wenn im Falle der Foundation-Modelle die benötigten materiellen Infrastrukturen für Computation und Speicherung so groß sind, dass die universitäre Forschung von der Spitzenforschung ausgeschlossen ist, werden die neuen Durchbrüche aus den Spitzenlaboren der Industrieunternehmen kommen, die sich zusätzlich zur Infrastruktur auch personell und kulturell einen echten Forschungsbiotop aufbauen und leisten können. Ein Kennzeichen solch erfolgreicher Strukturen ist der scheinbare Überfluss, gekennzeichnet durch Redundanz in Infrastruktur und personeller Kompetenz.

Der Begriff AI CERN für die von CLAIRE⁴ und anderen Akteuren geforderte und dringend benötigte Infrastruktur für die europäische KI-Forschung ist eine wirksame Metapher, aber im gewissen Sinne auch irreführend. Der infrastrukturelle Bedarf der Teilchenphysik unterscheidet sich stark vom Bedarf der KI und der gesamten Softwaretechnologieforschung: Während die Infrastruktur des CERN von Grundlagenforschungsprojekten gebucht werden kann, um in wenigen sehr großen Experimenten grundlegende Forschungsfragen zu beantworten, nutzen die KI-Forschungsgruppen der großen industriellen Labore die Computerressourcen fast durchgängig für große Zahlen von Experimenten für das schrittweise Trainieren großer Modelle und für deren Evaluation und Modifikation. Die Evolutionszyklen bestehend aus Modifikation der Technologie und Selektion sind sehr kurz, weil sie nicht wie in der Physik oder in den klassischen Ingenieurwissenschaften in der physikalischen Welt realisiert und

⁴ Die Confederation of Laboratories for Artificial Intelligence Research in Europe (CLAIRE) ist ein Verein mit dem Ziel, Forschung, Innovation und Zusammenarbeit im Bereich der KI zu verstärken. Mehr als 1000 KI-Expert:innen aus ganz Europa unterstützen die Bestrebungen.

getestet werden müssen. Auch die Wege vom Labor in die Anwendungen können in der KI sehr kurz sein. So konnte Google die großen Sprachmodelle zeitnah nach ihrer Fertigstellung bereits im Produktionsbetrieb für die Suche und für die Platzierung von Werbung testen.

Wegen der hohen infrastrukturellen Anforderungen gibt es bisher weltweit nur eine kleine Zahl von Forschungslaboren und Entwicklungs-zentren, die den heutigen Stand der KI-Technologie in ihrer vollen Komplexität beherrschen. Das heißt, es gibt nur ganz wenige Organisationen, die Training und Evaluation von großen multilingualen, multimodalen und multimedialen Foundation-Modellen selbst ausführen können. Mit der multimodalen und multimedialen Verknüpfung der Daten und der Kombination der assoziativen Modelle mit expliziten Daten- und Wissensrepositorien nimmt diese Komplexität noch weiter zu.

Mit dieser Komplexität wachsen natürlich auch die Einstiegsbarrieren für neue Akteure, denn zusätzlich zu gut ausgebildeten Expert:innen wird deren extensive individuelle und kollektive Erfahrung im Trainieren, Evaluieren und Einsatz großer Foundation-Modelle benötigt.

Bedarf der Wirtschaft an KI- Foundation-Modellen

.....●



4. Bedarf der Wirtschaft an KI-Foundations-Modellen

Eine grundlegende Frage, die diese Studie klären möchte, ist, welche Bedeutung Foundation-Modelle für die Wirtschaft haben und welche Anforderungen Unternehmen an die Modelle stellen.

Um dieser Frage auf den Grund zu gehen wurden zwei Methoden angewandt⁵. Diese sind:

- eine Umfrage an kleine und mittlere Unternehmen, die sich auf die Entwicklung von KI spezialisiert haben sowie
- leitfadengestützte Interviews mit den Leiter:innen KI, Data, o. ä. deutscher Großunternehmen.

Die an der Umfrage teilnehmenden Unternehmen lassen sich wie folgt zusammenfassen:

Anzahl Empfänger	373 Mitglieder des KI Bundesverbandes 25 Kontakte aus dem Netzwerk der Merantix Momentum GmbH
Anzahl Antworten	71 Teilnehmer (18 % Rückmeldequote)
Größe der Unternehmen	90 % weniger als 100 Mitarbeiter
Branche der Unternehmen (Mehrfachnennung möglich)	80 % Informationstechnologie 25 % Dienstleistungen je 11 % Industrie & Medien
Einsatz von KI	90 % setzen KI ein

Tabelle 3: Übersicht der wichtigsten Information zu der Umfrage für die Wirtschaft

⁵ Eine genauere Betrachtung der Methodik findet sich im Kapitel V.

Interviews wurden mit folgenden Personen durchgeführt. Ergebnisprotokolle der Interviews befinden sich im Anhang B.3:

Titel	Vorname	Name	Unternehmen
	Wolfgang	Hauner	Allianz SE
Dr.	Maik	Friedel	BASF SE
Dr.	Marion	Legler	Bayer AG
Dr.	Hans-Jörg	Vögel	BMW Group
Dr.	Michael	Fausten	Robert Bosch GmbH
	Jean-Paul	Schmetz	Burda Media
	Mario	Deng	BWI GmbH
Dr.	Corina	Apachițe	Continental AG
Dr.	Matthias	Dorner	DATEV eG
	Stephan	Kaulbach	Deutsche Bahn AG
Dr.	Frank	Säuberlich	EnBW Energie Baden-Württemberg AG
Dr.	Sebastian	Kaiser	Ergo Group AG
	Thomas	Wolf	Hugging Face, Inc.
Dr.	Sabine	Donauer	Infineon Technologies AG
	Nico	Kelling	Infineon Technologies AG
	Rainer	Sträter	Ionos SE
	Christian	Spannbauer	Lufthansa Group
Dr.	Stephan	Meyer	Munich RE
	Jochen	Kaiser	Mercedes-Benz Group AG
Dr.	Michael	Müller-Wünsch	Otto GmbH & Co KG
Dr.	Armin	Kurrle	Porsche AG
Dr.	Lorenz	Determann	Rewe Group
Dr.	Feiyu	Xu	SAP SE
Dr.	Andreas	Wierse	sicos BW GmbH
Dr.	Michael	May	Siemens AG
Dr.	Dirk	Schlesinger	TÜV Süd AG
Dr.	Sebastian	Hallensleben	VDE Verband der Elektrotechnik Elektronik Informationstechnik e. V.
Dr.	Patrick	van der Smagt	Volkswagen AG
Dr.	Alexander	Borek	Zalando SE

Tabelle 4: Befragte Expert:innen aus der Wirtschaft

Alle Erkenntnisse aus der Umfrage und den Interviews sind in die LEAM-Machbarkeitsstudie eingeflossen. Im Folgenden werden wir zusammenfassend einige Aspekte und Übereinstimmungen illustrieren.

Auswertung der Interviews und Umfrage

Ungefähr 66 % der befragten KI-Unternehmen setzen bereits Foundation-Modelle ein oder beabsichtigen, diese in der Zukunft einzusetzen. Dies ist ein beachtlicher Anteil, wenn man bedenkt, dass Foundation-Modelle eine relativ neue Entwicklung der Künstlichen Intelligenz sind. Daneben bestätigen auch viele der interviewten Großunternehmen, bereits KI-Foundation-Modelle im produktiven Einsatz zu haben oder aktuell an Anwendungen zu arbeiten. Dadurch wird eindeutig, wie essentiell Foundation-Modelle für die gesamte Wirtschaft bereits sind.

Die Interviewten geben außerdem an, dass die Bedeutung in den nächsten Jahren weiter steigen wird. Es sei aktuell noch nicht abzusehen, welche Disruption und neuen Geschäftsmodelle KI-Foundation-Modelle in den nächsten Jahren ermöglichen werden.

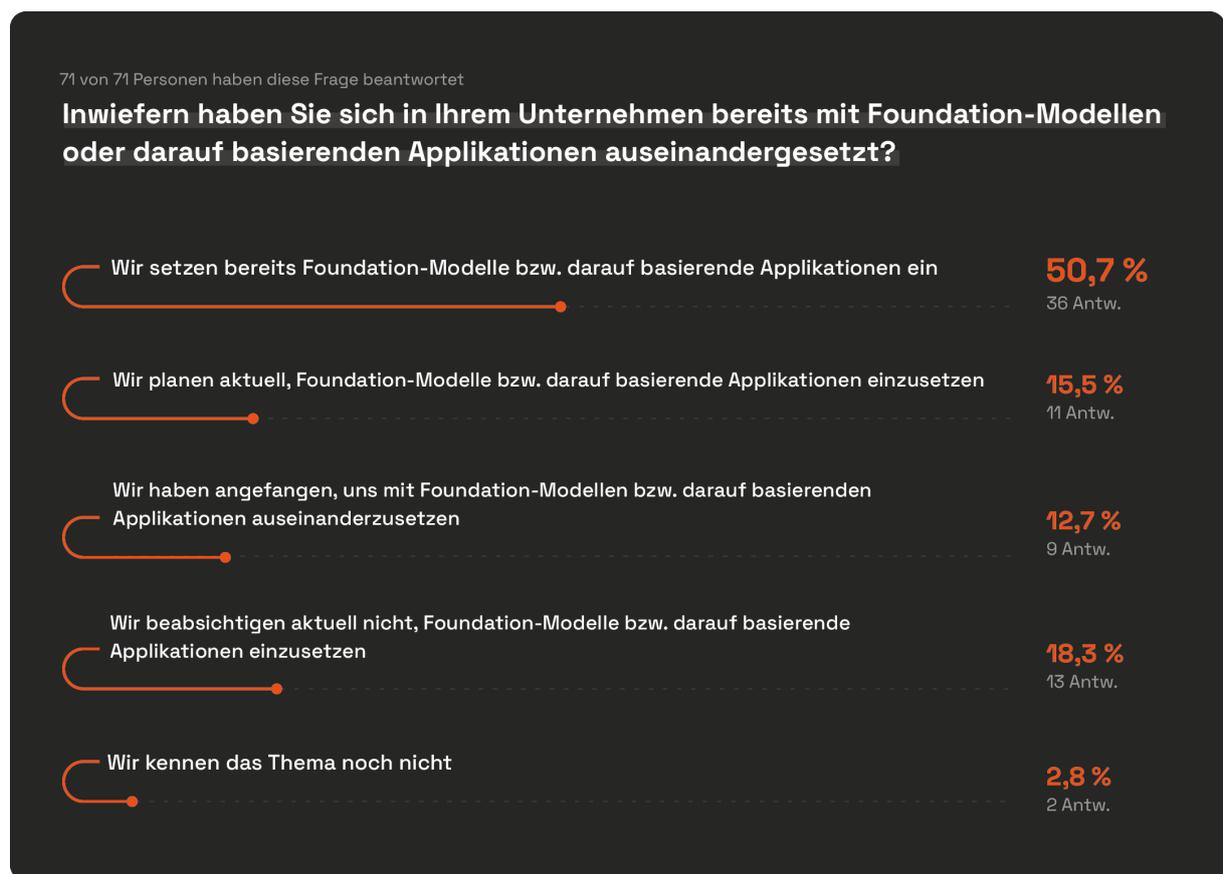


Abb. 14: Ergebnisse der Umfrage mit KMUs zu deren Einsatz von Foundation-Modellen

KI-Unternehmen, die aktuell noch nicht mit KI-Foundation-Modellen arbeiten, geben verschiedene Gründe an, die erfüllt werden müssten, damit sie Foundation-Modelle nutzen würden. Genannt wurden hier insbesondere niedrigere Kosten, die Bereitstellung von Open Source-Modellen sowie die Verfügbarkeit von Daten, die jeweils rund 58 % der Befragten angaben. Weitere Hürden, die Unternehmen als Gründe äußerten, Foundation-Modelle nicht zu nutzen, sind ein Mangel an Recheninfrastruktur (38 %), datenschutzrechtliche Hürden (33 %), ein Mangel an qualifizierten Mitarbeiter:innen (25 %) sowie verfügbaren europäischen Modellen (25 %).

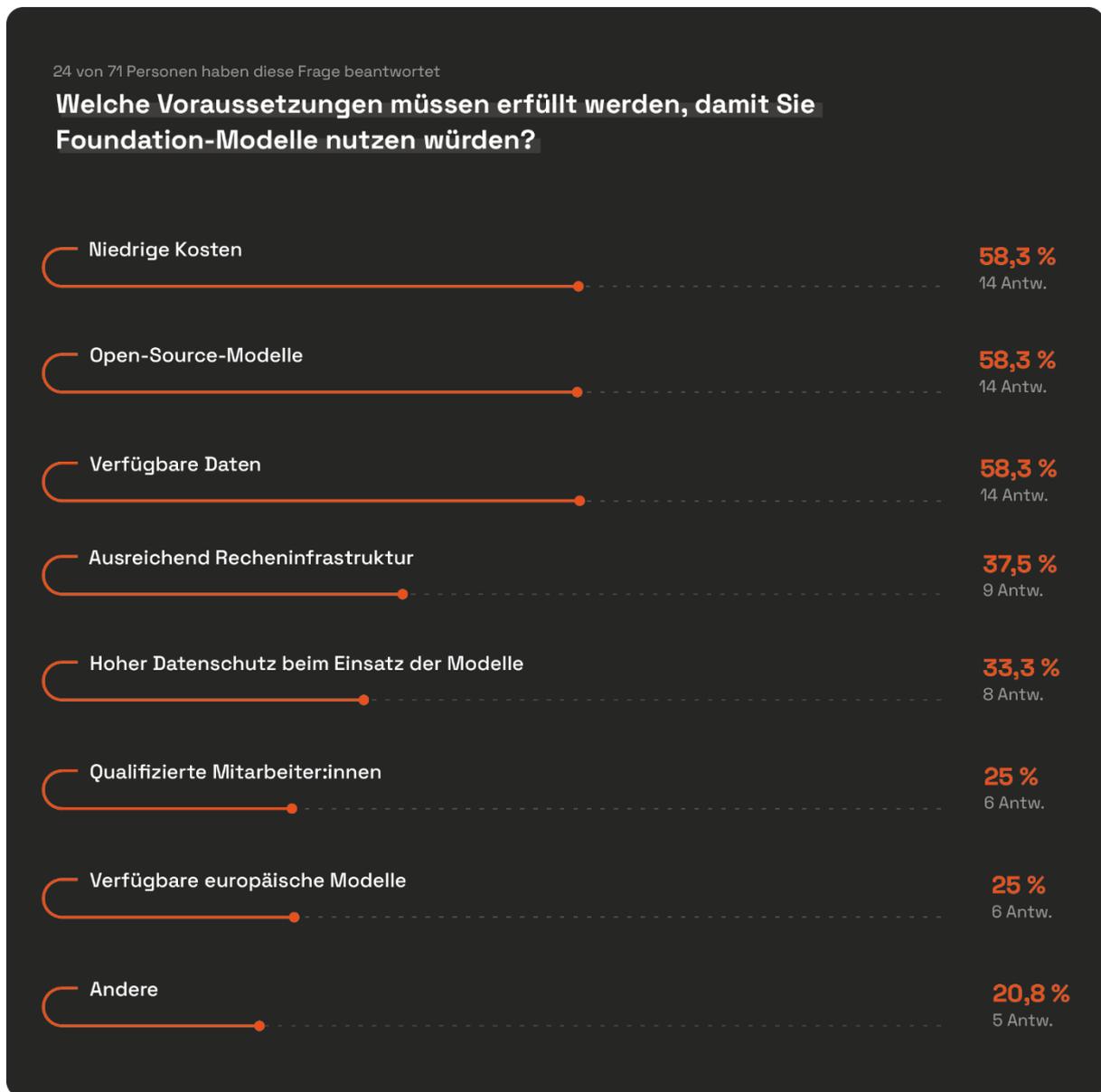


Abb. 15: Ergebnisse der Umfrage mit KMUs zu Hindernissen beim Einsatz von Foundation-Modellen

Auch Großunternehmen teilten diese Bedenken in den geführten Interviews. Dabei wurde deutlich: Großunternehmen verfolgen nicht das Ziel, eigene Foundation-Modelle zu entwickeln. Stattdessen wollen sie bestehende Modelle für ausgewählte Anwendungen anpassen. In der aktuellen Lage ist dies häufig aber aus Compliance-Gründen nicht möglich, da die Modelle nicht frei, sondern nur über Programmierschnittstellen zur Verfügung stehen und Daten für das Tuning aus dem europäischen Wirtschaftsraum heraus gesendet werden müssen. Vor allem mit sensiblen Datensätzen ist daher ein Anpassen der Modelle für Großunternehmen nicht möglich. Die befragten Großunternehmen sehen darin einen klaren Nachteil im internationalen Wettbewerb. Selbst im stark reglementierten und national ausgeprägten Versicherungsumfeld berichten Unternehmen, dass neue, digitale Geschäftsmodelle, die außerhalb Europas entstehen, eine Gefahr darstellen.

Darüber hinaus zeigten die Interviews mit Großunternehmen, dass generelle Foundation-Modelle häufig nicht ausreichen, um den hohen Qualitätsstandards der Unternehmen zu entsprechen. Stattdessen brauche es Foundation-Modelle, die auf die Bedürfnisse einzelner Branchen abgestimmt sind. Das Ziel sollte es also sein, bspw. ein Gesundheitsmodell, ein Industriemodell und ein Versicherungsmodell zu entwickeln, auf deren Basis die Unternehmen dann einzelne Anwendungen entwickeln können. Trotz bestehender Datensätze findet diese Entwicklung aktuell nicht statt, da die deutschen Unternehmen ihre Patienten-, Maschinen- und Versicherungsdaten nicht in die USA übertragen möchten. Dieser Wunsch nach speziellen Modellen spiele auch eine Rolle, da auf dem Datensatz einer bestimmten Population trainierte Modelle nicht direkt auf andere Populationen übertragbar seien. So berichtet beispielsweise das Chemie- und Pharmaunternehmen Bayer davon, dass sich amerikanische Patient:innendaten von europäischen, asiatischen oder afrikanischen unterscheiden.

Daneben fehle es in vielen Unternehmen an gut ausgebildeten Mitarbeiter:innen und es sei schwierig, KI-Modelle in den laufenden Betrieb zu integrieren.

Insgesamt wird die Bedeutung von Foundation-Modellen für die gesamtwirtschaftliche Entwicklung Deutschlands mit 73 % als sehr hoch eingeschätzt. Besonders der Aufbau eines europäischen KI-Ökosystems wird von 82 % der Befragten, die Berücksichtigung von Werten wie Transparenz, Reduktion von Bias und Nachhaltigkeit von 85 % der Befragten als relevant erachtet. Dabei wird die direkte Zusammenarbeit mit der Forschung etwa von der Hälfte (54 %) der Befragten als bedeutend eingeschätzt. Die Zusammenarbeit mit KMUs und Start-ups hingegen von 67 %. Neben der bereits hohen Nutzung von Foundation-Modellen zeigen diese Ergebnisse die gesamtwirtschaftliche Relevanz, die ihnen in der Industrie zugesprochen wird.

Anzahl der Antworten variiert

Welchen der folgenden Punkte messen Sie eine große Bedeutung zu?

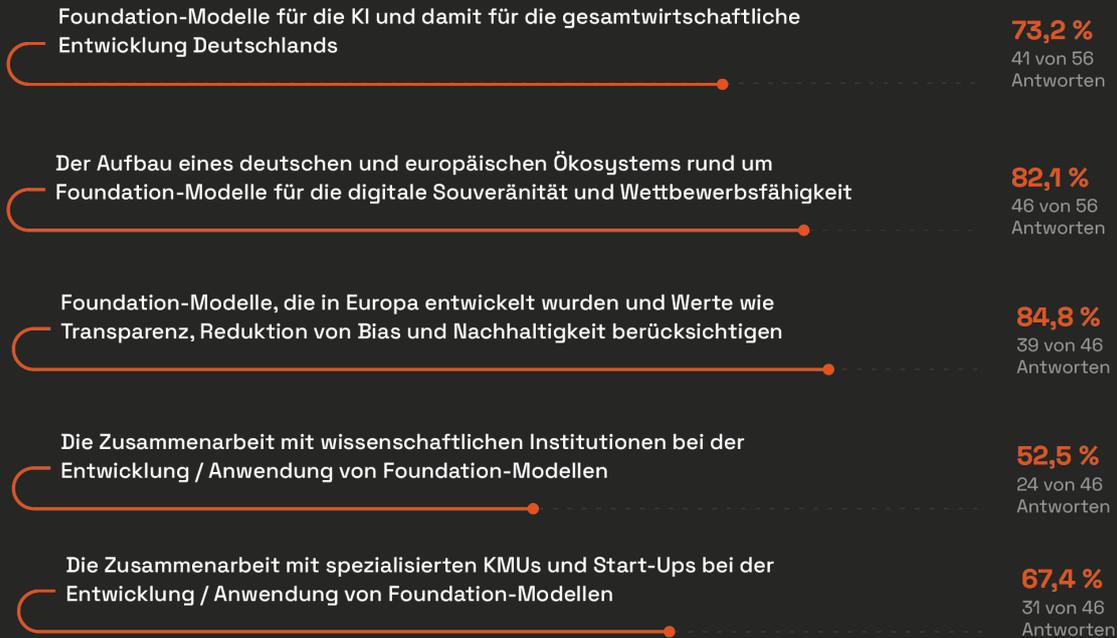


Abb. 16: Ergebnisse der Umfrage mit KMUs zur Bedeutung von unterschiedlichen Aspekten der Foundation-Modell-Entwicklung

Die Interviewten sehen in KI-Foundation-Modellen eine strategische Bedeutung für den Wirtschaftsstandort Europa. So betonen sie die Notwendigkeit eigener europäischer Modelle, um die Wettbewerbsvorteile der Technologie vollständig zu nutzen, die Qualität der Modelle umfänglich zu kontrollieren und Sicherheitsrisiken zu minimieren. Der letzte Punkt sei mit Hinblick auf die geostrategische Situation Europas und existierende Diskussionen rund um 5G und Huawei besonders wichtig.

Aktuell haben für die befragten Unternehmen besonders Sprachmodelle eine hohe Relevanz bei der Nutzung und Entwicklung von Foundation-Modellen (genannt von 71 % der Unternehmen). Dies reflektiert den hohen Erfolg und die Prominenz von Sprachmodellen, wie GTP-3. Ebenfalls als wichtig eingeschätzt werden multilinguale Sprachmodelle (52 %) und multimodale Modelle (38 %). Beide Arten von Modellen stellen logische nächste Schritte von Sprachmodellen dar und werden von Unternehmen wie OpenAI mit Dall-E 2 und ChatGPT bereits erfolgreich entwickelt. Wie oben angedeutet, gehen die Einschätzungen der Großunternehmen noch einen Schritt weiter. In den Interviews wurden zwar auch vor allem Sprach- und Multimodale Modelle genannt, mit denen viele Unternehmen bereits experimentieren, daneben brauche es aber speziellere Modelle für einzelne Anwendungsbereiche. In der aktuellen Situation seien nur inkrementelle Fortschritte und keine Disruption möglich.

Mit jeweils circa 34 % der Befragten finden auch Geschäfts- und Fertigungsprozesse sowie Robotik als bedeutsame Bereiche Beachtung. Diese werden in der aktuellen Foundation-Modell Entwicklung wenig fokussiert behandelt. Hier kann für LEAM eine Chance liegen, diese Nischen in der Modellentwicklung zu besetzen.

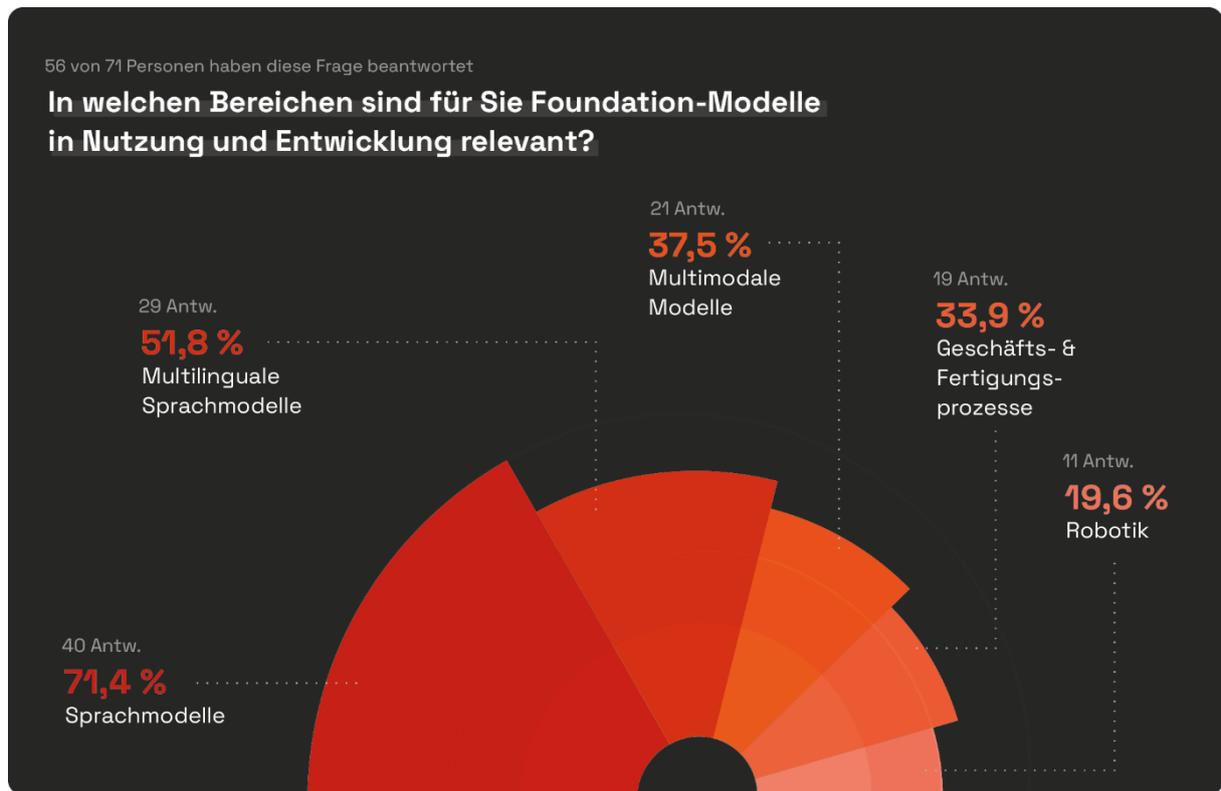


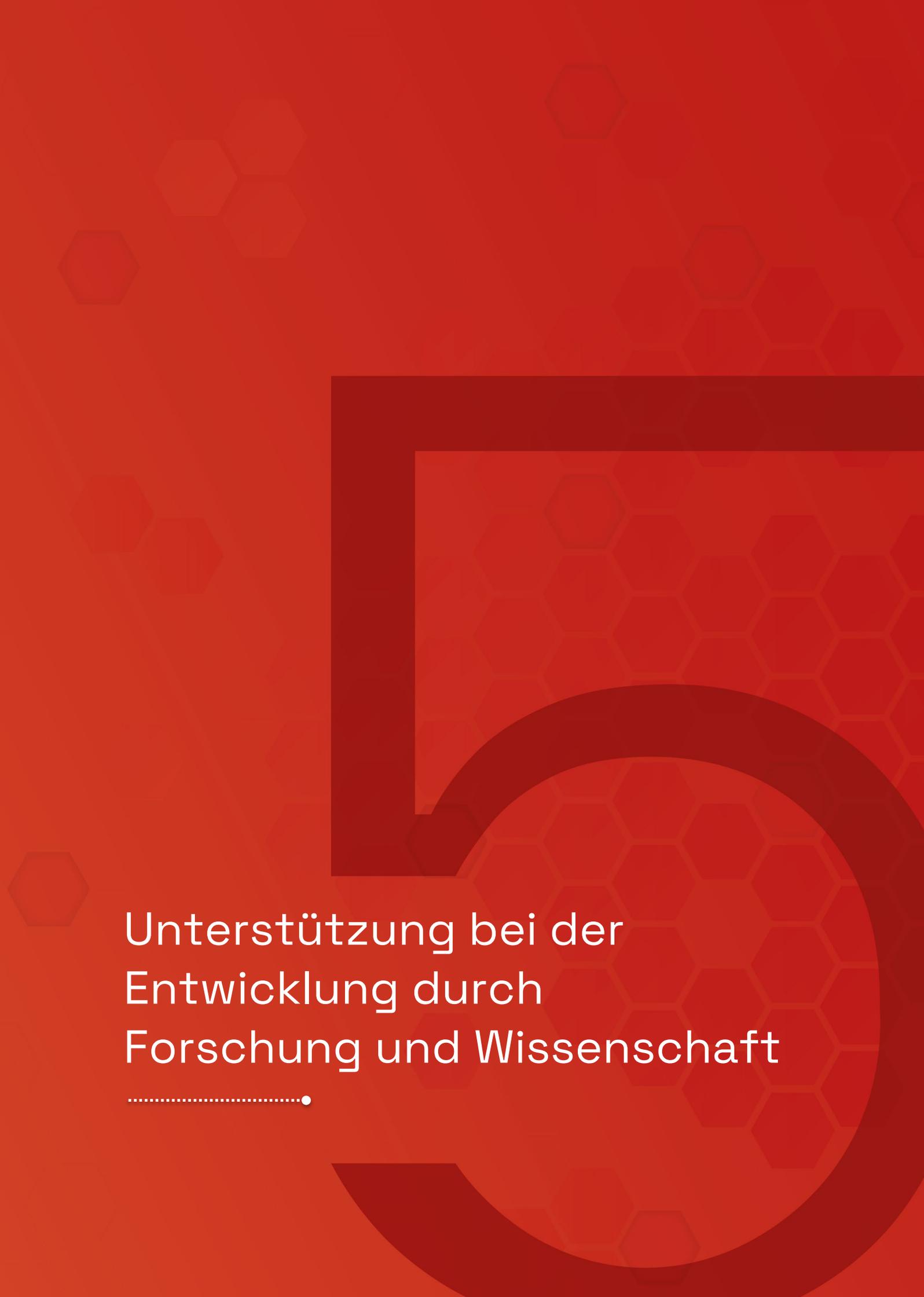
Abb. 17: Ergebnisse der Umfrage mit KMUs zur Relevanz von unterschiedlichen Arten von KI-Modellen bei der Foundation-Modell-Entwicklung (Antworten mit einer Antwortrate von weniger als 20 % wurden ausgelassen. Die vollständigen Antworten befinden sich in Anhang A.2)

Insgesamt zeigt die Auswertung den Bedarf an europäischen Foundation-Modellen. Die befragten Unternehmen hatten am Ende der Umfrage die Möglichkeit, weitere Kommentare zu hinterlassen. Ähnlich wie in den Expert:inneninterviews mit der Wissenschaft (s. Kapitel 5) wurde vermehrt angemerkt, dass „alle außer [den] große[n] Internet-Konzerne[n] darauf beschränkt sind, existierende Foundation-Modelle zu benutzen“ und somit in massive Abhängigkeit geraten. Biases können so bei down-stream Anwendungen nur schwer vermieden werden. Ein weiteres häufiges Thema in den Antworten war, dass die europäische Wirtschaft bereits jetzt im internationalen Vergleich zurückliegt. Diese Aussagen bestätigen auch die Interviews mit den Großunternehmen. Es brauche jetzt eine gemeinsame Aktion im Bereich Infrastruktur und Daten und es sei unbestritten, dass kompetitive europäische Modelle, amerikanischen vorzuziehen seien. Besonders im Bereich Datenschutz und -sicherheit würden europäische Modelle die Implementierung enorm vereinfachen.

Erfreulicherweise geben viele Interviewte an, die Initiative LEAM unterstützen zu wollen und einem Engagement im Rahmen eines Joint Ventures bzw. einer PPP generell offen gegenüberzustehen.

Die Auswertung der Umfrage sowie der Interviews lässt sich in drei Kernaussagen zusammenfassen:

- (1) KI-Foundation-Modelle werden elementarer Bestandteil der IT-Anwendungs-Architektur. Branchenübergreifend implementieren bzw. planen Unternehmen die Implementierung von Anwendungen auf Basis der Modelle.
- (2) Aktuelle Modelle haben häufig den Nachteil, dass sie nicht Open Source verfügbar und damit frei anpassbar sind. Es gibt erhebliche datenschutzrechtliche Bedenken bzgl. der Nutzung außereuropäischer Modelle. Europäische Open-Source-Modelle werden daher als Alternative gebraucht.
- (3) Die Aufgabe kann nicht von einem Akteur alleine bewältigt werden. Es braucht eine gemeinsame Aktion der Unternehmen in Deutschland. Die Wirtschaft erkennt diesen Bedarf und steht einem Engagement offen gegenüber.



Unterstützung bei der
Entwicklung durch
Forschung und Wissenschaft

.....●

5. Unterstützung bei der Entwicklung durch Forschung und Wissenschaft

Aus den Interviews mit führenden Industrie- und KI-Unternehmen wird deutlich: Die Nachfrage nach europäischen Foundation-Modellen ist groß. Eines der primären Ziele, die sich LEAM daher gesetzt hat, ist es Foundation-Modelle für die Wirtschaft bereitzustellen.

Die Entwicklung von Foundation-Modellen erfordert jedoch umfangreiche Maßnahmen im Bereich der Forschung und Entwicklung. Aus diesem Grund ist es wichtig, die spezifischen Anforderungen von Forschung und Wissenschaft zu kennen und mitzudenken, damit die Entwicklung von leistungsfähigen Foundation-Modellen gelingt.

Hierfür wurden im Rahmen dieser Studie Interviews mit führenden Wissenschaftler:innen aus Deutschland durchgeführt. Darin wird herausgestellt, wie der aktuelle Forschungsstand von KI-Foundation-Modellen in der Wissenschaft ist und welche Rahmenbedingungen nötig sind, um Wirtschaft und Wissenschaft bei der Entwicklung dieser entsprechend zu unterstützen.

Methodik der Interviews

Die Interviews wurden durch fünf Leitfragen strukturiert:

- Leitfrage 1 „Relevanz von Foundation-Modellen in der Wissenschaft“
- Leitfrage 2 „International führende Arbeitsgruppen“
- Leitfrage 3 „Wissenschaftliche und wirtschaftliche Defizite der Foundation-Modelle“
- Leitfrage 4 „Maßnahmen zur Förderung der Forschung und wirtschaftlichen Nutzung von Foundation-Modellen in Deutschland“
- Leitfrage 5 „Sonderstellung der Foundation-Modelle und zukünftige Entwicklungen“

Die Interviews wurden mit 21 Expert:innen zwischen Anfang Oktober und Mitte Dezember 2022 geführt (s. Tabelle im Anhang B.1).

Die Erkenntnisse aus den Interviews sind in die Planung von LEAM und an den passenden Stellen in den gesamten Text dieses Kapitels eingeflossen. Im Folgenden werden wir lediglich einige ausgewählte Aspekte der Interviews entlang der fünf Leitfragen illustrieren.

Auswertung der Interviews

In der **Leitfrage 1 „Relevanz von Foundation-Modellen in der Wissenschaft“** beschreiben sich die Expert:innen als Nutzer:innen der Modelle und erkennen die große Bedeutung der Modelle – auch für die Wirtschaft – als Forschungsparadigma der kommenden Jahre an. Für viele Expert:innen bestehen Hürden, um auf einer internationalen Bühne wissenschaftlich mitwirken zu können, da der Zugriff zu den notwendigen Technologien fehlt und sie somit in ihrer Forschung eingeschränkt sind. Foundation-Modelle können klimaverträglich gestaltet werden, wenn zentral verwaltete Modelle der KI-Community zur Verfügung gestellt werden können. Die Reproduktion der Modelle würde damit hinfällig.

Als Antwort auf **Leitfrage 2 „international führende Arbeitsgruppen“** wurden US-amerikanische Technologie-Konzerne wie Microsoft, Open AI, Google, Deepmind, und Meta genannt. In einzelnen Fällen wurden Universitäten wie Stanford oder die Carnegie Mellon University erwähnt, die bei den kleineren Sprachmodellen kompetitiv sind. Die chinesischen Modelle wurden hingegen selten erwähnt. Als Erfolgsfaktoren wurden der Zugang zu Infrastruktur, Daten und Talenten hervorgehoben.

Manche Interviewten betonten Deutschlands Stärke in anderen Gebieten wie dem Vorhandensein von gut ausgebildeten Wissenschaftler:innen, die allerdings aufgrund der internationalen Angebote oftmals nicht in Deutschland bleiben. In diesem Zusammenhang wurde auch von der Abhängigkeit von den überwiegend kommerziellen Anbietern der Modelle gewarnt.

Die **Leitfrage 3 „Wissenschaftliche und wirtschaftliche Defizite der Foundation-Modelle“** wurde mit der mangelnden Anpassbarkeit der Foundation-Modelle beantwortet. Es fehlt an deutschem und auch mehrsprachigem Vokabular, Fachwissen, Robustheit/Invarianz, sowie Erklärungen der Ergebnisse. Diese Hürden können nicht durch ein Nachtraining abgebaut werden, bzw. nur zu finanziellen Bedingungen der Hyperscaler. Probleme wie Bias, Privacy und Intransparenz bezüglich der Trainingsdaten und -Prozeduren erschweren die Nutzung. Die Vertrauenswürdigkeit der existierenden Modelle ist damit infrage gestellt und rechtliche Fragen bleiben unbeantwortet.

Das kontinuierliche Lernen und auch die Verbindung mit Domänenwissen, Unternehmenswissen oder Applikationswissen, wie etwa Faktenwissen aus Wissensgraphen ist ausbaufähig. Die Frage nach der Kontrolle und Validierung der Ergebnisse ist eine offene Forschungsfrage. Weitere Defizite sind fehlende Geschäftsmodelle und die geringe Effizienz der Modelle. An letzterer wird aktiv geforscht, aber der Vergleich mit sehr großen Foundation-Modellen (z.B. GPT-3) kann nicht gezogen werden, was im Gegensatz zur wissenschaftlichen Praxis steht.

Die Generalisierungsfähigkeit der Modelle wurde als unzureichend eingestuft. Derzeit werden zum größten Teil (Sprach-)Daten als einzige Wissensquellen genutzt, welche in zukünftiger Entwicklung ein limitierender Faktor sein können. Somit ist eine Verbesserung der Kuratierung von Trainingsdaten, die Anreicherung mit Wissen sowie mehr komplementäre Daten für die Modellentwicklung wie Ontologien, Sequenzdaten oder Bilder nötig.

Foundation-Modelle können als Kulturgut, bzw. ein öffentliches Gut für die Grundlagenforschung verstanden werden, die einem zentralem und transparenten Entwicklungsprozess unterliegen sollten, um Vertrauen zu schaffen. Andernfalls können Machtkonzentration und sinkende digitale Souveränität die Folge sein.

Als Antwort zur **Leitfrage 4 „Maßnahmen zur Förderung der Forschung und wirtschaftlichen Nutzung von Foundation-Modellen in Deutschland“** wurde hauptsächlich die Förderung einer Infrastruktur und der leichte Zugang zu jener für Wirtschaft und Wissenschaft genannt. Komplizierte Antragsverfahren und zu lange Wartezeit auf Rechenkapazität stellen eine erhebliche Hürde für beide Sektoren dar. Darüber hinaus ist die Kuratierung von geeigneten Trainingsdaten essentiell, wobei die europäische Sprachenvielfalt und Multimodalität fokussiert betrachtet werden sollten. Zusätzlich spielen Zeitreihendaten und Ontologien eine wichtige Rolle. In allen Fällen ist die Rechtssicherheit zu berücksichtigen und sicherzustellen, dass die Datenstrategie ethischen Prinzipien genügt.

Attraktive Forschungsbedingungen als starkes Ökosystem sind nötig, in dem Forscher:innen kollaborieren können. Es sollen Anreize geschaffen werden, noch nicht stark digitalisierte, alte Industrien interessanter für junge Forscher:innen zu machen.

Im Themenkomplex der **Leitfrage 5 „Sonderstellung der Foundation-Modelle und zukünftige Entwicklungen“** wurde das implizite Wissen der Foundation-Modelle hervorgehoben. Die Modelle sind damit u.a. in der Lage, Programmiersprachen zu erlernen, was bis vor einigen Jahren technologisch nicht möglich war.

Potential wird in Deutschland und Europa darin gesehen, Sprache als wichtigstes menschliches Kommunikationsmittel in vielen verschiedenen wirtschaftlichen B2B-Anwendungen durch Assistenzsysteme zu unterstützen (z.B. beim Einkauf, bei der Ansprache der Kund:innen, bei der Kommunikation mit den Kund:innen, bei der schnelleren Abarbeitung von Dokumenten, Rechnungen, Verkauf, Service, Ermittlung von Stimmungen etc.).

Wissenschaftliche Durchbrüche werden in der Effizienzsteigerung (mobile Anwendungen und Edge-Computing) und bspw. in der Verarbeitung längerer Texte gesehen, wozu auch das Erkennen dokumentübergreifender Beziehungen gehört. Weiteres Potential wird auch in einer kontrollierteren Generierung von Texten gesehen.

Einige Expert:innen betonten das Potential durch die Einbindung von Wissen bis hin zu (symbolischen) Subsystemen, die vom neuronalen System (assoziativ) angesprochen werden. Hier wird Modularität zur Anpassung führen, ohne Gelerntes „vergessen“ zu müssen.

SPOTLIGHT

Alexander Thamm GmbH

An verschiedenen Stellen in der Studie stellen wir ausgewählte deutsche Startups und Unternehmen vor, die KI-Foundation-Modelle bereits nutzen. Sie berichten über ihre Anwendungsfälle, Herausforderungen und die Vorteile europäischer Modelle.



Alexander Thamm,
Founder und CEO der
Alexander Thamm GmbH.

Die **Alexander Thamm GmbH [at]** ist ein führendes deutsches Beratungsunternehmen für Data & AI. Mit 350 Mitarbeitern unterstützt [at] seit mehr als 10 Jahren DAX-Konzerne sowie mittelständische Unternehmen und setzt innovative KI-Projekte um.

Wo setzt ihr Foundation-Modelle ein? Was ist euer Use-Case?

Der Einsatz von Künstlicher Intelligenz ist bei unseren Kunden ein wichtiges Thema. Wir entwickeln AI-Strategien, Konzepte und implementieren Projekte auf Basis der neuesten wissenschaftlichen Erkenntnisse. Dabei spielen Sprachverarbeitung und Foundation-Modelle eine immer wichtigere Rolle und

wir investieren in die Nutzung der Technologie in Deutschland, unter anderem in der aktiven Mitwirkung am Projekt OpenGPT-X

Unsere Teams sind auf vielfältige Bereiche der KI-Entwicklung spezialisiert und setzen unter anderem Projekte um in den Bereichen Bildverarbeitung, Natural Language Processing, Forecasting, Anomalie-Detection. Beispiele hierfür sind ein KI-gesteuertes System zur Unterstützung der Zug-Disposition bei der DB, Robotersystem zur Unterstützung der Altenpflege und neuartige Verfahren für das autonome Fahren.

Welchen Einfluss haben KI-Foundation-Modelle auf euer Geschäftsmodell bzw. eure Projekte?

Foundation-Modelle werden zentraler Bestandteil der KI-Anwendungen und der Infrastruktur in vielen Bereichen. Derzeit entwickeln wir individuelle KI-Applikationen oft von Grund auf unter Nutzung spezifischer Kundendaten. In der Zukunft wird es hier eine Verlagerung hin zum Transfer-Learning bzw. Tuning von existierenden, leistungsfähigen Foundation-Modellen geben.

Gleichzeitig werden durch die Nutzung von Foundation-Modellen neue Anwendungsgebiete erschlossen und wir werden für unsere Kunden Applikationen entwickeln, die derzeit noch schwer umsetzbar sind – vor allem im Bereich NLP. Der Markt wird wachsen und wir sehen hier eine große Chance für uns, aber vor allem auch für die Wettbewerbsfähigkeit der deutschen Wirtschaft.

Das hat intensive Auswirkungen auf unser Geschäftsmodell, vor allem, wenn wir auf die Nutzung und Lizenzierung von Foundation-Modellen angewiesen wären, auf die wir nur über APIs zugreifen können und auf die wir keinen direkten Einfluss haben. Wenn wir diese Modelle dann nur von nicht-europäischen Anbietern beziehen können, müssen wir uns zusätzlich noch intensiv mit Datenschutz- und Datensicherheitsaspekten auseinandersetzen.

Damit unsere Kunden und unser Unternehmen nicht in eine einseitige Abhängigkeit geraten, ist es enorm wichtig, dass wir auch auf Foundation-Modelle zugreifen können, die in Deutschland oder Europa entwickelt wurden, und wir diese nicht nur über APIs nutzen können. Gleichzeitig hat die Berücksichtigung von europäischen Werten, z.B. beim Thema Bias, für uns und unsere Kunden eine enorme Bedeutung.

Welche Schwierigkeiten und Probleme siehst du, dass nur USA und China derzeit KI-Foundation-Modelle umfassend bereitstellen?

US-amerikanische Internet-Unternehmen investieren derzeit intensiv in die Entwicklung und Verbreitung von Foundation-Modellen. Durch die Bereitstellung über APIs können die ersten am Markt befindlichen Services gleichzeitig eine Menge Daten z.B. über die Nutzungsschwerpunkte sammeln. Damit besteht die Gefahr, dass sich hier wieder – wie bei den Suchmaschinen – Monopole bilden und eine zunehmende technologische Abhängigkeit entsteht.

Wenn die zentralen KI-Anwendungen nur aus Übersee kommen, werden sich langfristig unsere Aktivitäten auf die Gestaltung von Frontends- und Workflows beschränken. Wir haben keinen oder nur noch geringen Einfluss auf die Modelle, was vor allem hinsichtlich Qualität und Bias problematisch ist. Damit könnte diese Entwicklung auch zu einer potenziellen Bedrohung unseres derzeitigen Geschäftsmodells werden – und zu unserer Unternehmens-Mission, die Wettbewerbsfähigkeit der europäischen Wirtschaft in diesem Bereich sicherzustellen.

Wie würden euch europäische Modelle - Open Source, alle europäischen Sprachen abdeckend, mit hohen Datenschutzstandards und minimalem Bias - helfen?

Europäische Foundation-Modelle, die wir als Basis für unsere KI-Entwicklungen nutzen könnten, würden uns befähigen, auch in Zukunft innovative Anwendungen zu entwickeln. Da wir nicht nur über APIs zugreifen, sondern die Modelle als Open Source zur Verfügung hätten, könnten wir in vielen Bereichen auch unsere eigenen Forschungsaktivitäten intensivieren und für unsere Kunden State-of-the Art KI-Systeme bauen. Damit wäre sichergestellt, dass wir auch in Zukunft bei unseren Kunden Prozesse optimieren und neue Produkte und Geschäftsmodelle ermöglichen können.

Methodik der Umfrage

Zusätzlich zu den 20 interviewten Expert:innen aus der industriellen und universitären Forschung wurden aus Kapazitätsgründen, und um das Meinungsbild noch detaillierter darzustellen, weitere vertiefende Stellungnahmen von 21 führenden KI-Forscher:innen in Form einer Umfrage eingeholt. Die Umfrage wurde zwischen Anfang Oktober und Mitte Dezember 2022 durchgeführt und umfasste 23 Fragen (siehe Tabelle im Anhang A.1). Hinzugekommen sind einige detailliertere Fragen, die aufgrund der begrenzten Interviewzeit in den Expert:inneninterviews nicht gestellt werden konnten.

Alle Erkenntnisse aus der Umfrage sind in die LEAM-Machbarkeitsstudie und in den Text des gesamten Kapitels eingeflossen. Im Folgenden werden wir in Anlehnung an die geführten Expert:inneninterviews einige Aspekte und Übereinstimmungen entlang ausgewählter Leitfragen illustrieren.

Auswertung der Umfrage

In der Umfrage haben sich überwiegend die Aussagen der Interviewten aus den Expert:inneninterviews bestätigt.

Die große Mehrheit aller befragten KI-Forscher:innen arbeiten bereits mit KI-Foundation-Modellen oder beabsichtigen, die Modelle zukünftig einzusetzen. Alle schätzen ihre Bedeutung für die Wissenschaft und Wirtschaft als hoch ein.

Die Art der verwendeten Daten ist durchaus unterschiedlich. So wurden u.a. öffentlich verfügbare Texte und wissenschaftliche Publikationen sowie medizinische und technische Daten aufgeführt. Das Thema fehlender Daten ist allerdings auch einer der Gründe, warum Foundation-Modelle noch nicht überall eingesetzt werden. Weitere Gründe waren eine mangelnde Transparenz im Datenschutz bei existierenden Modellen, fehlende Infrastruktur, technische Limitierungen, Fachkräftemangel, hohe Investitionskosten und Unklarheiten bei der Regulation. Dennoch besteht der Wille, Foundation-Modelle künftig einzusetzen.

Die Diversität der genannten möglichen Einsatzgebiete spiegelt dies ebenso wider, wobei auch die Art der als bedeutsam eingeschätzten Foundation-Modelle eine Rolle spielt – Bilder, Audio und Mobilitätsdaten wurden hier genannt. Keine größere Rolle spielen die vermeintlichen Gefahren, die von den Modellen ausgehen könnten. Lediglich potentieller militärischer Missbrauch, sowie die derzeit noch fehlende Erklärbarkeit wurden hier genannt. Außerdem wurden die Dominanz einzelner Big-Tech-Unternehmen und die fehlenden Antworten seitens der akademischen Forschung bemängelt. Auf dieses Ungleichgewicht wurde zudem häufiger in den Expert:inneninterviews hingewiesen.

Im internationalen Vergleich wird Deutschland lediglich in den Bereichen Kompetenz und Kreativität sowie Ausbildungsmöglichkeiten als wettbewerbsfähig angesehen. In anderen Punkten wie Datenlage, Hardware und Förderung fällt Deutschland hinter den USA und China zurück. Die schwache Digitalisierung und Kommerzialisierung werden zudem als fehlende Voraussetzungen gesehen. Daraus folgt auch die zunehmende Abwanderung von Talenten, die in Deutschland einfach nicht die gleichen Voraussetzungen finden, wie

anderswo. Dies betrifft nicht nur fehlende Investitionen, sondern auch die starke Regulierung neuer Technologien, ausgeprägte gesellschaftliche Ängste und die generelle Abneigung gegenüber dem Fortschritt, der aus der KI-Forschung entstehen kann.

Die Relevanz verschiedener Modellarchitekturen sowie wissenschaftlicher Fragestellungen bzgl. Daten und Datenverarbeitung wurde von den Befragten durchschnittlich als „eher hoch“ eingestuft. Gleiches gilt für die domänenagnostischen Pre-Trainings und das nachfolgende domänenspezifische Fine-Tuning. Auch hier sahen die Befragten die Relevanz als „eher hoch“ an. Für diese grundlegenden Themen besteht demnach größerer Forschungsbedarf innerhalb Deutschlands.

Bzgl. der inhaltlichen Einsatzgebiete besteht wie erwartet keine einheitliche Meinung, da die Befragten aus unterschiedlichen wissenschaftlichen Richtungen kommen. Die Notwendigkeit von Foundation-Modellen in den Anwendungen der verschiedenen Disziplinen wurde generell als „mittel“ bis „eher hoch“ eingestuft. Erklärbarkeit, Aufbau von Common Sense Wissen, Einbezug weiterer Datenquellen (außer Text), hybride Verarbeitung (symbolisch und subsymbolisch), Kausalität, uvm. wurden ebenso genannt.

Die erwarteten Kosten der Umfrageteilnehmer:innen stehen im Einklang mit den tatsächlich bekannten Kosten für große Foundation-Modelle. Die Befragten waren in dieser Hinsicht bereits gut informiert. Bzgl. der Regulation von Foundation-Modellen fühlte sich jedoch der überwiegende Teil der Befragten als nicht ausreichend informiert, wenngleich sie dies als Voraussetzung für ein erfolgreiches Einsetzen der Technologie sahen.

Bezüglich möglicher Mängel der Daten wie Bias, Diskriminierung und Misrepresentation fielen die Antworten dichotom aus. Nur eine kleine Mehrheit der Befragten gab an, sich mit diesen zu beschäftigen und sah sich in der Lage, diese angemessen zu adressieren. Diese Ergebnisse stehen im Einklang mit den Ergebnissen der Interviews, in denen die Hälfte aller Befragten auf diese Problematik hinwies. Die Mehrzahl der Befragten empfand ihr Wissen über die Regulation der Entwicklung von Foundation-Modellen als Voraussetzung für einen möglichen Einsatz der Modelle.

Zusammenfassend zeigten sich die Erkenntnisse der Interviews größtenteils kongruent mit den Ergebnissen der Umfrage. Demnach sind die momentan führenden Nationen auf dem Gebiet der Foundation-Modelle die USA und China. Faktoren, die diese Entwicklung begünstigten, sind die Datenlage, Hardware, Ressourcen und Förderungen. Im internationalen Vergleich wird Deutschland lediglich in den Bereichen Kompetenz und Kreativität sowie Ausbildungsmöglichkeiten als vergleichbar angesehen. Faktoren, welche in Deutschland den Einsatz von Foundation-Modellen noch hindern, seien fehlende Infrastruktur und Zugang, technische Limitierungen, Fachkräftemangel, Bedenken bzgl. Datenschutzes und mangelnde Information über die Regulation der Entwicklung von Foundation-Modellen.

Zudem wurden auch mögliche Gefahren der Foundation-Modelle genannt, diese bezogen sich hauptsächlich auf die Monopolisierung der Technologie sowie mögliche Mängel der Modelle wie Bias, Diskriminierung, Toxizität, Misrepresentation und Erklärbarkeit. Die meistgenannten Interessen für Foundation-Modelle bezogen sich auf Multimodalität, die europäische Sprachenvielfalt, Erklärbarkeit und den Aufbau von allgemeinem Wissen.

Abschließend lässt sich festhalten, dass die Mehrheit aller befragten KI-Forscher:innen bereits mit Foundation-Modellen arbeitet oder beabsichtigt, die Modelle zukünftig einzusetzen und ihre Bedeutung für die Wissenschaft und Wirtschaft als hoch eingeschätzt werden.

Die deutsche Wissenschaft und Forschung sind also bestens in der Lage leistungsfähige Foundation-Modelle umzusetzen, um der Nachfrage der Wirtschaft gerecht zu werden.

Chancen und Pläne bei der Entwicklung europäischer KI-Foundation-Modelle



6. Chancen und Pläne bei der Entwicklung europäischer KI-Foundation-Modelle

Bei der Frage welche Modelle am dringendsten von der Wirtschaft gebraucht werden, zeichnet sich ein deutliches Bild ab. Multilinguale Sprachmodelle wurden in den Befragungen der Expert:innen am häufigsten genannt. An zweiter Stelle folgten multimodale Modelle, die ebenfalls Sprache beinhalten.

Obwohl die wirtschaftliche Bedeutung von Foundation-Modellen in der Biomedizin, in der autonomen Steuerung von Fahrzeugen und Robotern und nicht zuletzt auch im Management von Produktions- und Geschäftsprozessen als sehr hoch eingeschätzt wurde, wurden Sprachmodelle durchweg als wichtigster Anwendungsbereich genannt.

Es ist wenig überraschend, dass großen multilingualen Sprachmodellen die höchste Priorität gegeben wurde. Die meisten der bereits erfolgreichen Foundation-Modelle sind entweder Sprachmodelle oder multimodale Modelle mit einem hohen Anteil an Sprachdaten und sprachlicher Funktionalität. Durch die zentrale Rolle der Sprache in der menschlichen Intelligenz und in nahezu allen Teilbereichen der menschlichen Gesellschaft lassen sich die Sprachmodelle für eine Vielzahl von Aufgaben in fast allen Sektoren der Wirtschaft und Gesellschaft einsetzen.

Bei der Entwicklung von eigenen europäischen Foundation-Modellen empfehlen wir daher, mit multilingualen Sprachmodellen zu beginnen. Dafür gibt es mehrere Gründe, die wir in diesem Kapitel näher erläutern möchten.

Unmittelbarer Bedarf

Sprachtechnologien werden bereits heute in unzähligen Anwendungen genutzt: Dazu gehören Chatbots und Voice Assistants, automatische Übersetzung, Texterzeugung, Textkorrektur, Text-zusammen-fassung, Textvereinfachung, semantische Suche, Verschlagwortung, Tutorensysteme, Informationsextraktion, Entdeckung von Fake News, automatische Klassifikation und Beantwortung von Emails, Sprachlernsoftware, und forensische Textanalyse. In allen diesen Anwendungsbereichen ist die heutige Sprachtechnologie noch stark verbesserungsfähig. Die bestehenden Defizite in der Verlässlichkeit schränken die Märkte für die Sprachtechnologien noch stark ein. Mit den immensen Leistungssteigerungen durch Foundation-Modelle können in allernächster Zukunft größere Märkte erschlossen werden, denn Basisanwendungen mit Produkten, Dienstleistungen und Vertriebskanälen existieren bereits. Deutschland hat Hunderte von Unternehmen, die von dem Technologiefortschritt profitieren würden. Andererseits wären die Produkte und Dienstleistungen dieser Firmen bedroht, wenn stattdessen nur ausländische, insbesondere US-amerikanische Anbieter mithilfe der neuen Technologie leistungsfähigere Produkte auf den Markt bringen. Die betroffenen deutschen Firmen sind als KMUs nicht in der Lage, selbst große Foundation-Modelle zu trainieren.

Wirtschaftliches Potential

Das wirtschaftliche Potential liegt nicht nur in der Verbesserung der Qualität, der Funktionalität und der Marktchancen von bestehenden Anwendungen. Durch die zusätzliche Leistungsfähigkeit werden viele neue Anwendungen möglich. Dazu gehören z.B. leicht zu bedienende Zugänge für Bürger:innen, Patient:innen, Mitarbeiter:innen zu praktischer Information, zu Wissen und zu Vorgängen, die sie betreffen. Das gilt für öffentliche Verwaltungen, Gesundheitssysteme, Unternehmen und jedes andere Teilsystem der Gesellschaft. Als Folge der Digitalisierung wird der direkte passive und aktive Zugriff auf all diese Daten und Prozesse zwar möglich, ist jedoch oft zu schwer zu bedienen. Erst wenn der digitale Zugang so einfach ist, wie zu der Zeit, als man mit kooperativen menschlichen Ansprechpartner:innen kommunizierte, wird die Digitalisierung ihr ganzes Potential entfalten können und allseits akzeptiert werden. Neueste Chatbots wie Googles LaMDA und OpenAIs ChatGPT demonstrieren, dass Foundation-Sprachmodelle mächtig genug sind, um solche Schnittstellen zu digitalen Diensten zu realisieren. Schon sehr bald werden wir alle in der Lage sein, in unserer Muttersprache, gesprochen oder geschrieben, ohne Behördentermine, Warteschleifen oder komplexe Eingabemasken mit allen Diensten zu kommunizieren, die unseren Alltag bestimmen. In der Wirtschaft werden solche Systeme die Kommunikation mit anderen Unternehmen und mit Endkunden:innen revolutionieren. Multilinguale Modelle gestatten es Organisationen mit wenig Aufwand, Akquise und Kundenkommunikation auf andere Länder auszudehnen. Im Bildungswesen werden natürlichsprachliche Lern- oder Tutorsysteme den Wissensstand der Lernenden ermitteln, Defizite erkennen und diese gezielt durch geeignetes Lernmaterial oder personalisierte Wissens Elemente und Erklärungen überwinden.

Gesellschaftliche Relevanz

Unsere Sprache ist ein so wichtiges Element der menschlichen Kultur, dass wir die Technologien, die den Gebrauch der Sprache erleichtern, beeinflussen und für verschiedenste Zwecke einsetzen, selbst beherrschen, und sie für unsere Ziele anpassen müssen. Die Sprachtechnologie wird schließlich in der Zukunft eine noch wichtigere Rolle in der Kommunikation zwischen Menschen und zwischen Mensch und Technik einnehmen.

Alle gesellschaftlichen Entwicklungen spiegeln sich auch in der Sprache wider, das zeigt sich z.B. in den aktuellen Bemühungen um eine Sprache, die unseren ethischen Werten, unserem Geschichtsverständnis und unserem wissenschaftlichen Weltbild entspricht.

Von hoher gesellschaftlicher Relevanz ist auch die Multilingualität. Deutschland sieht seine eigene Zukunft als Teil einer multikulturellen und multilingualen europäischen Gesellschaft. Nur so können wir im geopolitischen Kräftespiel unsere Werte bewahren und wirtschaftlich eine Rolle in der Welt einnehmen, die es uns ermöglicht, unseren Lebensstandard zu erhalten. Eine zentrale Komponente der europäischen Integration ist das Prinzip der Gleichheit unter den beteiligten Sprachen. Selbst wenn das Englische in der wissenschaftlichen, technischen, wirtschaftlichen Kommunikation weltweit eine besondere Rolle als Lingua Franca einnimmt, haben wir allen Mitgliedern der Europäischen Union garantiert, dass ihre Sprachen erhalten und geschützt werden, unabhängig von der Zahl ihrer muttersprachlichen Sprecher:innen.

Eine große Barriere für den gemeinsamen digitalen Binnenmarkt der europäischen Gemeinschaft ist die Mehrsprachigkeit dieses Marktes. Im Unterschied zu unseren größten globalen Wettbewerbern, den USA und China, müssen die Endkunden der Produkte und Dienstleistungen in vielen Sprachen erreicht werden. Während US-amerikanische und chinesische Anbieter kleinere europäische Sprachen gefahrlos vernachlässigen können, dürfen europäische Anbieter das nicht. Das gilt natürlich nicht nur für digitale Produkte, ein Beispiel sind die Beipackzettel der Arzneimittelhersteller.

2018 verabschiedete das europäische Parlament eine Resolution mit dem Titel „Gleichstellung von Sprachen im digitalen Zeitalter“ (Parliament, 2018), die nicht nur die Gleichheit der Sprachen betont, sondern darüber hinaus auch Forderungen an die Sprachtechnologie richtet, um diese Gleichstellung auch in der Praxis zu erreichen.

Wissenserwerb

Foundation-Sprachmodelle zeichnen sich gegenüber Modellen in anderen Datendomänen, wie Proteinen oder DNA-Sequenzen, dadurch aus, dass über die probabilistische Modellierung sprachlicher Texte Wissen zu vielen, wenn nicht sogar zu allen Sachgebieten gelernt wird. Ein Sachgebiet könnte nur dann völlig unberücksichtigt bleiben, wenn es überhaupt nicht in digitalen Texten repräsentiert ist. Durch diese Form des Wissenserwerbs wird auch ein ungelöstes Problem der Wissensmodellierung über Knowledge Engineering in der symbolischen KI angegangen, nämlich die Schwierigkeit, die Verbindungen zwischen den Sachgebieten herzustellen, z.B. zwischen Werkstoffen und Verfahren, zwischen Werkstoffen und Marktpreisen oder zwischen synthetisierbaren Proteinen und deren Beschaffungsquellen. Durch die Vielseitigkeit des erworbenen Wissens in klassischen oder auch neu definierten Wissensgebieten ist das Anwendungspotential der Foundation-Sprachmodelle besonders hoch.

Entwicklungsstand und Vergleichbarkeit

Für die Wahl von Foundation-Sprachmodellen spricht auch der Stand der Entwicklung. Durch das Vorhandensein von mehrjähriger gut dokumentierter Erfahrung im Training der Sprachmodelle und umfangreichen und vielseitigen Benchmarks ist es leichter, zur Spitzenforschung aufzuschließen und den eigenen Fortschritt zu kontrollieren als bei neueren Arten von Foundation-Modellen.

Vorhandene Kompetenz

Auf den Gebieten der neuronalen NLP und der multilingualen Sprachtechnologie gibt es in Deutschland eine anerkannte Forschungstradition und eine starke wissenschaftliche Community mit weltweit anerkannten Spitzenforscher:innen und exzellent ausgebildeten Nachwuchskräften. Es gibt auch eine Vielzahl von älteren Sprachtechnologie Unternehmen und neuen KI-Start-ups, die die NLP-Komponenten in marktfähige Produkte integrieren können und die Anforderungen der bestehenden Märkte kennen.

Verfügbare Datenbestände

Durch mehrere erfolgreiche EU-Projekte zur Sammlung von Sprachdaten für europäische Sprachen haben wir Zugriff auf große Mengen von geeigneten Daten für europäische Foundation-Sprachmodelle. Auch sogenannte kleinere Sprachen sind in den Datensammlungen vertreten. Viele der Daten haben eine hohe Qualität, weil sie von öffentlichen Verwaltungen und gemeinnützigen Organisationen der EU-Mitgliedstaaten eingebracht wurden. Unter den Daten sind auch viele parallele bilinguale und multilinguale Textkorpora. Die dadurch verfügbaren Sprachdaten erhöhen die Erfolgswahrscheinlichkeit für Sprachmodelle, die den Anspruch haben, den speziellen europäischen Anforderungen zu genügen, und sie lassen hoffen, dass diese Foundation-Modelle zumindest nach diesem Kriterium deutliche Vorteile gegenüber den großen bestehenden Modellen bieten.

Potential für künftige Entwicklungen

Neuere Entwicklungen von multimodalen und multimedialen Sprachmodellen verbinden die Daten der repräsentierten Datendomänen untereinander. Das bewirkt, dass man sich mit Hilfe von textueller Eingabe auch thematisch passende Bilder oder Videos generieren lassen kann. Diese Funktionalität wird als cross-modal (transmodal) bezeichnet. Der Sprache kommt hierbei jeweils eine besondere Bedeutung zu, denn sie ist geeignet um die Konzepte und Kriterien der Benutzer:innen auszudrücken. In der Zukunft wird die Zahl der Datendomänen zunehmen, wobei die Sprache voraussichtlich immer die semantische Basis für die Kombinationen darstellen wird. Daher können neue Kombinationen durch die Erweiterung von Sprachmodellen hergestellt werden.

Aus diesen Gründen empfehlen wir, die Entwicklung von Foundation-Modellen mit großen europäischen Sprachmodellen zu beginnen, denn hier liegt das größte Potential für die deutsche Wirtschaft und die Nachfrage ist besonders hoch. Schrittweise sollen die Foundation-Modelle dann auch um weitere Datenarten erweitert werden.

6.1 Erste europäische multilinguale Foundation-Sprachmodelle

Mit Blick auf Sprachmodelle ergeben sich eine Vielzahl an Fragen für die konkrete Umsetzung und Weiterentwicklung. Im Folgenden werden wir ausführen, welche Architekturen und Daten den Sprachmodellen zugrunde gelegt werden sollen und welche Prioritäten die Entwicklung und die Evaluation der Modelle bestimmen sollen.

Architektur der Basis Foundation-Modelle

Für die Basisarchitektur gibt es derzeit keine ernsthafte Alternative zum Transformer-Ansatz. Fast alle großen Foundation-Sprachmodelle folgen diesem Ansatz, der sich auch für Bild-, Video- und Proteindaten als tragfähig erwiesen hat und sich somit auch für die geplanten Erweiterungen um multimodale und multimediale Daten eignet.

Nun gibt es allerdings verschiedene Ausformungen des ursprünglichen Transformer Modells, die grob in drei Klassen fallen: (i) Encoder-Decoder Modelle, die mehr oder weniger dem Ausgangsmodell entsprechen, (ii) Encoder Modelle und (iii) Decoder Modelle.

Am erfolgreichsten sind bisher Modelle der Klassen (i) und (iii), wobei die allergrößten Modelle häufiger in die Klasse (iii) fallen, was aber sicher daran liegt, dass der Trainingsaufwand für große Decoder Modelle sehr viel geringer ist.

Die Fachwelt geht aber mehrheitlich davon aus, dass Modelle der Klasse (i), also volle Transformer Modelle, Vorteile für die bestimmte anspruchsvolle Aufgabenstellungen bieten, die die tiefere Analyse der Eingabe und die dadurch erzeugte abstrakte Repräsentation der semantischen Beziehungen nutzen können. Dazu gehören die Beantwortung von Fragen und andere Aufgaben, die Schlussfolgerungen erfordern, sowie die automatische Übersetzung.

Bisher konnten Decoder-Modelle die Nachteile der einfacheren Architektur durch Größe, eine größere Menge von Lerndaten und eine höhere Zahl von Parametern ausgleichen. Besonders die Performanz von ChatGPT, des neuesten Modells der GPT-3 Klasse, demonstriert auf beeindruckende Weise, dass mit dem Decoder-Modell (zusammen mit geeigneten Verfahren des Nachtrainierens) Funktionalitäten möglich sind, die einen hohen Grad an semantischer Abstraktion erfordern.

Für die erste Generation an Modellen wird daher empfohlen, die Decoder Architektur der GPT-Modelle zu verwenden. Das bietet die folgenden Vorteile:

- schneller Kompetenzaufbau durch die Verwendung der einfacheren Architektur
- bessere Vergleichbarkeit mit den neuesten GPT-Modellen und mit europäischen Modellen (BLOOM, Luminous, GPT-SW3)
- durch die Vergleichbarkeit eine leichtere Evaluation der Beiträge der zusätzlichen Daten
- kürzere Zeit bis zu ersten verwendbaren Ergebnissen

Versionen der ersten Modelle sollen sich von den bestehenden Vorbildern aber auch durch die Auswahl der Lerndaten unterscheiden. Zusätzlich zu den bisher eingesetzten gecrawlten Webdaten werden spezielle europäische Korpusdaten verwendet, die sowohl die europäischen Sprachen als auch die für die Wirtschaft relevanten Gegenstandsbereiche besser abdecken. Einen besonderen Effekt für die angestrebte Multilingualität erwarten wir vom Einsatz paralleler bilingualer und multilingualer Korpora.

Daneben werden die Trainingsdaten auch um Wissensdaten erweitert, die aus großen Wissensgraphen stammen.

6.1.1 Lerndaten

Lerndaten für Foundation-Sprachmodelle sind große Mengen von digitalen Textdaten, in denen die europäischen Sprachen bestmöglich repräsentiert sind.

Die zentrale Komponente der Trainingsdaten für die größten Sprachmodelle sind sehr große Textvolumina, die durch Webcrawling kuratiert wurden. Die gemeinnützige Organisation Common Crawl sammelt und archiviert seit 2011 große Teile des World Wide Webs. Diese werden kostenlos für Forschung, Entwicklung und andere Nutzungen zur Verfügung gestellt, seit 2013 im Web-Archivformat WARC.

Aufbauend auf den Datenbeständen von Common Crawl wurden die Web-Crawl Corpora zusammengestellt und gereinigt, die den größten Foundation-Sprachmodellen zugrunde liegen. Eine Auswahl dieser Datensätze soll hier vorgestellt werden.

C4 - Colossal Clean Crawled Corpus

Das ist insbesondere der Korpus C4 (Colossal Clean Crawled Corpus), der sowohl exklusiv für das Englische, aber auch in multilingualen Varianten existiert (Raffel et al., 2020). Das Attribut „Clean“ bezieht sich auf mehrere Verfahren zur Datensäuberung, die eingesetzt wurden, um echte monolinguale Texte von Mischdaten zu trennen.

Zu den Säuberungsverfahren gehörte auch der Einsatz von Blockierlisten (Blocklists), Listen von Wörtern, an denen man obszöne, rassistische und anderweitig anstößige Texte zu erkennen hoffte. Beispiele sind die Lists of Dirty, Naughty, Obscene, and Otherwise Bad Words (LDNOOBW), die auf GitHub für das Englische und ca. 25 weitere Sprachen angeboten werden. Im C4 Korpus wurden nun alle Webseiten ausgefiltert, auf denen sich mindestens eines der anstößigen Wörter in der jeweiligen Sprache fand. Man nahm an, dass es bei der großen Menge an Daten weniger Probleme geben würde, wenn nach diesem groben Kriterium mitunter auch ungerechtfertigt gefiltert würde, als durch die Aufnahme anstößiger Inhalte in die KI-Modelle.

Es konnte dann aber von Kritiker:innen gezeigt werden, dass die Löschung von Texten mit sexuellen Bezügen oder mit Slangausdrücken dazu führte, dass wichtige Bereiche der Gesellschaft in den so „gereinigten“ Texten unterrepräsentiert waren, unter anderem Teile der LGBTQ-Gemeinschaft oder ethnische Minderheiten (Dodge et al., 2021). Zum Glück gibt es aber auch „noblocklist“ Versionen des C4 Korpus, auf die diese lexikalischen Filter nicht angewandt wurden.

mC4

Für multilinguale Foundation-Sprachmodelle wurde das Webkorpus mC4 geschaffen, eine spezielle multilinguale Fassung des C4 Korpus, dessen Vorteile durch die Performanz des mit mC4 trainierten Modells mT5 demonstriert werden konnten (Xue et al., 2021).

Das Korpus mC4 enthält 27 TB Textdaten für 101 Sprachen. Die best repräsentierte Sprache ist natürlich Englisch mit 10401 GB, während das westafrikanische Yoruba mit nur mehr 0,158 GB in diesem Korpus das Schlusslicht bildet.

Das Englische macht mit 2733 Mrd. Token 5,67 % der Daten aus. Deutsch liegt mit 1404 GB oder 347 Mrd. Token nach Russisch und Spanisch an vierter Stelle, was 3,05 % entspricht. Kleinere europäische Sprachen wie Litauisch, Estnisch und Lettisch sind noch mit jeweils 11, 6,9 und 7,9 Mrd. Tokens vertreten. Die kleinsten im Korpus präsenten europäischen Sprachen sind Luxemburgisch mit 1 Mrd. und Irisch mit 0,5 Mrd. Tokens. Im Gegensatz zum Luxemburgischen ist das Irische seit dem 1. Januar 2022 sogar eine der Arbeitssprachen der Europäischen Union.

6.1.2 Europäische Projekte

Glücklicherweise gab es in Europa seit 2010 eine Reihe von Projekten, die europäische Sprachdaten für die Forschung und Wirtschaft kuratieren und verfügbar machen. Diese von der EU in verschiedenen Programmen geförderte Vorhaben sollten die europäische KI, insbesondere die Sprachtechnologie, in die Lage versetzen, die europäische Sprachenvielfalt technologisch beherrschbar zu machen und somit für die europäische Wirtschaft und den gemeinsamen digitalen Binnenmarkt Wettbewerbschranken abzubauen.

Dazu gehören insbesondere die Vorhaben ParaCrawl und European Language Resource Coordination (ELRC), gefördert im Programm Connecting Europe Facility (CEF), und das European Language Grid (ELG), gefördert im Programm Horizon 2020.

ParaCrawl

ParaCrawl hat von 2017 bis 2021 in drei Phasen parallele Texte in den europäischen Sprachen durch Web-Crawling kuratiert. Das Ergebnis ist ein paralleler Datensatz mit 41 Sprachpaaren, die hauptsächlich Englisch mit einer anderen Sprache verbinden (39 von 41). Neben den europäischen Sprachen enthält ParaCrawl auch Daten für neun ressourcenarme, außereuropäische Sprachen.

ELRC - European Language Resource Coordination

Der Schwerpunkt des Infrastrukturvorhabens ELRC, das im Januar 2023 abgeschlossen wird, lag auf der Stärkung der europäischen Übersetzungstechnologie. Die europäische Wirtschaft und Gesellschaft soll auch unabhängig von Google Translate und den anderen großen Übersetzungsdiensten der amerikanischen Hyperscaler in der Lage sein, Sprachbarrieren zu überwinden, ohne dass dabei die Stellung von europäischen Sprachen, auch nicht der kleineren und kleinsten Sprachen, leidet. Daher lag der Schwerpunkt von ELRC auf der Sammlung oder Erzeugung von bilingualen und multilingualen parallelen Korpora, die dann zum überwachten Trainieren von maschinellen Übersetzungssystemen eingesetzt werden können. Eine Herausforderung bestand darin, hochqualitative und wirtschaftlich bzw. gesellschaftlich relevante Korpora in den Staaten der EU zu kuratieren, und das von Organisationen in allen EU-Mitgliedsländern.

ELRC hat insgesamt über 200 Milliarden Wörter an hochqualitativen Sprachdaten in Europa gesammelt und aufgearbeitet und ein europaweites ELRC Netzwerk in allen EU-Mitgliedsstaaten etabliert. Die Daten enthalten 5600 Ressourcen: bilinguale und

multilinguale parallele Korpora sowie monolinguale Textkorpora. Es gibt mehr als 880 parallele Korpora mit mehr als 1 Mio. Wörtern, darunter mehr als 230 mit mehr als 10 Mio. Wörtern und mehr als 50 mit über 100 Mio. Wörtern. Die europäischen Nationalsprachen sowie die Sprachen der größten Minderheiten sind in den Daten vertreten. Über 950 Ressourcen enthalten deutsche Daten und selbst kleine Sprachen wie das Irische und das Maltesische sind noch mit über 300 bzw. 200 Ressourcen recht gut vertreten.

ELG - European Language Grid

Das Projekt ELG European Language Grid hat unter seinem Namen eine Plattform geschaffen, auf der zehntausende sprachtechnologische Ressourcen für die Nutzung bereitgestellt werden, die meisten davon unter nichtkommerziellen Lizenzbedingungen. Unter diesen Ressourcen finden sich auch tausende von Textkorpora, darunter auch sehr viele für europäische Sprachen, die in den großen Webkorpora zu wenig repräsentiert sind.

EFNIL - European Federation of National Institutions for Language

Im Dachverband EFNIL (European Federation of National Institutions for Language) haben sich nationalen Sprachinstitutionen der europäischen Staaten zusammengeschlossen. Zu diesen Institutionen gehören Sprachinstitute wie das Institut für Deutsche Sprache in Mannheim oder das Institut für die Tschechische Sprache in Prag, aber auch Einrichtungen der Regierungen wie der Dänische Sprachrat oder die Generaldelegation für die Französische Sprache. Viele dieser Einrichtungen verwalten und pflegen nationale Korpora, große, gut gepflegte und mehr oder weniger repräsentative Textsammlungen für ihre jeweilige Sprache oder Sprachen. In der Vergangenheit waren diese Daten für die Sprachtechnologie meist nicht zugänglich, weil die jeweils geltenden Eigentums- oder Urheberrechte dieser Nutzung im Wege standen.

OPUS

Neben diesen EU-Projekten gibt es bereits seit 2004 die Initiative OPUS des Nordic Natural Language Processing Lab, die jetzt über drei Millionen parallele Texte mit über 100 Millionen Sätzen frei zur Verfügung stellt.

eTranslation

Die meisten europäischen Organisationen in Wirtschaft, Politik und Zivilgesellschaft schrecken aus Sicherheitsgründen davor zurück, die Übersetzungsdienste großer multinationaler Unternehmen für Übersetzung, Suche und Dolmetschen zu verwenden. Für die reine Textübersetzung hat die Europäische Kommission bereits 2017 den Übersetzungsdienst eTranslation geschaffen, der europäischen Organisationen die kostenlose Übersetzung zwischen allen Arbeitssprachen der EU anbietet.

Dieser Dienst der Generaldirektion Übersetzung der Europäischen Kommission setzt bereits neuronale Übersetzungssysteme für viele Sprachpaare ein, wobei in der Regel immer zwischen Englisch und einer weiteren europäischen Sprache übersetzt wird und alle anderen Übersetzungen indirekt über das Englische als Zwischensprache vorgenommen werden. Die Europäische Kommission verfügt über keine Foundation-Modelle.

Obwohl eTranslation in der Qualität nicht an Google Translate heranreicht, wird der Dienst von vielen Organisationen für eine Rohübersetzung bzw. für Informationsübersetzungen verwendet, was nicht nur an den Sicherheitsanforderungen der Nutzerorganisationen liegt, sondern auch an den Stärken des europäischen Systems im Hinblick auf europaspezifische Sprache und Inhalte. Pro Jahr werden ca. 300 Mio Seiten übersetzt.

6.1.3 Die Bedeutung paralleler Sprachdaten

Es ist eine offene Forschungsfrage, welche Rolle parallele Sprachdaten längerfristig in der Welt der Foundation-Modelle spielen werden.

Wenn parallele Sprachdaten bereits vorhanden sind, kann man diese als Übersetzungspaare direkt in die Daten für das Pretraining übernehmen, wobei das Modell dann ganz alleine lernt, dass es sich um Paare von bedeutungsgleichen Sätzen in den jeweiligen zwei Sprachen handelt. Man kann aber auch das überwachte Trainieren der Übersetzungsfähigkeit in das Pretraining integrieren, ohne dass zusätzlich Kosten für die Datenproduktion oder -annotation anfallen. Die parallelen Daten können dann natürlich auch noch als monolinguale Daten für die Ergänzungs- oder Ersetzungsaufgaben des selbstüberwachten Pretrainings verwendet werden.

Es wurde aber beobachtet, dass sich die Übersetzungsfähigkeit als eine emergente Funktionalität einstellt, sobald nur hinreichend große Textvolumina für die Einzelsprachen in den Pretrainingsdaten vorhanden sind. Textübersetzung wird somit zu einer Zero-Shot oder höchstens zu einer Few-Shot Anwendung des multilingualen Foundation-Modells. Für Sprachen, die nicht hinreichend in den Lerndaten repräsentiert sind, müsste das Modell dann weiterhin durch überwachtes Lernen nachtrainiert werden. Somit verringert sich der Bedarf an parallelen Korpora, die in ihrer bestehenden Menge begrenzt und teuer zu produzieren sind.

In ihrem Forschungsbericht *nmT5 - Is parallel data still relevant for pre-training massively multilingual language models?* zeigen Kale et al. (2021) jedoch, dass parallele Korpora wegen der Knappheit der Daten für viele Sprachen immer noch eine wesentliche Bedeutung für die Anwendbarkeit der Modelle in diesen Sprachen und für die Qualität der Übersetzungen haben.

Dieses Alleinstellungsmerkmal hat großes wirtschaftliches Potential, das über Anwendungen für Textübersetzung und multilinguale parallele Texterzeugung weit hinausgeht. Ein weiteres Anwendungsgebiet ist die translinguale Suche (crosslingual search), und zwar sowohl die Suche nach Dokumenten als auch die Suche nach Informationen und Wissensinhalten. Wenn die Bürger:innen oder Kund:innen, insbesondere Menschen, die nicht Englisch als Muttersprache haben, in ihrer eigenen Sprache in vielsprachigen Inhalten suchen können, erleichtert das ihr Leben. Eine andere Anwendung ist die Unterstützung von bilingualer und multilingualer Konversation, wie z.B. durch eine simultane Dolmetscherfunktion für Beratungen, Besprechungen und Verhandlungen. Zoom bietet, aufbauend auf deutscher Übersetzungstechnologie, bereits

die fast-simultane Übersetzung zwischen neun großen Sprachen an, darunter fünf EU-Sprachen, aber eine Erweiterung auf die Breite der europäischen Sprachen ist schon alleine wegen des Mangels an geeigneten Trainingsdaten nicht in Sicht.

6.1.4 Empfehlung

Trotz der Tatsache, dass die europäischen Sprachen, insbesondere die sogenannten kleineren Sprachen, in den Korpora noch nicht hinreichend repräsentiert sind, sollten zuerst die bestehenden Korpora C4 bzw. mC4 für das Training der multilingualen europäischen Foundation-Modelle eingesetzt werden. Dies erlaubt zum einen eine Vergleichbarkeit mit anderen großen Sprachmodellen und verhindert, dass Daten fehlen, die essentiell zur Performanz der bekannten Modelle beigetragen haben. Zum anderen erlaubt es, die Beiträge der speziellen europäischen Daten für relevante Anwendungen besser beurteilen zu können.

Es gibt bisher keine Untersuchungen darüber, zu welchen Anteilen sich die Daten in den durch die europäischen Projekte bereitgestellten Korpora mit den Webdaten der großen amerikanischen Foundation-Modelle überlappen. Dadurch, dass Common Crawl aber keine Daten einsammelt, die nur nach Registrierung zugänglich sind und zudem die Crawling-Bestimmungen (nofollow, robots.txt) der originären Websites respektiert, ist anzunehmen, dass über die europäischen Projekte große Volumina an zusätzlichen Daten für das Training verwendet werden können.

ELRC und ELG werden vom DFKI koordiniert. Die Datenhaltung und -bereitstellung für beide Projekte wird vom Projektpartner ILSP in Athen verantwortet. Die Projektkoordinatoren und der Direktor des ILSP haben ihre Bereitschaft bekundet, die Entwicklung europäischer Modelle bei der Datenkuratierung aus ihren Beständen zu unterstützen.

Im September 2021 wurden auf der EFNIL Jahrestagung den nationalen Sprachinstitutionen die Möglichkeiten und Absichten bezüglich europäischer Foundation-Modelle vorgestellt. Die erste Kommunikation mit EFNIL Mitgliedsorganisationen ergab deren grundsätzliche Bereitwilligkeit, ihre Korpusdaten unter kontrollierten Bedingungen für das Training von Foundation-Modellen zur Verfügung zu stellen.

Das Institut für Deutsche Sprache hat seine Bereitschaft erklärt, die hochqualitativen Textkorpora des Instituts von rund 40 Milliarden Wörtern für das Training europäischer Modelle verfügbar zu machen und das Vorhaben zu Aspekten der Repräsentanz der deutschen Sprache in den europäischen Foundation-Modellen bei Bedarf auch wissenschaftlich zu unterstützen.

Darüber hinaus sammelt das Projekt OpenGPT-X Daten für das Training eines europäischen Sprachmodells. Die dort gemachten Erfahrungen und verwendeten Daten können auch über das Projekt hinaus Anwendung finden.

Diese Daten, zusammen mit dem Vorhandensein der parallelen Sprachdaten für europäische Sprachen, ermöglichen das Training erster europäischer Foundation-Modelle.

Durch die Verwendung der parallelen Daten im Pretraining gibt es die Möglichkeit, mehrere Ziele zu erreichen:

- (1) die Erfüllung der besonderen europäischen Anforderungen auf Überwindung der Sprachgrenzen
- (2) eine bessere Berücksichtigung der kleineren europäischen Sprachen
- (3) die Schaffung eines besonderen Alleinstellungsmerkmals der Modelle gegenüber den bestehenden Foundation-Modellen

SPOTLIGHT

Bayer AG

An verschiedenen Stellen in der Studie stellen wir ausgewählte deutsche Startups und Unternehmen vor, die KI-Foundation-Modelle bereits nutzen. Sie berichten über ihre Anwendungsfälle, Herausforderungen und die Vorteile europäischer Modelle.



Dr. Marion Legler, Head of Decision Science & Advanced Analytics, Bayer Pharma.

Bayer ist ein weltweit tätiges Unternehmen mit Kernkompetenzen auf den Life-Science-Gebieten Gesundheit und Ernährung. Mit seinen Produkten und Dienstleistungen will das Unternehmen Menschen nützen und die Umwelt schonen, indem es zur Lösung grundlegender Herausforderungen einer stetig wachsenden und alternden Weltbevölkerung beiträgt.

Wo setzt ihr Foundation-Modelle ein? Was ist euer Use-Case?

Wir verwenden KI-Foundation-Modelle, um große Textmengen automatisch zu verarbeiten und unseren Experten zu helfen, schnell relevante Informationen zu finden. Zum Beispiel, die Modelle:

- Sie konvertieren die von den Ärzten während der klinischen Studien verfassten Texte in standardisierte medizinische Diagnosecodes, was die anschließende manuelle Überprüfung durch unsere Experten vereinfacht.
- Lesen Sie die gesamte medizinische Literatur und die Berichte über klinische Studien, um diejenigen zu identifizieren, die für die Therapiegebiete und Behandlungen von Bayer besonders relevant sind.
- Tausende von Dokumenten der Zulassungsbehörden (EMA, FDA usw.) werden durchgesehen und die Themen in jedem Teil der Dokumente automatisch klassifiziert, so dass die Informationen leicht auffindbar sind. Scannen Sie die von Patienten erhaltenen Mitteilungen auf unerwünschte Ereignisse.
- Gruppierung der von Bayer-Vertretern nach Besprechungen mit Ärzten verfassten Erkenntnisse, um aufkommende Diskussionsthemen zu entdecken.

Wir verwenden auch vortrainierte Computer-Vision-Modelle für Anwendungsfälle, in denen nur begrenzt kommentierte Bilder zur Verfügung stehen, wie z. B. im Zusammenhang mit bestimmten Krebstumoren.

Welchen Einfluss haben KI-Foundation-Modelle auf euer Geschäftsmodell bzw. eure Projekte?

Die meisten Anwendungsfälle, die durch diese Modelle ermöglicht werden, waren vorher nicht möglich, insbesondere solche, bei denen es darum geht, große Mengen von Dokumenten zu erkennen. In anderen Fällen, wie der Kodierung in klinischen Studien oder der Erkennung von unerwünschten Ereignissen in Texten, unterstützen die KI-Modelle die manuelle Arbeit der menschlichen Experten und sparen etwa 50 % der Zeit, die diese für die sich wiederholenden Aufgaben aufwenden.

Wie würden euch europäische Modelle - Open Source, alle europäischen Sprachen abdeckend, mit hohen Datenschutzstandards und minimalem Bias - helfen?

Da die meisten großen Sprachmodelle nur in englischer Sprache verfügbar sind, müssen wir oft erst aus anderen Sprachen ins Englische übersetzen und dann die Modelle verwenden. Der Inhalt und die Feinheiten des Textes gehen bei der Übersetzung sicherlich verloren.

Open-Source-Modelle, die alle europäischen Sprachen abdecken, könnten dazu beitragen, dass Patient:innen, Ärzt:innen oder Behörden, die sich in verschiedenen Sprachen äußern, gleichermaßen berücksichtigt werden. Zwar gibt es solche Modelle bereits (BLOOM), doch werden sie derzeit kaum genutzt, da sie auf allgemeine, nicht-medizinische Texte trainiert sind. Die Herausforderung, jemals ein brauchbares mehrsprachiges Modell für Gesundheitsanwendungen zu haben, besteht darin, dass die wissenschaftliche Literatur und die Dokumente, die zum Trainieren medizinischer Sprachmodelle verwendet werden, ursprünglich nur auf Englisch geschrieben sind.

6.2 Vermeidung von Falschaussagen, Bias und Toxizität

Wie in Kapitel 2.3 beschrieben, wurden bei neuronalen Modellen, die auf großen Mengen von Webdaten trainiert wurden, mehrere Formen von inadäquatem Verhalten beobachtet. Um die Foundation-Modelle so zu gestalten, dass sie den Kriterien für Anwendungen in Wirtschaft und Gesellschaft genügen, muss solches Verhalten verhindert oder minimiert werden. Deshalb sollte die Verhinderung des Fehlverhaltens eine zentrale Priorität bei Entwurf, Training und Evaluation der Modelle sein.

Falschaussagen

Die deutlichste Form von Falschverhalten sind fehlerhafte Aussagen. Das Modell wird solche Aussagen tätigen, wenn trotz des Pretrainings Wissenslücken verbleiben oder Wissensinkonsistenzen durch widersprüchliche Trainingsdaten erst entstehen. Zudem können Foundation-Modelle nicht zwischen faktisch korrekten Aussagen und plausiblen inkorrekten Aussagen unterscheiden, die ja auch im Bereich der Literatur oft zu finden sind. Die Methoden zur Behebung sind eine bessere Datenauswahl und die Hinzunahme von Wissen aus nichtsprachlichen Wissensbeständen (s. Kapitel 2.3).

Bias

Schwieriger zu entdecken und auch schwerer zu verhindern sind Unausgewogenheiten bis hin zu Voreingenommenheiten, falschen Verallgemeinerungen und Fehlurteilen mit ethisch verwerflichen Konsequenzen. Diese werden oft unter dem Sammelbegriff Bias subsumiert.

Häufig kommentierte Formen von Bias sind Voreingenommenheiten in Bezug auf Geschlecht und ethnische oder soziale Herkunft, die wir auch bei Menschen finden. (ref)

Allerdings muss nicht jeder Bias negativ sein, es kann z.B. vorkommen, dass im Zusammenhang mit einem Produkt die häufigere Erwähnung von seltenen Gefahren im Vergleich zur Erwähnung der Vorteile des Produkts zu Aussagen führt, die einen vorsichtigen Gebrauch nahelegen.

Es ist auch nicht möglich, alle Formen von negativen Bias vorherzusehen, weil diese von bestimmten gesellschaftlichen Entwicklungen von Problembewusstsein und zeitabhängigen Sensibilisierungen bestimmt werden.

Versuche, präventiv alle Quellen von negativem Bias aus den Trainingsdaten zu tilgen, sind daher unrealistisch. Stattdessen versuchen Entwickler:innen, durch gezieltes Nachtrainieren bereits bekannte Formen von Bias durch adäquates Antwortverhalten zu überschreiben. Einen ähnlichen Ansatz haben die Entwickler von ChatGPT gewählt, die damit wirksam auf Kritiken bezüglich der früheren GPT-Modelle reagiert haben.

Zum adäquaten Antwortverhalten gehört auch eine Pluralität und Ausgewogenheit in der Nennung von alternativen Antworten. Die Ergänzungsaufgaben des Pretrainings reichen nicht aus, um ein solches Verhalten zu erreichen. Wenn ein Modell bei mehreren möglichen Antworten nicht einfach die wählen soll, die statistisch durch die Lerndaten präferiert ist, muss ein Sprachmodell das angemessene Antwortverhalten durch dediziertes Training erlernen. Auch hier hat ChatGPT vorgemacht, wie dieses Ziel durch Nachtrainieren erreicht werden kann. OpenAI hat diese Verbesserungen im Wesentlichen

durch Bestärkungslernen (Reinforcement Learning) erreicht, das heißt in diesem Fall die systematische Korrektur der Gewichte durch die Reaktion von Testbenutzern.

Aus den Testeingaben zusammen mit den Antworten und Reaktionen der Testbenutzer:innen lassen sich annotierte Lerndaten generieren, so dass man spätere Modelle dann überwacht nachtrainieren kann und so die Kosten für die manuelle Leistung spart.

Zusätzlich zum Nachtrainieren bietet sich auch eine bewährte einfachere Methode für die Verbesserung von Antwortverhalten an: die Einbettung der Benutzer:inneneingaben in Prompts, die in einem oder mehreren Sätzen spezifizieren, welche Form der Antwort erwartet wird. Diese Methode wird auch vielfach eingesetzt, um Inhalt und Form der erwarteten Antworten auf die Anforderungen spezifischer Anwendungen anzupassen.

Toxizität

Das Phänomen der Toxizität reicht von der Verwendung obszöner oder ethisch anstößiger Ausdrücke bis hin zu Äußerungen, die von Menschen als Ausdruck von Hass oder Verachtung interpretiert werden können oder auf andere Weise als beleidigend oder verletzend empfunden werden.

Versuche, solche Ausgaben durch Zensur der Lerndaten zu erreichen, also durch das automatische Ausfiltern von Texten, die bestimmte anstößige Wörter oder Ausdrücke enthalten, sind nicht das geeignete Mittel, um Toxizität zu verhindern. Viele Wörter, die auf die Listen der anstößigen Wörter (Blocklists) gelangt sind, haben auch Verwendungen, die durchaus akzeptabel sind. Das Ausfiltern der harmlosen Verwendungen würde auf der Datenseite ganze Themenbereiche schwächen. Zum anderen können nicht alle Beleidigungen oder Verächtlichmachungen an den verwendeten Wörtern alleine erkannt werden. Es zeigt sich aber, dass die neueren Sprachmodelle selbst eine abstrakte Zuordnung von sprachlichen Ausdrücken zu sprachlichen Stilen und Register erlernen. Viele offensichtliche Formen der Toxizität kann das Modell nach Bestärkungslernen eigenständig vermeiden. Der Sprachgebrauch von ChatGPT erscheint im Vergleich zu früheren Foundation-Modellen sehr vorsichtig, ja fast schon konservativ.

Ob sich aber verlässlich alle Äußerungen vermeiden lassen, deren Aussagen oder Präsuppositionen von sensibilisierten Benutzer:innengruppen als beleidigend oder verletzend empfunden werden können, ist eine offene Frage. Dieses Thema ist ja auch eine Herausforderung für menschliche Textproduzent:innen. Das Problem wird sich aber ohnehin auch nur bei ganz speziellen Anwendungen stellen.

Empfehlung

Für alle Formen von inadäquatem Antwortverhalten gilt: Selbst wenn es nicht sofort möglich ist, alle Formen dieses Fehlverhaltens für alle Anwendungen zu 100 % auszuschließen, so ist es doch für den Erfolg der Foundation-Modelle und deren Akzeptanz in Wirtschaft und Gesellschaft essentiell, Art, Grad und Häufigkeit von potentiellen Fehlverhalten empirisch zu bestimmen bzw. vorherzusagen.

Es sollten deshalb schon während der Entwicklungszeit der ersten Modelle eine Testbatterie von Eingabeprompts für alle bekannten Arten von Bias und Toxizität zusammenstellen, die geeignet sind, Ausdrücke des Fehlverhaltens hervorzurufen und in ihrer Häufigkeit zu messen. Die ständige Evaluation und Verbesserung der Modelle in Bezug auf diese Probleme sollte ein zentrales Forschungsthema sein.

6.3 Verbindung von Foundation-Modellen mit großen Wissensbeständen

Neben dem aus Texten gewonnenen Wissen können auch bereits explizit formalisiertes Wissen für Foundation-Modelle verfügbar gemacht werden. Explizites Wissen kann in Datenbanken, Ontologien oder Wissensgraphen kodiert sein. Heutzutage werden für die Repräsentation großer Wissensbestände hauptsächlich Wissensgraphen (engl. Knowledge-Graphen, kurz KG) eingesetzt, weil sie die Vorteile von Ontologien und Datenbanken verbinden und sich im großvolumigen Einsatz bewährt haben.

Die Nutzung von KGs soll mehrere Probleme der Foundation-Modelle lösen oder reduzieren:

- **Falsche Antworten bei unzureichendem Wissen oder unzureichender Konfidenz:** Neuronale Modelle geben in solchen Situationen manchmal ganz falsche Antworten. Auch ohne direkt nach ihnen gefragt zu werden, können Modelle mitunter in Ausgaben auch Fakten behaupten, die nicht der Wahrheit entsprechen. Das gehört in den Bereich der sogenannten Halluzinationen.
- **Dynamik des Wissens:** Änderungen im Wissen oder gänzlich neue Fakten sind oft in den Trainingsdaten noch nicht repräsentiert oder im Vergleich zum überholten Wissen statistisch unterrepräsentiert.
- **Lücken im Detailwissen:** KGs enthalten auch Details, die wegen mangelnder allgemeiner Relevanz im frei zugreifbaren Internet nicht zu finden sind, z.B. gewisse Teile von Produktspezifikationen, Mitgliederlisten, Messwerte usw.

Zur Nutzung der Wissensrepositorien gibt es drei vielversprechende Ansätze:

- (1) Die Aufnahme der Wissensbestände in die Trainingsdaten (z.B. KELM)
- (2) Der Zugriff des Modells auf die Wissensbestände als Teil der Inferenz
- (3) Die Berücksichtigung von Wissen in großen Textkorpora durch Retrieval

Beim ersten Ansatz kann man die Wissens Elemente des KG, sogenannte RDF Triple, automatisch in Sätze einer natürlichen Sprache umwandeln, wobei sich wegen der Benennung in den KGs das Englische anbietet. Alternativ kann man die RDF-Triples aber auch in der RDF-Syntax wie Sätze einer eigenen Sprache zu den multilingualen Trainingsdaten hinzufügen.

Die Hinzunahme des Wissens zu den Trainingsdaten hat den Vorteil, dass sie bereits beim Pretraining das Weltwissen des Modells verbessern. Außerdem besteht die Hoffnung, dass die Projekte erlernen können, zwischen einfachen sprachlichen Aussagen und den entsprechenden Wissensgraph-Fragmenten zu übersetzen ähnlich der Übersetzung zwischen natürlichen Sprachen.

Beim zweiten Ansatz erlernt das Modell, durch SPARQL Queries auf benötigtes Wissen zuzugreifen und dieses für die Berechnung der endgültigen Beantwortung zu berücksichtigen. Dieser Ansatz hat den Vorteil, dass alles neue Wissen im KG unmittelbar ohne Nachtraining verfügbar wird.

Beide Ansätze werden gegenwärtig bereits im Projekt Open GPT-X getestet. Folgeprojekte werden in der Realisierung dieses Ziels auf Erkenntnissen und praktischen Resultaten dieses Vorläufer-Vorhabens aufbauen können.

Geeignete Kandidaten für Wissensrepositorien sind dabei DBpedia und WikiData zwei große generische Wissensgraphen, die von großen Wissensgemeinschaften (Knowledge Communities) gepflegt werden und kostenfrei nutzbar sind. In diesen Ressourcen ist allerdings das Wissen über die europäischen Wirtschaftsunternehmen nicht vollständig und auch nicht immer für alle Firmen aktuell. Ins Auge gefasst sollte daher auch die Einbeziehung von Spezialwissensquellen wie OpenCorporates oder alternative kommerzielle Angebote.

Ein dritter Ansatz sind Retrievalverfahren zur Nutzung zusätzlicher großer Textdatenbestände. Hierbei kann es sich um aktuelle Texte (z.B. Suchmaschinenergebnisse) handeln, die noch nicht in die Trainingsdaten eingeflossen sind. Zum anderen können das auch interne Daten (z.B. Servicereports) sein, die als Zusatzinformation für das Modell wertvoll sein können. Über ein embedding-basiertes Retriever-Reader-Modell können sie in die Answererzeugung einfließen und so Fehler in den erzeugten Texten reduzieren und aktuelle Informationen verwenden. Multilinguale Retriever-Reader-Modelle können dabei die Information in unterschiedlichen Sprachen nutzen.

6.4 Kombination von Sprache mit anderen Modi und Medien

Auf die Frage nach dem Faszinierenden an Foundation-Modelle bemerkte Prof. Andreas Dengel: „Von Forschungsseite ist natürlich Multimodalität sehr spannend.“ Als zusätzliche Medien kommen insbesondere Bilder, Videos, gesprochene Sprache, Audio, und 3D-Modelle in Betracht, die mit Text kombiniert werden können. Yann LeCun, Forschungschef von Meta, geht noch einen Schritt weiter: „Anstelle von Sprache oder Bildern wird die nächste KI-Generation jedoch direkt aus Videos lernen. Meta unternimmt derzeit große Anstrengungen, um Videodaten aus der Ich-Perspektive für diese neue KI-Generation zu sammeln, aber auch YouTube-Videos sind als Trainingsmaterial geeignet“ (Schreiner, 2022).

Die europäische Wirtschaft und Wissenschaft sollte diese Entwicklung ernst nehmen. Eine Implementierung und Training dieser Modelle für europäische Sprachen und Bildinhalte sowie ihre Erweiterung auf größere Dauer sind ein lohnendes Ziel für europäische Modelle.

Die erfolgreiche Nutzung der Modelle erfordert die Lösung einer Reihe von Problemen:

- Verlängerung der zeitlichen Dauer von Videos um etwa eine Größenordnung.
- Mögliche Verwendung einer autoregressiven Architektur.
- Effiziente Nutzung von bestehenden Modellen zur Einzelbilderzeugung (z.B. Stable Diffusion).
- In der Regel reicht ein einfacher Satz nicht mehr zur Spezifikation der Inhalte aus, sondern es muss eine Storyline mit mehreren Punkten angegeben und berücksichtigt werden.
- Neue Strategien um längere Eingabesequenzen verarbeiten zu können, z.B. nach dem Muster von S4.
- Erweiterte Ansätze zur temporalen und räumlichen Disaggregation bei der Videogenerierung durch Diffusions-Modelle.
- Einbeziehung weiterer Modalitäten, wie etwa gesprochene Sprache und Geräusche.
- Einbindung von existierenden Bildern und 3D-Modellen, welche mit den Video-Techniken animiert werden können.
- Gleichzeitige Behandlung verschiedener Objekte der gleichen Art.
- Berücksichtigung von Kamerabewegungen, Morphing und Szenenwechsel

Für synthetisch generierte Videos gibt es einen riesigen Anwendungsbereich:

- Bei Ausbildung und Lehre können Zusammenhänge direkt visualisiert werden. Ein Beispiel ist: „Zeige wie der indische Subkontinent das Himalayagebirge auffaltete.“
- Im Sprachunterricht können Szenen und Abläufe visualisiert werden, die dann die/der Schüler:in beschreiben muss.
- Bei der personalisierten Werbung für ein neues Produkt: „Zeige wie Karl Müller mit seinem neuen Elektromobil über den Gotthardpass fährt“.
- Anleitungen zum Gebrauch eines Produktes können „on the fly“ für eine neue Umgebung erstellt werden.
- Animationsfilme lassen sich auf kostengünstige Art und Weise produzieren.

Das Erlernen von Text-Video-Modellen erfordert in der Regel einen Datensatz von Videoclips mit manuell eingegebenen Untertiteln. Die Erstellung solcher Datensätze ist jedoch teuer und zeitaufwendig und daher nur schwer in großem Umfang möglich. Stattdessen lassen sich diese Modelle mit Hilfe von Videos mit natürlichsprachlichen Annotationen in Form von automatisch transkribierten Sprachdaten trainieren. Ein Beispiel ist der HowTo100M Datenbestand mit 136 Millionen Videoclips aus 1,22 Millionen kommentierten Lehrvideos im Internet (Miech et al., 2019), die Menschen bei der Ausführung und Beschreibung von über 23.000 verschiedenen visuellen Aufgaben zeigen. Allerdings hat dieses Vorgehen auch einige potentielle Nachteile. Einerseits sind die durch Spracherkennung erzeugten Texte nicht fehlerfrei und die zeitliche Zuordnung ist nicht perfekt. Zudem handelt es sich um die eingeschränkte Domäne von Lehrvideos. Die neueren Datenbestände sind umfangreicher und erfassen einen größeren Themenbereich. HD-VILA-100M enthält 100 Millionen Videos in 720p-Auflösung, welche gleichzeitig Audiodaten, Untertitel und Video-Frames enthalten und durchschnittlich 13.4 Sekunden dauern (Zellers et al., 2022). Die Autor:innen zeigen, dass Audiodaten den Trainingserfolg signifikant verbessern. (Nagrani et al., 2022) übertragen Untertitel aus Bild-Text-Daten auf Videoclips ohne zusätzlichen manuellen Aufwand. Mit dieser Pipeline erstellen sie unscharf annotierte Audio-Video-Daten mit Millionen von gepaarten Clips und Beschriftungen. Sie zeigen, dass mit diesen Daten sehr leistungsfähige Modelle zur Videosuche und Videountertitelung trainiert werden können.

Insgesamt gibt es auf Youtube, Shutterstock, Dreamstime und Reddit viele Millionen Videos, die mit Audio und teilweise mit Untertiteln verfügbar sind. Eine weitere wertvolle Ressource sind die Archive der Rundfunkanstalten, die für den barrierefreien Zugang zu ihren Sendungen routinemäßig Transkripte und Untertitel produzieren müssen.

SPOTLIGHT

Continental Automotive Technologies

An verschiedenen Stellen in der Studie stellen wir ausgewählte deutsche Startups und Unternehmen vor, die KI-Foundation-Modelle bereits nutzen. Sie berichten über ihre Anwendungsfälle, Herausforderungen und die Vorteile europäischer Modelle.



Dr. Corina Apachițe,
Head of AI, Continental
Automotive Technologies

Continental entwickelt wegweisende Technologien und Dienste für die nachhaltige und vernetzte Mobilität der Menschen und ihrer Güter. Das 1871 gegründete Technologieunternehmen bietet sichere, effiziente, intelligente und erschwingliche Lösungen für Fahrzeuge, Maschinen, Verkehr und Transport. Continental erzielte 2021 einen Umsatz von 33,8 Milliarden Euro und beschäftigt aktuell mehr als 190.000 Mitarbeiterinnen und Mitarbeiter in 58 Ländern und Märkten.

Wo setzt ihr Foundation-Modelle ein? Was ist euer Use-Case?

Wir setzen Foundation-Modelle insbesondere im Bereich „Natural Language Processing“, also der Analyse von Texten ein. Dort sind diese Modelle momentan das Maß aller Dinge. Unsere Anwendungen sind dabei die Analyse von „Requirements-Dokumenten zur Unterstützung unserer Entwickler, oder im Bereich „Conversational AI“, also generell bei Mensch-Maschine-Schnittstellen. Eine weitere Anwendung, die zukünftig eine Rolle im automatisierten Fahren spielen kann, ist die Analyse und Formalisierung von Verkehrsregeln mit Hilfe von Sprachmodellen. Im Bereich Bildverstehen oder Bildgenerierung können Foundation-Modelle auch für das Erstellen und Analysieren von Bild-Datensätzen verwendet werden, um beispielsweise Trainingsdaten für AI-Modelle automatisch zu annotieren.

Welchen Einfluss haben KI-Foundation-Modelle auf euer Geschäftsmodell bzw. eure Projekte?

Für qualitativ hochwertige Anwendungen im Bereich Sprachverstehen sind Foundation-Modelle unerlässlich. Somit ermöglichen uns erst die KI-Foundation-Modelle unsere Anwendungen. Im Bereich Textanalyse von Requirements-Dokumenten spart uns die Technologie erheblich Zeit, macht die Analyse einfacher und effizienter und entlastet damit unsere Mitarbeiter:innen. Die Nutzbarkeit von Chat-Bots hängt primär von der Qualität der KI-Modelle ab. Zukünftige Foundation-Modelle, wie beispielsweise multi-modale Varianten, die sowohl Text als auch Bilddaten verstehen, werden auch für eine Vielzahl weiterer Aufgaben interessant sein, die bisher nur schwer umzusetzen sind.

Welche Schwierigkeiten und Probleme siehst du, dass nur USA und China derzeit KI-Foundation-Modelle umfassend bereitstellen?

Generell haben KI-Modelle eine direkte Abhängigkeit von den Daten, mit denen sie trainiert wurden. Ihre Qualität und Funktion der Modelle hängen primär von diesen Daten ab. Somit müssen Modelle primär auf Daten der entsprechenden Regionen trainiert werden. Europäische Besonderheiten, wie beispielsweise die Sprachdiversität sowie regulatorische oder kulturelle Unterschiede im Straßenverkehr, werden von nicht-europäischen KI-Modellen möglicherweise vernachlässigt. Qualität und Funktionen verlangen also nach regionalen Lösungen.

Zudem sorgen die hohen datenschutzrechtlichen Standards in Europa dafür, dass sensible Daten kaum an amerikanische oder chinesische KI-Modelle gesendet werden können. Des Weiteren bleiben die Modelle geschlossene Systeme („Black Box“). Detailliertere Analysen oder Weiterentwicklungen der Modelle werden nicht ermöglicht.

Eine weitere Schwierigkeit ist die Abhängigkeit von den Besitzern der KI-Modelle in kommerzieller Hinsicht. Zugangshürden könnten aufgebaut oder spezielle Eigenschaften der Modelle designt werden, die nicht im Interesse europäischer Nutzer:innen sind. Und nicht zuletzt geht es auch um Standortnachteile durch Abwanderung von Talenten dorthin, wo die „besten“ Foundation-Modelle erstellt werden.

Wie würden euch europäische Modelle - Open Source, alle europäischen Sprachen abdeckend, mit hohen Datenschutzstandards und minimalem Bias - helfen?

Jeder dieser Aspekte wäre ein großer Vorteil für die Verwendung von KI-Foundation-Modellen. Die Veröffentlichung als Open Source würde die Untersuchung der Inhalte von Foundation-Modellen vollumfänglich ermöglichen. Es gäbe mehr Gestaltungsmöglichkeiten, gerade auch im Hinblick auf die hohe Diversität in Europa. Foundation-Modelle können angepasst und beliebig modifiziert werden, wie z.B. das Zuschneiden von Modellen auf regionale Besonderheiten, auf die Bedürfnisse verschiedener Industrien oder Anwendungsdomänen.

In hohen Datenschutzstandards sehen wir einen kompetitiven Vorteil, der zusammen mit Transparenz die Akzeptanz von Foundation-Modellen erheblich erhöhen wird.

6.5 Fragestellungen und Weiterentwicklungen

Gerade auch für multimodaler Modelle gilt die Interviewaussage von Prof. Dr. Wrobel: „Das größte Defizit ist das mangelnde Verständnis der Fähigkeiten und Grenzen.“ Dies wird von Prof. Schütze in seinem Interview unterstützt und konkretisiert: „Es wäre nochmal ein Paradigmenwechsel, wenn man andere Modalitäten vollumfänglich in die Modelle integrieren könnte und wenn man auch tatsächlich echtes Grounding berücksichtigen könnte.“

Gegenüber reinen Sprachmodellen weisen Foundation-Modelle, die Texte mit Bildern bzw. Videos kombinieren, noch eine zusätzliche Komplexitätsdimension auf. Daher sollten Ansätze entwickelt werden, um die Modellergebnisse zu erklären und die Zuverlässigkeit der Resultate für bestimmte Eingaben abzuschätzen.

Videos enthalten sehr viele unterschiedliche Informationstypen, die bei der Analyse und Generierung integriert werden können:

- OCR kann verwendet werden, um im Video sichtbare Schriftzüge zu erkennen. Dazu kann man z.B. ein Texterkennungsmodell integrieren, welches Texte oder Tokeneinbettungen liefert.
- Gesprochene Sprache lässt sich über eine Spracherkennung gewinnen, die Texte oder Tokeneinbettungen produziert.
- Gesichter und die zugehörigen Personen sind besonders wichtige Merkmale zum Verständnis eines Videos. Hier lassen sich Modelle zur Gesichtserkennung in die Video-Pipeline einbeziehen.
- Audio, zum Beispiel Motorengeräusche, können bei der Interpretation von Videos verwendet werden. Hier lässt sich beispielsweise ein Audio-Erkennungssystem nutzen, welches auf den YouTube-8M Daten trainiert wurde, bei denen Objekte in den Videos annotiert wurden.
- Szenen (z.B. Bäume, Berge, Friedhof) können über ein Szenen-Modell erkannt werden, welches mit dem PLACE365 trainiert wurde.
- Objekte, wie sie etwa im ImageNet annotiert sind.
- Bewegungen, die durch Modelle zur Aktionenerkennung erkannt werden können, welche mit den Kinetics-Daten trainiert werden können.

Ein erstes Modell in diese Richtung ist Merlot Reserve, das Audio, Untertitel und Video einbezieht (Zellers et al., 2022).

Die gleiche Szene kann auch durch mehrere Videos erfasst werden, z.B. binokulare Kameras oder mehrere Kameras mit unterschiedlichem Gesichtsfeld. Hier besteht die Aufgabe, diese Abläufe zu integrieren und daraus eine einheitliche Interpretation zu gewinnen. Eine derartige Anordnung erleichtert die Rekonstruktion der 3D-Szene.

Modelle zur Erzeugung von Videos können jedoch auch missbraucht werden, um beispielsweise gefälschte, hetzende, herabsetzende oder bösartige Inhalte zu erzeugen.

Diese Gefahren können z.B. durch interne Filter für die Eingabetexte und für die Prüfung der erzeugten Videoinhalte reduziert werden. Hier gibt es aber einen großen Forschungsbedarf, um die Modelle abzusichern und schwer erkennbare soziale Vorurteile und Stereotypen herauszufiltern. Google und Meta nennen diese Gefahren als Grund, warum sie ihre Modelle noch nicht freigeben.

Auch in diesem Bereich der KI-Foundation-Modelle zeichnet sich bereits eine amerikanische Dominanz ab. Europäische Entwickler:innen muss die Möglichkeit gegeben werden, Videodaten zu nutzen und auf deren Basis fortschrittliche KI-Foundation-Modelle zu entwickeln. Nur so kann sichergestellt werden, dass europäische Werte und Standards auch in diesen Modellen Beachtung finden. Dafür müssen Entwickler:innen und Forscher:innen bestmögliche Voraussetzungen haben.

6.6 Foundation-Modelle in anderen Datendomänen

Die vorgeschlagene Strategie der Modellentwicklung, die multilingualen Foundation-Sprachmodellen schrittweise um Wissensbestände und Multimodalität zu erweitern, bietet auch eine gute Ausgangsbasis für die Entwicklung von Foundation-Modellen in ganz anderen Daten- oder Wissenschaftsdomänen. Erfolgreiche Anwendungen von Modellen in Genetik und Proteomik zum Beispiel beruhen auf der Einsicht, dass auch in diesen Bereichen Muster in Symbolsequenzen gelernt werden können, die neue Erkenntnisse zu Eigenschaften, Wirkung oder Veränderung der Moleküle ermöglichen. Es ist sehr wahrscheinlich, dass sich in Chemie, Werkstoffwissenschaften, Biowissenschaften oder anderen Natur- und Ingenieurwissenschaften weitere solche Einsatzgebiete finden werden.

Große zukünftige Anwendungsbereiche der Technologie liegen in der Betriebswirtschaft und in allen anderen Bereichen der Gesellschaft, in denen Prozesse gestaltet, gesteuert, optimiert, automatisiert und überprüft werden müssen. Weil Prozesse sequentielle Abfolgen von Teilprozessen und Einzelhandlungen sind, liegt der Einsatz von Transformermodellen zum Entwurf, zur Verbesserung und zum Monitoring solcher Prozesse nahe. Durch die Digitalisierung der Unternehmen und Verwaltungen wird es hinreichend Daten zu den tatsächlichen Abläufen in Geschäfts-, Produktions- und Verwaltungsprozessen geben, aus denen Foundation-Modelle ein generelles Wissen über die Eigenschaften und Elemente dieser Prozesse erwerben können. Solch ein Ansatz zur Gewinnung von neuronalen Prozessmodellen wird wahrscheinlich sehr schnell multimodal werden, indem er die sprachlichen Benamungen und Beschreibungen von Prozesselementen sowie die Informationsobjekte der Prozesse wie Formulare, Datentransaktionen, Geolokationen und Transportwege in die Lerndaten einbezieht. GATO (Reed et al., 2022) ist ein erstes multimodales Modell, welches Sequenzen von Texten, Bildern und Messwerten verarbeiten und daraus Steuerungsstrategien ableiten kann. Es erzielte auf mehr als 600 Benchmarks gute Steuerungsergebnisse.

Das in dieser Studie diskutierte KI-Rechenzentrum würde den Einstieg der deutschen Wirtschaft und Forschung zu solchen neuen Modelltypen sehr erleichtern.

6.7 Zusammenfassung

Bis hierhin hat die Studie gezeigt, dass das wirtschaftliche Potential und die gesellschaftliche Relevanz von KI-Foundation-Modellen immens sind. Um aber das volle wirtschaftliche Potential auszuschöpfen, müssen europäische Entwickler:innen dazu befähigt werden, eigene Foundation-Modelle zu entwickeln. Ansonsten besteht die Gefahr, von amerikanischen Modellen abhängig zu werden oder die Modelle gar nicht zu nutzen. Beide Fälle bedeuten einen erheblichen Wettbewerbsnachteil für die Wirtschaft.

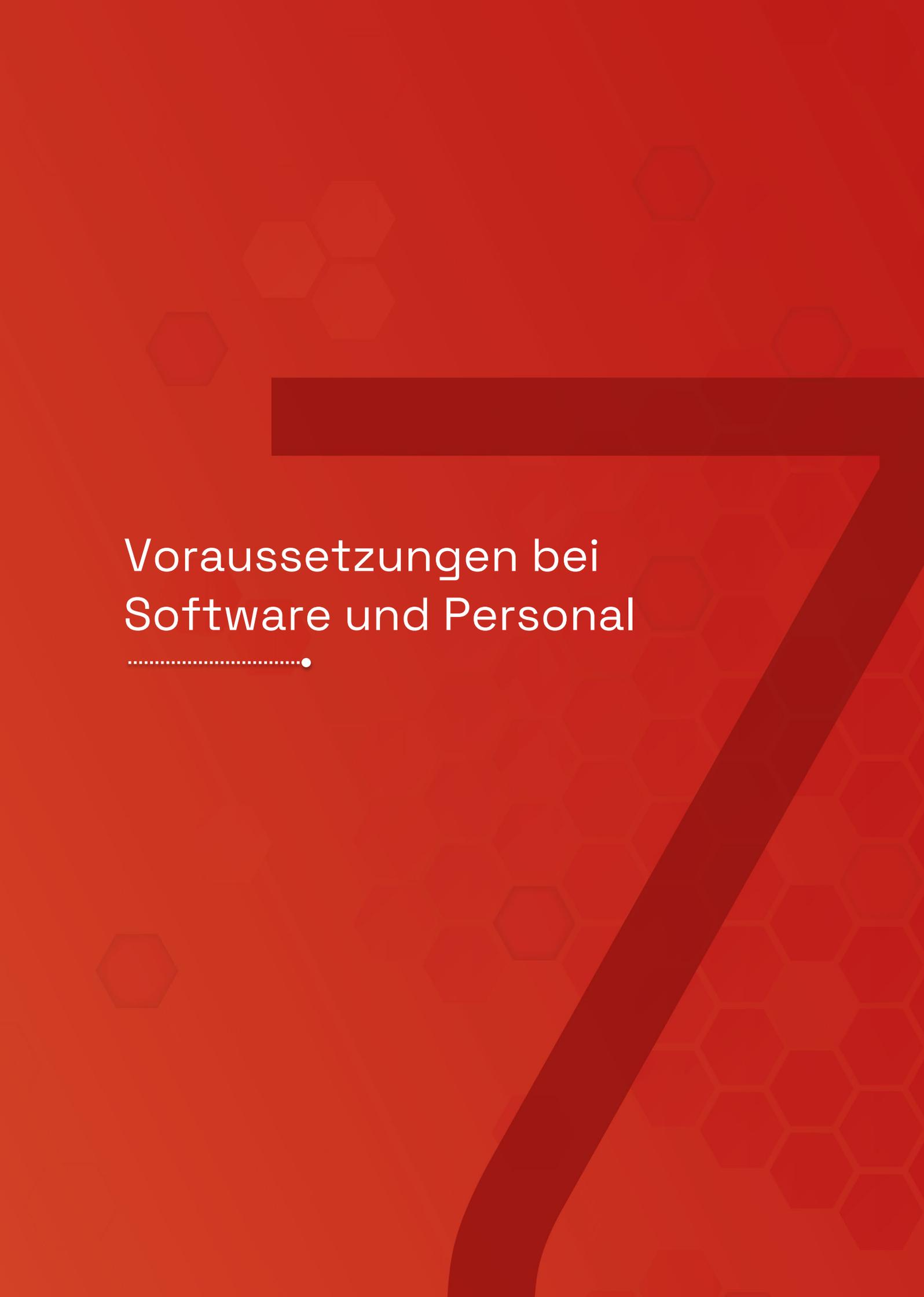
Um qualitativ hochwertige Foundation-Modelle zu trainieren, müssen Bias, Toxizität und Falschaussagen reduziert oder bestenfalls komplett ausgeschlossen werden. Auch wenn es hier bereits große Fortschritte gibt, hilft nur die Arbeit mit und die Forschung an KI-Foundation-Modellen, diesen Themenkomplex anzugehen. Die deutsche und die europäische Gesellschaft sollten alles daransetzen, diese Entwicklungen nicht den Amerikanern oder Chinesen zu überlassen.

Dabei sollte der Fokus zunächst auf Sprachmodellen liegen. Die Multilingualität Europas ist eine Herausforderung, aber zugleich auch eine große Chance für europäische Sprachmodelle. Eine adäquate Abbildung dieser Multilingualität in der Funktionalität der Foundation-Modelle erhöht deren Akzeptanz und den wirtschaftlichen Nutzen. Sie ist aber auch von hoher Relevanz für die gesamte Technologieentwicklung, denn der größte Teil der Welt ähnelt in Bezug auf Sprachenvielfalt eher Europa als den USA oder China.

Verschiedene Projekte bereiten bereits vielsprachige Datensätze auf und stellen diese zur Verfügung. Dahinter steht bereits die nächste Generation der Foundation-Modelle in den Startlöchern, welche z.B. Videos erzeugen könnten. Hier ist der Entwicklungsbedarf noch größer, das Potential immens und viele Fragestellungen weiter ungelöst.

Deutsche und europäische Entwickler:innen und Forscher:innen müssen jetzt befähigt werden, nach besten Standards an KI-Foundation-Modellen zu arbeiten und zu forschen. Der erste Schritt dafür sollte der Aufbau einer kompetitiven Infrastruktur für das Training an KI-Foundation-Modellen sein.

Wie das technisch gelingen kann, darin gibt nun das nächste Kapitel einen Einblick.

The background is a solid red color with a faint, repeating pattern of hexagons. A large, dark red, abstract shape resembling a stylized '7' or a thick diagonal line is positioned on the right side. A horizontal bar of the same dark red color is located above the main text.

Voraussetzungen bei Software und Personal

.....●

7. Voraussetzungen bei Software und Personal

Die in LEAM angedachte Entwicklung und das Training von KI-Foundation-Modellen (Bommasani et al., 2021) setzt eine integrierte und leistungsfähige Hard- und Software-Infrastruktur voraus, wie sie es bisher und in öffentlichen Investitionsplänen in Deutschland und der EU noch nicht gibt. Während sich diese Infrastruktur in ihrer Grundstruktur von etablierten Strukturen in High-Performance-Computing-Systemen (HPC) ableiten lässt, setzt die Arbeit mit KI-Foundation-Modellen besondere Voraussetzungen an deren Organisation und der dabei eingesetzten Software. Dieses Kapitel erklärt sowohl diese besonderen KI-Elemente und deren Zusammenwirken mit der HPC-Grundstruktur, als auch die dafür notwendigen Voraussetzungen und Investitionen, um LEAM möglich zu machen.

Hierfür ist es notwendig, gleichzeitig die technischen Voraussetzungen und Möglichkeiten der LEAM-Initiative wie auch ihrer Chance für das KI-Ökosystem und die deutsche und europäische Gesellschaft zu betrachten. Während in diesem Kapitel gezeigt werden kann, dass die Herausforderungen an LEAM auf der Software-Seite lösbar sind, werden auch ihre sozialen und ökonomischen Mehrwerte aufgezeigt. Ein Ziel von LEAM sollte es sein, das monolithisch simplifizierte Berufsbild der „IT-Fachkraft“ aufzubrechen, um die vielfältigen, spezialisierten und voneinander abhängigen Rollen, die für ein wettbewerbsfähiges KI-Ökosystem notwendig sind, zu differenzieren und zu stärken. Als technologisches, soziales und ökonomisches Leuchtturmprojekt kann LEAM dafür den richtigen Impuls setzen, damit Deutschland und Europa zu einem globalen KI-Wettbewerber avanciert, wenn die Chancen des KI-Ökosystems richtig eingeschätzt und genutzt werden.

In diesem Kapitel wird eine Variante für den Software-Stack eines KI-HPC-Systems und die dafür benötigten hochspezialisierten Berufe und Fähigkeiten skizziert. Hierbei ist es möglich, strukturell auf existierenden und praxisbewährten HPC-Systemen aufzubauen und um KI-spezifische Komponenten zu ergänzen.⁶ Besonderes Augenmaß wird darauf gelegt, den Stack mit Hilfe von Open-Source-Software (OSS) zu konstruieren, da das globale KI-Ökosystem einerseits auf nicht-proprietärer Software aufgebaut ist und andererseits so eine Abhängigkeit von Software-Konzernen vermieden werden kann. Dieser Aspekt der Unabhängigkeit fördert auch die Souveränität und Resilienz des KI-Ökosystems, da auf Zulieferer im Software-Bereich weitestgehend verzichtet werden kann, wie es z.B. auch die Digitalstrategie der Bundesregierung vorsieht (Bundesministerium für Digitales und Verkehr, 2022). In der tatsächlichen Implementierung können die Details abweichen, für diese Machbarkeitsstudie ist jedoch

⁶ Wir betrachten in diesem Kapitel bevorzugt zentralisierte und homogen organisierte Infrastrukturen, obwohl B. Yuan et al., (2022) prototypisch gezeigt haben, dass KI-Modelle auch dezentralisiert und heterogen auf vernetzten Rechnern trainiert werden können. Zentralisierte Infrastrukturen haben den Vorteil, dass der eingesetzte Software-Stack einfacher und weniger fehleranfällig gehalten werden kann, die Datenquellen zuverlässiger und schneller verfügbar sind sowie Nutzerrechte und Sicherheitsvorkehrungen (DMZ, VPN) schlichter zu handhaben sind. Darüber hinaus können mit einem zentralisierten KI-Hochleistungszentrum zuverlässigere Statistiken über Energieverbrauch und Effizienz erhoben werden.

entscheidend, dass LEAM auf der Software-Ebene mit bereits existierenden Technologien umgesetzt werden kann.

Bei der Betrachtung dieser Infrastruktur lassen sich vier aufeinander aufbauende Ebenen (Layer) unterscheiden, die mit einzelnen Modulen interagieren, um die Entwicklung und das Training von KI-Modellen möglich zu machen (Abb. 18). Diese Layer setzen jeweils verschiedene spezialisierte Berufe voraus, um deren komplexe Funktionalität bereitstellen zu können. Diese einzelnen Hard- und Software-Layer sind dabei jeweils für bestimmte Aufgaben optimiert, um dedizierte, rechenintensive Anwendungen in einem effizienten Rechenprozess zu organisieren. In der Grafik sind die für LEAM notwendigen und für die Arbeit mit KI-Foundation-Modellen besonderen KI-Elemente im Trainings- und Applikations-Layer gebündelt, welches auf den System- und Framework-Layern der High-Performance-Computing-Infrastruktur aufsitzt und von den Nutzer:innen durch ein Service-Layer angesteuert werden kann.

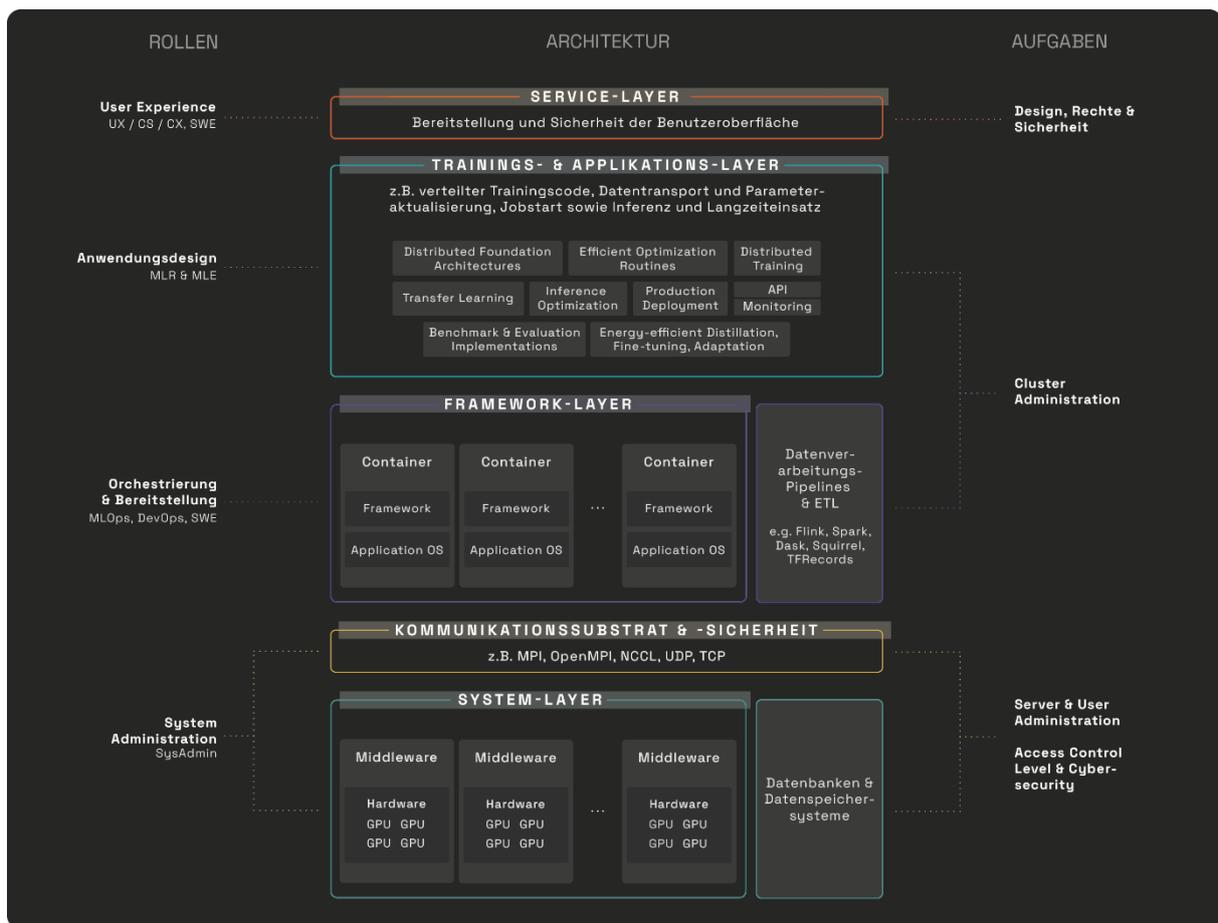


Abb. 18: Simplifizierte Darstellung der Hard- und Software-Infrastruktur von HPCs

Das System-Layer bildet dabei mit seinen Recheneinheiten ein Hardware-Fundament, welches durch das Framework-Layer in betriebsfähige Software-Systeme gebündelt wird. Das Trainings- und Applikations-Layer nutzt diese Systeme, um KI-Anwendungen durchzuführen, während es vom Data-Storage & -Loading-Modul mit Daten beliefert wird. Der zentrale Unterschied zum Aufbau von traditionellen HPC-Systemen liegt hier darin, dass für die Entwicklung von KI-Modellen anwendungsspezifische KI-Beschleuniger wie

z.B. GPUs, FPGAs, Cerebras-Chips und andere (siehe System-Layer) statt CPUs eingesetzt werden. Diese sind notwendig, damit die auf den System- und Framework-Layern aufbauenden, rechenintensiven Trainings- und Inferenzanwendungen innerhalb akzeptabler Zeiträume abgeschlossen werden können.⁷ Während CPUs für die Datenaufbereitung über eine Cloud-Lösung gemietet werden könnten, sollten die mehreren tausend GPUs und unterstützenden CPUs, die für den Betrieb von LEAM notwendig wären, lokal verfügbar sein. In Kapitel 8 wird für diese Machbarkeitsstudie mit einer Anzahl von 4480 GPUs kalkuliert.

Pre-Processing	Training	Tuning	Inference
10 - 20 k CPUs Cloud gemietet	3,5 - 4,5 k GPUs lokal verfügbar	30 - 50 GPUs Cloud / lokal	30 - 50 GPUs ...
Daten-Speicherplatz: mehrere PetaBytes			30 - 50 GPUs ...
			30 - 50 GPUs

Tabelle 5: Im Betrieb von LEAM werden für Training, Tuning und Inference Tausende GPUs benötigt.

LEAM ist für den Auf- und Ausbau des Innovationsstandorts Deutschland und Europa daher unerlässlich, um dieses komplexe und aufeinander abgestimmte Geflecht an Hard- und Software aufzubauen und der Wissenschaft und Wirtschaft zur Verfügung zu stellen. Diese massive Recheninfrastruktur ist für die Entwicklung von KI-Foundation-Modellen notwendig. So erfordert das Training großer KI-Modelle zum Beispiel eine Vielzahl leistungsfähiger und optimierter KI-Beschleuniger, auf denen sowohl Daten als auch Modelle parallel geschaltet werden, was eine deutliche Leistungssteigerung hiesiger Rechenzentren erfordert. Dabei ist die Wiederverwendbarkeit dieser rechenintensiven Modelle von zentraler Bedeutung, um die dabei verbrauchten Ressourcen und investierten Kosten amortisieren zu können.

Diese Leistungssteigerung ist durch die Erfüllung technologischer Voraussetzungen erreichbar, wie z.B. durch das effiziente und schnelle Laden und Speichern von Daten. Gleichzeitig ist sie aber auch von sozialen Veränderungen abhängig, wie z.B. die gezielte Anziehung von Fachkräften, um das Wachstum des KI-Ökosystems aufrechtzuerhalten. Hier sollte LEAM sowohl ein Anstoß und eine Inspiration sein, in Technologien zu investieren, die den Innovations- und den Wirtschaftsstandort Deutschland und Europa stärken sowie Wege zu finden, um Fachkräfte differenzierter und erfolgreicher zu umwerben und auszubilden.

⁷ Mit einer einzelnen NVIDIA V100 GPU braucht man 355 Jahre, um GPT-3 zu trainieren (Li, 2020). CPU-basierte Frameworks sind mindestens 5-10 mal langsamer als GPU-basierte Frameworks. Ohne den Einsatz von GPUs braucht es also mehrere tausend Jahre Rechenzeit, um GPT-3 zu trainieren.

Aus einer für diese Machbarkeitsstudie durchgeführten Umfrage mit 71 Teilnehmer:innen aus dem KI-Ökosystem Deutschlands ging hervor, dass die Berufe DevOps, MLOps, Machine Learning Engineer und Machine Learning Researcher für die Entwicklung und das Training von KI-Foundation-Modellen besonders wichtig sind. Des Weiteren braucht es für den Betrieb von LEAM Software Engineers und System-Administrator:innen, während Site Reliability Engineers unter den Befragten als weniger wichtig erachtet wurden. 56 % der Befragten gaben an, dass die Entwicklung eines KI-Foundation-Modells mehr als 25 Mio. Euro kosten würde, während 37 % davon ausgehen, dass sich diese Kosten innerhalb von fünf bis zehn Jahren amortisieren.

Dieses Kapitel der Machbarkeitsstudien zeigt also auf, dass die Herausforderungen an Software für die Entwicklung von KI-Foundation-Modellen in LEAM mit den heute zur Verfügung stehenden Technologien und Mitteln bereits lösbar sind, wenn sich Deutschland und Europa bereit zeigen, in die Integration von LEAM zu investieren. Durch den Einsatz von Open-Source-Software in Verbindung mit wenigen, ausgewählten proprietären Applikationen lassen sich auch hierzulande große KI-Modelle entwickeln und trainieren, obgleich die dafür notwendige Hardware zum größten Teil aus den USA und dem nicht-europäischen Ausland stammt.

Eine der größten Herausforderungen und Chancen für LEAM stellen die dafür notwendigen spezialisierten und umworbene Fachkräfte dar. Für das Training der in LEAM vorgesehenen KI-Foundation-Modelle wird ein Team von etwa 20 dedizierten Expert:innen entlang der oben beschriebenen Rollen vorausgesetzt. Während das KI-Ökosystem in Deutschland und Europa bereits heute von hochqualifizierten Expert:innen betrieben wird, übersteigt die Nachfrage das Angebot noch deutlich (Streim, 2022). LEAM kann dazu beitragen, die Ausbildungsqualität und die Attraktivität des hiesigen KI-Ökosystems zu verbessern, und vor allem bewirken, dass KI-Fachkräfte in Deutschland und Europa bleiben wollen, weil sie hier mit LEAM Voraussetzungen wiederfinden, wie sie es derzeit nur außerhalb Europas gibt.

7.1 Applikations-Layer: Trainings- & Inference-Technologien

Das Training und die Entwicklung von großen KI-Foundation-Modellen bergen im Vergleich zu klassischen, verteilten Systemen ohne naiven Parallelismus einige Besonderheiten: Deep-Learning-basierte KI-Entwicklungen sind für die Konzeption und das Training von KI-Foundation-Modellen einzigartig befähigt, weil deren Algorithmen parallel geschaltet werden können. Das bedeutet, dass eine Vielzahl von Datenpunkten gleichzeitig verarbeitet werden, anstatt sie nacheinander, also sequentiell, zu bearbeiten. Diese Eigenschaft bildet einen wesentlichen Vorteil gegenüber klassischen HPC-Anwendungen und erlaubt es, spezifische algorithmische Vorteile auszunutzen, um das Training großer KI-Foundation-Modelle überhaupt erst in endlicher Zeit zu ermöglichen.

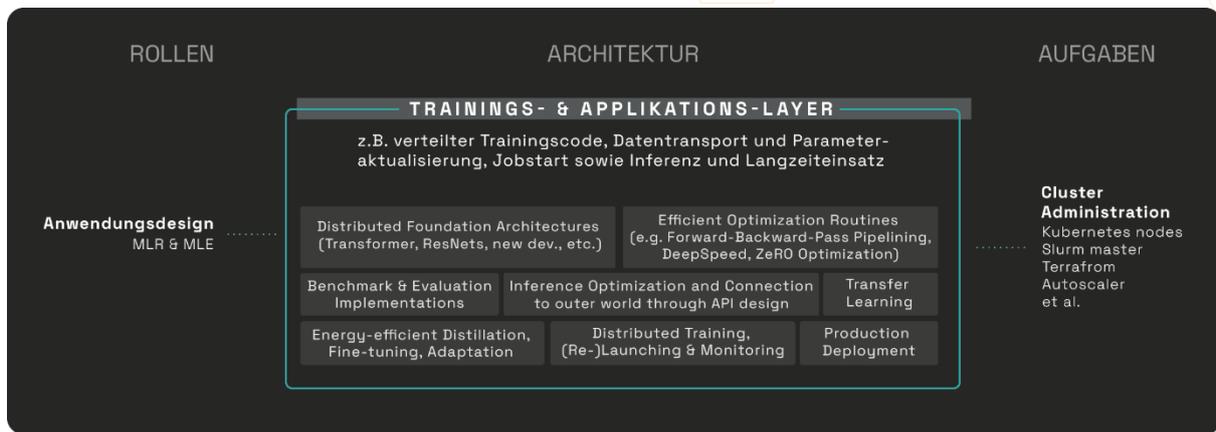


Abb. 19: Die Architektur des Trainings- & Applikations-Layers im Detail

Daten- & Modell-Parallelismus

Diese Eigenschaften sind im „Daten-“ bzw. „Modell-Parallelismus“ zusammengefasst und erklären, warum dafür eine massive Recheninfrastruktur notwendig ist (Hugging Face, o.D.). Während im Daten-Parallelismus Trainingsdaten zu gleichen Anteilen auf die KI-Beschleuniger zwar aufgeteilt werden kann, ist die Anzahl der zu berechnenden Parameter in KI-Foundation-Modellen so groß, dass die Modelle selbst im Modell-Parallelismus⁸ auf mehrere Beschleuniger aufgeteilt werden müssen. Das bedeutet, dass nicht jeder Beschleuniger eine identische Kopie des Modells parallel berechnet und deren Ergebnisse iterativ an alle parallelen Beschleuniger übermittelt, sondern die zu berechnenden Daten sequentiell von einem Beschleuniger zum nächsten weitergegeben werden müssen. Damit man dabei die Modellparameter aktualisieren kann, überträgt der letzte Beschleuniger in einer Reihe das Fehlersignal (Loss) wieder an den ersten Beschleuniger zurück. Daten- und Modell-Parallelismus⁹ können gleichzeitig verwendet werden, was die Komplexität der Entwicklung weiter erhöht.

Bei solch einer zirkulären Sequenzierung bleiben allerdings die meisten KI-Beschleuniger ungenutzt, wenn diese gerade keine Operation in der Reihenschaltung durchführen. Diese Stillzeit kann im Modell-Parallelismus über raffiniertes Programmieren (Kosson et al., 2021) und durch Optimierungssysteme minimiert werden, die eine höchstmögliche Auslastung aller Beschleuniger gewährleisten (s. Kapitel 8.3). Diese Voraussetzungen machen das Programmieren und das Betreuen des Trainings von KI-Foundation-Modellen kompliziert und herausfordernd (Bommasani et al., 2021).

⁸ Narayanan et al. (2021) beschreiben Modell- bzw. Pipeline-Parallelismus im Detail.

⁹ Mudigere et al. (2022) beschreiben Best Practices im Training großer Recommender-Modelle.

Trainingsmanagement, Evaluation & Benchmarking

Im Applikations-Layer wird auch das zuverlässige Management von einzelnen Trainingsjobs gesteuert, um bei einem Hardware-Versagen keine Resultate zu verlieren. Anders als bei dem Training von kleineren KI-Modellen kommt es beim Training von KI-Foundation-Modellen häufig zu einem Ausfall einzelner KI-Beschleuniger.¹⁰ Ein zuverlässiges Launch- und Relaunch-System beugt dabei durch fortlaufendes Monitoring Ergebnis-Verluste vor, um von derselben Stelle aus weiter zu trainieren, bei der der Fehler aufgetreten ist.

Des Weiteren werden auch die Evaluation und das Benchmarking von KI-Modellen in diesem Layer implementiert. Dazu gehören sowohl das Monitoring während des Trainingsprozesses, als auch das Testen der Modelle an anwendungsrelevanten Datensätzen, um deren Performanz festzustellen und diese später in Data & Model Card-Dokumentationen überführen zu können (Pushkarna et al., 2022). Für das Training und die Evaluation ist auch eine Anbindung an die ETL-Infrastruktur¹¹ notwendig, die hier implementiert, getestet und versioniert wird.

Wiederverwendbarkeit in Deployment-Infrastruktur

Aufgrund des rechenintensiven Trainings von KI-Foundation-Modellen steht die Wiederverwendbarkeit von kostspielig trainierten KI-Modellen im Mittelpunkt, um die dafür notwendigen Ressourcen und Kosten zu amortisieren. Dabei werden u.a. Destillierungsmechanismen, Adaption- oder Finetuning-Verfahren eingesetzt (s. Kapitel 2), welche wiederum anwendungsrelevante Benchmarks voraussetzen.

Dabei spielt die Entwicklung und Betreuung einer optimalen Deployment-Infrastruktur eine zentrale Rolle, da so große Effizienzsteigerungen erreicht werden können. Hier muss auf eine Implementierung Wert gelegt werden, um die KI-Foundation-Modelle bestmöglich mit der darunterliegenden Hardware betreiben zu können und externe Anfragen per API schnell und zeitnah zu beantworten.¹² Es muss hierbei auf komplexe Produktionsaspekte geachtet werden, um einen reibungslosen Prozess auch unter Stress zu gewährleisten. So sollte die Deployment-Infrastruktur z.B. resilient gegenüber einer Häufung von Anfragen sein, mit einer stabilen API ausgestattet sein sowie robuste Zugangsbeschränkungen und Sicherheitschecks beinhalten.

¹⁰ Dies ist ein bekanntes Phänomen aus der Datacenter-Branche: Je mehr Festplatten betrieben werden, desto häufiger fallen diese aus. Gleiches gilt dementsprechend auch für HPC-Beschleuniger.

¹¹ ETL steht für Extract-Transform-Load und beschreibt die Aggregation von Rohdaten aus einer Produktionsdatenbank in ein Format, das zur Analyse der Daten verwendet werden kann.

¹² Beispielfhaft können hier Freeware-Bibliotheken wie TensorRT und Triton genannt werden, welche von NVIDIA zur Verfügung gestellt werden.

Ressourceneffizienz des Trainings

Das ressourcenintensive Training von KI-Foundation-Modellen kann zu einem Problem werden, wenn der Trainingsvorgang nicht kontinuierlich kontrolliert wird und dabei der Energieverbrauch mit geeigneten Maßnahmen verringert wird. Die Nachhaltigkeit muss also bei der Konzeption eines LEAM-KI-Hochleistungsrechenzentrum und beim Entwickeln und Trainieren von KI-Modellen von Beginn an mitgedacht werden.¹³

Die in der KI-Entwicklung verbrauchten Ressourcen können auf dem Software-Level durch die regelmäßige und fortlaufende Messung von CO₂-Äquivalenten festgehalten werden. Dabei werden die gemessenen CO₂-Äquivalenten an drei verschiedenen Komponenten kontrolliert, um sie der richtigen Quelle zuzuordnen (Hintemann, 2020).

Während des Trainings und der Inferenz benötigen die KI-Beschleuniger Energie zum Ausführen der Rechenoperationen, sogenannter MACs.¹⁴ Diese lassen sich auf dem Applikations- und Framework-Level feststellen (Bannour et al., 2021; CodeCarbon, 2020). Aufgrund der hohen Datenmenge verbraucht der Data-Storage & -Loading-Layer für das Bereitstellen, die Verarbeitung und den Transport von Daten ebenfalls einen signifikanten Anteil des Energieverbrauchs. Die zuverlässige Kühlung des Gesamtsystems und andere, kleinere Komponenten benötigen darüber hinaus eine ununterbrochene Stromzufuhr. Wie das Hochleistungszentrum an sich ressourcenarm/klimaneutral konzipiert werden kann wird im Kapitel 8.6 erläutert.

Die verschiedenen Messungen können direkt an den Service-Layer übergeben werden, um den Nutzer:innen die Einsicht und Kontrolle über den Energieverbrauchs des Gesamtsystems und des Trainingsprozesses zu gewährleisten. Über Warnsignale kann ein unkontrollierter Mehrverbrauch verhindert werden. Solche Kontroll-Angaben werden teilweise bereits heute von Hyperscalern zur Verfügung gestellt.

Voraussetzungen

Die Lingua Franca für die Entwicklung moderner KI-Systeme ist Python (van Rossum, 1995), wobei KI-Anwendungen zunehmend auch in anderen Programmiersprachen entwickelt werden. Für die Entwicklung des Service-Layers kommen viele Sprachen in Frage, die hier nicht aufgelistet werden sollen. Die Softwarebibliotheken, die beim Trainieren und der Inferenz zum Einsatz kommen, sind großteils, wenn nicht gänzlich, durch Open-Source-Software abdeckbar. Beispielhaft sind hier auf dem Framework-Level PyTorch (Paszke et al., 2019), TensorFlow (Abadi et al., 2016), Keras (*Keras*, 2015/2022), und Jax (Frostig et al., 2018) zu nennen.

Bei der Orchestrierung nutzt der Applikations-Layer teilweise dieselbe Software wie der Framework-Layer, bspw. Ray (Moritz et al., 2018), Slurm (Yoo et al., 2003), kubeflow (*Kubeflow*, 2017/2022), hydra (*Hydra*, 2019/2022), Abseil (*Abseil Python Common Libraries*, 2017/2022) und andere (s. Kapitel 7.4). Weiterhin werden auch numerische Bibliotheken

¹³ Vgl. bspw. (*CSTB Releases Report Fostering Responsible Computing Research*, 2022; Patterson et al., 2022) und die darin enthaltenen Referenzen.

¹⁴ MAC steht für Multiply-Add-Compute, der zentralen Rechenoperation von Computerprozessoren.

wie BLAS („An Updated Set of Basic Linear Algebra Subprograms (BLAS)“, 2002), NumPy (Harris et al., 2020) und andere Verwendung finden. Beim ML-spezifischen Monitoring kann auf Bibliotheken wie AimStack (Arakelyan et al., 2020) oder MLFlow (*MLflow*, 2018/2022) zurückgegriffen werden. Diese Liste ist ausschließlich exemplarisch und dient lediglich der Darstellung der Wichtigkeit von Open-Source-Software für die Entwicklung moderner KI-Applikationen.

Die Aufgaben im Applikations-Layer erfordern eine hohe Spezialisierung, welche üblicherweise von Informatiker:innen und Software-Entwickler:innen in geteilten Rollen übernommen wird. Diese kann man in vier Rollenprofile unterscheiden: Machine Learning Researcher konzipieren und entwickeln neue Algorithmen, während sich Machine Learning Engineers mit der Optimierung des Trainings- und Inferenzcodes und der Implementierung des verteilten Lernens beschäftigen. Zusammen bilden sie ein Team, um KI-Modelle zu skalieren.

Full-Stack und Backend-Ingenieur:innen entwickeln den Service-Layer und arbeiten u.A. mit Systemadministrator:innen und Dev-Ops-Expert:innen zusammen, um die Benutzeroberfläche mit den verschiedenen Funktionalitäten des Gesamtsystems zu koppeln. Data Engineers und Data Scientists arbeiten an dem Data-Storage & Data-Loading-Layer und seiner Infrastruktur. Das beinhaltet das Design der Datenbanken oder anderer Speichereinheiten, die Entwicklung der ETL-Pipelines sowie das Kontrollieren der Daten-Lade-Pipelines und das korrekte Design der Trainings-Datensätze.

SPOTLIGHT

Fyrfeed GmbH

An verschiedenen Stellen in der Studie stellen wir ausgewählte deutsche Startups und Unternehmen vor, die KI-Foundation-Modelle bereits nutzen. Sie berichten über ihre Anwendungsfälle, Herausforderungen und die Vorteile europäischer Modelle.



Fyrfeed-Gründerteam:
Ehud Alexander Avner,
Dr. Thomas Lindemann,
Benjamin Zengler

Fyrfeed erstellt Content für B2B-Unternehmen durch eine Kombination aus Fachautoren und Künstlicher Intelligenz.

Kunden müssen dadurch bloß 5 Minuten pro Monat investieren und sparen 80% der Kosten gegenüber Agenturen.

Wo setzt ihr Foundation-Modelle ein? Was ist euer Use-Case?

Wir setzen, neben anderen Technologien und Tools, auch große, allgemeine Sprachmodelle ein, um Fachautoren bei der Erstellung von hochwertigem Content – z. B. Beiträgen für soziale Medien, Blogartikeln oder Whitepapers – zu unterstützen.

Dieser Ansatz, bei dem Mensch und KI zusammenarbeiten, nennt sich Human-in-the-Loop.

Welchen Einfluss haben KI-Foundation-Modelle auf euer Geschäftsmodell bzw. eure Projekte?

KI-Foundation-Modelle ermöglichen uns, qualitativ hochwertige Texte schnell und kostengünstig anzubieten. Statt Modelle von Grund auf selbst neu trainieren zu müssen, was mit erheblichem Kosten- und Zeitaufwand verbunden ist, können wir allgemeine, vortrainierte Modelle sofort einsetzen. Dadurch werden Weiterentwicklungen des Produkts sowie das Testen neuer Anwendungsmöglichkeiten um ein Vielfaches einfacher.

Wie würden euch europäische Modelle - Open Source, alle europäischen Sprachen abdeckend, mit hohen Datenschutzstandards und minimalem Bias - helfen?

Da wir Kunden aus der ganzen Welt bedienen, ist es denkbar, dass wir für verschiedene Sprachen verschiedene Modelle einsetzen. Obwohl bestehende State-of-the-Art-Sprachmodelle multilingual einsetzbar sind (und dies tun wir auch sehr erfolgreich), kann die Qualität zwischen Sprachen (und Fachbereichen) variieren. Modelle, die auf europäische (und außereuropäische) Sprachen spezialisiert sind – und gleichzeitig auch alle Vorteile von Open Source anbieten – wären ein großer Gewinn.

7.2 Data-Storage & -Loading-Layer

Der Erfolg vieler KI-Foundation-Modelle hängt wesentlich von einem effizienten und schnellen Aufbereiten der zu ladenden Daten ab. Sowohl in der ETL als auch während des Trainings der Modelle ist Effizienz ausschlaggebend. Damit kommt dem Data-Storage & Loading-Layer eine besondere Bedeutung zu, wie sie es so in anderen HPC-Systemen oft nicht gibt.

Beim Training von KI-Foundation-Modellen wird das Laden der Daten mithilfe eines mehrschichtigen Prozesses beschleunigt: Der ETL-Vorgang sowie das Filtern und Aufbereiten der Rohdaten werden sequentiell ausgeführt, sodass diese möglichst latenzarm geladen und im Daten-Parallelismus verteilt werden können. Für ETL werden dabei häufig große, klassische CPU-Server eingesetzt. Diese müssen in der Lage sein, Peta- und Exabyte an Daten zu verarbeiten, z.B. beim Scrapen von Webseiten und bei Map-Reduce-Jobs (Dean & Ghemawat, 2004) und ähnlichen, massiven Vorgängen. Dabei kommen spezialisierte Tools¹⁵ in der Daten-Infrastruktur zum Einsatz, wie z.B. Spark (*Apache Spark*, 2014/2022), Flink (*Apache Flink*, 2014/2022), oder Dask (*Dask*, 2015/2022).

Die per ETL bereitgestellten Rohdaten werden dann entweder in einer Datenbank oder einem Cloud-basierten Storage-Layer abgelegt. Dabei ist es möglich, aus diesen Rohdaten einen vorgefilterten Datensatz zu generieren und ebenfalls lokal oder in der Cloud abzulegen. Für das Vorfiltern werden ähnliche Tools und Ressourcen wie auch für ETL benötigt, diese sind also bereits vorhanden. Diese Art der Datenverarbeitung beschleunigt später den Trainingsprozess, setzt aber einen erhöhten Speicheraufwand voraus.

Während des Trainings von KI-Foundation-Modellen ist es wichtig, die Daten effizient und möglichst ohne Redundanzen zum KI-Beschleuniger zu transportieren. Dazu ist nicht nur eine entsprechende Netzwerk-Architektur notwendig (s. Kapitel 8), sondern auch ein dedizierter Software-Stack, welcher für das Laden von Daten in verteilten Systemen optimiert wurde. Dabei ist essentiell, dass Daten aufgrund ihrer Größe nicht auf einzelne Rechner geladen und gespeichert werden können, sondern durch ein effizientes Streaming-System bereitgestellt werden müssen, was die Entwicklung deutlich erschwert.

Dabei werden die Daten periodisch immer wieder in zufälliger Reihenfolge gestreamt (Nguyen et al., 2022), wenn sie von variierenden KI-Beschleunigern angefragt werden. Der Software-Stack muss in der Lage sein, sowohl von lokalen Datenbanken oder Festplatten, als auch von unterschiedlichen Cloud-Storage-Systemen lesen zu können, um etwaigen Benutzeranforderungen entsprechen zu können.¹⁶ Solche Systeme werden z.B. von Amazon, Google und Microsoft vertrieben oder werden von Konzernen privat betrieben. Ebenso müssen beim Laden Filterfunktionen, Datenaugmentierungen und -

¹⁵ Diese und folgende Listen haben keinen Anspruch auf Vollständigkeit und es werden nur exemplarische Elemente jeder Kategorie erwähnt. Die genaue Wahl der Tools hängt am Ende von den Implementierungsdetails ab und soll zu diesem Zeitpunkt nicht festgelegt werden

¹⁶ Erwartungen an die Datenhoheit sind erfahrungsgemäß sehr heterogen. Um eine breite Akzeptanz und Benutzung zu erreichen, muss das System also mit möglichst vielen Szenarien kompatibel sein.

transformationen unterstützt werden, wie es beim Training von Deep Neural Networks üblich ist. Dies ist erforderlich, um Modelle robust und generalisierbar zu trainieren. Der Software-Stack muss auch dazu fähig sein, mehrere Datensätze miteinander zu kombinieren, um neue ETL-Jobs zu vermeiden und somit Ressourcen zu sparen.

Der Data-Storage- & Loading-Layer muss auch mit lokalen Benutzerrechten sowie mit geltendem Datenrecht kompatibel sein. Als Bibliothek muss die Daten-Infrastruktur Access Control Levels (ACL¹⁷) berücksichtigen, um den Zugang zu Daten abzusichern und so (un-)beabsichtigte Zugriffe auf Daten von Dritten zu vermeiden. Dies wird über eine Schnittstelle zum Service-Layer gesteuert, in dem die Benutzerverwaltung organisiert ist.

Internationales Datenrecht sieht vor, dass das Speichern und Verarbeiten von Daten streng regulierten Praktiken zugrunde liegen muss, wenn es die Verarbeitung von personenbezogenen Daten betrifft, wie es z.B. in der Datenschutzgrundverordnung geregelt ist (*Data Protection in the EU*, o.D.). Darüber hinaus sind in verschiedenen geographischen Räumen auch bestimmte Zertifikate verpflichtend oder werden vom Markt erwartet, wie etwa SOC-2/SOC-3 (*System and Organization Controls*, o.D.) in den USA und Nordamerika oder ISO 27001 (*ISO - ISO/IEC 27001 and Related Standards — Information Security Management*, o.D.) bzw. ISO 27017/27018 (*ISO 27017 and ISO 27018 Certification | DEKRA*, o.D.) im europäischen Raum. Daten mit bestimmten Eigenschaften, bspw. Informationen über Gesundheit oder Kreditwürdigkeit unterliegen weiteren, lokal regulierten Bestimmungen, wie etwa die in den USA geltenden Regulierungen HIPAA oder FCRA.

Darüber hinaus kann es auch notwendig sein, dass der Daten-Layer innerhalb einer Demilitarized Zone (DMZ) oder in einem Virtual Private Network (VPN) auf die Daten zugreifen kann, falls die Daten nicht innerhalb des Hochleistungszentrums temporär abgelegt werden können. Dies sollte allerdings in der Praxis vermieden werden, da die damit verbundene Latenzzeit die Geschwindigkeit beim Laden stark beeinträchtigen kann.

Voraussetzungen

Für die Entwicklung und Betreuung der Daten-Systeme sind besondere Fähigkeiten von Data Engineers und Machine Learning Engineers notwendig, um das ETL und das Stream-basierte, verteilte Laden der Daten auf die Beschleunigerknoten zu gewährleisten. Eingesetzte Software sind z.B. Spark, Flink, oder Dask für ETL oder torchdatasets (Maszke, 2019/2022), Squirrel (Sohofi et al., 2022/2022), Deep Lake (*Deep Lake*, 2019/2022), ffcv (FFCV, 2021/2022) und andere (Ofeidis et al., 2022) für das verteilte Laden der Daten. Es handelt sich hier in der Regel um frei verfügbare Open-Source-Software, deren Beschaffung keine Herausforderung darstellt. Zum Speichern der Daten können klassische verteilte Datenbanken wie HDFS (*Apache Hadoop*, 2014/2022) oder GlusterFS (GlusterFS, 2011/2022), Cloud-Storage Systeme, ähnlich Google Cloud Storage, Azure Blob Storage oder auch NAS-Systeme verwendet werden.

¹⁷ Man unterscheidet hierbei zwischen Read, Write, Execute und Discoverable. Also die Erlaubnis, Daten zu lesen, zu schreiben, Anwendungen auszuführen oder - weniger bekannt - die Fähigkeit, Informationen über das Vorhandensein von Daten zu finden, ohne diese jedoch lesen zu dürfen.

7.3 System-Layer

Im System-Layer werden die tatsächlichen KI-Beschleuniger und die für deren Einsatz notwendige Hardware-Struktur organisiert, um hochleistungsfähige und verlustarme Rechenleistung zur Verfügung zu stellen. KI-Beschleuniger („AI Accelerator“, 2022) sind typischerweise spezialisierte Prozessoren, welche für die Entwicklung und das Training von KI-Modellen optimiert wurden (Reuther et al., 2022). Der Einsatz dieser Prozessoren ist unerlässlich, um den zeitlichen Rechenaufwand für KI-Foundation-Modelle zu beherrschen sowie wettbewerbsfähig zu sein (Khan & Mann, 2020). Weltweit führend in der Herstellung von High-Performance-Prozessoren sind die US-amerikanischen Firmen NVIDIA (Campa et al., 2020), AMD (AMD, 2021) und Intel (Intel, 2022), aber auch Start-ups wie z.B. Tachyum (Tachyum, 2022). Edge- und Embedded-Systems, wie z.B. NVIDIA Jetsons oder FPGAs, kommen hier nicht in Betracht, da sie nicht für das Training von KI-Foundation-Modellen geeignet sind.

Diese Prozessoren werden mithilfe spezialisierter Bibliotheken gesteuert, die oft von Hardware-Herstellern selbst entwickelt werden. Beispielhaft zu nennen sind hier CUDA (NVIDIA Developer, 2013) und ROCm (AMD, o.D.), welche in der Regel als Freeware¹⁸ mit der Hardware zur Verfügung gestellt werden. Das darüber liegende Server-Betriebssystem steuert die physikalische Hardware mittels dieser Bibliotheken an, wobei fast immer Linux-Betriebssysteme, wie Distributionen der Debian- (Ubuntu), Redhat- (Fedora, CentOS) oder Arch-Familien, eingesetzt werden (Joseph et al., 2022). Die einzelnen Server-Einheiten, welche mehrere KI-Beschleuniger bündeln, werden dann mittels entsprechender Middleware in das Kommunikations-Substrat eingebunden, wie es in Abbildung 20 näher beschrieben ist.

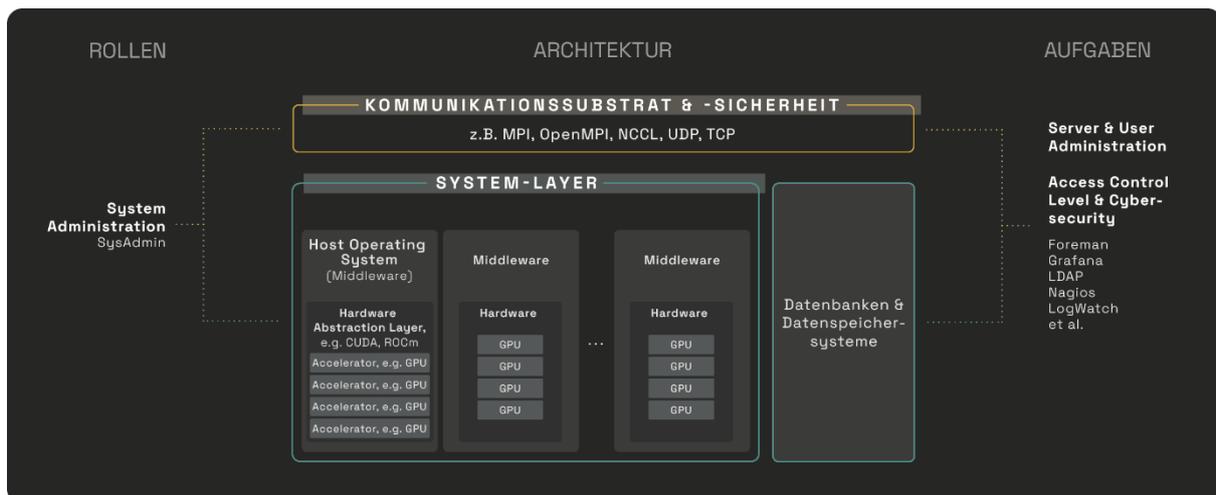


Abb. 20: Die Architektur des System- und Data-Storage & Loading-Layers im Detail

¹⁸ Im Vergleich zu OSS stehen bei Freeware-Bibliotheken nur die kompilierten Binärdateien zur Verfügung und nicht der gesamte, menschenlesbare Quellcode.

Kommunikations-Substrat

Das Kommunikations-Substrat dient einerseits der nahtlosen Verbindung von System- und Framework-Layer, aber auch dazu den KI-Beschleunigern Datenbanken und Datenspeicher zur Verfügung zu stellen. Dabei kann das Kommunikations-Substrat mehrere Server-Einheiten miteinander verknüpfen, um große Datenmengen effizient verarbeiten zu können. Die dabei für KI-Anwendungen typischerweise eingesetzten Kommunikationsbibliotheken basieren auf spezialisierten Technologien, wie MPI („Message Passing Interface“, 2022), OpenMPI („Open MPI“, 2022) oder NCCL (NVIDIA, o.D.).

Die genaue Topologie des Substrates ist durch die tatsächliche, physikalische Verkabelung der Hardware gegeben und muss beim Bau eines Hochleistungszentrums mitgedacht werden. Dabei muss bereits zu Anfang klargestellt werden, welche Anwendungen die KI-Beschleuniger berechnen werden.

Voraussetzungen

Für die Konzeption, den Aufbau und die Betreuung des System-Layers und des Kommunikations-Substrats bedarf es als Systemadministrator:innen ausgebildete Fachkräfte. Für die Betreuung wird spezialisierte Kontroll-Software eingesetzt, welche aus proprietärer und Open-Source-Software (OSS) bestehen kann. Generell besteht die Hardware- und Software-Infrastruktur für KI-Hochleistungsrechenanwendungen überwiegend aus OSS, was deren Bedeutung für das KI-Ökosystem unterstreicht (Sonnenburg et al., 2007).

Darüber hinaus übernehmen die Systemadministrator:innen auch die Rechtevergabe und Kontrolle der Nutzer:innen, bspw. über entsprechende Access Control Levels mittels LDAP- oder AD-Systeme, aber auch die Serveradministration und das -monitoring, um die dauerhafte Gesundheit des Systems zu gewährleisten. Des Weiteren fällt auch die Administration von DMZs und VPNs sowie Cyber-Sicherheits-Vorkehrungen in den Aufgabenbereich der Systemadministrator:innen.

7.4 Framework- & Service-Layer

Framework-Layer

In dem über dem Kommunikations-Substrat liegenden Framework-Layer werden die spezialisierten KI-Technologien eingesetzt (Abb. 21). Dazu werden die im System-Layer als Server-Einheiten zusammengefassten KI-Beschleuniger nochmals durch Containerlösungen wie Docker (Merkel, 2014) abstrahiert, um die Umgebung auf die entsprechenden KI-Systeme zu normalisieren (Carpintero, 2021) und zu homogenisieren. Damit lassen sich unterschiedlichste KI-Systeme entwickeln, obwohl die darunterliegende Hardware dieselbe bleibt.

In der Container-Umgebung wird ein weiteres, eigenes Linux-basiertes Betriebssystem eingesetzt, welches mit Hochleistungsbibliotheken wie BLAS oder cuBLAS („Basic Linear Algebra Subprograms“, 2022) sowie für die Entwicklung von KI-Modellen notwendigen Programmiersprachen wie Python, C und C++ samt ihrer Compiler ausgestattet sind. Weiterhin befinden sich hier auch die für Deep Learning-Anwendungen spezifischen

Frameworks, mit welchen sich KI-Anwendungen programmieren lassen. Darunter zählen z.B. PyTorch, TensorFlow, Keras oder Jax (Gopani, 2021).

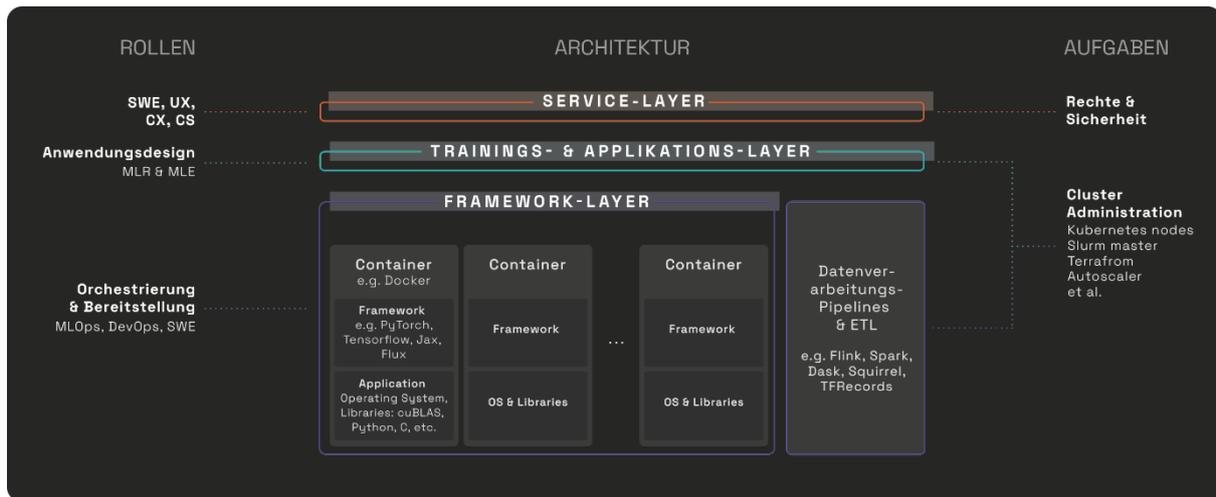


Abb. 21: Die Architektur des Framework- & Service-Layers im Detail

Service-Layer

Über die für die operative Entwicklung und das Training von KI-Foundation-Modellen notwendige Hard- und Software-Infrastruktur werden manche der darin enthaltenen Services im Service-Layer über Benutzeroberflächen abgebildet. Zum Beispiel werden hier unveränderte KI-Modelle für eigene Use-Cases eingesetzt oder KI-Modelle automatisch an bereitgestellte Datensätze adaptiert. Im Service-Layer können aber auch administrative Prozesse wie die Kontrolle von Nutzern oder die Anfragen von weiteren Ressourcen abgewickelt werden.

Voraussetzungen

Im Framework-Layer wird die Cluster-Orchestrierung gesteuert, um gezielt Aufträge im System starten zu können und verfügbare System-Ressourcen effizient einsetzen zu können. Hierfür wird ein Job-Management-System eingesetzt, welches typischerweise auf SLURM, Kubernetes (*Kubernetes (K8s)*, 2014/2022), Terraform (*Terraform*, 2014/2022), DockerHub oder anderen Komponenten basiert (Mujkanovic et al., 2020). Damit lassen sich Aufträge aneinanderreihen und individuell priorisieren. Darüber hinaus gehört zum Framework-Layer auch eine für Deep Learning-Anwendungen spezifische Komponente, um Daten passgenau und effektiv aus den Datenspeicher-Systemen in die Rechnerumgebung zu laden (s. Kapitel 7.2).

Für diese Fülle an spezialisierten Aufgaben werden Fachkräfte benötigt, die in der teilweise automatisierten Operationalisierung der Entwicklungs- (DevOps) und Machine-Learning-Umgebung (MLOps) ausgebildet sind. Dabei ist es notwendig, sowohl die dauerhafte Gesundheit der Umgebungen zu gewährleisten, als auch über Weiter- und Neuentwicklungen von Software-Lösungen für das Framework-Layer informiert zu bleiben, um bei eingehender Prüfung Verbesserungen am System vorzunehmen. Für die Orchestrierung und Priorisierung der Cluster werden vor allem Software Engineers und DevOps-Expert:innen gebraucht.

7.5 LEAM als Leuchtturmprojekt für die Zukunft des KI-Ökosystems

Eine für die Konzeption, das Training und die Unterhaltung von KI-Foundation-Modellen notwendige und leistungsfähige KI-Software-Umgebung setzt aber nicht nur technische und infrastrukturelle Bedürfnisse voraus, sondern stellt auch vielfältige und zahlreiche Voraussetzungen an gut ausgebildetes Personal, welche derzeit in Deutschland und der EU noch kaum oder gar nicht erfüllt werden. So belegen Studien zum Fachkräftemangel in der IT-Dienstleistungsbranche einen Mangel an spezialisiert ausgebildeten Fachkräften, aber auch eine fehlende Differenzierung zwischen einzelnen hochspezialisierten Rollen, die für die Entwicklung und das Training von KI-Foundation-Modellen notwendigerweise zusammenwirken müssen (Hickmann & Koneberg, 2022). Dies erschwert eine passgenaue Ansprache und Werbung für den Nachwuchs.

In der Praxis wird die Arbeit an KI im Allgemeinen und insbesondere an KI-Foundation-Modellen in Teams organisiert, die verschiedene Fähigkeiten und Ausbildungen einbringen. Spezialisierte Teams müssen verschiedene Programmiersprachen sowie Mathematik, Datenwissenschaft und Datentechnik beherrschen, aber auch vertiefte Kenntnisse in Informatik, Statistik und Wissen über Software-, Hardwarekomponenten und -architekturen vereinen: „Machine Learning-Spezialisten arbeiten in Teams – mit ML-Spezialisten an der Spitze, Softwareentwicklern in großer Zahl an der Basis und Datenwissenschaftlern und -ingenieuren dazwischen – das ist die typische Personalstruktur fortschrittlicher Technologiecluster wie dem Silicon Valley“ (Philippe Lorenz & Kate Saslow, 2019, eigene Übersetzung).

Heute sind diese vielfältigen Teams in Deutschland allerdings unterbesetzt, sodass es „für neun von zehn offenen Stellen [in der Informatik] zuletzt keine passend qualifizierten Arbeitslosen“ gab (Hickmann & Koneberg, 2022). Bitkom-Präsident Achim Berg sagte dazu: „Der sich verschärfende Mangel an IT-Spezialistinnen und -Spezialisten wächst sich zu einer ganz realen Bedrohung für Deutschlands große Transformationsaufgaben aus.“ (Bitkom e.V., 2022a).

In einer dreiteiligen Studie hat das Center for Security and Emerging Technology (CSET) an der Georgetown University festgestellt, wie schnell der Bedarf an KI-Expert:innen wächst: Zwischen 2015 und 2019 ist der Anteil der gesamten Arbeitnehmer:innen, die in den USA direkt am KI-Ökosystem beteiligt sind, von sechs auf neun Prozent gewachsen – zu 14 Millionen Arbeitnehmer:innen (Gehlhaus et al., 2021). Innerhalb der nächsten zehn Jahre soll diese Berufsgruppe dabei sogar doppelt so schnell wachsen wie der Bundesdurchschnitt (Gehlhaus et al., 2021). Doch anders als in den USA, können bürokratische Hürden, ausländische Fachkräfte einzustellen und an den Arbeitsmarkt zu binden, hierzulande deutlich höher sein.

In der schulischen und beruflichen Bildung ist es dabei notwendig, stärker zu differenzieren und die unterschiedlichen Rollen im KI-Ökosystem gezielter auszubilden. Hier können neben Universitäten insbesondere Fachhochschulen und Hochschulen für Angewandte Wissenschaften mehr in attraktive Angebote für spezialisierte Aufgabenfelder investieren, während Universitäten mehr Mittel für die Ausbildung von ML-Spezialisten zur Verfügung haben sollten (Wannemacher & Bodmann, 2021).

Das Bildungsangebot und der Arbeitsmarkt der USA sind uns dabei mehrere Schritte voraus. Dazu stellte CSET weiterhin fest, dass viele Berufswege im KI-Ökosystem keine mehrjährige Ausbildung an einer Universität oder gar einen Dokortitel benötigen. In der Studie, „Training Tomorrow’s AI Workforce,“ plädiert das CSET für eine neue Strategie, nicht-universitäre Bildungseinrichtungen besser zu befähigen, KI-Expert:innen auszubilden, in dem diese auf föderaler und regionaler Ebene gefördert werden sowie besser mit den lokal verorteten Unternehmen verknüpft werden (Gehlhaus & Koslosky, 2022). So sollen Student:innen und Arbeitgeber:innen bessere Voraussetzungen vorfinden, regional integrierte KI-Ökosysteme zu betreiben.

Darüber hinaus plädiert CSET für eine Abkehr von weitreichenden, langjährigen Studiengängen in der Informatik hin zu einer größeren Vielfalt an modular aufgebauten Weiterbildungskursen, um sich schneller zu spezialisieren und so für den Arbeitsmarkt verfügbar zu sein. Dieser Strategiewechsel muss laut CSET aus der US-Regierung heraus gesteuert werden, indem eine Stabstelle für die Ausbildung von KI-Expert:innen im Bundeskabinett geschaffen wird (Gehlhaus & Koslosky, 2022). Eine ebenso zentralisierte, befähigte Stabstelle wäre auch in Deutschland und der EU ein wichtiges Signal, um den Fachkräftemangel im Detail richtig beurteilen zu können und das Angebot an KI-Weiterbildungen in der Masse und in der Tiefe zu verbessern.

Gerade diese wirtschaftliche Entwicklung des KI-Ökosystems hin zu einer zentralen, alle anderen Wirtschaftsbereiche durchdringenden Querschnittsbranche untermauert die Wichtigkeit von LEAM als Befähiger der deutschen und europäischen Wissenschaft und Wirtschaft. LEAM kann neben seinen wirtschaftlichen Möglichkeiten auch ein wirtschaftspolitischer Leuchtturm sein, um Expert:innen auszubilden, anzuwerben und langfristig zu binden, wie auch junge Menschen dazu zu inspirieren, IT-Berufe zu erlernen. LEAM kann also nicht nur ein starkes Zeichen für die Wettbewerbsfähigkeit des Innovationsstandorts Deutschland und Europa setzen, sondern auch über die konkrete Anwendung hinaus einen für die Zukunftsfähigkeit unserer Wirtschaft notwendigen Bildungsauftrag in der Gesellschaft erfüllen.

7.6 Zusammenfassung

Die Herausforderungen des von LEAM vorgeschlagenen KI-Hochleistungsrechenzentrum zur Entwicklung und des Trainings von KI-Foundation-Modellen sind auf der Software-Ebene technisch lösbar. Die dafür notwendigen Technologien wurden bereits entwickelt und stehen größtenteils als Open-Source-Software zur Verfügung. Die wesentliche Herausforderung liegt also nicht in der Entwicklung der Software, sondern im Betrieb eines solchen KI-Hochleistungsrechenzentrum, wofür eine Vielzahl an hochspezialisierten Expert:innen in Teams zusammenarbeiten muss.

Falls eine Umsetzung gelingt, kann LEAM wissenschaftliche Durchbrüche aus Deutschland und Europa lancieren, indem mit komplexer, heute bereits verfügbarer Software auf moderner, massiver Hardware KI-Modelle robust skaliert werden. So könnte LEAM die deutsche und europäische Wissenschaft und Wirtschaft nicht nur zukunfts- und wettbewerbsfähig machen, sondern auch für unsere Gesellschaft über die tatsächliche, technische Umsetzung Mehrwerte in der Bindung von Fachkräften und der Lösung von gesellschaftlichen Aufgaben bieten

Aufbau eines KI-Hochleistungsrechenzentrums



8. Aufbau eines KI-Hochleistungsrechenzentrums

Als einer der Hauptgründe, warum in Deutschland bzw. Europa kein regelmäßiges Training großer KI-Foundation-Modelle à la GPT-3 möglich ist, nennen die Expert:innen, die für diese Studie interviewt wurden, die fehlende High Performance Computing (HPC) Infrastruktur für die Berechnung solcher Modelle. Denn KI-Foundation-Modelle stellen insbesondere an die Rechenkapazitäten besondere Anforderungen. Anders als bei herkömmlichen HPC-Systemen, die Central Processing Unit (CPU)-basiert arbeiten, werden im Bereich Künstlicher Intelligenz und speziell bei der Berechnung großer KI-Foundation-Modelle außergewöhnlich hohe Rechenkapazitäten von bis zu 4500 Graphics Processing Unit (GPU) benötigt, die trotz dieser hohen Kapazität sehr lange Laufzeiten benötigen, um die nötigen Rechenaufgaben zu bewältigen. Diese Art von konzentrierten Rechenkapazitäten sucht in Deutschland und EU bisher noch ihresgleichen. Indem Industrie und Wissenschaft die Zugänge zu diesen Rechenkapazitäten fehlen, werden Sprunginnovationen im Bereich KI-Foundation-Modelle stark erschwert.

8.1 Definition Rechenzentrum

Ein Rechenzentrum (RZ) ist eine Infrastruktureinrichtung, in der Computer, Server, Speichersysteme und andere Technologiekomponenten zusammengefasst sind, um eine große Menge an Daten und Anwendungen zu verarbeiten und zu speichern (Hintemann & Clausen, 2018). Rechenzentren dienen in der Regel als zentrale Ressource für die Verarbeitung und Speicherung von Daten und Anwendungen in Unternehmen, Organisationen und Institutionen. Sie können auch für verschiedene Zwecke eingesetzt werden, von der Verarbeitung von Transaktionen und der Bereitstellung von IT-Diensten bis hin zur Ausführung von KI-Anwendungen.

Es gibt verschiedene Arten von Rechenzentren, die sich in ihrer Größe, ihrem Zweck und ihrer technischen Ausstattung unterscheiden (vgl. Tabelle 6).

Rechenzentrum	Beschreibung
Enterprise Rechenzentren	Enterprise Rechenzentren sind Rechenzentren, die von Unternehmen und Organisationen betrieben werden, um die Verarbeitung und Speicherung von Daten und Anwendungen für ihre internen Zwecke zu ermöglichen. Enterprise Rechenzentren können in verschiedenen Größen und Formen auftreten, von kleinen Räumen mit wenigen Servern bis hin zu großen Anlagen mit tausenden Computern und Servern.
Collocation Rechenzentren	Collocation Rechenzentren sind Rechenzentren, die von Dritten betrieben werden und an Unternehmen und Organisationen vermietet werden. Collocation Rechenzentren bieten den Mietern Platz für ihre eigenen Server und andere Computerkomponenten und stellen die erforderliche Infrastruktur wie Strom, Kühlung und Netzwerkverbindungen zur Verfügung.
Hyperscaler Rechenzentren	Hyperscaler Rechenzentren sind Rechenzentren, die von Hyperscaler-Unternehmen betrieben werden, um die Verarbeitung und Speicherung von Daten und Anwendungen für ihre Kunden zu ermöglichen. Hyperscaler Rechenzentren sind oft sehr groß und bieten eine hohe Rechenleistung und Speicherkapazität, um große Mengen an Daten schnell und effizient zu verarbeiten und zu speichern.

Tabelle 6: Beispiele für Rechenzentren

Neben dem Geschäftsmodell können Rechenzentren auch nach ihrer Größe unterschieden werden. Folgende Größenangaben können hier als Referenz dienen (*The role of data centers in an interconnected world*, o.D.):

Art	Größen	Leistungs- aufnahme	Schwerpunkt
Micro Data Center	Ab 1 Server-Rack aufwärts; passende Konfigurationen für einen Container	100 kW	EDGE-Anwendungen, die von der Nähe zur IoT- Quelle profitieren
Kleine Rechenzentren (auch EDGE- Rechenzentren)	ca. 500 m ²	1 MW	Oftmals ein Unternehmens-eigenes Rechenzentrum für kritischen Datenbestand
Mittlere Collocation/Hostin- g Rechenzentren	ca. 10.000 m ²	10 MW	Multi Tenant Collocation RZs sowie Ausrichtung auf Hosting und Managed Services
Große Collocation Rechenzentren	ca. 50.000 m ²	50 MW	Multi und Single Tenant Collocation sowie große Hosting/Managed Service Anbieter mit internationaler Ausrichtung
Hyperscaler	100.000 m ² und mehr	100 MW und mehr	Große, global operierende Cloud-Anbieter, die oftmals an mehreren Standorten (10-200) weltweit tätig sind

Tabelle 7: Größen von Rechenzentren

Abgrenzung von KI-Hochleistungsrechenzentren

LEAM plant die Berechnung von großen KI-Foundation-Modellen, hierzu ist spezielle Hardware notwendig. Die benötigten KI-Hochleistungsrechenzentren unterscheiden sich in erster Linie durch ihren Fokus auf die Verarbeitung von KI-Anwendungen. Andere Rechenzentren sind meist auf die Verarbeitung allgemeiner Daten und Anwendungen ausgerichtet und können für eine Vielzahl von Zwecken verwendet werden. Ein KI-Hochleistungsrechenzentrum hingegen ist speziell für die Verarbeitung von KI-Anwendungen entwickelt und ausgestattet und bietet die erforderliche Rechenleistung und -umgebung, um KI-Modelle und -Algorithmen schnell und effizient zu trainieren.

KI-Hochleistungsrechenzentren grenzen sich von anderen Rechenzentren auch in ihrer technischen Ausstattung und Funktionsweise ab. KI-Hochleistungsrechenzentren können zum Beispiel über eine höhere Rechenleistung und spezielle Hardware wie Grafikprozessoren (GPUs) verfügen, die für die Verarbeitung von KI-Anwendungen besonders geeignet sind. Sie können auch über spezielle Software-Tools und -Umgebungen verfügen, die es ermöglichen, KI-Modelle und -Algorithmen zu entwickeln und zu trainieren.

8.2 Anforderungen an ein KI-Hochleistungsrechenzentrum

Insgesamt sind KI-Hochleistungsrechenzentren also auf die speziellen Anforderungen von KI-Anwendungen ausgerichtet und bieten die erforderliche Infrastruktur und Ressourcen, um KI-Anwendungen effektiv zu verarbeiten. Im Folgenden werden die konkreten Anforderungen an ein KI-Hochleistungsrechenzentrum zur Berechnung großer KI-Foundation-Modelle beschrieben.

Prozessoren

Für die Berechnung von KI-Anwendungen sind Grafikprozessoren (Graphic Processing Units/GPUs) geeignet. GPUs sind speziell für die Bearbeitung von Graphikaufgaben konzipiert. Sie verfügen über eine Vielzahl von Rechenkernen, wodurch sie große Datenmengen schnell und parallel verarbeiten können. Dadurch können sie Aufgaben wie das Rendern von 3D-Grafiken oder das Trainieren von Modellen für maschinelles Lernen und Künstliche Intelligenz schnell bewältigen. In vielen Rechenzentren werden jedoch anstatt von GPUs, CPUs (Central Processing Units) verwendet. CPUs haben eine geringere Anzahl an Rechenkernen (6-12) und können ein breites Spektrum an Aufgaben bewältigen, sind aber nicht besonders gut für Aufgaben geeignet, die ein großes Maß an paralleler Verarbeitung erfordern.

Netzwerkanforderungen

Neben der Leistung der Prozessoren ist es für die Berechnung von großen KI-Foundation-Modellen von entscheidender Bedeutung, welche Verbindungstechnologie und welche Bandbreite zwischen den GPUs eingesetzt wird und mit welcher Bandbreite die einzelnen Knoten verbunden sind. Es ist zwingend notwendig, die Arbeitslast auf sehr viele GPUs zu verteilen. Dafür sind laut der befragten Expert:innen eine InfiniBand oder schnelle Ethernet Verschaltung und eine starke Bandbreite innerhalb des Clusters, aber auch zwischen Cluster und Storage notwendig. InfiniBand ist besonders gut für den Einsatz im HPC-Bereich geeignet und bis zu 10-mal leistungsfähiger als das gängige Interconnect PCIe. Außerdem sollten hier Storage-Umgebungen mit geringer Latenz eingebunden werden (Hensel & Ostler, 2020).

Leistungsdichten

Als Leistungsdichte wird die elektrische Aufnahme der IT-Komponenten pro Server-Rack bzw. pro m² Whitespace (Flächenbedarf pro Rack) verstanden. Laut der befragten Expert:innen steigen die Leistungsdichten pro Rack im High Performance Computing Bereich stetig und liegen aktuell zwischen 20-30 kW pro Rack und vereinzelt auch höher. Die für die Berechnung von Foundation-Models benötigte Leistungsdichte liegt zwischen 30 und 45 kW (z.B. NVIDIA Superpod) pro Rack und somit im High Performance Computing-Bereich. Die Leistungsdichte hat ebenfalls Einfluss auf die Wahl der eingesetzten Kühlung der Systeme.

Kühlung

Die von der Server-Hardware aufgenommene elektrische Leistung wird zu fast 100 % in Wärme umgewandelt. Je höher die Leistungsdichte eines Server-Racks, desto höher die abgegebene Wärmemenge. Computer dürfen die für sie zugelassenen Betriebstemperaturen jedoch nicht überschreiten, um einen sicheren Betrieb zu gewährleisten. Hersteller von IT-Equipment schreiben daher in der Regel die einzuhaltenden Temperaturbereiche und Luftfeuchtebedingungen vor, an die wiederum die Garantieleistungen geknüpft sind. Viele klimatechnische Vorgaben für den Betrieb von IT-Equipment in Serverräumen finden sich in den „Data Center Power Equipment Thermal Guidelines and Best Practices“ des amerikanischen Verbandes ASHRAE wieder, an denen sich die IT-Hersteller ausrichten (ASHRAE, 2016). Um die Temperaturen im zulässigen Bereich zu halten, gibt es verschiedene Möglichkeiten der Kühlung, die in Rechenzentren zum Einsatz kommen. Generell wird zwischen Methoden der Luftkühlung sowie der Flüssigkeitskühlung (Wasserkühlung) unterschieden.

Da Kühlflüssigkeiten physikalisch dichtere Medien sind als Luft, können diese in der Regel auch deutlich höhere Wärmelasten abtransportieren.

Aufgrund der stetig steigenden Leistungsdichten in den Racks und der damit steigenden Abwärmemenge wird es herausfordernder, Systeme besonders im HPC-Bereich effizient mit einer Luftkühlung zu kühlen. Laut der befragten Expert:innen ist die Luftkühlung eines Racks mit einer Leistungsdichte bis maximal 20 kW pro Rack umsetzbar. So entspricht 300 m³/h Luft einem Wasserdurchsatz von 90 l/h (Dürr, 2018).

Bei der von LEAM angestrebten Leistungsdichte von 36 kW pro Rack sind nur eine direct-to-chip Flüssigkeitskühlung oder flüssigkeitsgekühlte Racks geeignet. Flüssigkeiten sind ein effizienteres Wärmeübertragungsmedium als Luft und eignen sich somit besser bei hohen Leistungsdichten. Expert:innen geben jedoch an, dass bisher nur sehr wenige Rechenzentren mit einer Flüssigkeitskühlung oder hybriden Lösung aus Luft- und Flüssigkeitskühlung ausgestattet sind. Die Gebäudeinfrastruktur wird heute bei neuen Rechenzentren von Rechenzentrumsbetreibern und Collocation-Anbietern so geplant, dass die Installation einer Flüssigkeits-Kühlung möglich ist.

Kühlmethode	Beschreibung
Luftkühlung	<p>Die gängigste Methode zur Luftkühlung ist die raumbasierte Kühlung mit Doppelboden in Kalt- und Warmgangkonfiguration. Hierbei wird die kühle Luft durch Auslassöffnungen im Doppelboden in den sog. Kaltgang vor die Racks geleitet und durch Ventilatoren der Server in den Serverschrank eingesaugt. Die erwärmte Luft wird auf der Rückseite des Racks über den Warmgang abgeleitet und einer erneuten Kühlung zugeführt.</p> <p>Des Weiteren finden sich auch Konzepte, bei denen auf einen Doppelboden verzichtet wird. In diesem Fall wird die kalte Luft von der Seite in den Serverraum eingeblasen und die erwärmte Abluft über eine spezielle „Doppelbodendecke“ aus dem Raum abgeführt.</p>
Direct-to-Chip/Direct-to-Plate <i>Flüssigkeitskühlung</i>	<p>Bei der Direct-to-Chip Kühlung wird eine Kühlflüssigkeit in kleinen Schläuchen oder Kupferrohren zu Kühlplatten geleitet, die sich direkt neben den zu kühlenden Komponenten (z.B. CPUs und GPUs) befinden. Die so erwärmte Flüssigkeit wird über einen Wärmetauscher und weitere nachgelagerte Kühlkreisläufe abgeführt.</p>
Tauchkühlung/Immersion Cooling <i>Flüssigkeitskühlung</i>	<p>Beim sog. Immersion Cooling werden die elektronischen Bauteile, in der Regel die komplette Platine eines Servers, in einem Behälter mit einer speziellen elektrisch nichtleitenden Flüssigkeit eingetaucht, die das Kühlmittel darstellt. Die Flüssigkeit nimmt die Wärme auf und wird dann über einen Kühler abgeleitet, um die Wärme abzuführen.</p>
Kühlung über Rücktüren <i>Flüssigkeitskühlung</i>	<p>Wassergekühlte Rücktüren von Serverschränken bieten den Vorteil, dass konventionelle Servertechnik in den Schränken verbaut werden kann. Die von den Ventilatoren abtransportierte, erwärmte Abluft wird dabei durch Wärmetauscher in den rückseitigen Rack-Tür aufgenommen. Durch die angeschlossene Verrohrung wird die Wärme aus dem Serverraum abgeführt. Da bei dieser Technologie sowohl ein Luftstrom zum Abtransport der Wärme in die Wärmetauscher der Rücktüren als auch ein Wasserkreislauf zum weiteren Transport aus dem Serverraum zum Einsatz kommt, stellen diese Systeme quasi einen hybriden Lösungsansatz dar.</p>

Tabelle 8: Übersicht über die Kühlmöglichkeiten in Rechenzentren

Gebäudeinfrastruktur

Die Anforderungen an die Gebäudeinfrastruktur eines KI-Rechenzentrums unterscheiden sich nicht grundlegend von den Anforderungen anderer Rechenzentren. Zu beachten ist jedoch, dass vor dem Bau eines neuen Rechenzentrums die Kühltechnologien geplant und die Bauweise des Gebäudes entsprechend angepasst wird. Die für das LEAM-Vorhaben benötigten Compute-Ressourcen und verwendeten Leistungsdichten von ca. 36 kW/Rack können nur durch eine Flüssigkeitskühlung gekühlt werden. Deshalb sollte die Gebäudeinfrastruktur für die Nutzung einer Flüssigkeitskühlung ausgestattet sein. Dies ist beim Bau eines Rechenzentrums sowie bei der Auswahl eines geeigneten Collocation/Housing Anbieters zu beachten.

Flächenbedarf

Entscheidenden Einfluss auf den benötigten Flächenbedarf hat die Leistungsdichte pro Rack. Wird bei einer vorgegebener Gesamtrechenkapazität die Leistungsdichte pro Rack erhöht, nimmt die Anzahl der benötigten Racks ab und der Flächenbedarf sinkt. Wird dagegen die Leistungsdichte pro Rack verringert, werden mehr Racks benötigt, um die angestrebte Gesamtleistung zu erreichen und somit steigt auch der Flächenbedarf. Laut der interviewten Expert:innen liegt der Flächenbedarf pro Rack (sog. Whitespace) in der Regel bei brutto 3,0-3,5 m² (die reine Standfläche eines Racks beträgt dabei ca. 1 m²). Für die von der LEAM-Initiative kalkulierten flüssigkeitsgekühlten 140 Racks (z.B. 4 NVIDIA DGX H 100 Knoten/Rack) mit einer Gesamtleistung von ca. 4,0 MW zur Berechnung und Training des KI-Foundation-Modells ergibt sich demnach ein Flächenbedarf von 345-525 m². Beim Einsatz von älteren Knoten (z.B. NVIDIA DGX A 100) in luftgekühlten Racks in ergibt sich eine höhere Anzahl von Racks und eine entsprechend größere Fläche, da weniger Knoten pro Rack verbaut werden können.

Stromversorgung

Die grundsätzlichen Anforderungen von KI-Hochleistungsrechenzentren sind laut der interviewten Expert:innen mit den Anforderungen an die Stromversorgung anderer Rechenzentren gleichzusetzen. Gleichwohl zeichnen sich Höchstleistungsrechenzentren durch eine höhere Leistungsdichte (kW/Rack) und damit verbunden eine deutlich höhere Abwärmelast pro Serverraum aus. Die Stromversorgung eines Rechenzentrums besteht in der Regel aus mehreren Komponenten, die zusammenarbeiten, um sicherzustellen, dass das Rechenzentrum mit ausreichend Strom versorgt wird.

Zu diesen Komponenten gehören in der Regel:

1. Netzeinspeisung: Hier liegt der Übergabepunkt für den Strom, der für den Betrieb des Rechenzentrums benötigt wird. In der Regel wird der Strom aus dem öffentlichen Stromnetz bezogen. Kleinst- und Kleinrechenzentren werden meist über den normalen Hausanschluss mit 400 V Drehstrom versorgt. Bei größeren Abnahmemengen erfolgt die Einspeisung in der Regel mit 10 kV bzw. mit 20 kV (Mittelspannung).
2. Stromverteilung: Die Hauptstromversorgung wird über kaskadierende Abstufungen von MSHV (Mittelspannungshauptverteilung) und NSHV (Niederspannungshauptverteilung) bis zu den jeweiligen Anschlusspunkten auf Netzteilenebene im Rechenzentrum verteilt. Die Stromverteiler verteilen den Strom an die verschiedenen Bereiche des Rechenzentrums und sorgen dafür, dass der Strom zu den benötigten Stellen gelangt.
3. Unterbrechungsfreie Stromversorgung (USV): Um sicherzustellen, dass das Rechenzentrum auch bei kurz- und langfristigen Stromausfällen weiter betrieben werden kann, muss nach INF.2 des BSI eine USV installiert werden. Die USV stellt über Batteriepuffer oder Schwungmassen-Systeme eine kontinuierliche Stromversorgung des Rechenzentrums sicher, bis die Hauptstromversorgung wiederhergestellt wird. Weitere Aufgaben einer USV sind das sog. Glätten von Spannungstößen (Surge; <4ms), Abfedern von Oberschwingungen oder die galvanische Trennung des internen vom externen Stromkreislauf. Der Einsatz einer USV ist in Deutschland Pflicht.

Auf das Thema Nachhaltigkeit in Bezug auf die Stromversorgung wird im Kapitel 8.6.3 eingegangen.

Energieverbrauch & Effizienzparameter

Der PUE-Wert (Power Usage Effectiveness) ist eine vom Industriekonsortium *The Green Grid* eingeführte technische Kennzahl, welche die von der IT im Rechenzentrum verbrauchte Energie ins Verhältnis zum Gesamtenergieverbrauch setzt. In der Theorie beträgt der optimale PUE-Wert 1,0 (Gesamtstromverbrauch entspricht dem reinen IT-Verbrauch ohne sonstige energetische Aufwände wie z.B. für Kühlung oder für die USV-Verlustleistung). Der durchschnittliche PUE-Wert neu gebauter Rechenzentren betrug im Jahr 2015 unter 1,5, während der durchschnittliche PUE-Wert der luftgekühlten Bestandsrechenzentren in Deutschland im selben Jahr bei 1,8 lag. Im Jahr 2010 lag der durchschnittliche PUE-Wert der deutschen Rechenzentren noch bei 1,98 (Stobbe et al., 2015).

Einfluss auf den PUE-Wert eines Rechenzentrums hat in erster Linie die Klimazone, in der das Rechenzentrum betrieben wird, sowie die eingesetzte Kühltechnologie, gefolgt von den energetischen Aufwänden zum Betrieb einer USV. In warmen Klimazonen (z.B. Mittelmeerraum) sind die energetischen Aufwände für die Kühlung naturgemäß höher als in kälteren Klimazonen (z.B. Skandinavien). Im Hinblick auf die Strompreisentwicklung ist die Senkung des PUE die vorrangige Möglichkeit, um die Betriebskosten zu senken (Lamonica, 2014). Rechenzentren mit einer Luftkühlung liegen durchschnittlich bei einem PUE-Wert zwischen 1,5 und 1,2 währenddessen flüssigkeitsgekühlte Rechenzentren einen

PUE-Wert von bis zu 1,06 erreichen können (PUE-Werte im Google-Rechenzentrum). Deutsche Rechenzentrumsbetreiber:innen bestätigten dies in der Befragung und gaben an, dass der PUE-Wert bei neuen, wassergekühlten Systemen Werte bereits bis zu 1,1 beträgt. Aktuell werde hier konzentriert die Nutzung der Abwärme optimiert, um insgesamt klimaneutral zu werden.

Die vom Umweltbundesamt entwickelte Berechnungsmethode KPI4DCE (Key Performance Indicators for Data Center Efficiency) ist ein ganzheitlicher Ansatz für die Berechnung der Energieeffizienz von Rechenzentren, welcher auch den Lebenszyklus des IT-Equipment und der technischen Versorgungsstruktur einbezieht. Diese Berechnungsmethode ist in der Theorie aussagekräftiger als der PUE-Wert allein, jedoch ist sie auch aufwendiger. Nicht allen Rechenzentren stellen die geforderten Messwerte zur Verfügung. Das Ziel ist eine automatisierte Messwertaufnahme (Schödwel et al., 2018). Auf EU-Ebene gibt es jedoch im Rahmen der Initiative *Climate Neutral Data Centre* Überlegungen, Rechenzentren mit diesen Kennzahlen zukünftig zu überwachen (*Climate Neutral Data Centre Pact – The Green Deal Need Green Infrastructure*, o.D.).

Skalierbarkeit / Modularer Aufbau

Für die Skalierbarkeit eines KI-Hochleistungsrechenzentrums gelten die gleichen Gesetzmäßigkeiten wie für die Skalierbarkeit anderer Rechenzentren. Die Skalierbarkeit wird maßgeblich von drei Faktoren beeinflusst: ausbaubare Leerflächen, die maximale Leistungsdichte pro Rack und die maximal zugesicherte Stromleistung (im Gegensatz zum aktuellen Verbrauch). Hat ein Rechenzentrum beispielsweise die vorhandene Fläche bereits vollständig ausgeschöpft, kann die Leistungsdichte in den Racks unter Prüfung der vorhandenen Kühlung und Stromversorgung erhöht werden. Sind die zur Verfügung stehenden Serverräume bisher nicht vollständig genutzt, kann durch die Inbetriebnahme weiterer Räume die Gesamtleistung des Rechenzentrums erhöht werden. Seit 2021 ist eine leichte Abnahme der Gesamt-IT Flächen von 2,1 Mio. m² in Deutschland durch eine Erhöhung der Leistungsdichten festzustellen. Diese Konsolidierungsbestrebungen machen sich am stärksten durch die Flächenabnahmen im Bereich der traditionellen, häufig unternehmenseigenen, Rechenzentren erkennbar. Hier ist eine Migration der installierten IT-Leistung hin zu Cloud- und Edge-Betriebsmodellen erkennbar (Hintemann et al., 2022). Die IT-Fläche als Maß zur Beschreibung der Entwicklung der Rechenzentrumskapazitäten ist somit nur noch sehr bedingt aussagekräftig.

Bei entsprechender Grundstücksfläche ist laut der befragten Expert:innen auch ein modularer Rechenzentrumsaufbau denkbar. Modularer Aufbau bedeutet, dass zunächst nicht die gesamte Grundstücksfläche bebaut wird, sondern zunächst nur die Flächen, die aktuell gebraucht werden. Flächen für absehbares zukünftiges Wachstum des Rechenzentrums werden freigehalten, sofern dies in entsprechenden Bauanträgen vorgesehen ist.

Latenzen

Der Begriff beschreibt das Zeitintervall zwischen dem Moment, in dem eine Anfrage an ein System gestellt wird, und dem Zeitpunkt, an dem die Antwort des Systems empfangen wird. Latenzen werden in Millisekunden oder Mikrosekunden gemessen.

Den befragten Expert:innen zufolge sind die Latenzen für das Training eines KI-Foundation-Modells nicht von großer Bedeutung, da dies lokal auf einem geeigneten Rechencluster erfolgt. Für den Betrieb und die Entwicklung von Inference Anwendungen, die später auf Basis des Foundation-Modells entstehen, sollte ein Rechenzentrum je nach Use Case über eine Bandbreite von mindestens 100 Gbit/s verfügen und Latenzanforderungen von unter 10 ms RTD (Round Trip Delay) erfüllen.

Anforderungen an Zertifizierungen, Datenschutz und Compliance

KI-Hochleistungsrechenzentren sind in Bezug auf ihre Anforderungen im Bereich Zertifizierungen und der Datenschutzgrundverordnung (DSGVO) mit anderen Rechenzentren gleichzustellen. Die gängigsten Zertifizierungen sind u.a. laut Rechtsexpert:innen ISO 27001 (Zertifizierung auf der Basis von IT-Grundschutz), ISO 9001 (Qualitätsmanagementsystem) und ISO EN 50600 (Bau und Betrieb sicherer Rechenzentren). Der Kriterienkatalog C5 des BSI beschreibt die Mindestanforderungen für sicheres Cloud Computing und muss berücksichtigt werden.

8.3 Nachhaltigkeitsaspekte

Abwärme

Die Nutzung der entstehenden Abwärme gehört zu den wesentlichsten Nachhaltigkeitsaspekten in Rechenzentren und wird aufgrund verschiedener Herausforderungen rege diskutiert. Bislang weitgehend ungenutztes Potenzial liegt in der Einspeisung CO₂-freier Abwärme von Rechenzentren in Nah- und Fernwärmenetze. Die vorhandene Kühlungstechnologie ist dabei der ausschlaggebende Faktor, wie energieeffizient die entstandene Abwärme eines Rechenzentrums weiter genutzt werden kann. Ist eine Luftkühlung im Rechenzentrum installiert, erreicht die Abwärme laut Expert:innen Temperaturen von 30-35°C, in Spezialfällen auch bis zu 50°C. Diese Temperaturen sind jedoch zu niedrig, um die Abwärme direkt in ein Wärmenetz einspeisen zu können. Dies bedeutet, dass vor der Weiterleitung der Abwärme eine Wärmepumpe eingesetzt werden muss, um die Temperatur den Anforderungen des Wärmenetzes anzupassen. Ist eine Wasserkühlung im Rechenzentrum verbaut, kann die Abwärme mit Temperaturen von 60-70° C direkt einem Wärmenetz der vierten Generation zur Verfügung gestellt werden. Technologisch ältere Wärmenetze erfordern allerdings höhere Einspeisetemperaturen. Durch das Hochverdichten der zu niedrigen Abwärme entstehen hier zusätzlich energetische Aufwände durch den Betrieb der Wärmepumpen. Moderne Methoden der Wasserkühlung werden aufgrund ihrer Neuheit noch selten genutzt (vgl. Kühlung).

Eine weitere Herausforderung ist die lokale Abnahme der Abwärme über Nahwärmenetze. Auch wenn es möglich wäre, die Abwärme für die Beheizung umliegender Gebäude zu nutzen, fehlt es häufig vor Ort an Abnehmern (z.B. Wohn- und Büro- und Industriegebäuden), die genügend Abwärme aufnehmen können, sowie den politischen und regulatorischen Rahmenbedingungen (Bitkom e.V., 2022b). Es können deshalb bislang nur kleine Teile der Abwärme der Rechenzentren genutzt werden.

Strom

Ein weiterer Nachhaltigkeitsaspekt ist die Stromversorgung eines Rechenzentrums mit grünem Strom. Entscheidend ist hierbei eine konsequente und erfolgreiche Umsetzung der Energiewende, die den Ausbau und insbesondere die Verfügbarkeit von Strom aus erneuerbaren Energien in Deutschland beschleunigt (Bitkom e.V., 2022b). Grüner Strom wird aus erneuerbaren Energiequellen wie Sonne, Wind, Wasser oder Biomasse gewonnen.

Zukünftig wird das Forschungsprojekt ESCADE des Bundesministeriums für Wirtschaft und Klimaschutz (BMWK) durch weltweit führende Hard- und Software-Technologien prüfen, wie die Nachhaltigkeitsbilanz von KI-Anwendungen verbessert werden kann.

8.4 Infrastrukturanforderungen im Detail

Die kalkulierten Infrastrukturanforderungen zur Berechnung des KI-Foundation-Modells erfolgten auf Basis der öffentlichen Informationen zu dem GPT-3 Sprachmodell von OpenAI.

Compute Anforderungen für die Berechnung eines Foundationmodells	
Anzahl GPU	4480 GPU (z.B. 560 NVIDIA DGX H 100)
Zeit pro Durchlauf, in Std.	694
Durchläufe gesamt	4
Laufzeit gesamt, in Std.	2777
Internet Bandbreite	>100 Gbit/s
Interconnect	InfiniBand oder vergleichbar
Storage	2000 - 5000 TB
Preprocessing	
CPU	10 - 20 k CPU-Cores
Zeit	840 - 1000 Std.
Gesamt IT-Leistung	ca. 4,0 MW
kW pro Rack	36
Anzahl des Racks (bei 4x DGX H100 / Rack)	ca. 140

Tabelle 9: Compute Anforderungen für die Berechnung eines Foundationmodells

Nach Einschätzung der Expert:innen werden für das Training des LEAM Foundation-Modells folgende Compute-Ressourcen benötigt. Dabei wird exemplarisch für die Berechnung von einem Einsatz von 560 der leistungsfähigsten GPU-Knoten (z.B. NVIDIA DGX H 100, AMD Instinct MI200 oder Intel Ponte Vecchio Data Center GPU) ausgegangen. Für einen Trainingsdurchlauf benötigt das System 694 Stunden, insgesamt werden für das Training des Foundation-Modells vier Durchläufe und 2777 Stunden bzw. 115 Tage benötigt. Für die Installation des Systems würden 140 Racks benötigt, in denen jeweils 4 DGX-Knoten à 8 GPU verbaut sind. Pro DGX entsteht Abwärme von bis zu 9 kW pro Rack, sodass insgesamt Werte von 36 kW pro Rack erreicht werden können. Die benötigte Kühlung ist bei diesen Leistungsdichten nur durch Flüssigkeitskühlung bzw. direct-to-Chip Kühlung erreichbar. Bei der neuesten Generation von GPU-Systemen (z.B. NVIDIA DGX H100) ist die Möglichkeit zur direct-to-Chip Kühlung gegeben.

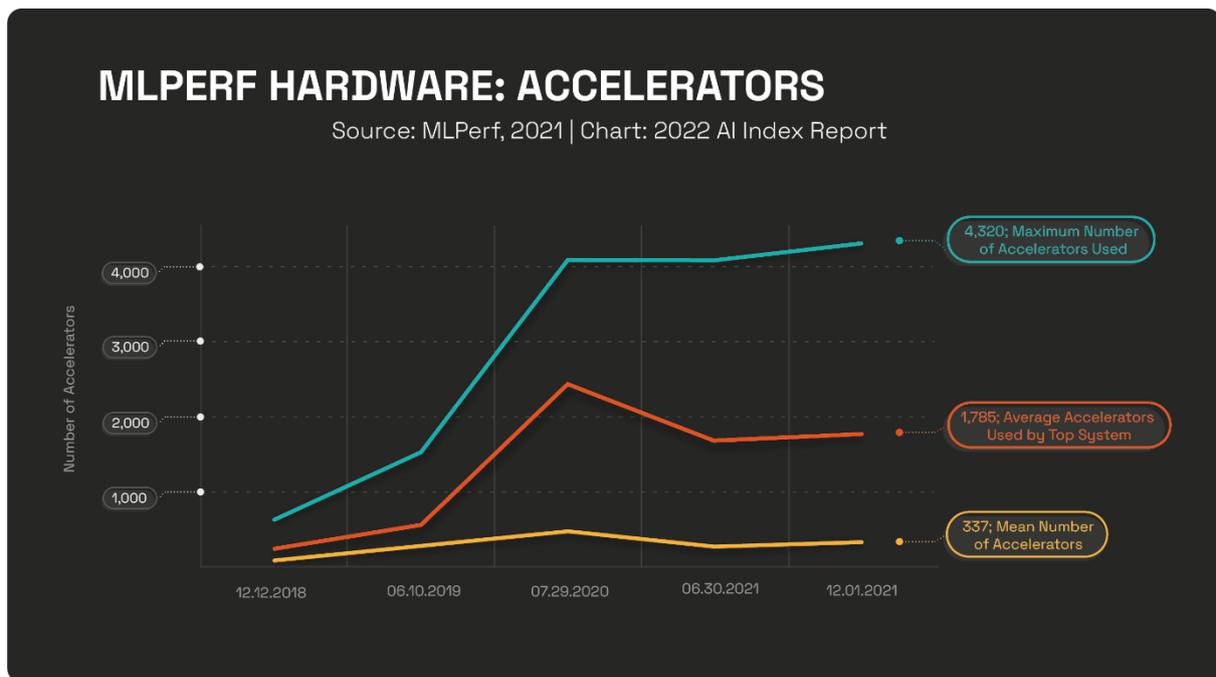


Abb. 22: MLPerf hardware: accelerators (Zhang et al., 2022, S.18)

Der Artificial Intelligence Index Report 2022 der Stanford University beschreibt die Entwicklung der Anzahl der genutzten GPU in den Top HPC-Systemen für das Training von Machine Learning Algorithmen und zeigt auf, dass für das Training die schnellsten KI-Algorithmen Rechencluster mit einer sehr hohen Anzahl von GPU eingesetzt werden (Zhang et al., 2022). Die maximale Zahl der genutzten GPU-Beschleuniger ist vermutlich seit der Erhebung im Januar 2021 nochmal gestiegen.

8.5 Standortauswahl

Was sind die Parameter für eine qualifizierte Auswahl eines Standortes für ein KI-Rechenzentrum? Neben der regionalen Verortung wird untersucht, was in Bezug auf die Energieversorgung sowie die Anbindung an vorhandene Infrastrukturen zu beachten ist.

8.5.1 Vorhandene HPC-Kapazitäten in Deutschland

Forschungsbereich

Das Angebot an High-Performance-Computing für die Wissenschaft wird in Deutschland entsprechend der Leistungsfähigkeit der HPC-Zentren auf den Ebenen 0-3 strukturiert.

In Deutschland gibt es drei Hochleistungsrechenzentren der Ebene 1. Das Gauss Centre for Supercomputing (GCS) vereint die drei bedeutendsten Hochleistungsrechenzentren unter einem Namen. Die Gesamtleistung des Jülich Supercomputing Centre (JSC) in Nordrhein-Westfalen, dem Leibniz-Rechenzentrum (LRZ) in Garching bei München und dem Höchstleistungsrechenzentrum Stuttgart (HLRS) beträgt 130 Peta-FLOPS (Stand November 2021). Die HPC-Zentren haben unterschiedliche Ausrichtungen und können so unterschiedlichen Nutzeranforderungen gerecht werden. Außerdem können sie zusammen oder arbeitsteilig agieren.

Ebene 2 umfasst 12 überregionale Hochleistungsrechenzentren an Forschungseinrichtungen und Hochschulen. Zum Verbund Nationales Hochleistungsrechnen (NHR) haben sich acht der 12 universitären HPC-Zentren der Ebene 2 zusammengeschlossen. Dazu gehören:

- IT-Center - RWTH Aachen
- Zuse-Institut Berlin - Berlin University Alliance
- Hochschulrechenzentrum (HRZ) - Technische Universität Darmstadt
- Zentrum für Informationsdienste und Hochleistungsrechnen - Technische Universität Dresden
- Regionales Rechenzentrum Erlangen - Universität Erlangen-Nürnberg
- Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen - Universität Göttingen
- Steinbuch Centre for Computing (SCC) - Karlsruher Institut für Technologie
- Paderborn Center for Parallel Computing - Universität Paderborn

Zur Ebene 3 gehören regionale HPC-Zentren und Institutionen mit eigenen Clustern für Anwendungen, die eine geringere Leistungsfähigkeit benötigen. Exemplarisch werden hier zwei der leistungsfähigsten HPC-Zentren vorgestellt:

Jülich Supercomputing Centre (JSC)

Das Jülich Supercomputing Centre am Forschungszentrum Jülich gGmbH (JSC) ist ein Institut der Helmholtz-Gemeinschaft Deutscher Forschungszentren und ist durch den Bund (90 %) sowie das Land Nordrhein-Westfalen (NRW) (10 %) grundfinanziert. Das JSC verfügt u.a. über den Supercomputer JULES mit 86 Peta-FLOPS Leistung, einen der derzeit leistungsstärksten Rechner Europas. Der JUWELS Multi-Petaflop Supercomputer verfügt über ein Booster Modul, das mit NVIDIA Ampere GPUs ausgestattet ist und somit für die Berechnung von großen KI-Foundation-Modellen geeignet ist. Der Zugang zu HPC-Rechenressourcen erfolgt über halbjährliche Projektaufträge, die anhand eines objektiven Peer-Review-Verfahrens ausgewählt werden. Das FZ Jülich ist eine gemeinnützige GmbH des öffentlichen Rechts und die Finanzierung erfolgt hier zu 50 % über die EU über die Organisation PRACE - Partnership for Advanced Computing in Europe und zu 50 % über das Bundesministerium für Bildung und Forschung (BMBF) und das Land NRW über den Verein Gauss Centre for Supercomputing (GCS). Für Projekte (üblicherweise Simulationen im Bereich Klimaforschung und Quantenphysik) gilt eine maximale Berechnungszeit von 24 Stunden. Nur in Ausnahmefällen und für größere Projekte sind Jobketten und Reservierung möglich. Ab 2023 wird hier mit der Installation des ersten europäischen Exascale-Rechners JUPITER begonnen, der unter anderem genutzt werden soll, um rechenintensive Machine-Learning-Algorithmen der neuesten Generation zu trainieren (Jülich Forschungszentrum, 2022). Das JSC richtet sich vornehmlich an die Wissenschaft und vergibt die Rechenzeit in einem kompetitiven Verfahren zweimal jährlich nach dem Peer-Review-Verfahren.

Das FZ Jülich ist auch Partner des durch das Bundeswirtschaftsministerium geförderten Projekts OpenGPT-X, in dessen Rahmen ein Sprachmodell auf dem Supercomputer JUWELS trainiert wird. Aktuell nutzt das Projekt allerdings nur rund 320 GPUs. Dieser Wert wird sicherlich noch steigen, ist aber nicht mit den hier vorgeschlagenen 4480 GPUs vergleichbar.

Höchstleistungsrechenzentrum Stuttgart (HLRS)

Das Höchstleistungsrechenzentrum Stuttgart (HLRS) ist ein zentrales Institut der Universität Stuttgart, das u.a. den Supercomputer Hawk mit 26 Peta-FLOPS betreibt und seit 25 Jahren für Wissenschaft und Industrie zur Verfügung steht. Das HLRS ist Mitglied des deutschen Gauss Centre for Supercomputing (GCS), wodurch es eine teilweise Grundfinanzierung durch das Bundesministerium für Bildung und Forschung (BMBF) erhält. Der andere Teil der Grundfinanzierung wird durch das Land Baden-Württemberg bereitgestellt. Darüber hinaus finanziert sich das HLRS durch Forschungsmittel (Projektförderung) und Einnahmen aus der Nutzung der HLRS HPC-Rechenkapazitäten durch Unternehmen und die Industrie. Die Ressourcennutzung durch die Privatwirtschaft ist auf ca. 10 % der Rechenkapazität beschränkt und machte im Jahr 2021 rund 2 % der Drittmiteinnahmen aus. Kennzeichnend für das HLRS sind die sogenannten Solution Center, die als externe Gesellschaften den Transfer in die Wissenschaft und Wirtschaft organisieren und den Zugang zu Höchstleistungsrechnern fördern.

Nutzung von HPC-Rechenkapazitäten aus dem Bereich der Forschung am Beispiel HLRS*:

Das HLRS kann zurzeit **maximal 192 GPUs (24 GPU-Knoten)** gleichzeitig für die Berechnung eines Foundation-Modells anbieten. Hier würden dann entsprechend **16.192 Knotenstunden** für einen Trainingsdurchlauf anfallen. Das entspricht in etwa **675 Tagen Dauerbetrieb** von **24 KI-Knoten** mit jeweils **8 GPUs**.

Eine exklusive Nutzung sämtlicher GPU-Knoten in dieser Form wäre am HLRS aktuell nicht realisierbar und nicht mit den zeitlichen Anforderungen an die Innovationszyklen bei der Entwicklung eines Foundation-Modells vereinbar (vgl. 1.3 Anforderungen an ein KI-Hochleistungsrechenzentrum).

Für die GPU/CPU Rechenleistung fallen die folgenden Kosten für die Berechnung eines Foundation-Modells an:

1. GPU-Nutzung

Die 24 KI-Knoten des HLRS benötigen ca. $23 * 694$ Knotenstunden für die Berechnung ($560/24 = \text{ca. } 23$) --> 16.193 Knotenstunden = 4 Durchläufe ergeben dann 64.772 Knotenstunden.

64.772 KI-Knotenstunden x 19,50 EUR pro KI-Knotenstunde entsprechen einem Preis von 1.263.054 EUR

Für das Preprocessing der Daten, das bis zu 20.000 CPU-Cores und eine geschätzte maximale Laufzeit von 1.000 Stunden in Anspruch nehmen wird, würden folgende Zielkosten für die Nutzung der CPU-Cluster am HLRS entstehen:

2. CPU-Nutzung

20.000 CPU-Cores für 840 Stunden Laufzeit --> $168.000.000$ Core-Stunden
Die aktuellen HAWK-Knoten am HLRS besitzen 128 CPU-Cores. Da hier pro Knotenstunde abrechnet wird, ergeben sich:

131.250 Knotenstunden x Forschungspreis (Stand: 2022) in Höhe von 1,13 EUR/Knotenstunde = ca. 80.000 EUR

Bei der Nutzung durch die Industrie wird am HLRS zudem ein Zuschlag in Höhe von **10% bis 30%** veranschlagt.

* Die Kostenabschätzung erfolgte auf Basis einer öffentlich zugänglichen Entgeltordnung mit Stand 2022.

An allen untersuchten Standorten mit HPC-Rechenkapazitäten stehen die benötigten GPU-Hardwareressourcen für LEAM nicht in ausreichendem Umfang zur Verfügung, sodass diese nicht als Bereitsteller von Infrastruktur in Frage kommen. Diese kommen eher als Nutzer von zukünftig verfügbaren spezialisierten KI-Recheninfrastrukturen in Frage. Kooperationen sind hier ebenfalls denkbar.

Kommerzielle Anbieter

Neben den Forschungseinrichtungen gibt es auch kommerzielle Anbieter, die HPC-Kapazitäten am Markt bereitstellen. Von diesen sind insbesondere die sogenannten Hyperscaler relevant. Als Hyperscaler werden Unternehmen bezeichnet, die sehr große Rechenzentren betreiben und ihren Kunden Clouddienste auf Basis von hochskalierbaren Infrastrukturen zur Verfügung stellen. Rechenleistung, Speicherkapazität und andere Ressourcen können hierbei nahezu verzögerungsfrei auf Anfrage bereitgestellt werden und die Abrechnung der verwendeten Ressourcen erfolgt in der Regel nach tatsächlicher Nutzung. Hyperscaler unterscheiden sich von anderen Anbietern durch die schiere Größe der jeweiligen Rechenzentren (teilweise deutlich über 100 MW Leistung) und ihre Fähigkeit, große Mengen an Daten und Rechenleistung schnell und effizient zu verarbeiten und die Dienste weltweit, hoch skalierend anbieten zu können. Um die gestiegenen Kundenanforderungen an niedrige Latenzen bedienen zu können, betreiben sie in der Regel mindestens ein Rechenzentrum auf den relevanten Kontinenten und bieten eine garantierte Datenspeicherung und -verarbeitung über verschiedene geografische Verfügbarkeitszonen an.

Die größten und bekanntesten Hyperscaler sind Amazon Web Services (AWS), Microsoft Azure und Google Cloud Platform (GCP) aus dem amerikanischen Raum sowie zunehmend die chinesischen Unternehmen Tencent Cloud und Alibaba. Daneben gibt es auch noch weitere große Unternehmen wie IBM, Oracle oder HPE, die HPC-Kapazitäten bereitstellen und einige kleinere, hochspezialisierte KMUs und Start-ups wie z.B. Lambda. Als deutsche Unternehmen bieten bspw. IONOS, Northern Data und auch Aleph Alpha HPC-Lösungen an.

Generell stellen alle diese Unternehmen spezielle Systeme zur Verfügung, die für den gedachten Einsatzzweck der Verarbeitung von KI-Anwendungen konzipiert wurden und Zugriff auf leistungsfähige GPUs ermöglichen. Unterschiede gibt es jedoch bei der verwendeten Hardware und der Bereitstellung. Bei den meisten Anbietern sind z.B. NVIDIAs A100 GPUs verfügbar. Der Einsatz der neuesten Generation der H100 GPUs wird aktuell evaluiert und soll in Kürze z.B. bei Microsoft Azure zur Verfügung stehen. Ein weiterer Unterschied ergibt sich in der Art der Bereitstellung der gewünschten Ressourcen. Als Cloudspezialisten bieten naturgemäß alle Unternehmen die Möglichkeit von virtualisierten Systemen an. Hierbei laufen auf einem Hostsystem ein oder mehrere virtualisierte Systeme. Manche Anbieter wie IBM und Microsoft Azure bieten darüber hinaus auch den Zugriff auf dedizierte Systeme an, die sich noch flexibler konfigurieren lassen. Im Fall von Microsoft Azure kann sogar ein dedizierter Supercomputer des Herstellers Cray integriert werden.

Auch wenn somit prinzipiell der Aufbau eines geeigneten Clusters möglich wäre, stehen zumindest in Europa nicht ausreichend Ressourcen in Form von GPUs zur Verfügung. Laut Aussage der befragten Expert:innen, gibt es keinen Anbieter, der die geforderte Anzahl von ca. 4.500 GPUs der neuesten Generation aus einem Rechenzentrum heraus zur Verfügung stellen kann. Die verteilte Nutzung von Ressourcen aus mehreren Rechenzentren scheitert aktuell an der notwendigen Bandbreite des internen Netzwerks. In diesem Zusammenhang muss jedoch erwähnt werden, dass die Anbieter vermutlich bereit wären, entsprechende Kapazitäten bei langfristigen Nutzungsverträgen

aufzubauen. Die Kosten hierfür dürften jedoch höher ausfallen als bei den anderen hier aufgezeigten Betriebsmodellen. Eine beispielhafte betriebswirtschaftliche Vergleichsrechnung für verschiedene HPC-Angebote für den Bezug von Rechenleistung aus der Cloud erfolgt in Kapitel 10.1.

Die Nutzung von Kapazitäten außerhalb Europas wäre zwar denkbar, steht aber dem Ziel des Aufbaus eines deutschen bzw. europäischen KI-Ökosystems zur Entwicklung, Bereitstellung, Betriebs sowie der Integration und Validierung besonders leistungsfähiger KI-Modelle entgegen.

Ein weiterer Punkt, der gegen den Aufbau eines cloudbasierten Systems unter Verwendung eines Hyperscalers spricht, ergibt sich aus der Zielgruppe der KMU. Auch wenn das Niveau und die Standards in Bezug auf Datensicherheit bei den Hyperscalern eher über dem Durchschnitt liegen, gibt es nach wie vor Bedenken bezüglich des Schutzes vor dem Zugriff Unbefugter, insbesondere staatlicher Stellen, auf die eigenen Daten. Je sensibler die eigenen Daten eingeschätzt werden, desto größer ist die Skepsis. Dies könnte unter Umständen dazu führen, dass notwendige Trainingsdaten nicht zur Verfügung gestellt werden. Hinzu kommt, dass viele Hyperscaler so genannte Lock-in Effekte zur Kundenbindung nutzen, die einen leichten Einstieg ermöglichen und einen späteren Wechsel zu einem anderen Anbieter erschweren.

SPOTLIGHT

Merantix Momentum GmbH

An verschiedenen Stellen in der Studie stellen wir ausgewählte deutsche Startups und Unternehmen vor, die KI-Foundation-Modelle bereits nutzen. Sie berichten über ihre Anwendungsfälle, Herausforderungen und die Vorteile europäischer Modelle.



Das Leadership-Team:
Dr. Johannes Otterbach,
Nicole Büttner-Thiel,
Dr. Thomas Wollmann.

Merantix Momentum ist ein in Berlin ansässiges KI-Startup, das auf die Einführung und Skalierung von KI-basierten Lösungen in verschiedenen Branchen spezialisiert ist. Mit einem erfahrenen Team lösen wir als KI-Service-Anbieter die Herausforderungen unserer Kunden durch maßgeschneiderte Machine-Learning-Lösungen und sichern so deren zukünftige Wettbewerbsfähigkeit in digitalen und datengetriebenen Märkten. Mit einer eigenen Forschungsabteilung unterstützen wir gleichzeitig aktiv den Transfer von Machine-Learning-Methoden in die produktive Anwendung bei Firmen und Organisationen in Deutschland und Europa.

Wo setzt ihr Foundation-Modelle ein? Was ist euer Use-Case?

Merantix Momentum spezialisiert sich auf die Implementierung und Umsetzung riskanter KI-Innovationsprojekte. Dabei begleiten wir unsere Kunden und Partner von der initialen Use Case Entwicklung, über die Datenstrategie bis hin zur Entwicklung und auch dem Produktionsbetrieb der KI-Lösungen.

Foundation-Modelle bieten dabei den Vorteil, schnell und effizient auf limitierten Datensätzen unserer Kunden, neue Deep-Learning Lösungen zu entwickeln. Hierbei stellt es sich als vorteilhaft heraus, dass Foundation-Modelle auf einer breiten Datenbasis trainiert wurden, die mittelbar auf die Kundendaten übertragbar sind. Mit einer breiteren Verfügbarkeit verschiedener Foundation-Modelle, das heißt trainiert auf diversen Daten, lässt sich somit die Anwendung moderner KI-Methoden auf bisher unerschlossene Anwendungen realisieren.

Welchen Einfluss haben KI-Foundation-Modelle auf euer Geschäftsmodell bzw. eure Projekte?

KI-Foundation-Modelle unterstützen uns bei der Projektentwicklung an mehreren Stellen. Zum einen helfen sie bei der Use Case Entwicklung, da sie ein schnelles Prototypisieren ermöglichen und damit die Kreativität unserer Kunden und Partner, aber auch unserer Entwickler:innen entfesseln. Zum anderen bieten sie aber auch die Möglichkeit, durch Distillation kleine und effiziente Modelle zu entwickeln, die durch das Trainieren eines neuen Modells von Grund auf gar nicht erst möglich gewesen wären, da die Datenlage oftmals nicht ausreichend ist. Damit schlagen KI-Foundation-Modelle gleich zweifach durch: In der Innovationsphase durch Unterstützung im Kreativprozess und in der anschließenden Entwicklung, die sonst nicht möglich wäre.

Wie würden euch europäische Modelle - Open Source, alle europäischen Sprachen abdeckend, mit hohen Datenschutzstandards und minimalem Bias - helfen?

Zunächst einmal sollte klargestellt werden, dass bestimmte Biases gewollt sind. Hierbei handelt es sich um ethische und moralische Grundeinstellungen der Modelle, eben basierend auf europäischen Werten. Dies wird bedingt durch die verschiedenen Kulturkreise, in welchen die Modelle entwickelt werden. Abgesehen davon sind Europa-basierte Modelle hilfreich, um Anwendungen schneller in die Praxis bringen zu können. Der europäische Bias ermöglicht es uns, voraussichtlich schneller die Modelle anpassen zu können mit weniger Daten. Zum anderen müssen wir uns weniger mit dem Output der Modelle und deren Untersuchung auf ungewollten Output oder mögliche Schwachstellen beschäftigen, da eine gewisse minimale Operationslinie angenommen werden kann, die man bei nicht-europäischen Modellen so nicht annehmen darf.

8.5.2 Erforderliche Standortparameter

Geopolitische und datenschutzrechtliche Verortung

Im europäischen Kontext stellen die sogenannten FLAP-Märkte (Frankfurt, London, Amsterdam und Paris) – in den vergangenen Jahren häufig durch Dublin als FLAP-D ergänzt – die historisch gewachsenen Gravitätszentren für die großen Internet Hubs in Europa dar. Durch diese digitalen Internetzentren verlaufen, ähnlich der Entwicklung der historischen Seidenstraße, die weltweiten Datenübertragungsleitungen für den globalen Internetverkehr. Mit Ausnahme von London liegen diese Zentren im Geltungsbereich der Europäischen Union und stellen damit einen datenschutzrechtlich sicheren EU-konformen Rechtsrahmen für ihre Nutzer:innen dar.

Historisch betrachtet folgt diese Entwicklung dem Aufbau der internationalen Telekommunikationshubs in der ersten Hälfte des 20. Jahrhunderts bzw. schon den Entwicklungen der ersten industriellen Revolution ab der zweiten Hälfte des 18. Jahrhunderts. Diese verdichtete Ansiedlung digitaler Infrastrukturelemente ist aus volkswirtschaftlicher Sicht im Bereich der sogenannten „Blauen Banane“, einer dicht besiedelten Kette von Ballungsräumen angefangen von Manchester, dem Großraum London über die Amsterdamer „Randstad“, das Ruhrgebiet, die Rhein-Main-Region hin zu den Industriestandorten in Mannheim, Ludwigshafen und Basel bis schlussendlich in die Industriezentren Norditaliens mit Mailand und Turin zu verorten.

Mit einer Übertragungskapazität von mehr als 50 Tbit/s zählt allein die West-Ost-Datentrasse von Dublin über Amsterdam Richtung Frankfurt zu den größten transeuropäischen Trassenführungen für das IP-Routing, entlang derer sich die großen Volumina an Datenverkehr im Internet bewegen. Dies entspricht mehr als dem fünffachen Durchsatz des heute weltweit größten Internetknotens in Frankfurt (Simons & Frese, 2021). Ein weiterer Datenkorridor (Nord-Süd) verbindet die skandinavischen Rechenzentrumsansiedlungen mit den europäischen und amerikanischen Content-Anbietern via Stockholm, Kopenhagen, Düsseldorf, Frankfurt und Paris.

Entlang dieser Trassen hat sich in den vergangenen 20 Jahren eine digitale Ökonomie mit allen Ausprägungen der Wertschöpfungskette herausgebildet. Um diese großen Rechenzentrumsansiedlungen haben sich häufig digitale Ökosysteme aus den Bereichen Software-Entwicklung, Content, KI oder IT-Dienste angesiedelt. Den Gravitätsanker für diese Ansiedlungen bilden häufig kurze Latenzzeiten, breitbandige Anbindung an die transkontinentalen Backbone-Trassen, verdichtete Metropolstrukturen mit der entsprechenden Anzahl potenzieller Nutzer:innen sowie ein ausreichend zur Verfügung stehender Markt an gut ausgebildeten Fachkräften.

Sowohl auf europäischer als auch auf nationaler Ebene ist der Wachstumstrend im Rechenzentrumsmarkt weiterhin ungebrochen. Zwischen 2016 und 2021 wuchsen die Kapazitäten gemessen in IT-Anschlussleistung um 30 % (Hintemann et al., 2022). Die Profiteure dieses Trends sind allen voran die Rechenzentrumsanbieter mit installierten Leistungen ab 5 MW und deutlich darüber hinaus. Kleinere Installationen unter 5 MW befinden sich aktuell in einem stagnierenden oder absteigenden Trend.

Gemessen an der Bedeutung nimmt die Rhein-Main-Region als Rechenzentrums-Standort in Deutschland unverändert eine Spitzenposition ein. Neben Frankfurt wird künftig Berlin für Rechenzentrums-Entwicklungen eine immer stärkere Rolle spielen. Gemessen an der IT-Anschlussleistung pro Einwohner kommt Hessen auf einen mehr als dreimal höheren Wert als die Stadtstaaten Hamburg oder Berlin. Neben Berlin werden künftig auch Standorte wie München, Hamburg sowie die Regionen Köln/Düsseldorf und Leipzig/Dresden für Rechenzentrumsentwicklungen immer wichtiger (Hintemann et al., 2022).

Im Bereich High Performance Computing (HPC) sind in Deutschland folgende bedeutende Standorte zu nennen:

Stadt	Einrichtung	Peak Performance	Memory	CPU-Cores
Eggenstein-Leopoldshafen	Steinbuch Centre for Computing - ForHLR - bwUniCluster	1171 TFlop/s 444 TFlop/s	136 TB 86 TB	34800 Cores 18304 Cores
Frankfurt	Center for Scientific Computing - LOEWE CSC - FUCHS	823 TFlop/s 41 TFlop/s	70 TB 18 TB	18960 Cores 6456 Cores
Offenbach	Deutscher Wetterdienst - Cray XC40 - Cray CX40 II	1073 TFlop/s 1073 TFlop/s	125 TB 125 TB	29552 Cores 29952 Cores
Hamburg	Deutsches Elektronen Synchrotron - Standort Zeuthen - Maxwell	16 TFlop/s -	- 402 TB	2288 Cores 26732 Cores
Hamburg	German Climate Computing Center Mistral (HLRE-3)	3590 TFlop/s	266 TB	101196 Cores
Kaiserslautern	Competence Center HPC - Beehive - Seislab - Ocean 1	67 TFlop/s 35 TFlop/s -	14 TB 6 TB 23 TB	3224 Cores 1584 Cores 11600 Cores
Kaiserslautern	Reg. Hochschulrechenzentrum - Elwetritsch - Elwetritsch II	134 TFlop/s	17 TB 53 TB	5624 Cores 10520 Cores
Köln	Reg. Rechenzentrum Uni Köln - CHEOPS	100 TFlop/s	36 TB	9712 Cores
Göttingen	Ges. f. wissen. Datenverarbeitung - Scientific Compute Cluster - Emmy	2883 TFlop/s 8261 TFlop/s	92 TB 498 TB	16640 Cores 116152 Cores
Darmstadt	Hochschulrechenzentrum - Lichtenberg II	3148 TFlop/s	251 TB	61824 Cores
Stuttgart	Höchstleistungsrechenzentrum - Hawk - Cray CS-Storm - vulcan	26000 TFlop/s - -	1 TB 9 TB 102 TB	702896 Cores 608 Cores 13856 Cores
Aachen	IT Center RWTH - CLAIK-2016 - CLAIK-2018	678 TFlop/s 4965 TFlop/s	88 TB 251 TB	16152 Cores 62736 Cores
Jülich	Jülich Supercomputing Centre - JUWELS	12000 TFlop/s	286 TB	123088 Cores
Berlin	Konrad-Zuse Zentrum - Lise	7907 TFlop/s	455 TB	110016 Cores
Garching	Leibniz Rechenzentrum - Super MUC - Super MUC NG	3580 TFlop/s	197 TB	86016 Cores
Garching	Max Planck Computing & Data Facility - COBRA	12720 TFlop/s	530 TB	136960 Cores
Paderborn	Paderborn Center for Parallel Comp. - XCL FPGA Cluster - Noctua 1 - Noctua 2	- 835 TFlop/s 7100 TFlop/s	512 GB 53 TB 355 TB	32 Cores 10960 Cores 143488 Cores
Mainz	Zentrum für Datenverarbeitung - Mogon - Clover - Mogon 2	379 TFlop/s 106 TFlop/s 3125 TFlop/s	90 TB 10 TB 194 TB	35760 Cores 5120 Cores 52248 Cores
Dresden	Center for Information Services / HPC - Taurus - Alpha Centauri	2621 TFlop/s -	279 TB 35 TB	64536 Cores 1632 Cores
Erlangen	Erlangen National Center for HPC - Emmy - Meggie - TinyGPU - TinyVec - TinyARM - TinyFAT - Alex - Fritz	232 TFlop/s 511 TFlop/s - - - - - -	36 TB 47 TB 5 TB 96 TB 128 GB 22 TB 65 TB 242 TB	11088 Cores 14560 Cores 1392 Cores 12 Cores 64 Cores 2484 Cores 8960 Cores 67968 Cores

Tabelle 10: HPC-Standorte in Deutschland

Die oben aufgeführten Rechenzentrumsstandorte für HPC-Anwendungen lassen sich in unmittelbarer Nähe zu Universitäten und Forschungseinrichtungen verorten und folgen nicht zwingend der vorab beschriebenen Entwicklungslogik gewerblicher Collocation und Hyperscale-Rechenzentren entlang historisch entwickelter Datentrassen. Im wissenschaftlichen und universitären Bereich haben sich diese Rechenzentren in der Regel über eigene Netze (z.B. das Wissenschaftsnetz X-WiN, welches vom Deutschen Forschungsnetz DFN betrieben wird) untereinander verbunden. Die Anbindung des X-WiN an externe Netze erfolgt an dedizierten Standorten über lokale Internet-Knotenpunkte, wie z.B. am DE-CIX in Frankfurt am Main und Hamburg, am ECIX in Düsseldorf und am BCIX in Berlin.

Regionale Verortung: Flächenbedarf, Kubatur, Gebäudeinfrastruktur

Bei der regionalen und lokalen Verortung von Rechenzentren rücken andere Kriterien in den Vordergrund als bei einer geopolitischen oder nationalen Betrachtung. Als generelle Ansiedlungskriterien von Rechenzentren wären hier zu nennen:

- Die Grundstücke liegen nicht in direkter Nachbarschaft zu oder in Wohngebieten.
- Eine einfache Erreichbarkeit, u.a. durch öffentlichen Personennahverkehr oder Straßenanbindung ist gegeben.
- Ausschluss von ansiedlungsbehindernden Bedrohungs-/Gefahrenlagen (siehe auch Seveso II/III-Gebiete): Nähe zu Flughäfen, chemischer Industrie, Güterverkehrsstrecken, Elektromagnetische Exposition, Schwingungsquellen etc. Aber auch mögliche Gefahrenlagen durch Naturereignisse (Hochwasser, aktive seismische Zonen, Nähe zu Küstenlinien, etc.) sind zu vermeiden.
- Es existieren Ansätze für lokale Nutzungen des Rechenzentrums, u.a. ein vielfältiges Nutzerspektrum an datenzentrierten Unternehmen und Forschungsinstitutionen; idealerweise im Umkreis von 50 km.
- Eine zuverlässige und redundant ausgebaute sowie skalierbare Stromversorgung wird bereitgestellt. Eine räumliche Nähe zum Umspannwerk wird hierbei angestrebt. Bei redundanter Versorgung über zwei Umspannwerke wird häufig die geographische Mitte zwischen zwei Umspannwerken bevorzugt.
- Eine gute Anbindung an überregionale Glasfasertrassen ist gegeben. Idealerweise sind mindestens zwei überregionale (sog. Longhaul) Glasfasertrassen-Anbieter vorhanden, welche in Summe drei schleifen- und kreuzungsfreie Wegeführungen mit mindestens 20 nutzbaren Glasfaserpaaren je Weg zur Trassenanbindung realisieren können; die maximale Distanz zur Trasse beträgt 5 km, bezogen auf eine potenzielle Ansiedlungsfläche.
- Es ist ein Zugang zu einem Internetknoten-Anbieter vorhanden, der mit einer diskriminierungsfreien, verteilten Plattformkonzeption eine Vielzahl von Interconnection-Diensten auf Enterprise Niveau realisieren kann.
- Ein zukünftig wichtiger werdender Faktor wird die Nähe zu möglichen Abnehmern von Abwärme (z.B. Quartiers-Konzepte, Schwimmbad, Vertical Farming etc.) oder die direkte Einspeisemöglichkeit in ein Nah- oder Fernwärmenetz bilden.

Flächenbedarf

Für Großrechenzentren und Hyperscaler spielt eine ausbaufähige Flächen-Skalierbarkeit eine entscheidende Rolle für das jeweilige Ansiedlungskonzept. Ansiedlungen von mittelständischen Betreibern gehen mit entsprechend kleineren Grundstücksgrößen ins Rennen. Grundsätzlich zu unterscheiden sind:

- „Solitär“-Rechenzentrum: Bau eines einzelnen in sich geschlossenen Rechenzentrums (häufig für unternehmensinterne Verwendung). Größenordnung: 0,5 bis ca. 3 ha Grundstücksfläche, Leistungsaufnahmen in der Regel bis 10 MW
- Campus Rechenzentrum: Es stehen mehrere jeweils für sich autark betriebene Rechenzentrums-Betriebseinheiten auf einem größeren Rechenzentrums-campus. Das Betriebsmodell ist häufig auf Collocation, Shell & Core oder Hyperscale ausgerichtet. Größenordnung: 10 ha Grundstücksfläche und mehr, Leistungsklasse: 50-120 MW
- Großer Hyperscale Campus: Ab ca. 15 ha aufwärts. Das aktuell größte Ansiedlungsprojekt in Europa befindet sich in der Nähe von Esbjerg, Dänemark und umfasst ca. 200 ha Arealfläche (die Größe Monacos); Leistungsklassen: >100 MW bis hin zu 300 MW und darüber hinaus.

Kubatur / Baukennzahlen

Das Maß der baulichen Nutzung ist als Angabe über die Intensität einer Grundstücksausnutzung Bestandteil des öffentlichen Baurechts und stellt somit ein wichtiges städtebauliches Steuerungsinstrument des BauGB dar. Bei der Errichtung von Rechenzentren sind aus baurechtlicher Sicht verschiedene Vorgaben einzuhalten. Unter anderem gibt die Grundflächenzahl (GRZ) vor, wie groß der Flächenanteil eines Grundstücks sein darf, der überbaut werden darf. Die Kubatur beschreibt den umbauten Raum eines Gebäudes. Analog zur Erhöhung der Leistungsdichte in einem Serverrack lässt sich die Gesamtanzahl betreibbarer Server mit ansteigender Zahl der Geschosse (Geschossflächenzahl GFZ) eines Gebäudes erhöhen. Die Baumassenzahl (BMZ) schließlich gibt an, wie viele Kubikmeter Baumasse je Quadratmeter Fläche eines Grundstücks zulässig sind.

Diese Vorgaben sind im lokalen Baurecht entsprechend verankert und können vom Errichter nicht verändert werden. In Deutschland sind mehrgeschossige Rechenzentrumsbauten in Industriegebieten üblich, jedoch sind Beispiele von Rechenzentren in Hochhäusern, wie die *60 Hudson Street* in Manhattan oder das *Royal Bank of Canada Data Centre* in Toronto in Deutschland bisher nicht anzutreffen.

Energieversorgung: Anbindung und Energieversorgungssicherheit

Rechenzentren erfordern für einen (ausfall-)sicheren Betrieb ein redundantes und durchdachtes Stromversorgungskonzept. Um die erforderliche Redundanz zu erreichen, kann, sofern möglich, auf die getrennte Einspeisung über zwei Umspannwerke zurückgegriffen werden. Ist dies nicht der Fall, empfiehlt sich eine Ringeinspeisung über zwei getrennte Zuführungen eines Umspannwerkes.

Wird das Rechenzentrum über eine Mittelspannung (10 kV oder 20 kV) versorgt, ist in der Regel auf dem Grundstück eine kundeneigene oder energievorsorgereigene Mittelspannungsstation errichtet. Transformatoreinspeisung und Messeinrichtungen sind in der Regel in diese Anlagen integriert. Um die Mittelspannung entsprechend heruntertransformieren zu können, werden üblicherweise in Rechenzentren Gießharztransformatoren eingesetzt. Gegenüber herkömmlichen Trockentransformatoren wird hierbei eine höhere Teilentladungs-, Kurzschluss- und Spannungsfestigkeit erreicht und die Empfindlichkeit gegenüber Umweltbelastungen wie Feuchtigkeit und Staub ist ebenfalls geringer (Dürr, 2018). Vereinzelt kommen auch Öltransformatoren zum Einsatz. Diese sind bauartbedingt verlustarmer und weisen eine höhere Energieeffizienz auf. Öltransformatoren weisen jedoch andere Brandschutzerfordernisse auf, die ggf. einen größeren Raumbedarf mit sich bringen. Eine weitere notwendige Komponente zur Sicherstellung der Energieversorgung stellen die Unterbrechungsfreien Stromversorgungen (USV) sowie die Netzersatzanlagen (NEA; z.B. Dieselgeneratoren) dar.

Die Aufgabe einer USV besteht vorrangig in zwei Punkten: a) der kurzfristigen Überbrückung bei Stromausfällen durch Umschaltung auf Batteriespeicher oder Schwungmassenspeicher sowie b) der Eliminierung von Spannungsschwankungen und -spitzen sowie Frequenzabweichungen. Ausgehend vom Wirkungsprinzip lassen sich dabei drei grundsätzliche Typen von USV-Anlagen unterscheiden:

- VFD (Voltage and Frequency Dependent): Der Strom wird bei diesem USV-Typ direkt vom Eingang zum Ausgang durchgeleitet. Hierbei erfolgt keine „galvanische Entkopplung“ mit der vorgeschalteten Last. Über den Eingang der USV wird kontinuierlich der Batteriespeicher mit Energie versorgt. Die Umschaltung im Falle eines Stromausfalls auf Batteriebetrieb ist jedoch unterbrechungsbehaftet und kann bis zu 10 ms benötigen. Dies ist ein Wert, der sich bei empfindlicher IT-Hardware u.U. bereits bemerkbar machen kann.
- VI (Voltage Independent): USV-Geräte vom Typ VI arbeiten mit einem AC/DC-Wandler als zentrale Komponente, der sowohl als Stromrichter als auch für die Aufladung der Batterien zuständig ist. Die USV läuft im Gegensatz zum vorgenannten Typ auch dann „aktiv“ mit, wenn der Strom über die Netzspannung zur Verfügung steht. Es werden jedoch Spannungsspitzen herausgefiltert, welche die IT-Hardware schädigen könnten. Die Umschaltzeit liegt hierbei zwischen 2,5 und 10 ms.
- VFI (Voltage and Frequency Independent): Im Gegensatz zu den beiden oben aufgeführten USV-Typen stellt eine USV vom Typ VFI zusätzlich noch sicher, dass es neben Spannungsspitzen und -schwankungen nicht auch zu unerwünschten Frequenzabweichungen kommt. Der Strom an der Ausgangsseite ist hier komplett entkoppelt vom Strom an der Eingangsseite. Da die USV dieses Typs im Dauerbetrieb arbeitet, fallen keine zusätzlichen Umschaltzeiten an.

Bei der Auslastung der USV-Systeme ist darauf zu achten, dass mit sinkendem Auslastungsgrad der Wirkungsgrad der Anlage kontinuierlich abnimmt. Die erhöhte Verlustleistung macht sich so bei der Energieeffizienz negativ bemerkbar. Die Aufgabe eines Systems aus USV und Batteriespeicher besteht jedoch immer nur darin, eine relativ kurze Zeit (wenige Minuten) zu überbrücken, bis eine Netzersatzanlage (NEA; = Notstromdiesel) für den längerfristigen Notbetrieb angelaufen ist. Notstromersatzanlagen sind nach ISO 8528 genormt.

Dieselbetriebene Notstromersatzanlagen müssen regelmäßigen Tests zur Sicherstellung des Betriebs unterzogen werden. Hierbei werden verschiedene Testbetriebsarten unterschieden:

- Netzparalleler Lastprobetrieb: Die NEA wird gestartet und mit der Sinuswelle der Netzversorgung synchronisiert. Im Anschluss wird der Generatorschalter eingekuppelt und die NEA parallel zum Netz betrieben. Nach vorhergehender Rücksprache mit dem Energieversorger kann sogar Last in das Netz zurückgespeist werden.
- Lastprobetrieb im Inselbetrieb: Die NEA wird im Netzparallelbetrieb wie oben beschrieben gestartet. Nach Hochlaufen der Last wird jedoch der Netzschalter ausgekoppelt, sodass die NEA nun die volle Last für den Betrieb der Server erbringen kann. Diese Methode ist der vorangestellten vorzuziehen, da sie die realen Bedingungen im Falle eines Netzausfalles besser abbilden kann.
- Netzausfalltest („Back Building Test“): Hierbei handelt es sich um einen „echten“ Netzersatztest. Der zentrale Netzschalter wird vor dem Anlaufen der NEA ausgekoppelt, so dass USV und NEA spontan einspringen müssen. Viele Betreiber schrecken vor dieser Art des Netztests zurück, da sie das Risiko zu hoch einschätzen, dass Anlagen nicht anlaufen und es zu einem „echten“ Ausfall im Rechenzentrum kommen kann.

Die Kunden im Rechenzentrum werden in der Regel über anstehende Tests von USV und NEA im Vorfeld durch den Betreiber informiert. Testläufe von Dieselaggregaten müssen (u.a. auch aus emissionsrechtlichen Erfordernissen) bei den zuständigen Behörden genehmigt werden. In der Regel wird ein Stundenkontingent pro Jahr (z.B. 30 h/a) für den Test-Betrieb genehmigt. Die Dieselgeneratoren werden im Standby-Betrieb elektrisch vorgewärmt, um im Einsatzfall möglichst kurze Anlaufzeiten bis zum Erreichen der vollen Last gewährleisten zu können. Somit verbrauchen NEAs auch offline einen gewissen Betrag elektrischer (Heiz-)Energie.

Verfügbarkeitsklassen

Rechenzentren werden nach dem Grad vorhandener redundanter Komponenten in sogenannte Verfügbarkeitsklassen (VK1 - VK4) unterteilt. Dabei unterscheidet man:

- Verfügbarkeitsklasse 1 (N): Bezeichnet man einzelne Komponenten eines Rechenzentrums (z.B. eine NEA, eine USV, ein Klimaschrank) mit der Variablen „N“, so liegt bei diesem Konzept keine zusätzliche Redundanz vor. Bei Ausfall einer Komponente muss diese zuerst gewartet/repariert werden, bevor diese wieder in Betrieb geht.
- Verfügbarkeitsklasse 2 (N+1): Der Ausfall einer einzelnen Komponente führt hier nicht zum Ausfall des kompletten Versorgungspfades, da eine zusätzliche Ersatzkomponente (+1) einspringen kann. Beispiel: Zur Kühlung des Serverraums werden 5 Umluftklimageräte benötigt. Das sechste, im Raum verbaute Gerät springt im Fall des Ausfalls eines anderen Geräts ein.
- Verfügbarkeitsklasse 3 (2N): Bei dieser Redundanzkonzeption sind sämtliche Versorgungspfade „doppelt“ ausgelegt. Beispiel: Sämtliche Server sind mit zwei Netzteilen ausgerüstet, die über zwei verschiedene Stromphasen mit Strom versorgt werden. Durch Abschaltung einer kompletten Stromphase (z.B. bei Wartungsarbeiten) ist der operative Betrieb im Serverraum weiterhin gewährleistet.
- Verfügbarkeitsklasse 4 (2N+1): Dies stellt die höchste Verfügbarkeitsklasse dar. Im Gegensatz zur 2N Konzeption kann bei Wartungsarbeiten eines kompletten Versorgungspfades zusätzlich noch die Ausfallsicherheit einer Einzelkomponente sichergestellt werden.

Den Verfügbarkeitsklassen sind entsprechende maximale Ausfallzeiten pro Jahr zugeordnet:

Verfügbarkeitsklasse	Bezeichnung	Mindestverfügbarkeit	Max. Ausfallzeit pro Monat	Max. Ausfallzeit pro Jahr
VK 0	Ohne zugesicherte Verfügbarkeit	--	--	--
VK1	Normale Verfügbarkeit	99,0 %	< 8 h	< 88 h
VK2	Erhöhte Verfügbarkeit	99,9 %	<44 min	<9 h
VK3	Hochverfügbarkeit	99,99 %	<5 min	<53 min
VK4	Höchstverfügbarkeit	99,999 %	<26 sek.	< 6 min

Tabelle 11: Verfügbarkeitsklassen (VK1 - VK4)

Mit einer höheren Redundanzauslegung und der damit verbundenen Verfügbarkeitsklasse steigen auch die Investitionskosten an. Da die zusätzlichen Komponenten oftmals nicht situativ im Falle einer Wartung oder eines Ausfalls hinzugeschaltet werden, sondern vielmehr im sogenannten Halblastparallelbetrieb (mit)laufen, sind bei höherer Redundanzauslegung auch die Effizienzwerte geringer als im Betrieb ohne Redundanzauslegung. Dies wird von den Betreibern jedoch bewusst in Kauf

genommen, um einen höheren Grad an Ausfallsicherheit zu gewährleisten. Für LEAM wird für die Inference Anwendungen eine Verfügbarkeit von 99 % benötigt.

Glasfaserversorgung: Backbone-Netze, Redundanzen, diskriminierungsfreier und Carrier-neutraler Zugang, Nähe und Zugang zu Internetaustauschknoten

Ähnlich wie die Stromversorgung von der externen Übertragungsinfrastruktur des Energieversorgers über Mittelspannungs- und Niederspannungsverteilung auf dem Rechenzentrums Gelände zum Server-Netzteil geleitet wird, erfolgt auch die Datenanbindung einer vergleichbaren Verteilstruktur.

- **Übergabepunkt:** Über nationale und internationale Carrier-Anbindungen wird die externe Datenanbindung zu den Rechenzentren über interne Meet-Me Räume (MMR) sichergestellt. Meet-Me Räume sind hierbei der zentrale Ort innerhalb eines Collocation-Rechenzentrums, an dem sich Telekommunikationsunternehmen und Carrier sowie die Kunden des Collocation-Betreibers physisch miteinander verbinden und Daten austauschen können. Oftmals befinden sich die Hochleistungsrouter eines dezentral aufgestellten Internetknotenpunkts in den Meet-Me Räumen der Rechenzentrumspartners („Enabled Sites“) und ermöglichen so den Zugang zu den ‘Connected Networks‘ des Knotenbetreibers.
- **Standortverteilung (SV):** Vom Meet-Me Raum als zentralen Übergabepunkt in das externe Netz gelangen die Daten über eine Primär/Campusverkabelung zu den einzelnen Stockwerken/Serverräumen.
- **Gebäudeverkabelung (GV):** Ggf. unterteilt sich die Netzwerkverkabelung in einzelne Stockwerke über entsprechende Etagenverteiler (vertikale Verteilung).
- **Tertiärverkabelung (EV):** Über eine Tertiär- oder Etagenverteilung erfolgt dann die Zuleitung der Verkabelung an die Serverschränke in den einzelnen Serverräumen.

Bei der Verteilung auf Serverschrankebene lassen sich zwei Konzepte voneinander unterscheiden:

- **End of Row (EoR):** Der erste und/oder der letzte Schrank einer Rackreihe ist mit dem Zugangsswitch für die Anbindung aller anderen Schränke der jeweiligen Rackreihe ausgestattet. Der Schrank mit dem EoR Switch muss dabei eine große Anzahl von Patchkabeln für die horizontale Verkabelung über seine Patchpanels unterbringen. Der Vorteil dieser Anordnung besteht im vereinfachten Change-Management, da sämtliche Patchkabel einer Rackreihe an diesem zentralen Ort zusammenlaufen.
- **Das Middle of Row Konzept (MoR)** ist mit dem des EoR vergleichbar, nur dass hierbei der zentrale Schrank mit den Zugangsswitchen in der Mitte der Rackreihe positioniert ist.
- **Top of Row (ToR):** Bei diesem Konzept befinden sich in jedem Schrank (zumeist oben) eigene Switches. So können die Patchkabel in der Regel sehr kurzgehalten werden. Bei hoher Rack Anzahl sind jedoch viele kleinere (Edge-)Switches erforderlich. Schrankübergreifende Change-Requests sind beim ToR Konzept nicht so leicht zu realisieren. Das Konzept ist kostenintensiver, da in der Regel mehr Switches benötigt werden.

Schutzbedarf/Risiko-Analyse

Viele Rechenzentren stellen systemrelevante Dienstleistungen zur Verfügung. Collocation- Cloud- und Hosting-Rechenzentren mit einer vertraglich vereinbarten Leistung von mehr als 3,5 MW fallen unter die Verordnung zur Bestimmung Kritischer Infrastrukturen (BSI-KritisV). Unabhängig davon haben alle Rechenzentrumsbetreiber mehr oder weniger strenge Sicherheitskonzepte für ihre Anlagen vorgesehen. Hierzu zählt die Etablierung eines Zonenkonzeptes. Die Gebäudeteile und Abschnitte eines Rechenzentrums(Campus) werden entsprechend ihrer Kritikalität in verschiedene Zonen und Sicherheitsbereiche eingeteilt.

Von außen nach innen lassen sich nach einem „Zwiebelschalen-Prinzip“ die folgenden fünf Zonen/Bereiche unterscheiden:

- Zone I: Das Grundstück oder Firmengelände ist entweder frei zugänglich oder mit einem Zaun und einer Videoüberwachungsanlage gesichert.
- Zone II: Halböffentlicher Bereich innerhalb des Betriebsgelände, z.B. mit normalen Büroarbeitsplätzen für Mitarbeiter:innen
- Zone III: IT-Nebenträume und administrative Steuerung. Spätestens dieser Bereich ist nur noch einem bestimmten Personenkreis vorbehalten. Oftmals existieren hier Zugangspunkte in Form von Schleusen und Personenvereinzelanlagen.
- Zone IV: Direkter Zugang zu technischen Anlagen des Rechenzentrums. Oftmals ist der Zugang in diesem Bereich ausschließlich für das technische Wartungspersonal vorgesehen.
- Zone V: Zugang zu den Serverräumen, dem eigentlichen „Herz“ des Rechenzentrums. In Hochsicherheitsrechenzentren sind die Zugänge für die Serverräume sowie die Technikräume für Klima/Strom so ausgelegt, dass sich die entsprechenden Mitarbeiter:innen in separaten Gängen bewegen und eine direkte Begegnung ausgeschlossen wird. Dieser Aspekt ist besonders wichtig für den Fall, dass externes Wartungs- und Technikpersonal zum Einsatz kommt.

Ergänzend zum Zonenkonzept werden in der DIN EN 50600-1 vier unterschiedliche Schutzklassen definiert:

Schutzart	Art des Zugangs
Schutzklasse 1	Öffentlicher oder halböffentlicher Bereich.
Schutzklasse 2	Bereich, der allen autorisierten Personen (Mitarbeiter:innen und Besucher:innen) zugänglich ist.
Schutzklasse 3	Bereich, der festgelegten Mitarbeiter:innen und Besucher:innen vorbehalten ist. Andere Personen mit Zugang zu Schutzklasse 2 müssen von Personen begleitet werden, die Zugang zu Bereichen der Schutzklasse 3 haben.
Schutzklasse 4	Bereich, der festgelegten Mitarbeiter:innen, die einen nachgewiesenen Bedarf für den Zugang haben, vorbehalten ist. Andere Personen mit Zugang zu Schutzklasse 2 und 3 müssen von Personen begleitet werden, die Zugang zu Bereichen der Schutzklasse 4 haben.

Tabelle 12: Schutzklassen nach DIN EN 50600-1

Die unterschiedlichen Schutzklassen 1 – 4 sind dabei durch geeignete Maßnahmen und Prozesse entsprechend abzusichern. Die wichtigsten Punkte hierbei sind u.a.:

- **Schutzklasse 1:**

- Identifizierbare physische Sperre an der externen (Grundstücks)Grenze.
- Türen, Fenster, Gitter müssen der Widerstandsklasse 2 (DIN EN 1627:2011) entsprechen.
- Physische Trennung des Zugangs von Fußgängern und Fahrzeugen der Schutzklasse 1 und 2.
- Ausgewiesene Parkplätze für nicht autorisierte Fahrzeuge (Besucher:innen).

- **Schutzklasse 2:**

- Identifizierbare physische Barriere an der Grenze zur Schutzklasse 2.
- Türen, Fenster, Gitter müssen der Widerstandsklasse 3 (DIN EN 1627:2011) entsprechen. Fenster und Türen müssen so konstruiert sein, dass sie im geschlossenen Zustand außerhalb der Schutzklasse 2 nicht geöffnet werden können.
- Physische Trennung des Zugangs von Fußgängern und Fahrzeugen der Schutzklasse 2 und 3.
- Maßnahmen zur Erkennung und Verhinderung unerwünschten und unnötigen Zugangs.
- Jedes Öffnen einer Notausgangstüre muss einen Alarm auslösen, der eine geeignete Reaktion auslöst.

- **Schutzklasse 3:**

- Identifizierbare physische Sperre an der externen Grenze.
- Türen, Fenster, Gitter müssen der Widerstandsklasse 4 (DIN EN 1627:2011) entsprechen.
- Begrenzungen im Bereich der Schutzklasse 3 dürfen nicht mit denen der Schutzklasse 1 örtlich zusammen angeordnet werden.
- Begrenzungen im Bereich der Schutzklasse 3, die mit Begrenzungen von Bereichen der Schutzklasse 2 zusammen angeordnet sind, müssen der Summe des Widerstands für Schutzklasse 2 und 3 entsprechen.
- Folgende Maßnahmen müssen vorhanden sein, um folgende Ereignisse zu erkennen und zu verhindern: Unerwünschter oder unnötiger Zugang zwischen Flächen der Schutzklasse 3 und 4; nicht autorisierter Zugang von einer Fläche der Schutzklasse 3 in Schutzklasse 4; Erkennung aller Personen sowie Materialien und Geräte (z.B. IT-Equipment), die Schutzklasse 3 betreten oder verlassen.

- **Schutzklasse 4**

- Identifizierbare physische Sperre an der externen Grenze.
- Türen, Fenster, Gitter müssen der Widerstandsklasse 4 (DIN EN 1627:2011) entsprechen.
- Begrenzungen im Bereich der Schutzklasse 4 dürfen nicht mit denen der Schutzklasse 1 örtlich zusammen angeordnet werden.
- Begrenzungen im Bereich der Schutzklasse 4, die mit Begrenzungen von Bereichen geringerer Schutzklassen zusammen angeordnet sind, müssen der Summe des Widerstands für alle Schutzklassen entsprechen.
- Alle Durchbrüche der physischen Begrenzung müssen den Zugang für nicht autorisierte Personen verhindern. Darin eingeschlossen sind z.B. auch Druckentlastungsklappen für Gaslöschanlagen.

Zugangskontrolle

Um den kontrollierten Zugang in bzw. aus dem Rechenzentrumsbereich zu gewährleisten sind in den Rechenzentren Zugangskontrollsysteme mit entsprechenden dahinterliegenden Prozessen installiert. Grundsätzlich lassen sich Online- und Offline-Zugangssysteme unterscheiden. Online-Anlagen sind permanent mit einer Zutrittskontrollzentrale verbunden. Alle Zutrittsbewegungen werden hierbei direkt an eine zentrale Stelle gesendet. Bei Offline-Anlagen ist die Zutrittsberechtigung auf entsprechenden Zutrittskarten gespeichert und können z.B. über das Zeiterfassungsterminal oder Onlineleser tagesaktuell übertragen werden (Dürr, 2018).

Zur Identifikation und Autorisierung sind unterschiedliche Systeme im Markt erhältlich. Folgende Grundtypen können unterschieden werden:

- Physikalische Erkennung: mittels Leser und Ausweisen oder Schlüsselanhänger.
- Logische Erkennung: mittels Eingabe von Zahlencodes. (Diese Art des Zugangs ist nicht personalisiert und birgt die Gefahr, dass Codes unberechtigt weitergegeben werden können).
- Biometrische Erkennung: Eindeutige Identifikation einer zugangsberechtigten Person z.B. mittels Fingerabdrucks, Iris-Scan, Handrückenerkennung, Venenerkennung.

Zur Erhöhung der Zugangssicherheit lassen sich die beschriebenen Methoden in der Regel auch miteinander kombinieren.

8.6 Betrieb eines KI-Rechenzentrums

Für den Betrieb eines KI-Rechenzentrums kommen generell drei Optionen in Betracht. Die Zusammenarbeit mit einem Collocation-Anbieter, der Aufbau eines eigenen HPC-Rechenzentrums sowie die Nutzung einer verteilten Infrastruktur. Die drei Optionen werden im folgenden näher beleuchtet.

Collocation

Ein Collocation-Betreiber ist ein Unternehmen, das Räume und Infrastrukturen für die Unterbringung von Rechenzentren und anderen IT-Anlagen bereitstellt. Collocation-Betreiber bieten ihren Kunden die Möglichkeit, ihre Rechenzentren und IT-Anlagen in Räumen unterzubringen, die speziell für den Betrieb von Rechenzentren ausgestattet sind (vgl. Tabelle 6). Diese Räume sind in der Regel mit Stromversorgungssystemen, Kühlungstechnologien, Netzwerkverbindungen und Sicherheitsmaßnahmen ausgestattet, um sicherzustellen, dass die Rechenzentren und IT-Anlagen ordnungsgemäß betrieben werden können. Collocation-Betreiber bieten ihren Kunden auch Dienstleistungen wie Wartung und Support für IT-Anlagen an. Sie tragen somit dazu bei, dass Unternehmen ihre IT-Anlagen sicher und zuverlässig betreiben können, ohne sich um die notwendige Infrastruktur und die Wartung der Geräte zu kümmern. Der Normalfall ist allerdings, dass die Wartung und Installation der Racks durch den Kunden oder von ihm beauftragten Personal in den Räumen des Collocation-Anbieters durchgeführt wird.

Bei der Auswahl eines Collocation Anbieters spielt auch die Erfüllung von Anforderungen an Kühlungssysteme eine wichtige Rolle. Weitere Kriterien sollten der Bezug von grünem Strom sowie ein schlüssiges Konzept zur Abwärmenutzung sein. Der Collocation Anbieter sollte demnach über entsprechende Infrastruktur zur Wasserkühlung verfügen. Die Anforderungen für das Training von Foundation-Modellen wurden bereits im Kapitel 2.2 erläutert.

[GSI - Helmholtzzentrum für Schwerionenforschung]

Das GSI Helmholtzzentrum für Schwerionenforschung in Darmstadt betreibt eine der weltweit führenden Teilchenbeschleunigeranlagen für die Forschung und den Green IT-Cube, ein fortschrittliches Rechenzentrum mit einer Kapazität von bis zu 12 MW. Dieses verfügt über hohe CPU Compute-Kapazitäten von mehr als 300.000 Cores und 400 GPUs. Es fungiert als Testrechenzentrum und verfügt auf sechs Etagen über eine Fläche von 4.645 m² und bietet Platz für 768 19" Racks à 2,2 Meter Höhe (4 MW und 256 Racks in der ersten Ausbaustufe). Dank eines speziellen Kühlsystems ist der Green Cube besonders energie- und kosteneffizient. Die Power Usage Effectiveness (PUE) beträgt hier weniger als 1,07 und wird über eine passive Wasserkühlung in den Rücktüren der Racks sowie Wärmetauscher-Verdunstungskühltürme erreicht. Dadurch, dass keine Raumluftkühlung nötig ist, sind hier hohe räumliche Leistungsdichten möglich, die für HPC-Systeme mit vielen GPU-Knoten benötigt werden. In einem KI-Cluster sind üblicherweise bis zu 4 NVIDIA DGX H100 Systeme pro Rack verbaut. Die entstehende Abwärme von 36 kW/Rack kann nach Aussage der Expert:innen des GSI mit dem Kühlsystem des Green Cube bewältigt werden.

Der Green Cube verfügt über eine N+1 Redundanz und bietet die Möglichkeit der Bereitstellung von Rackspace und Dienstleistungen im Rechenzentrum (Collocation). Laut dem Betreiber könnten in den geplanten Ausbaustufen ausreichende Flächen für einen HPC-Supercomputer mit 4 MW Leistung und entsprechender Kühlung (Wasserkühlung im Rack und perspektivisch auch direct-to-Chip Kühlung) im Rahmen eines Collocation Modells bereitgestellt werden. Als zusätzliche Möglichkeit könnten hier im Rahmen des Testbetriebs des HPC-Clusters wertvolle Daten und wissenschaftliche Erkenntnisse zum Betrieb eines solchen großen, GPU basierten HPC-Systems gesammelt und veröffentlicht werden. Der Green Cube bezieht zu 100 % grünen Strom und wurde, neben anderen Preisen, 2020 von der Bundesregierung mit dem Umweltzeichen Blauer Engel ausgezeichnet. Die Abwärme der Racks wird zum Beheizen des Büro- und Kantinengebäudes auf dem Campus verwendet. Die technische Umsetzbarkeit eines Collocationbetriebs des LEAM KI-Rechenzentrums im Green Cube wurde vom Betreiber GSI bereits bestätigt.

INFOBOX

Der europäische Collocation Markt lag 2021 bei einem Umsatzvolumen von ca. 10,7 Mrd. US-Dollar und wird von einer großen Anzahl global operierender Anbieter beherrscht, die europaweit bzw. weltweit eine Vielzahl von Rechenzentrumsstandorten unterhalten. Für den Zeitraum von 2021 bis 2027 wird aktuell ein jährliches Marktwachstum (CAGR) von 14.8 % prognostiziert (Research and Markets Ltd, 2021). So hat sich z.B. in der Rhein-Main Region die Leistung der jährlich neu hinzugekommenen Collocation-Rechenzentren von 2020 mit 62 MW Leistung im Jahr 2021 auf 139 MW Leistung mehr als verdoppelt (Hintemann et al., 2021).

Zu den führenden Unternehmen in dieser Branche zählen u.a.:

- **Digital Realty / Interxion** unterhält an weltweit über 300 Standorten Collocation-Rechenzentren. In Europa betreibt Digital Realty in 15 Metropolregionen, verteilt auf 13 Länder insgesamt 114 Rechenzentren
- **Equinix** betreibt weltweit 240 Rechenzentren in 31 Ländern auf 6 Kontinenten. Insgesamt unterhält das Unternehmen 2,6 Mio. m² Fläche weltweit. Die durchschnittliche Verfügbarkeit der Rechenzentren liegt bei 99.9999 %.
- **NTT Global Data Center** unterhält 600.000 m² RZ-Fläche in über 20 Ländern mit einer IT-Leistung von 1.500 MW. In Europa ist das Unternehmen in 15 Metropolregionen mit Collocation-Dienstleistungen tätig.
- **Cyxtera** betreibt mehr als 60 Rechenzentren in über 30 Märkten. In Europa ist das Unternehmen an den Standorten London, Amsterdam und Frankfurt tätig.
- **Cyrus One** unterhält in Nordamerika und Europa 50 Collocation Standorte. In Europa ist das Unternehmen in Deutschland, UK, Irland, den Niederlanden sowie Spanien mit 14 Collocation Rechenzentren vertreten.
- **KDDI / Telehouse** betreibt über 45 Collocation Rechenzentren weltweit. In Europa ist das Unternehmen an den Standorten London, Paris und Frankfurt mit 9 Rechenzentren präsent.
- **Vantage Data Centers** betreibt auf 5 Kontinenten 24 Campus-Standorte. In Europa ist das Unternehmen an den Standorten Berlin, Frankfurt, Mailand, Warschau und Zürich und Cardiff mit 8 Lokationen vertreten.
- **Iron Mountain** unterhält auf drei Kontinenten an insgesamt 21 Standorten über 370.000 m² Brutto RZ-Fläche. In Europa ist das Unternehmen an den Standorten Amsterdam, Frankfurt, London und Madrid präsent.
- **Global Switch** betreibt 13 Rechenzentren auf 2 Kontinenten. In Europa werden an den Standorten London, Amsterdam, Frankfurt, Paris und Madrid insgesamt 9 Rechenzentren betrieben.
- **COLT Data Center Services** betreibt in Asien (Tokyo, Osaka und Mumbai) sowie in Europa (Frankfurt, London, Paris, Rotterdam) 14 Rechenzentrumsstandorte.
- **Penta Infra** betreibt Rechenzentren in den Niederlanden, Dänemark und Deutschland. In Deutschland ist das Unternehmen u.a. mit Rechenzentren in Berlin, Hamburg, Düsseldorf, Köln und Leipzig präsent.
- **North C Datacenters** betreibt Rechenzentren an insgesamt 14 Standorten in Deutschland, der Schweiz sowie in den Niederlanden. In Deutschland ist das Unternehmen in Nürnberg und München vertreten.

Zu den vorrangig im deutschsprachigen Raum tätigen Collocation-Betreibern zählen weiterhin (Auswahl):

- **noris network AG** mit sieben Rechenzentren an fünf Standorten in Nürnberg, München und Hof.
- **StackIT (Schwarz IT)** mit zwei Standorten in Ellhofen sowie Ostermiething (AT).
- **Data Center One** mit Standorten in Düsseldorf, Leverkusen und Stuttgart.
- **MyLoc Managed IT** mit 3.500 m² Fläche an sechs Standorten in Düsseldorf.
- **Plusserver** betreibt in Köln, Düsseldorf und Hamburg eigene Rechenzentren.
- **ScaleUp Technologies** betreibt insgesamt sieben Hochleistungsrechenzentren an den Standorten Hamburg, Berlin und Düsseldorf.
- **Maincubes One** betreibt neben einem niederländischen Standort in Amsterdam in Deutschland drei Standorte in Frankfurt sowie einen weiteren Standort in Berlin.
- **Akquinet** betreibt vier Rechenzentren an den Standorten Hamburg, Norderstedt und Itzehoe.
- **CollocationIX** betreibt am Standort Bremen ein Hochsicherheits-Collocation-Rechenzentrum.
- **Grass Merkur** betreibt am Standort Hannover 3.500 m² RZ-Fläche

Darüber hinaus bieten viele Internet Services Provider ebenfalls Collocation Services neben ihrem klassischen IPS-Portfolio an (u.a. M-net, Pfalzkom, Telemaxx, NetCologne, Docom, EnviaTel).

Die Kosten liegen laut der befragten Expert:innen für Collocation Angebote in der geplanten Größenordnung zwischen 100-120 EUR/KW/Monat. Hinzu kommt der Strom, der im Beispiel mit einem Preis von 20 Cent/kWh berechnet wird. Dies entspricht der von LEAM benötigten Größenordnung von ca. 4 MW geschätzten monatlichen Kosten von ca. 400.000 bis 500.000 EUR für den laufenden Betrieb.

Bau und Betrieb eines eigenen HPC-Rechenzentrums

Eine weitere Option ist der Bau und Betrieb eines eigenen HPC-Rechenzentrums mit entsprechender Gebäude-Infrastruktur. Die Kostenstruktur für den Bau und den Betrieb von Rechenzentren richtet sich vorrangig nach der erforderlichen Verfügbarkeitsklasse. Grundstückskosten stellen oftmals eine untergeordnete Rolle in der Total Cost of Ownership (TCO)-Betrachtung dar. Die folgende Beispielrechnung ist als grober Richtwert zu verstehen und kann aufgrund lokaler baulicher Gegebenheiten sowie besonderer technischer Erfordernisse abweichen:

500m² Whitespace	ca. 4 MW IT-Leistung	36 - 45 KW / Rack Leistung / Rack	115 Rackanzahl	VK 3 Verfügbarkeit	1,2 PUE
Gesamtleistung: 4 MW * 1,2 = 5 MW				35 ct / kWh Berechnungsbasis Energiekosten	
Invest / Bau: 15.000 EUR pro installiertes KW = 75.000.000 EUR					

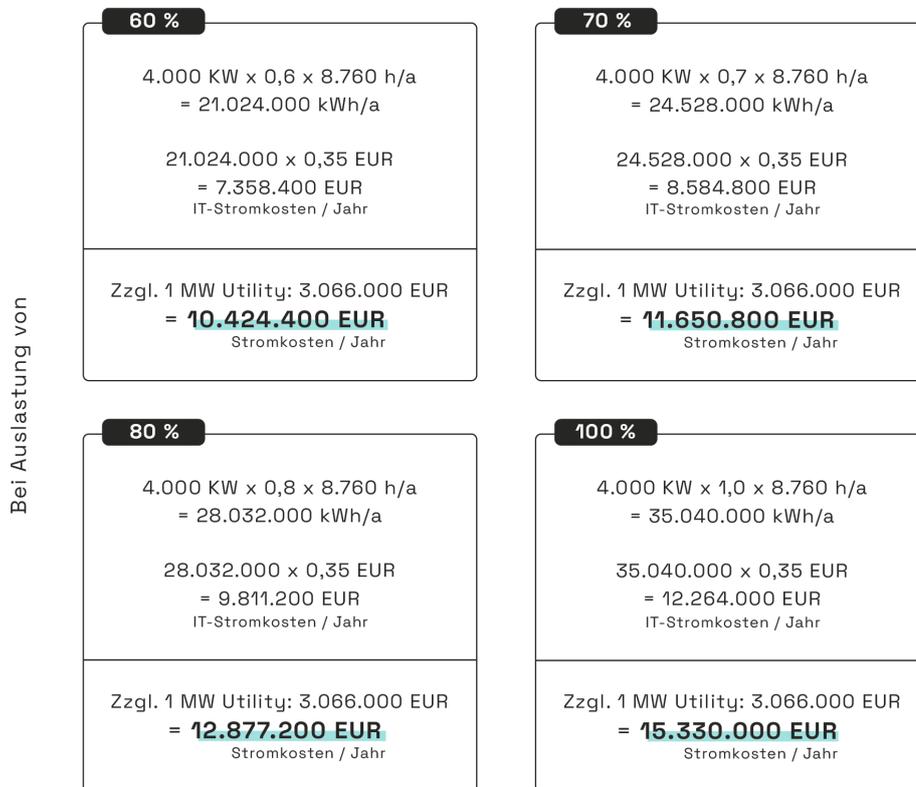


Abb. 23: Beispielrechnung Bau und Betrieb eines eigenen HPC-Rechenzentrums

Für die Inbetriebnahme eines Rechenzentrums mit eigener Gebäudeinfrastruktur rechnen die befragten Expert:innen mit zwei bis drei Jahren, je nachdem wie lange die baurechtlichen Genehmigungsprozesse dauern, die je nach zuständiger Kommune stark abweichen können. Die hier durchgeführte Betrachtung soll exemplarisch die verschiedenen Abhängigkeiten aufzeigen und eine Orientierung zur Planung eines Rechenzentrums geben.

Nutzung einer verteilten Infrastruktur

Fraglich ist, ob verteiltes Rechnen eines großen KI-Foundation-Modell mit mehreren zusammengeschalteten HPC-Rechenzentren möglich ist. Grundsätzlich ist verteiltes Rechnen bei der Erstellung eines großen KI-Foundation-Modells ein möglicher Ansatz, der aber noch weitgehend unerprobt ist. Außerdem stellt verteiltes Rechnen erhöhte Anforderungen an Infrastruktur, Netzwerk, Latenz (Ausfallsicherheit) und Sicherheit. Erste Studien zum Thema „Decentralized Training of Foundation-Models in Heterogeneous Environments“ kommen von der Stanford Universität (B. Yuan et al., 2022). Allerdings sind diese Systeme noch sehr neu und noch nicht ausreichend in der Praxis getestet. Somit würde ein solches verteiltes Rechenmodell für LEAM mit sehr hohem Risiko einhergehen, da beim Berechnen eines Foundation-Modells der Trainingsdurchlauf nicht unterbrochen werden sollte. Insbesondere scheitert das Rechnen auf verteilter Infrastruktur häufig noch an der nötigen Bandbreite bei der Vernetzung der einzelnen GPU-Knoten, die untereinander und mit dem Storage mit einer Bandbreite von bis zu 900 Gb/s verschaltet werden.

8.6.1 Strompreisentwicklung und Vertragsgestaltung

Die Stromkosten machen laut Béla Waldhauser CEO von Telehouse Deutschland GmbH sowie CEO von KDDI Deutschland GmbH mittlerweile in Deutschland 50 % der Kosten für Kunden der Rechenzentren aus (Weidmann & Krüger, 2020). Rechenzentren kaufen je nach Größe den notwendigen Strom entweder vom lokalen Energieversorger ein oder direkt an der Energiebörse. Hier sind die Preise aktuell volatil und betragen 0,20 EUR/KWh für den Normalbetrieb bis zu 0,50 EUR/KWh für den Einkauf bei Spitzenlasten. Im Vergleich dazu liegen die Strompreise für Anbieter in Norwegen, Finnland und Schweden bei ca. 0,05-0,10 EUR/KWh. Außerdem ergeben sich hier durch die niedrigeren nordischen Temperaturen weitere Vorteile im Bereich der Kühlung der Rechenzentren. Schwankungen des Strompreises können erheblichen Einfluss auf die Wirtschaftlichkeit des Geschäftsmodells haben. Dies wird auch durch die Ergebnisse einer Umfrage des Borderstep Instituts bestätigt, in dem die Befragten die Entwicklung des Strompreises als größtes Risiko für den Rechenzentrumsmarkt einstufen (Hintemann et al., 2022).

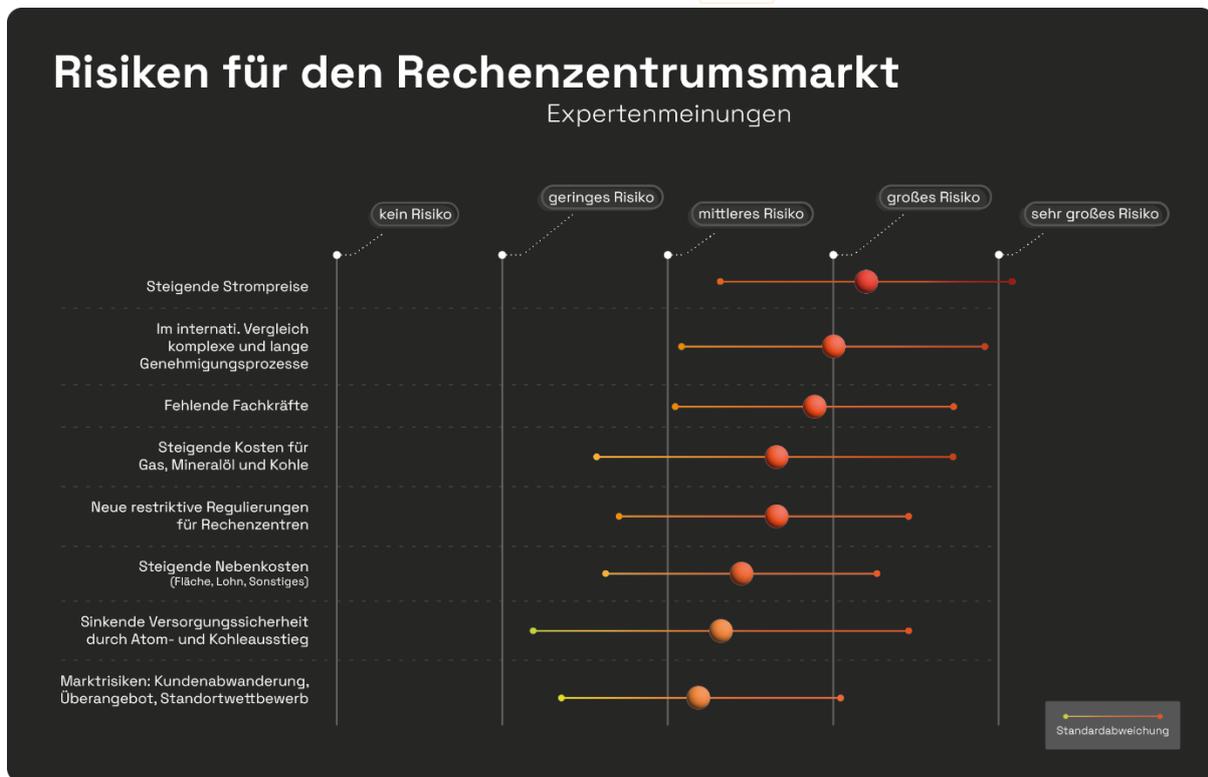


Abb. 24: Delphi-Befragung: Wie beurteilen Sie folgende Risiken für die Entwicklung des Rechenzentrumsmarktes in Deutschland? (Hintemann et al., 2022, S. 37)

Bei der Vertragsgestaltung zwischen Rechenzentren und ihren Kunden werden üblicherweise unterschiedliche Fristen für eine Strompreisbindung vereinbart. Dabei sichert der Rechenzentrumsbetreiber seinen Collocation-Kunden innerhalb der Laufzeit die Abgabe von Strom zu einem festen Kostensatz zu. Da der Rechenzentrumsbetreiber diesen Strom am Markt zu schwankenden Preisen einkaufen muss, stellen größere Schwankungen ein Risiko dar und gefährden unter Umständen die Wirtschaftlichkeit des Geschäftsmodells des Anbieters.

8.6.2 Verfügbarkeiten und Beschaffungszeitraum der erforderlichen IT-Ressourcen

Laut der befragten Expert:innen liegen die Lieferzeiten beispielsweise für NVIDIA DGX H 100 Systeme momentan bei unter sechs Monaten bei größeren Systemen. Der Hersteller NVIDIA sieht Lieferengpässen zurzeit eher bei den Netzwerkkomponenten, die ca. drei bis sechs Monate betragen.

8.6.3 Aspekte der Nachhaltigkeit

Diese Studie soll zur strategischen Ausrichtung der Planung nicht zuletzt auch die Nachhaltigkeit von Rechenzentren anhand des Beispiels einer wichtigen Initiative differenzierter thematisieren. Im Rahmen des *Climate Neutral Data Center Pact* setzen sich im Rahmen einer Selbstregulierungsinitiative zahlreiche Verbände und ca. 80 Unternehmen auf europäischer Ebene für die Schaffung von Standards im Bereich des nachhaltigen Betriebs von Rechenzentren ein (*Climate Neutral Data Centre Pact – The Green Deal Need Green Infrastructure*, o.D.). Im Rahmen einer Selbstverpflichtung haben sich die Mitglieder auf folgende Ziele geeinigt:

- **Energie Effizienz**
Bis 2025 sollen Rechenzentren mit 50 KW und mehr Energiebedarf in kühlen Klimazonen unter Volllast einen PUE-Wert von 1,3 erreichen. Rechenzentren haben in wärmeren Zonen einen Wert von 1,4. Außerdem sollen neue Effizienz-Metriken für Rechenzentren entwickelt werden.
- **Grüne Energie**
Die Rechenzentren werden ihren Strombedarf künftig durch den Einkauf von grünem Strom decken. 75 % CO₂-neutrale oder erneuerbare Energie bis Ende 2025; 100 % bis Ende 2030.
- **Wassernutzung**
Bis 2025 werden sich alle Rechenzentren, die in kühlen Klimazonen in Gebieten mit Wassermangel unter voller Auslastung betrieben werden und Trinkwasser für die Kühlung nutzen, einen maximalen WUE (Water Usage Effectiveness) Wert von 0,4 L/kWh erfüllen. Existierende Rechenzentren, die neue Kühlsysteme installieren, werden bis 2040 die angestrebten WUE-Werte erfüllen.
- **Kreislaufwirtschaft**
Die Wiederverwendung, Reparatur und das Recycling von Servern, elektrischen Geräten und anderen elektrischen Komponenten hat für die Betreiber von Rechenzentren Priorität. Hier werden hohe Standards an das Recycling gesetzt und angestrebt, zukünftig 100 % des eingesetzten Server-Equipments wiederzuverwerten.
- **Abwärmenutzung**
Die Rechenzentren planen den Ausbau der Einspeisung/Abgabe von Abwärme in die allgemeinen Energieversorgungsnetze und an andere Abnehmer voranzutreiben. Dies soll möglichst umweltfreundlich und kosteneffizient erfolgen.

Um die Entwicklung der Nachhaltigkeit von Rechenzentren weiter voranzutreiben und gleichzeitig Kosten zu senken, beobachtet und evaluiert die Rechenzentrumsbranche momentan außerdem die Themenfelder Refurbished IT, Remanufacturing und Re-Use sowie die Verwendung CO₂-armer Baustoffe oder die Wiederverwendung von Beton (Bitkom e.V., 2022b).

8.7 Zusammenfassung und Empfehlung

Zusammenfassend lässt sich feststellen, dass die zur Berechnung von großen Sprachmodellen benötigten HPC-Ressourcen in Deutschland/Europa derzeit kurzfristig nicht verfügbar sind. Die hohen Anforderungen an die GPU-Zahlen (ca. 4500) und die entsprechende schnelle Vernetzung der einzelnen GPUs untereinander können Stand heute nicht gewährleistet werden oder würden Berechnungszeiten erfordern, die die leistungsfähigsten deutschen HPC-Zentren für fast zwei Jahre komplett auslasten würden und außerdem die benötigten kurzen Innovationszyklen nicht gewährleisten könnten. Die einschlägigen Cloud-Dienste aus den USA und China können teilweise ausreichende KI-Rechenkapazitäten bereitstellen. Diese lassen sich aber nur schwerlich unter Wahrung der digitalen Souveränität und der europäischen Anforderungen an den Datenschutz nutzen. Zudem werden hier meist so genannte Lock-in Effekte wirksam, die einen späteren Wechsel zu einem anderen Anbieter erschweren.

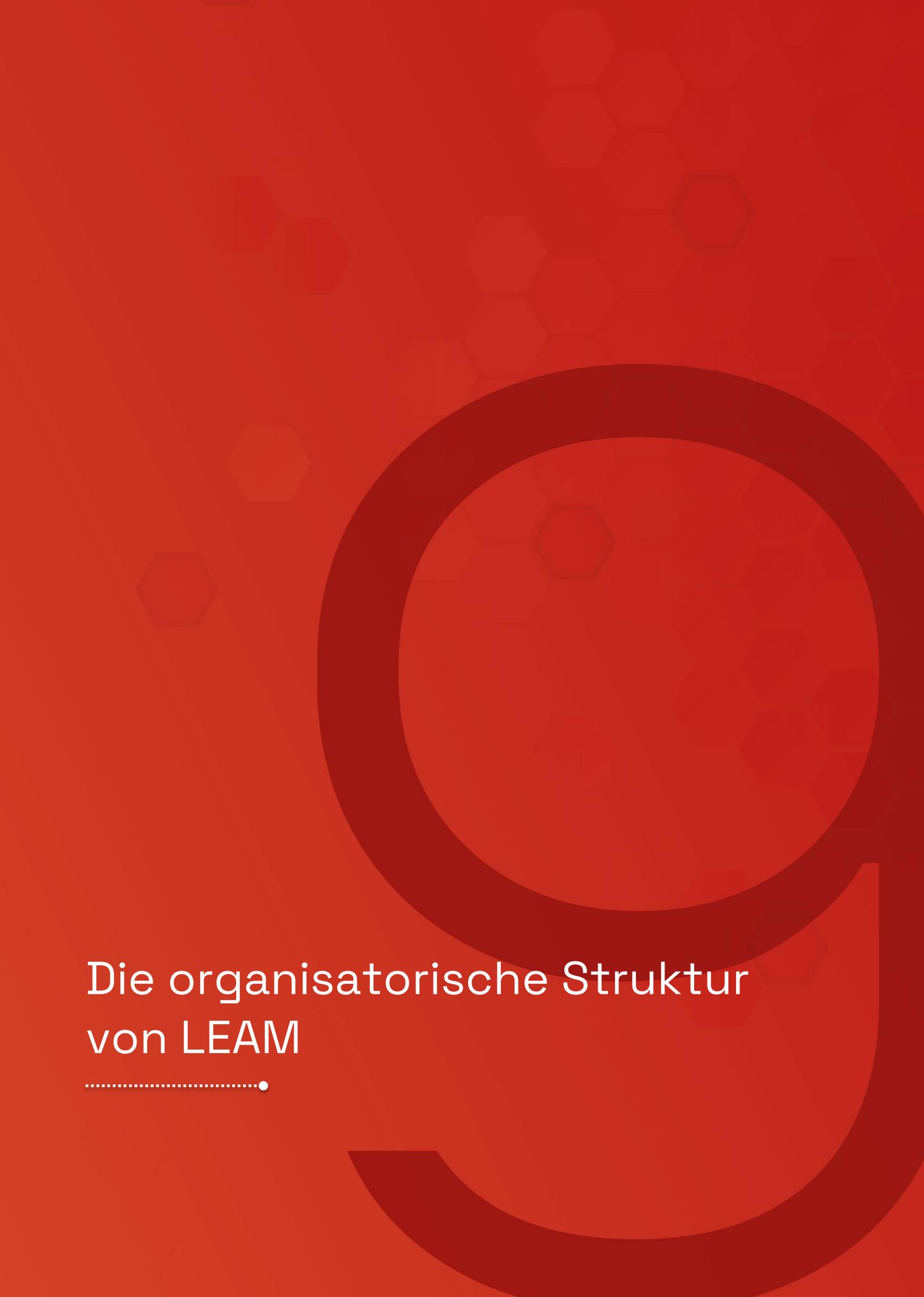
Aufgrund der vielen genannten Faktoren kann eine Standortempfehlung nur auf den konkreten Use Case bezogen gegeben werden. Diese Einzelfallbetrachtung kann im Rahmen der Studie nicht geleistet werden. Gespräche mit Vertreter:innen verschiedener Bundesländer sowie Regionalinitiativen haben aber gezeigt, dass es in verschiedenen Bundesländern eine generelle Bereitschaft für den Aufbau eines Rechenzentrums gibt.

Der Bau eines eigenen Rechenzentrums wäre generell deutlich teurer und würde einem schnellen Start der Entwicklungsaktivitäten entgegenwirken. Einschließlich der erforderlichen Planungs-, Genehmigungs- und Errichtungsphasen würden bis zu drei Jahre vergehen, bis ein entsprechendes KI-Rechenzentrum den operativen Betrieb aufnehmen könnte. Dies ist wegen des bereits beschriebenen Handlungsbedarfs zu lange. Von der Errichtung eines KI-Rechenzentrums mit eigener baulicher Infrastruktur für die sollte deshalb aus Zeit- und Kostengründen abgesehen werden.

Eine kurzfristig realisierbare Möglichkeit für den Betrieb eigener Rechenkapazitäten besteht im Rahmen eines Collocation Modells. Wie im Beispiel GSI Helmholtzzentrum für Schwerionenforschung aufgezeigt wurde, gibt es bereits Anbieter, die kurzfristig mit grünem Strom und nachhaltiger Abwärmenutzung betriebene Kapazitäten anbieten, die Anforderungen an die benötigte Wasserkühlung erfüllen und Skalierbarkeit ermöglichen. Die Angebote von Betreibern aus Norwegen, Finnland, Schweden und Island bieten hier zusätzlich über den Strompreis einen Betriebskostenvorteil wegen der niedrigeren durchschnittlichen Außentemperaturen. Juristisch und insbesondere datenschutzrechtlich orientieren sich diese Länder zudem an den europäischen Datenschutzstandards (Schweden und Finnland sind EU-Mitgliedstaaten, Island und Norwegen sind Teil des Europäische Wirtschaftsraums (EWR)) und bekommen deshalb eine Empfehlung von den Autor:innen. Für die Inference-Anwendungen können je nach Use Case jedoch Latenzen benötigt werden, die norwegische, isländische und schwedische Anbieter wegen der großen geografischen Entfernungen ggf. nicht erfüllen können.

Eine weitere Möglichkeit ist die Nutzung von neu zu schaffenden HPC-Kapazitäten, die von einem Anbieter, beispielsweise nach GPU-Stunden abgerechnet, bereitgestellt werden könnten. Auf diese Möglichkeit wird in Kapitel 10 im Rahmen der betriebswirtschaftlichen Betrachtungen näher eingegangen.

Nach Einschätzung der befragten Expert:innen evaluieren die einschlägigen, hier bereits genannten europäischen und internationalen Rechenzentren und Collocation Anbieter bereits größere Investitionen in HPC-Infrastrukturen in Europa, speziell in konzentrierte KI-geeignete Kapazitäten mit sehr großen Anzahlen von bis zu 20000 GPU. Hier werden parallel verschiedene Kooperationsmöglichkeiten evaluiert. Das Projekt Open GPT-X erforscht bereits heute die Entwicklung großer Sprachmodelle in Zusammenarbeit mit dem Jülich Supercomputing Centre (JSC). LEAM wird diese Ergebnisse im Rahmen der engen Zusammenarbeit mit dem Projekt in seine weitere Planung der Infrastruktur einfließen lassen.



Die organisatorische Struktur von LEAM

.....●

9. Die organisatorische Struktur von LEAM

Die vorangegangenen Kapitel haben gezeigt, dass der Aufbau eines KI-Hochleistungsrechenzentrums entscheidend für den Technologie- und Industriestandort Deutschland ist. Aufbau und Betrieb sind umsetzbar. Offen ist die Frage der wirtschaftlichen Machbarkeit eines solchen Zentrums. Die wirtschaftliche Machbarkeit wird in dem folgenden Kapitel näher betrachtet.

Auf Basis der Interviews mit Rechenzentrumsbetreibern und der Wirtschaft haben die Autor:innen die Idee eines LEAM KI-Servicezentrum (LKS) entwickelt. Welche Organisationsstruktur konkret hinter der Idee des LKS steckt wird im Folgenden genauer erklärt.

9.1 Zielgruppen des LEAM KI-Servicezentrums

Das in dieser Studie diskutierte LEAM-KI-Servicezentrum hat zum Ziel, an europäischen Werten orientierte KI-Foundation-Modelle zu entwickeln und insbesondere für die etablierte Industrie und junge Technologieunternehmen nutzbar zu machen. Dabei verfolgt das LKS zwei Zielsetzungen:

- (1) Kapazitäten des KI-Hochleistungsrechenzentrums bereitzustellen, die von Wirtschaft und Wissenschaft für das Training eigener Modelle zur Verfügung stehen.
- (2) KI-Foundation-Modelle Open Source anzubieten, die mit entsprechend Beratungs- und Serviceleistungen auf die individuellen Bedürfnisse von Industrieunternehmen angepasst werden können (Tuning).

Die Kapazitäten und Services sollen vier Zielgruppen angeboten werden. Die Zielgruppen sind:

- **Wirtschaft:** Unter die Zielgruppe Wirtschaft fallen privatwirtschaftliche Organisationen, die die Services der LKS für die Entwicklung von Modellen, Anwendungen oder Produkten in Anspruch nehmen.
- **KI-Start-ups:** Wie bei der Zielgruppe Wirtschaftliche Anwendungen handelt es sich hier um privatwirtschaftliche Organisationen. Sie unterscheiden sich aber von ersterer in zwei Punkten:
 - (1) KI-Startups arbeiten überwiegend an Künstlicher Intelligenz und
 - (2) es handelt sich um junge Unternehmen.

Start-ups sollten über spezielle Förderprojekte gezielt unterstützt werden. Ein Beispiel ist ein KI-Compute-Voucher, der Startups Zugang zu den Services des LKAS gewährleisten soll.

- **Public:** Öffentliche Institutionen, Behörden, Ministerien und Dienste können die Services des LKS nutzen, um die Verwaltung zu optimieren oder spezifische sicherheitsrelevante Insights auf Basis von großen Datenmengen zu erlangen. Für die Nutzung der LKS Services durch öffentliche Institutionen sind besondere Anforderungen hinsichtlich Datensicherheit, Datenschutz und allgemeine Richtlinien zur Verschwiegenheit zu erfüllen.
- **Wissenschaft:** Unter die Zielgruppe Forschung und Entwicklung fallen Hochschulen, außeruniversitäre Forschungsinstitute und staatliche Forschungseinrichtungen, die die Services der LKS für die Forschung in Anspruch nehmen und vor allem die Erstellung von Foundation-Modellen unterstützen.

9.2 Organisationseinheiten des LEAM KI-Servicezentrums

Das LEAM KI-Servicezentrum (LKS) ist in horizontalen und vertikalen Organisationseinheiten strukturiert (Abb. 25).

Die horizontalen Organisationseinheiten bilden die infrastrukturelle und kapazitive Grundlage des LEAM KI-Servicezentrums. Sie stellen den Betrieb der Infrastruktur und die Bereitstellung der Kernservices sicher. Hierzu gehören:

- Housing
- Infrastruktur-as-a-Service
- Training-as-a-Service

Die vertikalen Organisationseinheiten bauen auf den horizontalen Organisationseinheiten auf und gliedern sich in die vier verschiedenen Kernservices für Kunden und Nutzer:innen. Hierzu gehören:

- KI-Foundation-Model Development
- KI-Model Tuning
- Inference
- Consulting

Daneben ist die Organisationseinheit „Koordination“ als Managementeinheit verantwortlich für den Aufbau des LKS, hält engen Kontakt zum LEAM Board (s. Kapitel 9.3) und weiteren Stakeholdern aus Politik, Wissenschaft und Wirtschaft. Sie kann – je nach gewählter Gesellschaftsform – auch die Aufgaben eines gesellschaftsrechtlich erforderlichen Organs übernehmen, beispielsweise der Geschäftsführung oder des Vorstands.

Die einzelnen Organisationseinheiten können innerhalb einer Gesellschaft abgebildet werden oder entsprechend eines Governance-Konzepts in unterschiedliche juristische Einheiten aufgeteilt werden.

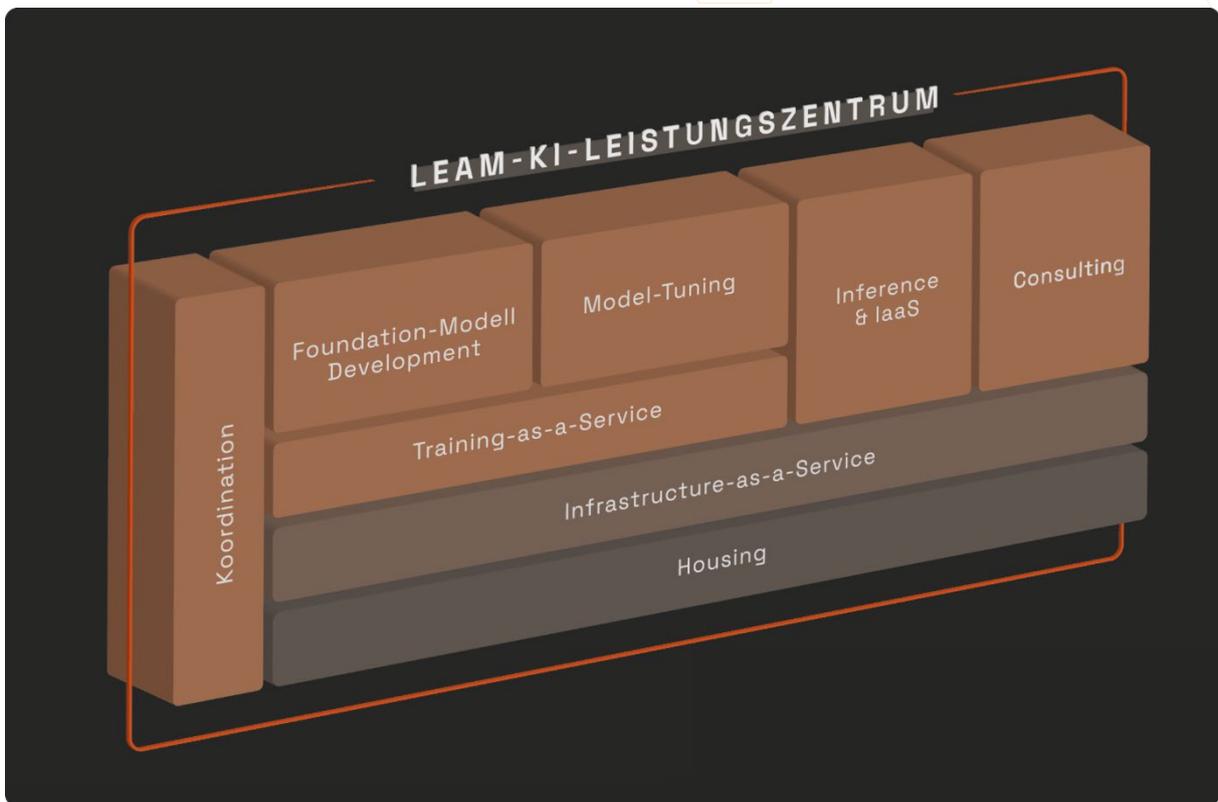


Abb. 25: Organisationseinheiten des LEAM-KI-Servicezentrums

9.2.1 Koordination

Die OE (Organisationseinheit) Koordination dient als Managementeinheit für das gesamte LKS. Sie stellt den kontinuierlichen Aufbau und Betrieb des LKS sicher und erweitert es innerhalb Deutschlands und der EU.

Die OE ist eine Anlaufstelle für interessierte Personen aus Wirtschaft und Wissenschaft und hält den Kontakt zur Politik. Sie koordiniert die Interessen der Stakeholder:innen, etabliert und steuert die horizontalen Organisationseinheiten, erstellt Marktstudien und akquiriert Förder- sowie Investitionsmittel. Als steuernde Einheit des LKS unterstützt sie die anderen Services und übernimmt die Vermarktung des LKS.

Dieser Bereich sollte unabhängig vom gewählten Organisationsszenario zusammen mit dem Bereich Consulting initial aufgebaut werden.

9.2.2 Housing

Die OE Housing stellt die Basisinfrastruktur für den Betrieb des Rechenzentrums zur Verfügung. Hierzu gehören bspw. die Gebäudeinfrastruktur, die Strom- und Telekommunikations-Versorgung sowie ein Kühlungskonzept. Das Kapitel 8 fasst die besonderen Anforderungen für die Ausstattung zusammen.

Da die Investitionen in die erforderliche Infrastruktur hoch sind und der Neuaufbau eines Rechenzentrums mehrere Jahre dauert, sollte zum jetzigen Zeitpunkt von einem Neubau abgesehen werden. Stattdessen bieten verschiedene Organisationen Housing Services an, die eingekauft oder angemietet werden können. Für den Aufbau des Rechenzentrums muss ein geeigneter Housing Partner gefunden werden.

Governance:	Außerhalb LEAM
Organisatorische Schnittstellen:	Koordination; Infrastruktur-as-a-Service
Kosten Collocation HW:	1.344.000 EUR per annum

Tabelle 13: Übersicht über die OE Housing

9.2.3 Infrastruktur-as-a-Service (IaaS)

Zentraler Bestandteil der OE IaaS ist der Aufbau und Betrieb eines KI-Supercomputers. Der Rechner benötigt zum Betrieb die Services der Housing-Infrastruktur. Die Anschaffung des Rechners ist einer der größten Posten des Gesamtbudgets und liegt im dreistelligen Millionen Euro Bereich (s. Kapitel 10). Nähere Infos zu den technischen Anforderungen und weiteren Aspekten des Rechners finden sich in Kapitel 8.

Für die organisatorische Einordnung der OE IaaS ergeben sich grundsätzlich zwei Szenarien: Einerseits ist die Anschaffung und Betrieb des Rechners durch das LKS denkbar, andererseits kann die Infrastruktur des LKS auch als Service durch ein externes Unternehmen bereitgestellt werden.

Aufgrund der hohen Relevanz des gewählten Szenarios für Investitions- und Betriebskosten des LKS sowie organisatorischen Gegebenheiten werden diese beiden Szenarien im Folgenden näher betrachtet.

9.2.4 IaaS innerhalb LEAM

Im ersten Szenario ist das LKS verantwortlich für die Anschaffung des Computers sowie den Betrieb und die Bereitstellung der Infrastruktur als Service. Dies birgt die folgenden Vor- und Nachteile.

VORTEILE EINER EIGENEN INFRASTRUKTUR

Umfangreiche Kontrolle und Gestaltungsmöglichkeiten

Bei der eigenen Anschaffung eines KI-Supercomputers können einzelne Komponenten, Bauteile sowie zugehörige Frameworks und Software selbstständig ausgewählt werden. Das macht eine flexiblere Auslegung und Ausrichtung auf die zukünftige Nutzung des Rechenzentrums möglich.

Unterstützung eines eigenständigen Geschäftsmodells

Die Investition in einen eigenen KI-Supercomputer macht es möglich, diesen nach der Abschreibungsdauer weiter zu nutzen. Die Abhängigkeit von eventuellen Kostenschwankungen und fremden Dienstleistern begrenzt sich dabei auf die zum Betrieb notwendigen Kosten.

Keine Datenweitergabe an Drittanbieter

Bei der Nutzung eines eigenen Rechenzentrums werden keine Daten über Nutzung etc. an Dritte weitergegeben. Das vereinfacht eventuelle Fragestellungen in Bezug auf Dritte (z.B. Sicherheitsdienste). Ebenso können auf diese Weise Anforderungen der Cybersicherheit (beispielsweise BSI-Anforderungen), deren Erfüllung für die Teilnahme an bestimmten öffentlichen Förderprogrammen oder öffentlichen Auftragsausschreibungen nachgewiesen werden müssen, bei der Nutzung eines eigenen Rechenzentrums besser nachgewiesen werden. Insbesondere bei einer staatlichen Finanzierung ist zudem Folgendes zu bedenken: Der Einkauf von Leistungen eines externen KI-Supercomputers bedarf einer öffentlichen Ausschreibung. Ein entsprechender Vertrag über den Einkauf kann zwar für einen längeren Zeitraum ausgestaltet sein, muss aber periodisch neu ausgeschrieben werden. Das kann die Investitionsbereitschaft eines externen Dienstleisters einschränken.

NACHTEILE EINER EIGENEN INFRASTRUKTUR

Hohe Investitionskosten

Die Anschaffung eines KI-Supercomputers ist mit einer sehr hohen initialen Investition verbunden. Als Abschreibungsdauer wird ein Zeitraum von vier Jahren angenommen.

Komplexe Finanzierung

Mit der Finanzierung und dem Aufbau des KI-Supercomputers ergeben sich komplexe Fragestellungen bezüglich der Finanzierung, der Besitzverhältnisse sowie den Nutzungsrechten.

Aufbau einer eigenen Betriebseinheit

Als Betreiber des Rechenzentrums ergeben sich hohe Kosten für Personal und Software. Die Verwaltung und Instandhaltung eines KI-Supercomputers ist komplex und muss durch entsprechendes Personal rund um die Uhr überwacht werden. Darüber hinaus muss man in der Lage sein, eine adäquate IT-Sicherheit herzustellen. Das bedeutet einen erheblichen administrativen Aufwand sowie hohe Kosten. Vor allem bei Förderprojekten oder Aufträgen aus der öffentlichen Verwaltung können zusätzliche Anforderungen und damit verbundene Aufwände beim Nachweis von Cybersicherheitsstandards entstehen.

9.2.5 IaaS über einen externen Partner

In diesem Szenario kauft das LKS die Rechenkapazitäten von einem externen Partner ein. Der Partner übernimmt dabei die komplette Installation und den Betrieb der Compute Infrastruktur und nutzt seine Erfahrung im Bereich des Hostings aus bestehenden Geschäftsmodellen.

VORTEILE EXTERNER INFRASTRUKTUR

Geringe Investitionskosten für eigenes Personal, Hardware und Software

Bei der Bereitstellung eines KI-Supercomputers über einen externen Partner ist dieser auch für die Verfügbarkeit und Funktionalität der Anlage verantwortlich. Daher entstehen vergleichsweise geringe Personal-, Hardware- und Softwarekosten, lediglich Unterstützungsfunktionen müssen übernommen werden.

Flexible Skalierbarkeit des KI-Supercomputers

Entsprechend der aktuellen Verfügbarkeit und des Bedarfs können die Kapazitäten, wenn vereinbart, flexibel dazu gebucht oder abgewählt werden. Das macht eine einfachere Anpassung auf den momentanen Bedarf möglich und fängt eventuelle Spitzen ab.

Zugriff auf fachliches Know-How und etablierte Prozesse

Expert:innen eines externen Partners können gegebenenfalls mittels fachlicher Expertise bei Fragestellungen unterstützen und die Nutzung der Services für Kunden so vereinfachen. Weiter kann bei der Abrechnung und dem Betrieb auf etablierte Prozesse und Erfahrungen - bspw. im Bereich IT-Sicherheitsstandards - des Drittanbieters zurückgegriffen werden.

NACHTEILE EXTERNER INFRASTRUKTUR

Weitergabe unternehmensbezogener Daten an Dritte

Bei der Ausführung von Services auf der Infrastruktur eines externen Anbieters werden evtl. Nutzungsdaten an diese weitergegeben. Dabei müssen Abwägungen über die Datensicherheit getroffen und Bestimmungen zur Weitergabe von Daten genau geprüft werden.

Abhängigkeit von externen Anbietern

Die Nutzung von Infrastruktur externer Anbieter steht und fällt mit der Verfügbarkeit von Kapazitäten dieses Anbieters. Wenn diese, sei es auch nur kurzfristig, nicht gegeben ist, muss mit Umsatzeinbußen und Compliance-Schwierigkeiten gerechnet werden. Um dies zu umgehen, müssen externe Anbieter die Verfügbarkeit von Kapazitäten garantieren können und mit dem Kunden vertraglich vereinbaren.

Rollierende, marktabhängige Kosten

Die Kosten zur Nutzung der Infrastruktur eines externen Anbieters fallen kontinuierlich an und sind höher als die Betriebskosten eines eigenen Rechenzentrums. Die Kosten können, je nach vertraglichen Regelungen, aufgrund von Marktbewegungen variieren.

Gefahr durch Übernahme oder Insolvenz

Die Infrastruktur des externen Anbieters steht unter dessen Kontrolle und könnte entweder in ein anderes Land verlagert werden, was sich möglicherweise nachteilig auf Datensicherheitsaspekte auswirkt, oder aufgekauft werden. Ebenfalls besteht die Möglichkeit, dass die Infrastruktur bspw. nach einer Insolvenz nicht mehr zur Verfügung steht

Empfehlung

Die befragten Expert:innen aus Industrie und Wissenschaft halten es für erstrebenswert, eine eigene, leistungsstarke und flexibel nutzbare Supercomputing-Infrastruktur aufzubauen, da die bisher verfügbare private und wissenschaftliche Infrastruktur diese Kriterien nicht erfüllt. Dies haben die Interviews im Rahmen dieser Machbarkeitsstudie, aber auch die Empfehlungen von wissenschaftlichen Initiativen, bspw. des European Language Equality Projekts, deutlich gemacht.¹⁹

Eine Entscheidungsgrundlage über die Finanzierung der Infrastruktur (eigene Hardware oder externe Infrastruktur) muss auch unter Berücksichtigung der hohen Investitionskosten weiter analysiert und vorbereitet werden.

9.2.6 Training-as-a-Service

Auf Basis der IaaS und damit bereitgestellten Compute-Kapazitäten werden Prozesse etabliert, um den Nutzer:innen und Kunden von LEAM umfangreiche Services für höher liegende Ebenen bereitzustellen. So ermöglicht das LKS potentiellen Nutzer:innen, die Infrastruktur zum Training ihrer Modelle zu nutzen.

Für das Training von KI-Foundation-Modellen sind umfangreiche Maßnahmen zum Aufbau und Betrieb der entsprechenden Prozesse und Software-Infrastrukturen erforderlich (s. Kapitel 7).

Die Aktivitäten können auf unterschiedlichen Ebenen (Layers) kategorisiert werden. Zu den Aktivitäten gehören u.a.:

- System Layer
 - Management der KI-Beschleuniger
 - Bereitstellung von (Open Source) Software und Softwarebibliotheken
- Data Layer
 - Sicherstellung von effizienten und stress-resilienten Datenspeicher- und Laderoutinen
 - Beachtung von gesetzlichen und regulatorischen Anforderungen an Datenschutz, Datenqualität und Datensicherheit
- Training and Application Layer
 - Aufbau eines effektiven Optimierungssystems zur optimalen Auslastung der GPUs während des Trainingsprozesses
 - Aufbau eines Systems zum Management der Trainingsjobs
 - Implementation von Evaluations- und Benchmarkingprozessen

¹⁹ "Current LT research also requires flexible access to High Performance Computing (HPC) facilities in the form of clusters of high capacity GPUs. There are many EU initiatives offering HPC: EuroHPC JU, PRACE, national computing facilities, etc. However, it is unclear if these initiatives are ready to provide the computing support that the European LT research community currently needs for developing state-of-the-art language models for all languages, domains, tasks and modalities."

(ELE Consortium 2022, S. 23).

- Framework & Service Layer
 - Normalisierung der Trainingsumgebung durch Bereitstellung von Container-Umgebungen
 - Implementierung von benutzerfreundlichen Interfaces für Entwickler:innen und Administrator:innen

Um diesen Service anbieten zu können, muss ein fachkundiges Team bestehend aus Data Science, Data Engineering, Machine Learning Engineering und DevOps Expert:innen aufgebaut werden. Dieses wird unter dem Team Services vereint und bildet Schnittstellen zum Team Consulting.

Governance:	Innerhalb LEAM
Organisatorische Schnittstellen:	Koordination; Housing, Training-as-a-Service, externe Kunden
Kosten Büroräumlichkeiten:	180.000 EUR per annum
Kosten Team Services:	3.000.000 EUR per annum

Tabelle 14: Übersicht über die Training-as-a-Service

9.2.7 KI-Foundation-Model Development

Der Service KI-Foundation-Model Development ist der Kernservice des LKS. Unter diesem Service werden alle Aktivitäten zusammengefasst, die direkt mit der Entwicklung neuer KI-Foundation-Modelle zusammenhängen.

LEAM wird den Service KI-Foundation-Model Development privatwirtschaftlichen, öffentlichen und wissenschaftlichen Einrichtungen anbieten. Es ist jedoch davon auszugehen, dass aufgrund der hohen Komplexität und Investitionskosten das grundlegende KI-Foundation-Modell Development überwiegend von Forschungs- und Kooperationsprojekten aus Forschung und Wirtschaft genutzt wird. Die Interviews mit der Forschung haben gezeigt, dass ein großes Interesse darin besteht, neue und konkurrenzfähige KI-Foundation-Modelle zu entwickeln. Darüber hinaus werden sie den Service nutzen, um wissenschaftliche Fragestellungen zu beantworten.

Sollte die LKS (teil-)öffentlich finanziert werden, müssen mindestens 80 % der Modelle Open Source verfügbar gemacht werden. Maximal 20 % können von Unternehmen mit Exklusivrechten genutzt werden. Prinzipiell eine Chance, denn unsere Befragung von KI-Unternehmen hat gezeigt, dass der Mangel von Open Source KI-Foundation-Modellen eines der Haupthindernisse für KI-Unternehmen ist, diese produktiv nutzen zu können. Zusammen mit der Verfügbarkeit von Daten und hohen Kosten wurde dies von 58 % der befragten Unternehmen als Hindernis angegeben.

Forschungsprojekte können sich dann über ein noch zu definierendes Verfahren auf Rechenzeit bewerben. Dabei soll auf die Erfahrungen der im Gauss Centre for Supercomputing organisierten Rechenzentren zurückgegriffen werden. Um allerdings den Besonderheiten von LEAM gerecht zu werden, sollten einige Punkte beachtet werden:

- ✓ Interessierten Projekten soll durchgehend die Möglichkeit gegeben werden, sich auf Projekte zu bewerben. Dies bietet Forschungsprojekten die notwendige Flexibilität, um in der schnelllebigen KI-Forschung zeitschonend zu forschen. Darüber hinaus ist dies eine Möglichkeit, Spitzen in der Nutzung der Infrastruktur abzuschwächen, da nicht alle Projekte zur gleichen Zeit mit der Berechnung ihrer Modelle starten.
- ✓ Die Bewerbungs- und Bewertungsverfahren sollen so einfach und flexibel wie möglich gestaltet werden. Ein häufiger Kritikpunkt an der aktuellen HPC-Landschaft ist, dass die Antragsphase zu lange dauert und wichtige Ressourcen bindet.
- ✓ Anwendungsbezogenen Forschungsprojekten soll ein Vorrang vor Grundlagenforschung gegeben werden.

Im Bereich KI-Foundation-Model Development werden folgende Services angeboten:

- **Erstellung und Bereitstellung allgemeiner Trainingsdatensätze:** Ein allgemeiner Trainingsdaten-Pool wird aufgebaut, entsprechend der Datenschutzvorgaben und Qualitätskriterien gepflegt und interessierten Organisationen zur Verfügung gestellt.
- **Bereitstellung von Basis-Algorithmen:** In einem Repository werden Code-Basen existierender (Open Source) Programme, erforderliche Hilfs-Tools und weitere Frameworks zur Verfügung gestellt.
- **Verwaltung und Bereitstellung von trainierten Foundation-Modellen:** Die trainierten Foundation-Modelle werden zur weiteren Nutzung in einem Repository abgelegt und verwaltet.

Um interessierten Organisationen den bestmöglichen Service anzubieten, müssen verschiedene Voraussetzungen erfüllt werden:

- **Einstellung von Mitarbeiter:innen:** Für die Begleitung und die Überwachung des Trainings werden Mitarbeiter:innen eingestellt und entsprechend ausgebildet.
- **Allokation von Compute-Ressourcen:** Damit Nutzer:innen den KI-Supercomputer zum Training von Foundation-Modellen nutzen können, koordinieren Mitarbeiter:innen von LEAM die Verwaltung und optimale Distribution der Compute-Ressourcen an Nutzer:innen.
- **Entwicklung eines Abrechnungsmodells:** Es wird ein Abrechnungsmodell für das Training der Modelle entwickelt. Hierbei erfolgt eine Orientierung an bestehenden Services im HPC-Bereich oder im kommerziellen Cloud-Services Umfeld.
- **Verwaltung des Trainingsdaten-Pools:** Trainingsdatensätze werden in Repositories gesammelt und Nutzer:innen zur Verfügung gestellt, um damit KI-Foundation-Modelle zu entwickeln. Mitarbeiter:innen des LKS unterstützen bei der Sammlung und Pflege von relevanten Datensätzen.
- **Support von Frameworks:** Für das Training von KI-Foundation-Modellen werden relevante Frameworks in Repositories gesammelt und den Nutzer:innen zur Verfügung gestellt. LEAM-Mitarbeiter:innen unterstützen die Nutzer:innen bei der Anwendung dieser im Zusammenhang mit dem KI-Supercomputer.

9.2.8 Model Tuning

Neben der Entwicklung von KI-Foundation-Modellen werden Ressourcen und Infrastruktur für das Tuning von Modellen bereitgestellt. Dies ist nötig, um die allgemeinen KI-Foundation-Modelle um domänenspezifisches Wissen zu ergänzen und so für konkrete Anwendungen zu nutzen und zu optimieren.

Der größte Teil der befragten Unternehmen ist an konkreten Anwendungen auf Basis von KI-Foundation-Modellen interessiert. 51 % der befragten KI-Unternehmen arbeiten bereits mit KI-Foundation-Modellen und 18 % planen die Nutzung von KI-Foundation-Modellen in der Zukunft (s. Kapitel 4).

Bereits vorhandene Modelle sollen durch Model-Tuning erweitert und für spezifische Zwecke nutzbar gemacht werden. Aktuell tun dies nur 27 % der befragten Unternehmen, die sich mit Foundation-Modellen auseinandersetzen. Somit ist es absehbar, dass die Nachfrage an Model-Tuning zeitnah steigen wird. LEAM kann dies der deutschen Industrie substantiell vereinfachen, indem es Expertise, Modelle, Daten und Infrastruktur bündelt und es Unternehmen erlaubt, Model-Tuning ohne großen Mehraufwand zu betreiben. Ein besonderer Fokus sollte hier darauf liegen, Start-ups aus dem Bereich KI, die ihre Modelle für Anwendungen in der Industrie weiterentwickeln möchten, Rechenkapazität zur Verfügung zu stellen.

Daneben sind auch wissenschaftliche Institute daran interessiert, den Model-Tuning-Service zu nutzen. Entsprechend ist auch für die Wissenschaft die Kombination an Expertise, Modellen und Daten, die das LKS bietet, von Interesse.

Im Bereich Tuning sollen folgende Services angeboten werden:

- **Tuning-as-a-Service:** Kunden können das Tuning von Modellen beim Rechenzentrum in Auftrag geben. In Kooperation mit den Kunden passen die Mitarbeiter:innen des Rechenzentrums die Foundation-Modelle an.
- **Beratung und fachliche Unterstützung** bei der Auswahl von Modellen, Daten und Algorithmen.

Um Kunden einen bestmöglichen Service anzubieten, müssen verschiedene Voraussetzungen erfüllt werden:

- **Einstellung von Mitarbeiter:innen:** Zur Beratung und Unterstützung der Nutzer:innen in der Bedienung der Infrastruktur und zur Auswahl des Modells sowie der Datensätze stellt die LKS Mitarbeiter:innen ein.
- **Allokation von Compute-Ressourcen:** Damit Nutzer:innen den KI-Supercomputer zum Tuning von Foundation-Modellen nutzen können, koordinieren Mitarbeiter:innen von LEAM die Verwaltung und optimale Distribution der Compute-Ressourcen an Nutzer:innen.
- **Entwicklung eines Abrechnungsmodells:** Es muss ein Abrechnungsmodell für das Tuning der Modelle entwickelt werden. Hierbei bietet sich beispielsweise ein Modell GPU/Stunde an.

- **Verwaltung des Trainingsdaten-Pools:** Trainingsdatensätze werden in Reopsitorys gesammelt und Nutzer:innen zur Verfügung gestellt, um damit KI-Foundation-Modelle zu entwickeln. Mitarbeiter:innen von LEAM sammeln relevante Datensätze und halten diese instand.

SPOTLIGHT

SAP SE

An verschiedenen Stellen in der Studie stellen wir ausgewählte deutsche Startups und Unternehmen vor, die KI-Foundation-Modelle bereits nutzen. Sie berichten über ihre Anwendungsfälle, Herausforderungen und die Vorteile europäischer Modelle.



Dr. Feiyu Xu, Vizepräsidentin
und Global Head of AI, SAP

Die **SAP SE** ist ein börsennotierter internationaler Softwarekonzern mit Sitz in Walldorf, Baden-Württemberg. Als ein Marktführer für Geschäftssoftware unterstützt SAP Unternehmen jeder Größe und Branche dabei, ihre Ziele bestmöglich zu erreichen: SAP-Kunden generieren 87 % des gesamten weltweiten Handels.

Wo setzt ihr Foundation-Modelle ein? Was ist euer Use-Case?

Foundation-Modelle werden eine sehr wichtige Rolle in der Enterprise AI spielen und zu disruptiven Innovationen im Prozessmanagement führen. Dazu gehören neue Anwendungen für business process mining, business

process prediction, business process composition und Prozessoptimierung.

Daneben werden die angepassten großen Sprachmodelle auch zur besseren Mensch-Maschine-Interaktion via Chatbot oder Digital-Assistenten beitragen und auch die semantische Verarbeitung der Business-Dokumente, insbesondere Informationsextraktion und Entity-Linking, erheblich verbessern.

Ein weiteres Anwendungsfeld ist die Generierung von Programmcode durch die Foundation-Modelle. Das steigert die Effizienz der Programmierer. Weil sich dann auch die Anwender neue Werkzeuge oder Erweiterungen von Softwareprogrammen von der KI erzeugen lassen können, führt das auch zu einer Demokratisierung der Softwareentwicklung.

Die potentiellen Use Cases kann man nach den jeweils benötigten Datenstrukturen ihrer I/Os klassifizieren:

1. Text2Text oder Speech2Speech:
 - Chatbots und Digitale Assistenten sind wichtige Anwendungen für SAP, durch die Benutzer:innen, z.B. Angestellte einer Kunden-Firma, natürlichsprachliche Fragen stellen können und Antworten über Fakten oder Transaktionen erhalten. Digitalassistenten lassen sich für Kundendienste einsetzen.
 - Übersetzung und Lokalisierung der Software, Dienste und Business-Dokumente
 - Zusammenfassungen von Textdokumenten oder Meeting-Transkripten

2. Text2Prozess und Prozess2Text
 - Die Integration der BPMN (Business Prozess Modelling Notation) in die Sprachmodelle ermöglicht die natürlichsprachliche Anfragen für die Generierung, Validierung und Ausführung der Business-Prozesse. In der anderen Richtung können zu bereits bestehenden Prozessmodellen textuelle Prozessbeschreibungen generiert werden, die z.B. für Schulungen oder Zertifizierungen benötigt werden.
3. Prozess2Prozess
 - Hierzu gehören Prozesskomposition, Prozessmodifikation, Prozessvalidierung und Prozessoptimierung.
4. Text2Code: low-code/no-code
 - Die automatische Generierung von Programmcode steigert die Effizienz und Produktivität der Programmierung

Welchen Einfluss haben KI-Foundation-Modelle auf euer Geschäftsmodell bzw. eure Projekte?

Foundation-Modelle führen einerseits zu großen inkrementellen Verbesserungen, da bessere Ergebnisse mit deutlich weniger (bis gar keinen) Trainingsdaten erzielt werden. Sie haben aber auch das Potential, neue und modularisierte End-to-End Geschäftsprozesse zu ermöglichen und so die Gesamtheit der Enterprise Resource Planning Landschaft zu revolutionieren.

Wir geben ein erstes Beispiel aus dem Bereich Businessdokumentverarbeitung. Die Anpassung bestehender Deep Learning Modelle erfordert momentan einen erheblichen Aufwand. Kunden müssen für ihre spezifischen Dokumentformate große Mengen an Trainingsdaten bereitstellen. Auch die Lokalisierung der Modelle in weiteren Sprachen ist ohne Trainingsdaten nicht machbar. KI-Foundation-Modelle haben das Potential, den Aufwand in beiden Bereichen erheblich zu reduzieren. Durch Foundation-Modelle können wir eine neue Art von Angebot an Kunden machen: Verarbeitung ihrer spezifischen Formate in den für sie relevanten Sprachen, mit wenig Trainingsdaten oder sogar out-of-the-box.

Welche Schwierigkeiten und Probleme siehst du, dass nur USA und China derzeit KI-Foundation-Modelle umfassend bereitstellen?

Europa benötigt eine eigene KI-Computing-Infrastruktur, um nicht in Abhängigkeit zu geraten. Zudem ist der Zugang zu Daten, inklusive mehrsprachiger Inhalte und explizit kodiertem Wissen, essenziell. KI in Europa kann nur langfristig erfolgreich sein, wenn es uns gelingt, KI-Talente auszubilden und bei uns zu halten. Eine wichtige Rahmenbedingung für Europa ist außerdem eine KI-freundliche Policy.

Wie würden euch europäische Modelle - Open Source, alle europäischen Sprachen abdeckend, mit hohen Datenschutzstandards und minimalem Bias - helfen?

Die Sprachabdeckung ist für uns essentiell, ebenso wie Transparenz zu verwendeten Daten. Nötig ist eine Infrastruktur, mit der KI-Modelle unter Gewährleistung der Transparenz und unter Einhaltung europäischer Datenschutzstandards generiert und Open Source bereitgestellt werden können. Bias ist in vielen Geschäftsbereichen ein kritisches Thema, etwa im Personalwesen und bei der Verarbeitung von Bewerbungen. Hier und auch in anderen KI-Anwendungen können KI-Modelle nur eingesetzt werden, wenn sie zu entsprechenden Regulierungen wie dem AI Act der Europäischen Union konform sind.

9.2.9 Inference

Das LEAM-KI-Servicezentrum soll auch dafür genutzt werden, die Modelle für potentielle Kunden bereitzustellen. Der Service wird zu marktüblichen Preisen angeboten, rund 5-10 % der Compute-Infrastruktur sollen für das Bereitstellen der Modelle genutzt werden.

Unternehmen sind vor allem am produktiven Einsatz von Anwendungen, die auf KI-Foundation-Modellen basieren, interessiert (s. Kapitel 4). Insbesondere Unternehmen, die keine eigenständige KI-Fachabteilung vorweisen und daher keine eigenen Modelle entwickeln bzw. anpassen können, werden auf diese Möglichkeit der Nutzung zurückgreifen. Da aktuell nur 50 % der befragten Unternehmen, die KI einsetzen, auch Foundation-Modelle nutzen, besteht hier eine relevante Zielgruppe für diesen Service. Für die Wissenschaft, die grundlegende Fragen zu KI-Foundation-Modellen beantworten möchte, ist dieser Service von geringerer Bedeutung.

KI-Unternehmen können auf Basis der Inference-Services eigene KI-Produkte und Anwendung entwickeln und anbieten und diese ihren Kunden z.B. über eine API und spezifische Abrechnungsmodelle zur Verfügung stellen.

Im Bereich Inference werden verschiedene Services angeboten:

- **Hosting-as-a-Service:** Kunden können das Bereitstellen von Anwendungen beim Rechenzentrum in Auftrag geben. Rund 5 bis 10 % der gesamten Infrastruktur werden für diesen Service reserviert. Mitarbeiter:innen des Rechenzentrums unterstützen und koordinieren die Vorhaben.

Um Kunden einen bestmöglichen Service anzubieten, müssen verschiedene Voraussetzungen erfüllt werden:

- **Einstellen von Mitarbeiter:innen:** Für die Einrichtung und den Betrieb von Inference-APIs stellt die LKS Mitarbeiter:innen ein.
- **Allokation von Compute-Ressourcen:** Damit Nutzer:innen den KI-Supercomputer zum Tuning von Foundation-Modellen nutzen können, koordinieren Mitarbeiter:innen von LEAM die Verwaltung und optimale Distribution der Compute-Ressourcen an die Nutzer:innen.
- **Entwicklung eines Abrechnungsmodells:** Es muss ein Abrechnungsmodell für die Bereitstellung der Modelle entwickelt werden. Hierbei bietet eine Berechnungsmethode des Entgelts basierend auf der Menge an genutzten Tokens (einzelne Anfragen an das Modell) an.
- **Schulung und Training von Nutzer:innen:** Um die Infrastruktur nutzen zu können, müssen potentielle Nutzer:innen geschult werden. Dafür muss ein Training vorbereitet und angeboten werden.

9.2.10 Consulting

Im Bereich Consulting erfolgt eine Beratung, die Kunden aus den verschiedenen Zielgruppen bei der Entwicklung, Optimierung sowie Implementierung von KI-Anwendungen unterstützt. Dieser Service erfolgt unabhängig von der Bereitstellung von Rechenzentrums-Infrastruktur-Leistungen.

Zielgruppe für Consulting Services sind alle wissenschaftlichen Institutionen, Start-ups und Unternehmen, die eigene Foundation-Modelle entwickeln oder existierende Foundation-Modelle für ihre spezifischen Anforderungen anpassen wollen.

Eine enge Zusammenarbeit mit Beratungs-Unternehmen aus der Wirtschaft und deren Befähigung, im Umfeld von KI-Foundation-Modellen Dienstleistungen anzubieten, gehört ebenfalls zu den Aktivitäten dieser OE.

Im Bereich Consulting werden folgende Services angeboten:

- **Beratung:** An KI-Anwendungen interessierte Kunden können eine Beratung in Anspruch nehmen. Diese wird die Organisationen bei der Entwicklung, Optimierung sowie Implementierung von KI-Anwendungen auf Basis von KI-Foundation-Modellen unterstützen.
- **Schulungen & Workshops:** Interessierte Unternehmen sowie Forschungseinrichtungen werden gezielt auf die Gegebenheiten der Entwicklung von Foundation-KI-Modellen mittels des LEAM-KI-Supercomputers vorbereitet.
- **Training von externen Beratungsunternehmen:** Um externe Beratungen mit den spezifischen Gegebenheiten und der Technologie vertraut zu machen, müssen diese zunächst geschult werden. Dies wird durch enge Zusammenarbeit zwischen externer Beratung und dem internen Personal des LKS oder bereits beratender externer Unternehmen erreicht.

Um Kunden einen bestmöglichen Service anzubieten, müssen verschiedene Voraussetzungen erfüllt werden:

- **Aufbau und Betrieb eines Beratungsteams:** Um diesen Service anbieten zu können, muss ein Team aus fachkundigen Berater:innen angeboten werden. Dieses sollte aus Fachexpert:innen für Data Science und Machine Learning bestehen. Alternativ können externe Beratung herangezogen werden.
- **Practice Work:** Das Consulting-Team des LKS entwickelt eigene Fähigkeiten im Rahmen der Entwicklung von Foundation-KI-Modellen stetig weiter, um im Rahmen der Beratung stets die aktuell wichtigsten Bereiche und Technologien abdecken zu können.

9.3 Das LEAM-Board

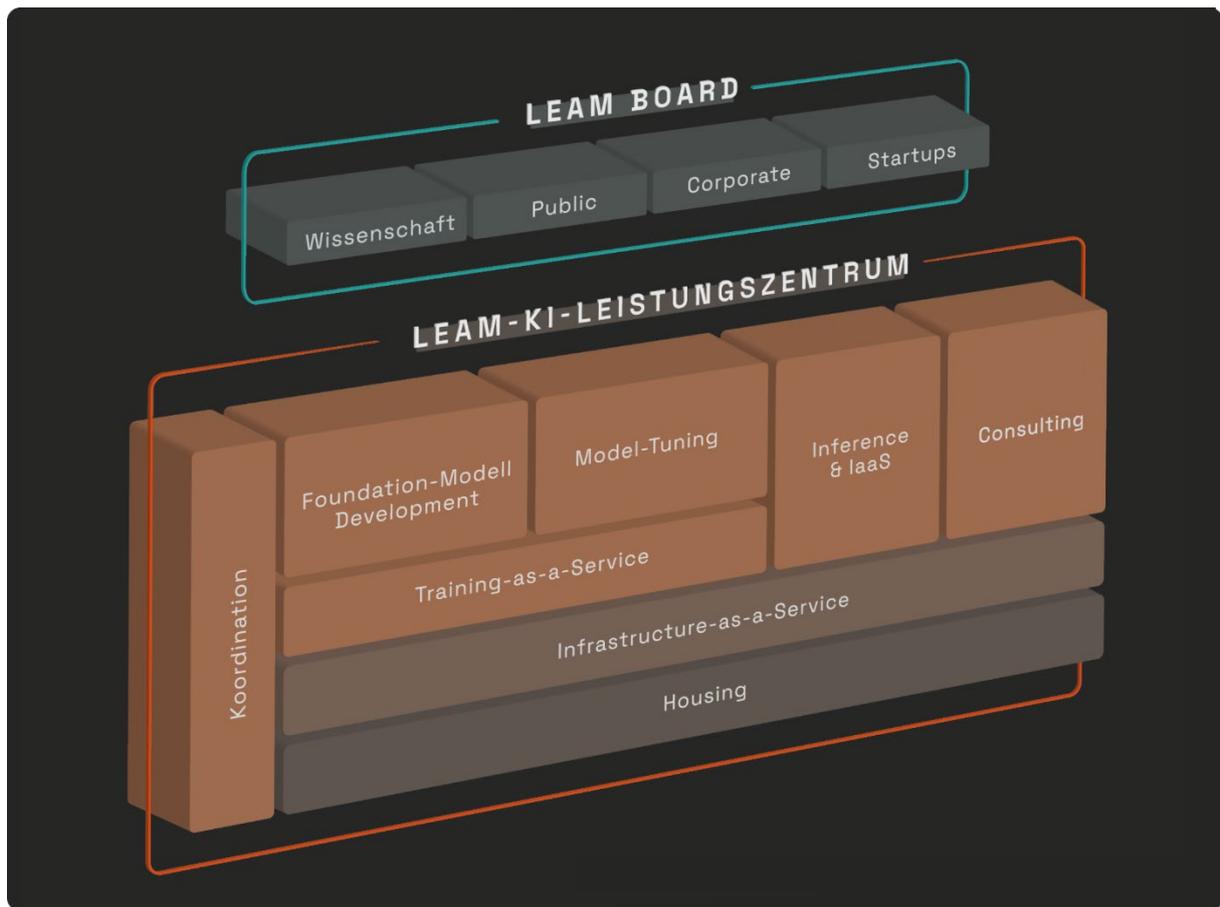


Abb. 26: Das LEAM-Board als zentrale Governance-Einheit des LKS

Das LEAM-Board überwacht und steuert die strategische Ausrichtung, Ziele sowie die Weiterentwicklung des LKS. Innerhalb des LEAM-Boards befinden sich Entscheidungsträger:innen und Expert:innen aus Wissenschaft, Politik, Unternehmen und Start-ups. Weiterführend entscheidet das LEAM-Board über Investitionen sowie die Vergabe von Compute-Ressourcen an Forschung und Entwicklung.

9.4 Zusammenfassung

Das LKS soll Kunden vier Kern-Services anbieten, die sich insbesondere an Unternehmen richten. Dabei kommt dem KI-Foundation-Modell Development die größte Bedeutung zu. Ziel muss es sein, Kunden optimal bei der Entwicklung und produktiven Nutzung von KI-Foundation-Modellen zu unterstützen. Dabei hilft der Service Consulting als Unterstützung bei der Entwicklung eines KI-Modells bis hin zum produktiven Einsatz mittels des Inference-Services. Interessierte Unternehmen sowie Forschung können jedoch auch einzelne Services wie Model Tuning, Inference oder Infrastruktur in Anspruch nehmen. Damit ist die Zielgruppe des KI-Rechenzentrums groß und der KI-Supercomputer wird flexibel genutzt. Ein Team aus LEAM-Mitarbeiter:innen steht dabei beratend und unterstützend zur Seite und begleitet Unternehmen auf ihrem Weg zur Entwicklung von KI-Foundation-Modellen.

Als Grundlage für diese Services dienen die Organisationseinheiten Housing, Infrastruktur-as-a-Service sowie teilweise Training-as-a-Service. Die Einheit Housing wird dabei größtenteils an externe Partner übergeben. Dies spart der LKS hohe Investitionskosten und reduziert die Zeit bis zum Start des LKS. Die Einheit Koordination unterstützt die anderen Einheiten und koordiniert den Betrieb des LKS.

Betriebswirtschaftliche Aspekte

.....●

10. Betriebswirtschaftliche Aspekte

Das LEAM-KI-Servicezentrum wird über die trainierten und bereitgestellten Foundation-Models einzigartige Wachstumsimpulse auslösen und branchenübergreifend enorme Effizienzgewinne in der Wirtschaft erreichen. Nach der anfänglich benötigten Investition, die entweder als öffentliche, private, gemischte Vollfinanzierung oder im Falle von externer IaaS mit Hilfe eines bestenfalls GAIA-X konformen Infrastrukturbetreibers erfolgt, wird die Recheninfrastruktur zu erheblichen Teilen der Wissenschaft und Unternehmen, die im Open-Source-Verfahren entwickeln, zur Entwicklung von Foundation-Modellen zur Verfügung gestellt. Die Entwickler:innen trainieren auf dieser Infrastruktur Foundation-Modelle der neuesten Generation und etablieren Wege zu standardisierten Fine-Tuning Prozessen. Die Modelle sowie die Verfahren und Prozesse zum Trainieren der Modelle werden entweder Open Source oder zu möglichst geringen Selbstkosten zur Verfügung gestellt.

Die etablierte Wirtschaft sowie Start-ups können diese Technologien schnell in ihre Produkte einbauen, da sie von verlässlichen Partnern stammen und nach europäischen Standards entwickelt wurden. Durch die Bereitstellung zum Selbstkostenpreis sind Unternehmen aus der Wirtschaft damit unabhängig von amerikanischen HyperScalern. Infolgedessen wird für Wirtschaftsunternehmen die Barriere reduziert, diese Modelle in ihre Produkte einzubauen oder neue Produkte auf dieser Basis anzubieten. Zudem profitiert die Wirtschaft von den entwickelten standardisierten Blueprint-Prozessen in den Bereichen Training, Tuning und Bereitstellung. Durch die Bündelung von Know-How innerhalb des LKS können Wirtschaft und Start-ups von erfahrenen Expert:innen entlang des kompletten KI-Foundation-Model-Lifecycles (Entwicklung, Training, Tuning, Bereitstellung, Anwendung) lernen und sich kompetent beraten lassen.

Damit leistet LEAM durch den Aufbau von Erfahrungswissen und den zu erwartenden Spillover-Effekten einen erheblichen Beitrag zur Gewinnung rarer Talente. Gleichzeitig erfolgt ein Wissenstransfer über die Schlüsseltechnologie „KI“ in die Wirtschaft, die vor dem Aufbau eigener Teams LEAM Services im Finetuning und Beratungskompetenz bei der Implementierung im Bereich Inference nutzen kann.

10.1 Kosten

Der Finanzierungsbedarf für das LEAM-KI-Servicezentrum setzt sich zusammen aus

- Investitionskosten (CapEx)
- Betriebskosten (OpEx)

Die jährliche Kostensituation ist u.a. abhängig von der Abschreibungsdauer der Investitionskosten, die im Folgenden mit vier Jahren angenommen wird.

Im Fall des Aufbaus der Infrastruktur über einen externen Partner und Einkauf von Rechenzentrumsleistung ist die Kostenkalkulation entsprechend anzupassen.

Die anfallenden Kosten sind abhängig von einer Vielzahl von Parametern, die in dieser Machbarkeitsstudie lediglich abgeschätzt werden können.

Die folgende Tabelle zeigt die für die Kostenkalkulation relevanten Parameter, die getroffenen Annahmen sowie die darüber hinaus gehenden möglichen Wertebereiche:

	Unit	Annahme	Kommentare - Range
Hardware			
Abschreibungszeitraum	Jahre	4	Generell 3-5 Jahre
Hardware (inkl. Speicher, Netw. Support, SW)	kEUR	260.000	Vorbild NVIDIA Selene
Anzahl DGXA100	Anzahl	560	Alternative HW: Cerebras, Graphcore, Intel, AMD, Nvidia H100
DGXA100 per Rack	Anzahl	2	gem. Info von NVIDIA
Betriebskosten Infrastruktur			
Miete Kolokation HW	[EUR/ Rack- Monate]	400	Durchschnittswert
Miete Büroraum	[EUR / qm]	30	15 - 40 EUR/m ²
Büroraum	qm	600	Durchschnittlich 10 qm / MA
Stromverbrauchs- Spitzenwert	kw	4.200	gem. Info von NVIDIA
Strompreis	EUR/kwh	0,35	aufgrund der aktuellen geopolitischen Lage ein hoher Unsicherheitsfaktor
Effektivität der Energienutzung	Faktor	1,50	relativ hoher Wert, kann in einem Green Data Center niedriger sein
Dauerlast von max Stromaufnahme	%	65	50 %-90 %
Team Operationen	FTE	20	Administration Hardware und SW- Infrastruktur
Kosten Organisation und Services			
Team Business Administration	FTE	20	Governance, Verwaltung, Vertrieb, Marketing
Team "Training-as-a- Service" & Consulting	FTE	20	Implementierung von Operationen, Prozessen, Dienstleistungen
Personalkosten	kEUR/FTE/ Jahr	150	Durchschnitt

Tabelle 15: Übersicht der Kosten des LEAM-KI-Servicezentrums

Anmerkung: Die in diesem Berechnungsmodell zugrunde liegende Architektur auf Basis von NVIDIA A100 wird durch die Hersteller in 2023 durch die neuere, leistungsfähigere Linie H100 ersetzt. Dadurch wird es auch zu einer Verbesserung des Performance- & Preisverhältnisses kommen.

Damit ergibt sich folgende Gesamtkostenstruktur auf der Basis eines Betriebes und einer Abschreibungsdauer von vier Jahren:

Invest	Jahr 1 kEUR	Jahr 2 kEUR	Jahr 3 kEUR	Jahr 4 kEUR	Summe kEUR
Infrastruktur	88.732	88.732	88.732	88.732	354.927
HW-Abschreibung	65.000	65.000	65.000	65.000	260.000
HW Collocation	1.344	1.344	1.344	1.344	5.376
Stromverbrauch	19.316	19.316	19.316	19.316	77.263
Team Operationen	3.000	3.000	3.000	3.000	12.000
Bürraum 1/3	72	72	72	72	288
Organisation und Services	6.144	6.144	6.144	6.144	24.576
Personal	6.000	6.000	6.000	6.000	24.000
Bürraum2/3	144	144	144	144	576
Total	94.876	94.876	94.876	94.876	379.503

Tabelle 16: Gesamtkostenstruktur des LEAM-KI-Servicezentrums bei einer Abschreibungsdauer von vier Jahren

Investitionskosten

Der zentrale Teil der Investitionskosten wird durch die Anschaffung eines KI-Supercomputers in Anspruch genommen.

Für die Berechnung dieser Infrastrukturkosten wird folgendes Szenario angenommen:

- Aufbau Betrieb eines KI-Supercomputers in der Größenordnung des NVIDIA Selene (Wikipedia Contributors, 2022).
- Die Dauer eines Trainingslaufs für ein Modell der Größenordnung GPT-3 beträgt auf Selene ca. 1-1,5 Wochen.
- Kernstück von Selene ist die NVIDIA Superpod-Architektur auf der Basis der DGX A100.
- Die Größenordnung einer NVIDIA Selene liegt bei 506 DGX A100 Nodes mit je 8 GPUs, in Summe.
- Die Anschaffungskosten liegen im Bereich von 260 Millionen Euro.

Auf dieser Basis wird hier lediglich eine Beispielrechnung durchgeführt. Zum Zeitpunkt der Anschaffung der Infrastruktur bzw. deren Ausschreibungen werden alternative Lösungen (z.B. Graphcore, Cerebras, AMD, Intel) bzw. die neueste Architektur von NVIDIA (H100) analysiert und evaluiert.

Betriebskosten

Die Betriebskosten des KI-Supercomputers setzen sich aus Collocation, Energiebedarf, Mietkosten und Personalkosten zusammen.

- **Miete Collocation HW:** Um den KI-Supercomputer zu betreiben, sind entsprechend ausgestattete Räumlichkeiten und Serverracks nötig. Diese müssen extern angemietet und mit der anzuschaffenden KI-Hardware ausgestattet werden. Pro Serverrack können zwei DGX A100 eingebaut werden. Bei Mietkosten in Höhe von 400 EUR pro Monat pro Rack ergeben sich jährliche Kosten in Höhe von 1.344.000 EUR.
- **Energiekosten:** Der Energiebedarf der Anlage beläuft sich auf circa 4.200 kWh. Er ist stark abhängig von der Auslastung der Anlage. Es ist von einer durchschnittlichen Dauerlast der Anlage von 65 % auszugehen. In der aktuellen Situation sind die Stromkosten höchst volatil. Auch wenn der Strompreis für Industriekunden aktuell noch niedriger liegt, wird aktuell mit 0,35 EUR/kWh geplant. Dadurch ergeben sich jährliche Kosten in Höhe von 19.316.000 EUR.
- **Team Infrastruktur:** Das Team Infrastruktur kümmert sich um den Betrieb der Hardware und SW-Infrastruktur. Dabei wird von einer Teamstärke von 20 FTE (Full-Time-Equivalents). Pro FTE sind pauschal 150.000 EUR Lohnkosten eingeplant.
- **Miete Büro:** Um Mitarbeiter:innen von LEAM einen Arbeitsplatz zu bieten, müssen Büroräumlichkeiten mit entsprechender Ausstattung angemietet werden. Es wird aktuell mit einem 600 m² großen Büro zu 30 EUR pro Quadratmeter geplant. Dies entspricht jährlichen Kosten in Höhe von 60.000 EUR .
- **Team Koordination:** Das Team Business Administration beschäftigt sich mit Governance, Sales, Marketing sowie administrativen Tätigkeiten rund um das Projekt LEAM. Dabei wird von einer Teamstärke von 20 FTE (Full-Time-Equivalents). Pro FTE sind pauschal 150.000 EUR Lohnkosten eingeplant.
- **Team Services und Consulting:** Dieses Team beschäftigt sich mit der Implementierung von Prozessen und Services rund um die Services, die LEAM anbietet. Dabei wird von einer Teamstärke von 20 FTE (Full-Time-Equivalents). Pro FTE sind pauschal 150.000 EUR Lohnkosten eingeplant.
- **Miete Büro:** Um Mitarbeiter:innen von LEAM einen Arbeitsplatz zu bieten, müssen Büroräumlichkeiten mit entsprechender Ausstattung angemietet werden. Es wird aktuell mit einem 600 m² großen Büro zu 30 EUR pro Quadratmeter geplant. Dies entspricht jährlichen Kosten in Höhe von 60.000 EUR.

Alternative Kostensituation bei Einkauf GPU-RZ-Leistungen

Beim Einkauf von Rechenleistung fallen die initialen Investitionskosten für LKS durch die fehlende Notwendigkeit der Beschaffung von KI-Hardware fast vollständig weg. Voraussetzung hierfür ist die Bereitschaft eines Unternehmens aus dem Bereich Cloud-Service-Providing, in die Infrastruktur zu investieren, wobei folgende Rahmenbedingungen berücksichtigt werden müssen:

- Es muss sichergestellt sein, dass ein ausreichend großes Compute-Cluster zur Verfügung gestellt wird (ca. 4500 GPUs).
- LKS wird über den Zeitraum von vier Jahren eine Mindest-Abnahmemenge von Rechenkapazität (z.B. 60 %) garantieren.
- Der Ankauf von Rechenleistung bzw. GPU-Stunden erfolgt nach marktüblichen Preisen (s.u.).
- Einzelheiten der Kalkulation und Vertragsgestaltung sind zu definieren.
- Die Auswahl eines Cloud-Service-Providers erfolgt evtl. im Rahmen einer öffentlichen Ausschreibung, deren Details zu definieren sind.

In diesem Szenario fallen folgende Kostenpositionen gemäß Tabelle 17 an:

	Annahme	Jährl Kosten in kEUR
Ankauf von Rechenzentrumskapazitäten / GPU Stunden	60 % der für den eigenen Betrieb errechneten Kosten	53.239
Organisation, Training-as-a-Service und Consulting	Entsprechend der Kalkulation mit eigenem RZ	6.144
Summe		59.383

Tabelle 17: Kosten des LEAM-KI-Servicezentrums bei einem Einkauf der GPU-RZ-Leistung

10.2 Einnahmen

Durch die angebotenen Services des LKS lassen sich verschiedene Einnahmequellen definieren:

- ✓ Verkauf von Rechenzentrums-Kapazität (GPU-Stunden)
- ✓ Services für das Training von Foundation-Modellen (Training-as-a-Service)
- ✓ Services für das Tuning von maßgeschneiderten Modellen (Training-as-a-Service)
- ✓ Inference Service (GPU-Stunden)
- ✓ Allgemeine Beratungs-Tätigkeiten (Consulting)

Die Möglichkeiten zur Generierung von Umsatz sind zielgruppenspezifisch zu differenzieren:

- **Wirtschaft:** Corporates und KMU werden Leistungen zu marktüblichen Preisen angeboten, wenn sie Modelle für den privatwirtschaftlichen Betrieb entwickeln. Bei Forschungsprojekten, die im Open-Source-Verfahren arbeiten, erhält die Wirtschaft Services kostenlos oder zu günstigen Preisen. Dies ist vor allem abhängig von der Governance und Finanzierung des LKS.
- **Wissenschaft:** Projekte, die im Open-Source-Verfahren arbeiten, erhalten Services kostenlos oder zu günstigen Preisen. Dies ist vor allem abhängig von der Governance und Finanzierung des LKS.
- **Public Sector:** Öffentliche Einrichtungen erhalten Services kostenlos oder zu günstigen Preisen. Dies ist vor allem abhängig von der Governance und Finanzierung des LKS.
- **Start-ups:** Junge Technologieunternehmen können für Leistungen von der öffentlichen Hand bereitgestellte Kontingente (z.B. KI-Compute-Voucher) zur Nutzung der LKS-Services beantragen.

Verkauf von Rechenzentrumskapazität

Basis für die Services KI-Foundation-Model Development, Tuning und Inference ist die Nutzung von Rechenzentrumskapazität nach GPU-Stunde.

Die Kosten für eine GPU-Stunde auf Basis des o.a. Kostenszenarios berechnet sich unter Annahme der Vollauslastung wie folgt:

Jährliche Kosten des Infrastrukturbetriebs gem. Tabelle 16:	88.732.000 EUR
Anzahl GPUs	4.480
Anzahl Stunden pro Jahr	8.765
Auslastung	100 %

Kosten pro GPU-Stunde **2,25 EUR**

Dies liegt im Bereich der derzeit marktüblichen Preise (s. Anhang C).

Unter der Annahme, dass 20 % der Rechenkapazität zu diesen Preisen am Markt verkauft werden können, ergibt sich ein möglicher Jahresumsatz (kostendeckend, ohne Marge) von:

Preis pro GPU-Stunde	2,25EUR
20 % von 4480 GPUs * 8765 h	7.848.960h

Jährlicher Umsatz ca.	14,464 Mio EUR
------------------------------	-----------------------

Diese grobe Kalkulation muss bei einer detaillierten Ausgestaltung des Szenarios angepasst werden, vor allem hinsichtlich der Parameter-Auslastung, Verfügbarkeit, Marge sowie angepassten Infrastrukturkosten.

KI-Foundation-Model-Training und -Tuning

Die Services KI-Foundation-Model-Development und -Tuning stellen eine Kombination aus Beratung und der Nutzung der Computer-Kapazitäten dar. Die beratenden Tätigkeiten sowie die tatsächliche Entwicklung und das Tuning werden dabei durch das Team Services und Consulting des LKS erbracht. Abgerechnet werden dabei projektspezifisch übliche Tagessätze zwischen 1.200 EUR und 2.500 EUR pro Berater:in/Entwickler:in.

	Annahme	
Verfügbare Kapazitäten	10 FTE bei 80 % Chargeability & 200 Tagen	1.600 Tage/Jahr
Durchschnittlicher Tagessatz	zwischen 1.200 & 2.500 Tagessatz	1.600EUR
Summe		2,56 Mio EUR/Jahr

Tabelle 18: Übersicht der Einnahmen durch das Model-Training

Beratung

Im Bereich Consulting sollen Beratungsleistungen nach marktüblichen Preisen abgerechnet werden. Diese werden anhand von Personentagen sowie der Seniorität und Expertise des Beratenden berechnet. Je nach der angefragten Leistung an das LKS wird dabei ein unterschiedlicher Umfang an Beratung nötig.

	Annahme	
Verfügbare Kapazitäten	10 FTE bei 80% Chargeability & 200 Tagen	1.600 Tage/Jahr
Durchschnittlicher Tagessatz	zwischen 1.200 & 2.500 Tagessatz	1.600EUR
Summe		2,56 Mio EUR/Jahr

Tabelle 19: Übersicht der Einnahmen durch die Beratung

Inference-Service

Der Inference-Service bietet Unternehmen die Möglichkeit, KI-Modelle mit den Compute-Ressourcen des LKS zu nutzen. Abgerechnet wird hier nach der Menge genutzter Compute-Ressourcen, ähnlich der Vermietung von GPU-Stunden. Eventuell fallen weitere Kosten an, insofern das Team Services oder Consulting Modelle ein Modell zunächst für die Nutzung mittels Compute-Ressourcen des LKS vorbereiten muss. Diese zusätzlich nötigen Aktivitäten werden analog zum KI-Foundation-Model-Development abgerechnet.

Finanzierungsmodelle von LEAM



11. Finanzierungsmodelle von LEAM

Zur Finanzierung des Aufbaus und Betriebs des LKS bieten sich drei verschiedene Modelle an: die öffentliche und private Finanzierung sowie die Finanzierung als Public-Private-Partnership. Abhängig von möglichen Förderungen und damit einhergehenden rechtlichen Auflagen an LEAM muss die genaue Finanzierungsstruktur weiter evaluiert werden. Im Folgenden werden die möglichen Finanzierungsmodelle zunächst überblicksartig dargestellt. In Kapitel 11.5 findet sich eine Betrachtung der spezifischen rechtlichen Überlegungen in Bezug auf die verschiedenen denkbaren Finanzierungsmodelle.

	Öffentliche Finanzierung	Private Finanzierung	Public-Private-Partnership
Quellen	<ul style="list-style-type: none"> • Bund • Länder 	<ul style="list-style-type: none"> • Corporates • Venture Capitals • Stiftungen 	<ul style="list-style-type: none"> • Kombination aus öffentlichen und privaten Quellen
Vehikel	<ul style="list-style-type: none"> • Staatliche Gesellschaft • Fördermittel (Förderung zur Gründung, Institutionelle Förderung, Projektförderung) • Garantierte Abnahme von Rechenleistung • (Kredite) 	<ul style="list-style-type: none"> • Eigenkapital • Fremdkapital • Garantierte Abnahme von Rechenleistung 	<ul style="list-style-type: none"> • Kombination aus den Vorgenannten
Bewertung	<ul style="list-style-type: none"> + maximale Unabhängigkeit + Konsens zwischen öffentl. Interesse und EU-konformer Entwicklung von KI-Modellen - Starre Strukturen und Prozesse - Einschränkungen in Agilität, Personalaufbau etc. 	<ul style="list-style-type: none"> + Flexibilität in der Unternehmensgestaltung + Einfaches und flexibles Anwerben von Personal - Komplexes Fundraising und eventuelle Governance - Erschwerte Kooperation mit Wissenschaft durch hohe Kosten 	<ul style="list-style-type: none"> + Public als Ankerinvestor + Flexibleres operatives Geschäft bei Einhaltung der Rahmenbedingungen für LEAM - Wettbewerbsrechtliche Beschränkungen - Kompliziertes Verfahren zur Gründung
Kommentar	<ul style="list-style-type: none"> • Berücksichtigung des EU-Beihilferechts mit Privilegierungen für Open-Source-KI-Projekte für Unternehmen und Wissenschaft 		<ul style="list-style-type: none"> • Beispiel: DFKI

Tabelle 20: Gegenüberstellung der drei Finanzierungsszenarien für das LKS

11.1 Öffentliche Finanzierung

In der Wissenschaftslandschaft Deutschlands existiert eine Reihe von öffentlich finanzierten Rechenzentren, die zum Teil zur Weltspitze gehören (s. Kapitel 8). In der Regel teilen sich hier Bund und das zugehörige Bundesland die Finanzierung der Investitions- und laufenden Kosten. Es wird ein Grundstock an Personal zur Erhaltung des Betriebs finanziert sowie Planstellen zur Forschung. Üblicherweise erhalten diese Rechenzentren erhebliche Anteile ihrer Finanzierung über Drittmittelprojekte, also kompetitiv bei Förderern (DFG, Bund, Länder, Industrie) eingeworbene Personal- und Sachkosten für spezifische Forschungsprojekte. Zu unterscheiden sind hier in der Regel die institutionelle Förderung und die Projektförderung: Die institutionelle Förderung wird wiederkehrend jährlich gewährt; sie mag in der Höhe abhängig von verschiedenen Faktoren schwanken, bildet aber eine sichere Grundlage für den wirtschaftlichen Betrieb der Einrichtung. Daneben tritt die bereits zuvor erwähnte Projektförderung, auf die sich die Antragsteller:innen selbst (oft im kompetitiven Wettbewerb mit anderen Einrichtungen) bewerben können.

Vorteile öffentlicher Finanzierung

Der Vorteil der öffentlichen Finanzierung für LEAM besteht in der engen Verzahnung zwischen öffentlichen Interessen eines an europäischen Werten orientierten Foundation-Models. Gleichzeitig bestehen nach erfolgter Finanzierungszusage eine hohe Verlässlichkeit und Planungssicherheit für die öffentlich geförderten Projekte. Damit könnte das Projekt über einen gewissen Zeitraum verlässlich wirtschaften und wäre unabhängig von der aktuellen Marktlage. Die entsprechenden Anreize für Start-ups könnten in dieser Finanzierungsform ebenfalls angeboten werden.

Nachteile öffentlicher Finanzierung

In der Regel unterliegen rein öffentlich finanzierte Rechenzentren aus EU-beihilferechtlichen, wettbewerbsrechtlichen oder steuerrechtlichen Gründen Beschränkungen in der Nutzung durch privatwirtschaftliche Akteure. So stellt bspw. das HLRS der Uni Stuttgart, das dezidiert Rechenzeit für die Industrie zur Verfügung anbietet, nur einen Bruchteil der möglichen Rechenzeit zur Verfügung (s. Kapitel 8.5). Eine Nutzung einer öffentlich finanzierten LKS durch die eher wirtschaftlich organisierten LEAM Services würde also Beschränkungen unterliegen und es könnte womöglich aus rechtlichen Gründen nicht genügend Rechenzeit zur Verfügung gestellt werden. Dies gilt auch für Start-ups, die ebenfalls um die knappen Rechenressourcen mit den anderen wirtschaftlichen Einheiten konkurrieren. Dies könnte den Aufbau von wirtschaftlichen Applikationen auf den Foundation-Modellen hemmen. Gleichzeitig unterliegen rein öffentlich finanzierte Projekte umfangreichen Genehmigungen und Auflagen und somit in der Regel langwierigen Abstimmungsprozessen sowie bei wechselnden politischen Mehrheiten auch sich ändernden politischen Gegebenheiten. Öffentliche Unternehmen sind in der Regel tarifgebunden oder lehnen sich an Tarifverträge an, was die Flexibilität bei der Gewinnung der für den Betrieb und die Services erforderlichen hochqualifizierten Mitarbeiter:innen erschweren kann.

11.2 Private Finanzierung

Für eine privatwirtschaftliche Finanzierung von LEAM kommen vor allem zwei Szenarien in Betracht:

- Joint-Venture von großen Unternehmen (Cloud-Service-Provider und Anwender). Hierfür wurde in den geführten Interviews von verschiedenen Unternehmen eine generelle Bereitschaft und Interesse signalisiert.
- Finanzierung über Risikokapital durch klassische Venture Capital (VC)-Gesellschaften oder Private-Equity-Investoren.

Da das Business-Modell einer rein privatwirtschaftlichen Gestaltung vom LKS mit hohen Risiken verbunden ist (u.a. auch durch die Dynamik der technologischen Entwicklungen und deren Auswirkungen auf den Markt), ist eine hundertprozentige Finanzierung durch private Unternehmen und Kapitalgeber nicht sehr wahrscheinlich.

Hier kann die öffentliche Hand durch Werkzeuge wie einer Anschubfinanzierung, Darlehen o.ä. helfen und das finanzielle Risiko abschwächen. Im Gegenzug kann vereinbart werden, dass z.B. ein gewisser Teil der Infrastruktur für nicht-wirtschaftliche Open-Source-Projekte von Unternehmen oder der Wissenschaft reserviert wird. Inwieweit diese Variante mit einer teilweisen oder gar überwiegenden wirtschaftlichen Nutzung auch in subventionsrechtlicher Hinsicht kompatibel ist, ist zu prüfen.

Die Wissenschaft und kooperierende Unternehmen, die bereit sind, im Open-Source-Verfahren zusammenzuarbeiten, würden in diesem Szenario im Rahmen von Auftragsforschung oder einer Projektförderung die Rechenzeit nutzen, um die Open-Source Modelle im LEAM KI-Servicezentrum zu trainieren. Im Falle der Auftragsforschung kämen als Auftraggeber für die Wissenschaft dabei sowohl LEAM selbst als auch die öffentliche Hand in Betracht.

Vorteile privater Finanzierung

Der Vorteil einer privaten Finanzierung liegt darin, dass Training und Verwertung ohne die regulatorischen Erfordernisse einer öffentlichen Finanzierung in einer Gesellschaft organisiert werden können. Die Investitionskosten könnten je nach Finanzierungsart auf viele Schultern verteilt werden, was das individuelle Risiko der Gesellschafter:innen minimieren würde.

Die Gesellschaft könnte zudem unbegrenzte Gewinne erwirtschaften und wäre unter Berücksichtigung der Interessen der Kapitalgeber frei bei der Wahl der Investitionen. Die Gesellschaft wäre durch frei wählbare Vergütungen auch flexibler darin, entsprechendes Personal anzuwerben und könnte so einen Vorteil beim Know-How Aufbau erlangen.

Nachteile privater Finanzierung

Die private Finanzierung birgt die Herausforderung, Geldgeber für einen erheblichen Investitionsbetrag zu finden. Die Wahrscheinlichkeit, dass ein einzelnes Unternehmen ins Risiko geht, die hohe Finanzierungslast auf sich zu nehmen, ist gering.

Es müssten Joint-Ventures aus mehreren Unternehmen gebildet werden, wodurch die Gestaltung der Gesellschaft und deren Governance komplex und langwierig werden kann.

Im Bereich der Venture Capital-Finanzierung übersteigen die erforderlichen Finanzierungssummen die zum Teil erheblich die Finanzierungspraxis der meisten deutschen und europäischen Fonds. So stellt der High-Tech-Gründerfonds (HTGF) nur Anschubfinanzierungen in einstelliger Millionenhöhe bereit.

Ein weiterer Nachteil privater Finanzierung liegt in der Kooperation mit den wissenschaftlichen Partnern, die in diesem Modell nicht direkt, sondern nur im Rahmen von Aufträgen und Projekten an LEAM beteiligt sind. Das Training der Foundation-Modelle, die selbst keinen Profit generieren, aber viel Rechenzeit benötigen, stünde in diesem Fall in enger Konkurrenz mit dem Kerngeschäft der Betreibergesellschaft Inference und Tuning. Rechenzeit würde somit prioritär für die kommerziellen Produkte von LEAM verwendet werden, wodurch das Training der Foundation-Modelle eher keinen experimentellen Charakter hätte. Zudem müssten Unternehmen und/oder die Wissenschaft, die bereit sind, im Open-Source-Modell zu arbeiten, Marktpreise für die Nutzung zahlen.

Der größte Nachteil ist jedoch, dass die LEAM Zielstellung, KI-Foundation-Modelle im öffentlichen Interesse bereitzustellen, mit dem Gewinnstreben der Organisation im Wettbewerb stehen würde. So würden Trainingsdaten für die Foundation-Modelle in diesem Szenario auch stärker unter der Perspektive einer ökonomischen Verwertbarkeit ausgewählt, weshalb seltene Sprachen auf Grund der schlechten Skalierbarkeit am Markt und des höheren Aufwands bei der Beschaffung von Trainingsdaten ähnlich schlecht repräsentiert würden, wie bei den bisher am Markt angebotenen Modellen.

11.3 Public-Private-Partnership

Bei einer Public-Private-Partnership erfolgt die Finanzierung durch die öffentliche Hand und private Unternehmen in jeweils zu definierenden Anteilen.

Ein Beispiel für eine erfolgreiche PPP im Bereich der Künstlichen Intelligenz ist das Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI). Das DFKI wurde 1988 als gemeinnützige Public-Private Partnership (PPP) gegründet. Es unterhält Standorte in Kaiserslautern, Saarbrücken, Bremen, Niedersachsen, Labore in Berlin und Darmstadt sowie Außenstellen in Lübeck und Trier. Die Finanzierung erfolgt über Zuwendungen öffentlicher Fördermittelgeber wie der Europäischen Union, dem Bundesministerium für Bildung und Forschung (BMBF), dem Bundesministerium für Wirtschaft und Klimaschutz (BMWK), den Bundesländern und der Deutschen Forschungsgemeinschaft (DFG) sowie durch Entwicklungsaufträge aus der Industrie.

Wichtig bei der Gestaltung der PPP ist, die jeweiligen Vorteile von öffentlicher/privater Finanzierung zu maximieren und deren Nachteile zu minimieren.

Die Herausforderung bei der Realisierung von LEAM als PPP steckt dabei im Charakter LEAMs als Infrastruktureinrichtung und den entsprechend hohen Investitionskosten, die für den Aufbau notwendig sind. Die Gesellschaft würde in diesem Szenario aus öffentlichen Mitteln (Bund und ggf. Land) die Infrastruktur beschaffen. Es müssten Mittel und Wege gefunden werden, die Mittel für die notwendige Infrastruktur zwischen den öffentlichen und privaten Partnern aufzuteilen. Grundsätzlich lassen sich in diesem Szenario auch die eher wirtschaftlich orientierten Services des Trainings-as-a-Service, Consulting und Inference/Tuning besser zusammen in einer Gesellschaft mit der Infrastruktur abbilden.

Vorteile einer Public-Private-Partnership

Eine PPP ist ein denkbare Modell, in dem die Interessen der Wirtschaft nach optimalen Verwertungsmöglichkeiten eines Open Source Foundation-Modells sowie die Interessen der Politik und Gesellschaft nach einem digital souveränen Europa miteinander in Einklang gebracht werden, da beide Seiten an einer Gesellschaft beteiligt wären. Darüber hinaus wird das Projekt finanziell leichter zu realisieren, da die Investitionskosten geteilt werden. Gleichzeitig wäre der öffentliche Haushalt entlastet, da die Privatwirtschaft ebenfalls für einen Teil der Investitionskosten für die Infrastruktur aufkommt. Die entsprechenden Anreize für Start-ups könnten in dieser Finanzierungsform ebenfalls angeboten werden.

Nachteile einer Public-Private-Partnership

Nachteil eines PPP besteht darin, dass die Gesellschaft nicht in derselben Weise privatwirtschaftlich agieren könnte, wie ein Wirtschaftsunternehmen, da der Bund bei seinen Beteiligungen wettbewerbsrechtlichen Beschränkungen unterworfen ist. Darüber hinaus ist sowohl mit Blick auf die genaue organisationstechnische Ausgestaltung der PPP als auch mit Blick auf die Spezifikationen der Foundation-Modelle mit einem intensiven Aushandlungsprozess zwischen den Vertretern der Privatwirtschaft und der öffentlichen Hand zu rechnen.

SPOTLIGHT

TUI Deutschland

An verschiedenen Stellen in der Studie stellen wir ausgewählte deutsche Startups und Unternehmen vor, die KI-Foundation-Modelle bereits nutzen. Sie berichten über ihre Anwendungsfälle, Herausforderungen und die Vorteile europäischer Modelle.



Henning von Roon,
Common Analytics Capabilities
Lead, TUI Deutschland

Die **TUI Group** ist einer der weltweit führenden Touristikkonzerne. Zum Konzern gehören über 400 Hotels und Resorts und 16 eigene Kreuzfahrtschiffe, außerdem europaweit führende Veranstaltermarken und Online-Vermarktungsplattformen, fünf Fluggesellschaften und über 1.000 Reisebüros. Neben dem Ausbau des Kerngeschäfts mit Hotels, Kreuzfahrten über erfolgreiche Joint Ventures und Aktivitäten in den Urlaubsdestinationen setzt die TUI verstärkt auf den Ausbau digitaler Plattformen.

Wo setzt ihr Foundation-Modelle ein? Was ist euer Use-Case?

Bei der TUI kommen vortrainierte Modell bei der Bilderkennung und Textklassifizierung zum Einsatz. Mittels dieser Modelle können wir beispielsweise die Produktpräsentation optimieren.

Welchen Einfluss haben KI-Foundation-Modelle auf euer Geschäftsmodell bzw. eure Projekte?

Durch den Einsatz von KI-Foundation-Modellen können viele Geschäftsprozesse leichter automatisiert, erweitert oder verbessert werden. Gerade bei der Personalisierung von Angeboten für unsere Millionen von Kunden pro Jahr können wir mit KI-Foundation-Modellen einen großen Mehrwert schaffen, ohne die Entwicklungskosten massiv zu steigern. Personalisierung auf dieser Größenordnung wäre ansonsten kaum umzusetzen.

Welche Schwierigkeiten und Probleme siehst du, dass nur USA und China derzeit KI-Foundation-Modelle umfassend bereitstellen?

Unsere hohen Ansprüche in Europa an Datenschutz und Datensicherheit werden oftmals durch außerhalb der EU bereitgestellte Modelle nicht erfüllt. Um keine Abstriche bei Datenschutz und Datensicherheit zu machen und gleichzeitig von den wirtschaftlichen Vorteilen des Einsatzes von KI-Foundation-Modellen zu profitieren, muss Europa deshalb eigene KI-Foundation-Modelle entwickeln. Zusätzlich ist die Übertragbarkeit häufig nicht gegeben, da die

Trainingsdaten nicht die europäischen Sprachen und Verhaltensweisen widerspiegeln.

Außerdem könnten Abhängigkeiten entstehen, welche sich in einer ungünstigen Preis- und Lizenzgestaltung abbilden.

Wie würden euch europäische Modelle - Open Source, alle europäischen Sprachen abdeckend, mit hohen Datenschutzstandards und minimalem Bias - helfen?

Open Source Modelle sorgen für Transparenz und erlauben eine bessere Interpretation der Ergebnisse. Zudem haben sie den Vorteil, dass wir sie entsprechend unserer spezifischen Bedürfnisse weiterentwickeln können. Europäische Modelle unterliegen von Anfang an den hiesigen Standards und bilden die lokalen Gegebenheiten deutlich besser ab, welches den Einsatz für uns attraktiver macht.

11.4 Rechtliche Rahmenbedingungen

Die Umsetzung der Finanzierung des LEAM KI-Supercomputers wirft zahlreiche Rechtsfragen auf, die im Rahmen dieser Machbarkeitsstudie weder abschließend noch vollständig behandelt werden können. Die tatsächlichen Gegebenheiten und Maßnahmen im Bezug auf rechtliche Rahmenbedingungen müssen daher im Nachgang tiefergehend evaluiert werden. Für die weitere Betrachtung sind jedoch grundsätzlich drei besondere Rechtsbereiche zu unterscheiden, die die geplante Struktur und den Betrieb des LKS mitprägen und bei der Strukturplanung mit in den Blick genommen werden müssen. Dabei handelt es sich hierbei um das EU-Beihilferecht, das (EU-)Vergaberecht sowie das öffentliche Dienst- und Vergütungsrecht. Nachfolgend sollen diese Rechtsmaterien zunächst überblicksmäßig dargestellt werden. Anschließend sollen die Rechtsmaterien und ihre Auswirkungen auf die drei grundsätzlich zu unterscheidenden Finanzierungsmodelle im Hinblick auf die Besonderheiten des LEAM-Projekts dargestellt werden, um hieraus in einer vergleichenden Rechtsbetrachtung die Vor- und Nachteile zusammenzufassen.

Die Bedeutung des EU-Beihilferechts

Eine staatliche Förderung für das Projekt LEAM sowie der Betrieb des LKS müssen mit dem EU-Beihilferecht vereinbar sein. Dieses ist bindend und steht dem deutschen Recht vor. Die Vorschriften über staatliche Beihilfen – Art. 107 bis 109 des Vertrages über die Arbeitsweise der Europäischen Union (AEUV) – zählen zu den wichtigsten europäischen Wettbewerbsregeln. Die beihilferechtlichen Regelungen verfolgen das Ziel, Wettbewerbsverzerrungen innerhalb des europäischen Binnenmarkts durch Beihilfen, die Mitgliedstaaten Unternehmen gewähren, zu verhindern. Daher normiert Art. 107 Abs. 1 AEUV ein grundsätzliches Beihilfeverbot, von dem jedoch Ausnahmemöglichkeiten bestehen.

Das EU-Beihilferecht wird – vereinfacht – wie folgt geprüft:

Auf der ersten Ebene wird geprüft, ob überhaupt begrifflich eine – nach dem soeben Gesagten grundsätzlich unzulässige – Beihilfe vorliegt. Häufig ist dies durch eine geschickte Ausgestaltung einer staatlichen Förderung bereits nicht der Fall – mit sehr günstigen Rechtsfolgen. Liegt begrifflich eine staatliche Beihilfe vor, so muss auf einer zweiten Ebene geprüft werden, ob diese aufgrund bestimmter Regeln der EU zulässig ist, ohne dass die EU der Beihilfe im Einzelnen ausdrücklich zustimmen muss. Eine Beihilfe kann insbesondere zulässig sein, wenn die Regeln der Allgemeinen

Gruppenfreistellungsverordnung (AGVO) greifen. Wenn solche generell geregelten Ausnahmen nicht einschlägig sind, muss auf der dritten Stufe eine Genehmigung der EU-Kommission für die Beihilfe eingeholt werden (sog. Notifizierung). Das ist oft langwierig und komplex und sollte für Projekte, die in besonderem Maße auf Schnelligkeit ausgerichtet sind, daher vermieden werden. Nicht mit dem Europarecht konforme Beihilfen darf der deutsche Staat nicht vergeben und ist verpflichtet, dennoch gewährte Gelder zurückzufordern. Daher muss das EU-Beihilferecht zwingend eingehalten werden.

Wichtig ist für das LEAM Projekt zunächst die Prüfung, ob überhaupt „begrifflich“ eine staatliche Beihilfe vorliegt. Eine solche ist gegeben, wenn die folgenden vier Merkmale kumulativ erfüllt sind. Eine Beihilfe ist

- eine staatliche Maßnahme,
- die eine Begünstigung,
- eines bestimmten Unternehmens darstellt und
- dadurch zum Eintritt einer (jedenfalls potenziellen) Wettbewerbsverfälschung und Beeinträchtigung des innergemeinschaftlichen Handels führt.

Eine staatliche Förderung der Infrastruktur des Projekts LEAM stellt, ohne dass hierauf im Einzelnen eingegangen werden soll, eine staatliche Maßnahme dar, die zu einer Begünstigung führen und auch Auswirkungen auf den europäischen Markt haben kann, da einer solchen großen Infrastruktur schon ihrem Sinn und Zweck nach europaweite Bedeutung zukommen kann.

Allerdings stellt sich – unabhängig von der Rechtsform – die Frage, ob das Projekt LEAM auch ein „Unternehmen“ im Sinne des EU-Beihilferechts darstellt. Der Begriff des Unternehmens im EU-Beihilferecht ist dabei nicht von der Rechtsform einer Einheit abhängig, er ist tätigkeitsbezogen. Maßgeblich ist immer die Frage, ob die Einheit eine wirtschaftliche Tätigkeit im Sinne des Beihilferechts ausübt. Nach ständiger EuGH-Rechtsprechung ist eine Tätigkeit wirtschaftlich, wenn sie darin besteht, Güter oder Dienstleistungen auf einem bestimmten Markt anzubieten,²⁰ unabhängig von ihrer Rechtsform, der Art ihrer Finanzierung sowie einer Gewinnerzielungsabsicht.²¹ Irrelevant ist daher, ob das Unternehmen in privater oder öffentlicher Trägerschaft betrieben wird. Ein Beispiel sind Forschungsinfrastrukturen: Wenn sie – vereinfacht ausgedrückt – für die Allgemeinheit forschen, sind sie kein Unternehmen im Sinne des EU-Beihilferechts. Wenn sie dagegen Auftragsforschung für andere Unternehmen als Auftraggeber erbringen, sind sie beihilferechtlich ein Unternehmen und es gelten die Beschränkungen des EU-Beihilferechts. Ein und dieselbe Einheit kann daher sowohl „Unternehmen“ wie auch „Nicht-Unternehmen“ im Sinne des EU-Beihilferechts sein, abhängig von der jeweils ausgeübten Tätigkeit.

²⁰ Vgl. EuGH, Urteil vom 16.06.1987, Rs. 118/85, Rn. 7 – Kommission/Italien; vom 12.09.2000, verbundene Rsen. C-180/98 bis C-184/98, Rn. 75 – Pavlov u.a. und vom 01.07.2006, Rs. C-49/07, Rn. 22 – MOTOE.

²¹ EuGH, Urteil vom 23.04.1991, C-41/90, Slg. 1991, I-1979 Rn. 21 – Höfer u. Elser/Macroton; von Wallenberg/Schütte, in: Grabitz/Hilf/Nettesheim, Das Recht der Europäischen Union, Werkstand: 67. EL Juni 2019, Art. 107 AEUV, Rn. 39.

Zusammengefasst bedeutet dies auch für das LKS:

- Soweit dieses nicht-wirtschaftliche Tätigkeiten im Sinne des EU-Beihilferechts ausübt, stellt die staatliche Förderung bzw. die Verwendung der staatlich geförderten Mittel keine Beihilfe dar.
- Wenn das LKS wirtschaftliche Tätigkeiten im Sinne des EU-Beihilferechts ausübt, ist es ein Unternehmen und die Beschränkungen des EU-Beihilferechts gelten für die staatliche Förderung bzw. die Verwendung der staatlich geförderten Mittel.

Für die Abgrenzung von nichtwirtschaftlichen und wirtschaftlichen Tätigkeiten, die, wie hier, einen Bezug zur Wissenschaft haben, ergeben sich Besonderheiten bei dieser Abgrenzung. Günstig ist, dass das Projekt LEAM einen großen Teil der Kapazitäten für die Entwicklung von KI-Foundation-Modellen zur Verfügung stellen will, die im Wege des Open-Source-Verfahrens der Allgemeinheit (bzw. diskriminierungsfrei der Wirtschaft) zur Verfügung gestellt werden sollen. Insbesondere ist bei der Auslegung und Anwendung des EU-Beihilferechts auf dem Gebiet der Entwicklung von KI-Modellen zur freien Nutzung der „Unionsrahmen für staatliche Beihilfen zur Förderung von Forschung, Entwicklung und Innovation“ vom 19.10.2022, 2022/C 7388 (Aktualisierung der Vorgängerversion vom 21.05.2014, 2014/C 198/01; nachfolgend „FuE-Rahmen“) maßgeblich. Dieser enthält die wichtigsten Unterscheidungen zwischen wirtschaftlichen und nichtwirtschaftlichen Tätigkeiten im Wissenschaftskontext sowie auch Privilegierungen für den Einsatz von Infrastruktur in der Wissenschaft, die dem LEAM-Projekt zugutekommen können.

Vereinfacht kann folgende Unterscheidung vorgenommen werden:

Soweit das LKS mit der Wissenschaft, im Open-Source-Verfahren arbeitenden Unternehmen oder Konsortien aus beiden Bereichen zusammenarbeitet und insbesondere KI-Foundation-Modelle entwickelt, die im Open-Source-Verfahren wiederum allgemein der Wissenschaft und Wirtschaft zur Verfügung gestellt werden, können solche allgemein die (europäische) Wissenschaft und Wirtschaft bereichernde Tätigkeiten und damit nicht-wirtschaftliche Tätigkeiten im Sinne des FuE-Rahmens sein (vgl. Rz. 20 FuE-Rahmen). KI hat dabei im neuen FuE-Rahmen 2022 eine ausdrückliche Erwähnung gefunden: Der Begriff der experimentellen Entwicklung wird in Rz. 16 lit. k) definiert als (verkürzt) den Erwerb, die Kombination, Gestaltung und Nutzung vorhandener wissenschaftlicher, technischer, wirtschaftlicher und sonstiger einschlägiger Kenntnisse und Fertigkeiten mit dem Ziel, in beliebigen Bereichen, Technologien, Branchen oder Wirtschaftszweigen neue oder verbesserte Produkte, Verfahren oder Dienstleistungen einschließlich digitaler Produkte, Verfahren oder Dienstleistungen zu entwickeln. Hierbei wird die Entwicklung Künstlicher Intelligenz ausdrücklich als ein Anwendungsbeispiel genannt. Damit dürften die KI-Foundation-Modelle (sowie sonstige KI-Modelle) unter den Begriff der Entwicklung im Rahmen des FuE-Rahmens zu subsumieren sein. Zudem wird die Entwicklung an Software mit dem Ziel, diese als Open-Source-Software zur Verfügung zu stellen, ausdrücklich als nicht-wirtschaftliche und damit nicht beihilferelevante Tätigkeit eingestuft (Ziff. 20 FuE-Rahmen).

Wie aus der deutschen Förderlandschaft bekannt, können sich Unternehmen auf solche Projekte bewerben, auch in Verbund mit der Wissenschaft, wenn sie bereit sind, die Software im Open-Source-Verfahren am Ende zur Verfügung zu stellen. Das LEAM-Projekt kann hierbei mehrere Wege wählen, um eine entsprechende beihilfekonforme Ausgestaltung zu erreichen; so kann sie beispielsweise die Software im Open-Source-

Verfahren zusammen mit Unternehmen oder der Wissenschaft entwickeln, wobei ihr wesentlicher Beitrag die Zurverfügungstellung der Infrastruktur und der technische Support sein kann (siehe beispielsweise Rz. 20, lit. a) ii) des FuE-Rahmens, "wirksame Zusammenarbeit"), oder eben allgemein die Entwicklung solcher KI-Foundations-Modelle im Wege einer Ausschreibung/diskriminierungsfreien Vergabe auch durch Unternehmen ermöglichen (siehe die Logiken in Rz. 20, lit. a) iii) und lit. b) des FuE-Rahmens, wonach die öffentliche Ausschreibung sogar von Dienstleistungen durch Unternehmen im Rahmen solcher Projekte den nichtwirtschaftlichen Charakter nicht berührt). Unternehmen können also an der Entwicklung von KI-Foundations-Modelle zusammen mit dem Projekt LEAM beteiligt werden, soweit sie im Open-Source-Verfahren arbeiten und sich damit bereit erklären, entwickelte Software offen und diskriminierungsfrei (jedenfalls im EU-Raum) zur Verfügung zu stellen. Wie dies letztendlich konkret ausgestaltet ist, hängt von der gewünschten Organisationsform ab.

Für das Projekt LEAM und seine Ziele kann dies eine erhebliche beihilferechtliche Privilegierung darstellen und eine öffentliche Finanzierung des Projekts begünstigen. Das bedeutet vereinfacht zusammengefasst: Eine staatliche Finanzierung der Infrastruktur des LEAM-Projekts ist jedenfalls dann beihilferechtlich möglich, wenn die Infrastruktur weit überwiegend der Wissenschaft und/oder kooperierenden, im Open-Source-Verfahren arbeitenden Unternehmen, in nicht-diskriminierender Weise zur Verfügung gestellt wird. Abzuklären wäre dabei im Rahmen der Umsetzung noch, ob eine zwingende Verpflichtung für die nutzende Wissenschaft bestehen muss, entwickelte KI-Modelle bzw.-Lösungen im Open Source-Verfahren wiederum der Allgemeinheit zur Verfügung zu stellen - dies wäre wohl jedenfalls beihilferechtlich der „sicherste“ Weg. Bei entwickelnden Unternehmen dürfte nach dem Beihilferahmen viel dafür sprechen, dass eine Open-Source-Nutzung verpflichtend ist, wobei Einzelheiten bei der konkreten Ausgestaltung festzulegen wären.

Selbst wenn beispielsweise generierte Modelle oder Wissen im Rahmen des Wissenstransfers in die Wirtschaft veräußert werden, kann dies eine nichtwirtschaftliche Tätigkeit sein, sofern der Gewinn aus der Veräußerung dem nichtwirtschaftlichen Betrieb zugute kommt (ebenfalls Rz. 20 FuE-Rahmen). Das bedeutet also, dass beispielsweise die Weitergabe von Lizenzen für entwickelte Software oder KI-Modelle durch das LEAM-Projekt gegen eine entsprechende Gebühr die öffentliche Förderung im Sinne des EU-Beihilferechts nicht gefährdet, wenn die Gewinne aus diesen Lizenzen wiederum dem nicht-wirtschaftlichen Bereich des LEAM-Projekts zugutekommen (also etwa mehr Rechenzeit oder bessere Infrastruktur für die Wissenschaft bzw. im Open-Source-Modell arbeitende Unternehmen zur Verfügung zu stellen). Insoweit ergibt sich dann kein beihilferechtliches Problem.

Andere geplante Tätigkeiten des Projekts LEAM, wie zum Beispiel die Consulting-Angebote oder die Vermietung der Infrastruktur, sind dagegen wirtschaftliche Tätigkeiten. Sie unterliegen den Beschränkungen des EU-Beihilferechts.

Sofern bei der öffentlichen Finanzierung des LEAM-Projekts danach doch das EU-Beihilferecht eingreift, kommt es zunächst auf der zweiten Ebene auf Ausnahmetatbestände an, die ausdrücklich geregelt sind und bei denen keine Genehmigung der EU-Kommission notwendig ist. Welche Ausnahmetatbestände dies sind, lässt sich nicht im Einzelnen vorab festlegen, sondern hängt von der konkreten

Ausgestaltung des Projekts und der gewünschten Finanzierung ab. Während die De-minimis-Verordnung keine Rolle spielen wird (die Fördergrenzen sind mit max. 200.000 Euro zu niedrig, um für das Projekt relevant zu werden), kann die sogenannte Allgemeine Gruppenfreistellungsverordnung eine Rolle spielen.

Die Allgemeinen Gruppenfreistellungsverordnung (Verordnung Nr. 652/2014 der EU-Kommission vom 17.06.2014, ABl. L 187/1 vom 26.06.2014; „AGVO“) enthält in größerem Umfang Tatbestände, bei denen eine begrifflich vorliegende Beihilfe an ein Unternehmen auch ohne Notifizierung (Genehmigung) durch die EU-Kommission beihilferechtskonform ist. Die Regeln sind dabei sehr dezidiert und müssen eingehalten werden, damit keine Notifizierung erforderlich ist. Sie sollen nachfolgend, soweit sie in Betracht kommen, bei den einzelnen Modellen vorgestellt werden.

Daneben können Maßnahmen zugunsten von Unternehmen, die mit sogenannten „Dienstleistungen von allgemeinem wirtschaftlichen Interesse“ befasst sind, auf Grundlage der Altmark-Trans-Rechtsprechung des EuGH bzw. des Freistellungsbeschlusses der EU-Kommission (Beschluss der Kommission (2012/21/EU) vom 20.12.2011, ABl. Nr. L 7/3 vom 11.01.2012) vom EU-Beihilfeverbot ausgenommen sein. Es ist jedoch unwahrscheinlich, dass das LEAM-Projekt (trotz seiner möglichen Bedeutung für die deutsche Wirtschaft) hierunter fallen kann, da diese Rechtsprechung sich eher auf soziale Bedürfnisse (Öffentlicher Nahverkehr, Gesundheitsversorgung o.ä.) bezieht.

Sofern dann auch Regeln wie die AGVO nicht greifen, um die Beihilfe zu rechtfertigen, kommt es auf eine Notifizierung (mithin ein Genehmigungsverfahren) für die Beihilfe bei der EU-Kommission an. Dieses ist rechtlich komplex und kann eine längere Zeit in Anspruch nehmen, wobei die EU-Kommission in der Regel auch Änderungswünsche an der Finanzierungsstruktur bzw. an sonstigen Aspekten der Verwirklichung der geplanten Struktur hat. Einzelheiten eines solchen Verfahrens können an dieser Stelle nicht dargestellt werden und würden den Umfang dieser Studie erheblich überschreiten. Zudem ist die Frage, nach welchen Regeln die EU-Kommission notifiziert, auch maßgeblich davon abhängig, wie die Finanzierungsstruktur des Projekts ausgestaltet wird und welche Zwecke es im Kern verfolgen soll. Sofern die notifizierungspflichtige Beihilfe nach der Zweckrichtung des LEAM-Projekts unter den FuE-Rahmen fallen kann, prüft die EU-Kommission, ob die staatliche Beihilfe zur Förderung von Forschung und Entwicklung als mit dem Binnenmarkt vereinbar angesehen werden kann. Dazu untersucht die Kommission, ob die Beihilfemaßnahme die Entwicklung eines bestimmten Wirtschaftszweigs fördert und ob sie die Handelsbedingungen in einer Weise verändert, die dem gemeinsamen Interesse zuwiderläuft (Rz. 38 FuE-Rahmen). Entscheidend ist hierbei, ob eine Beihilfe eine wesentliche Verbesserung bewirkt, die der Markt selbst nicht herbeiführen kann, insbesondere wenn es im Übrigen ein Marktversagen gibt. Dabei verlangt die Kommission andererseits wiederum, dass die Höhe und die Intensität der Beihilfe auf ein Minimum begrenzt sein muss, was in der Praxis regelmäßig umfangreich geprüft wird. Dies sind nur einige – überblicksmäßige – Zusammenfassungen des Prüfungsprogramms der EU-Kommission im Rahmen eines Notifizierungsverfahrens.

Die Bedeutung des Vergaberechts

Wenn die öffentliche Hand sich Güter und Dienstleistungen beschafft, ist sie in ihrer Auswahl, anders als private Unternehmen, nicht frei, sondern unterliegt den Beschränkungen des Vergaberechts. Dieses dient der Wirtschaftlichkeit und Transparenz staatlichen Handelns sowie einem fairen Wettbewerb um öffentliche Aufträge. Hierzu gibt das Vergaberecht vor, welches Verfahren bei der Auftragsvergabe einzuhalten ist und wie die Auswahl zwischen möglichen Vertragspartnern zu erfolgen hat.

Verpflichtet, öffentliche Aufträge auszuschreiben, sind insbesondere öffentliche Auftraggeber im Sinne von §§ 98, 99 des Gesetzes gegen Wettbewerbsbeschränkungen (GWB). Hierunter fallen nach § 99 Nr. 2 lit. a) GWB unter anderem juristische Personen des öffentlichen und des privaten Rechts, die zu dem besonderen Zweck gegründet wurden, im Allgemeininteresse liegende Aufgaben nichtgewerblicher Art zu erfüllen, sofern sie überwiegend von Gebietskörperschaften, das heißt, etwa von Bund und Ländern, durch Beteiligung oder in sonstiger Weise finanziert werden. Je nach konkreter Ausgestaltung der Finanzierung kann dies auf das LEAM-Projekt zutreffen. Diese Frage wird unten jeweils bezogen auf die einzelnen Finanzierungsmodelle erörtert.

Abgesehen von Bagatellgeschäften mit Wert von (je nach Bundesland bzw. Bund) 500 EUR bzw. 1000 EUR (netto) müssen Aufträge öffentlicher Auftraggeber ausgeschrieben werden, und zwar je nach geschätztem Auftragsvolumen entweder bundes- oder – bei Überschreitung bestimmter Schwellenwerte – europaweit. Derzeit liegt der Schwellenwert für Liefer- und Dienstleistungen für öffentliche Auftraggeber bei 215.000 EUR (netto). Es hat eine Veröffentlichung der Ausschreibung zu erfolgen und Bieter können ein Angebot abgeben. Nach der Angebotsauswahl, die bestimmten Vorgaben unterliegt, wird der Auftrag einem Bieter zugeschlagen.

Kommt das Vergaberecht zur Anwendung, sind seine Vorgaben unbedingt einzuhalten, da Verstöße hiergegen erhebliche finanzielle Konsequenzen verursachen und den zeitlichen Ablauf von Projekten stark verzögern können. Durch die der staatlichen Kontrolle unterliegenden vergaberechtlichen Vorgaben ist die Güter- und Leistungsbeschaffung durch öffentliche Auftraggeber zeitlich und inhaltlich deutlich weniger frei als für rein private Unternehmen.

Die Bedeutung des öffentlichen Dienst- und Vergütungsrecht

Bei der Entscheidung für ein Finanzierungsmodell sind auch Belange des öffentlichen Dienst- und Vergütungsrechts zu beachten. Im öffentlichen Dienst existieren spezielle Anforderungen an die Personalgewinnung sowie vergütungsrechtliche Bestimmungen. Aber auch für private Unternehmen, die staatliche Förderungen in Anspruch nehmen, können in diesem Bereich Einschränkungen gelten.

Für die Personalgewinnung im öffentlichen Dienst ist Art. 33 Abs. 2 Grundgesetz (GG) maßgeblich, der einen gleichen Zugang zu öffentlichen Ämtern garantiert. Daher hat die Stellenbesetzung im öffentlichen Dienst nach dem Prinzip der Bestenauslese zu erfolgen, was zu besonderen Anforderungen an die Ausgestaltung von Bewerbungsverfahren und die Personalauswahl führt. Auch bei der Personalgewinnung sind öffentliche Arbeitgeber daher weniger frei als private Unternehmen.

Weiter gilt für beim Bund Beschäftigte der Tarifvertrag für den öffentlichen Dienst (TVöD), für Beschäftigte der Länder außerhalb von Branchen mit eigenen Tarifverträgen der Tarifvertrag für den öffentlichen Dienst der Länder (TV-L). Die Gehälter für hochqualifizierte Kräfte nach diesen Tarifverträgen bleiben weit hinter ihren Verdienstmöglichkeiten in der freien Wirtschaft zurück und sind damit finanziell eher unattraktiv.

Für Empfänger hochvolumiger Beihilfen greift oftmals das sog. Besserstellungsverbot. Dabei wird der Bewilligungsbescheid für die Zuwendung mit einer Auflage versehen, die besagt, dass der Zuwendungsempfänger seine Arbeitnehmer:innen nicht besser entlohnen darf als vergleichbare Arbeitnehmer:innen des Zuwendungsgebers. Beispielhaft darf der Empfänger einer Beihilfe des Bundes in einem solchen Fall seine Angestellten nicht besser entlohnen, als es nach dem TVöD der Fall wäre.

11.5 Auswirkungen der Rechtsmaterien auf die Finanzierungsmodelle

Nachfolgend werden die Auswirkungen dieser Rechtsmaterien auf die einzelnen drei grundsätzlichen Finanzierungsmodelle dargestellt.

Das Modell der öffentlichen Finanzierung

Vorab ein Hinweis zur grundsätzlichen Zulässigkeit: Grundsätzlich ist der Bund befugt, Forschungsinfrastrukturen in eigener Verantwortung zu errichten oder sich an ihrer Finanzierung zu beteiligen. Neben der notwendigen Haushaltsermächtigung durch den Bundestag kommt es auf die konkrete Ausgestaltung der Struktur an, um dahinter liegende verfassungsrechtliche Fragen entsprechend beantworten zu können.

In Art. 91b des Grundgesetzes (GG) wurde verfassungsrechtlich abgesichert, dass Bund und Länder in Fällen überregionaler Bedeutung bei der Förderung von Wissenschaft und Forschung zusammenwirken können. Die Regelung wurde insbesondere für die Finanzierung großer Forschungsinfrastrukturen geschaffen und bietet dem Bund – dann in Zusammenwirken mit einem Sitzland – auch die Möglichkeit, das LEAM-Projekt öffentlich zu finanzieren.

Bei einem Modell der (rein) öffentlichen Finanzierung des LEAM-Projekts finden alle drei zuvor dargestellten Rechtsmaterien die stärkste Wirkung. Im Einzelnen:

EU-Beihilferecht

Bei einer rein oder jedenfalls weit überwiegenden öffentlichen Finanzierung gelten die oben genannten Darstellungen zum EU-Beihilferecht. Wie dort dargelegt, kann auf Basis des FuE-Rahmens voraussichtlich argumentiert werden, dass wesentliche geplante Einsätze des KI-Supercomputers nicht-wirtschaftliche Tätigkeiten darstellen und damit die öffentliche Finanzierung der geplanten Infrastruktur insoweit nicht den Beihilfetatbestand erfüllt. Das wäre insoweit vorteilhaft, als einerseits ein komplexes Notifizierungsverfahren bei der EU-Kommission entfällt und andererseits auch die Beschränkungen der AGVO, die in der Regel keine volle staatliche „Durchfinanzierung“ zulässt, nicht greifen würden. In einer öffentlichen Grundfinanzierung könnte, soweit es auch politisch gewollt ist, der Bund zusammen mit einem Sitzland damit grundsätzlich

Finanzmittel für den Aufbau der Infrastruktur sowie deren Betrieb (einschließlich Personal) zur Verfügung stellen.

Dabei sind auch Tätigkeiten aus dieser Infrastruktur heraus, die wirtschaftlicher Natur im Sinne des EU-Beihilferechts sind (z.B. das geplante Consulting-Angebot), nicht von vornherein ausgeschlossen. Der FuE-Rahmen erlaubt auch solche Tätigkeiten, wenngleich auch mit gewissen Anforderungen. Diese sollen nachfolgend überblicksmäßig dargestellt werden:

Das wesentlichste Gebot ist die Einführung einer sogenannten Trennungsrechnung. Das bedeutet Folgendes: Übt das LEAM-Projekt sowohl wirtschaftliche als auch nichtwirtschaftliche Tätigkeiten aus, so fällt die öffentliche Förderung der nichtwirtschaftlichen Tätigkeiten nicht unter das Beihilfeverbot, wenn die nichtwirtschaftlichen und die wirtschaftlichen Tätigkeiten und die Kosten, Finanzierung und Erlöse klar voneinander getrennt werden können, sodass keine Gefahr der Quersubventionierung der wirtschaftlichen Tätigkeiten besteht. Das Geld, das vom Bund bzw. einem Land stammt, darf nicht verwendet werden, um Verluste des wirtschaftlichen Bereichs auszugleichen (Verbot der Quersubventionierung und zugleich Gebot der sog. Trennungsrechnung, Ziff. 19 FuE-Rahmen). Eine solche Trennungsrechnung kann im Jahresabschluss durchgeführt werden, bedeutet aber auch, dass Einnahmen und Aufwand für nichtwirtschaftliche und wirtschaftliche Tätigkeiten in der Buchhaltung streng voneinander getrennt werden müssen und stets klar sein muss, ob ein Aufwand (gleich, ob Personal oder Infrastrukturnutzung) dem einen oder dem anderen Bereich zugeordnet werden kann. Das erfordert in der Praxis einiges an Verwaltungsaufwand, ist aber lösbar.

Deshalb darf der wirtschaftliche Bereich (also z.B. das Consulting) keinen Verlust erwirtschaften. Ebenso darf der wirtschaftliche Bereich nicht eine Hauptanwendung der öffentlich geförderten Infrastruktur werden, bei der der nichtwirtschaftliche Bereich in den Hintergrund gedrängt wird. Hierzu enthält der FuE-Rahmen noch eine Privilegierung für Forschungseinrichtungen bzw. -infrastrukturen: Wenn die Forschungseinrichtung oder Forschungsinfrastruktur fast ausschließlich für eine nichtwirtschaftliche Tätigkeit genutzt wird, kann ihre Förderung ganz aus dem Anwendungsbereich des Beihilferechts herausfallen, sofern die wirtschaftliche Nutzung eine reine Nebentätigkeit darstellt, die mit dem Betrieb der Infrastruktur unmittelbar verbunden und dafür erforderlich ist oder die in untrennbaren Zusammenhang mit der nichtwirtschaftlichen Haupttätigkeit steht und ihr Umfang begrenzt ist. Das ist der Fall, wenn für die wirtschaftlichen Tätigkeiten dieselben Inputs (wie Material, Ausstattung, Personal und Anlagekapital) eingesetzt werden, wie für die nichtwirtschaftlichen Tätigkeiten und wenn die für die betreffende wirtschaftliche Tätigkeit jährlich zugewiesene Kapazität nicht mehr als 20 % der jährlichen Kapazität der betreffenden Einrichtung bzw. Infrastruktur beträgt (Ziff. 21 FuE-Rahmen). Das würde, sehr grob vereinfacht, für das LEAM KI-Projekt beispielsweise bedeuten, dass 80 % der Rechenkapazität des KI-Supercomputers für nichtwirtschaftliche Aktivitäten benutzt werden dürfen (z.B. für die Entwicklung großer KI Foundation-Modelle mit Open Access für die Wirtschaft bzw. Wissenschaft), 20 % dann für wirtschaftliche Zwecke. Einzelheiten sind dann für den tatsächlichen Betrieb auszuarbeiten. Es zeigt sich aber, dass auch bei einer rein öffentlichen Finanzierung wirtschaftliche Tätigkeiten dem Projekt LEAM nicht untersagt sind und im Geschäftsmodell weiter eingeplant werden können. In

der Praxis gibt es solche Fälle zum Beispiel bei Hochschulen, deren Infrastruktur öffentlich finanziert ist, die aber gleichwohl wirtschaftliche Angebote wie Auftragsforschung oder forschungsnahе Dienstleistungen haben.

Nachdem eine Abgrenzung von wirtschaftlicher zu nichtwirtschaftlicher Tätigkeit vorgenommen wurde, muss sichergestellt werden, dass wirtschaftliche Tätigkeiten nicht zu einer Quersubventionierung führen. Dies bemisst sich am sog. Private-Investor-Test. Dabei wird das wirtschaftliche Handeln der staatlichen Stelle mit dem hypothetischen Verhalten eines Privatinvestors verglichen. Würde ein solcher den Vorteil nicht oder zu ungünstigeren Konditionen anbieten, liegt eine Beihilfe nach Art. 107 Abs. 1 AEUV vor. Der FuE-Rahmen enthält insoweit dezidierte Regelungen (vgl. Ziff. 26f. FuE-Rahmen), wonach die Preisermittlung sich entweder nach dem Marktpreis oder nach einer Vollkostenkalkulation mit Gewinnaufschlag richten muss. Gleichwohl können die Kosten, die das LEAM-Projekt für die Wirtschaft für solche wirtschaftlichen Tätigkeiten berechnet, betriebswirtschaftlich deutlich niedriger liegen als vergleichbare Angebote aus den USA, da sich auch beihilferechtlich das LEAM-Projekt bei diesen wirtschaftlichen Aktivitäten nicht an dem Gedanken der Profitmaximierung orientieren muss, sondern an dem Gedanken Vollkosten + angemessener Gewinnaufschlag. Letzterer dürfte deutlich niedriger liegen als bei Angeboten von US-Anbietern, die gegebenenfalls eine monopolähnliche Stellung haben.

Zusammengefasst bedeutet dies mithin, dass eine öffentliche Finanzierung des Projekts LEAM und deren Infrastruktur beihilferechtlich vor allem dann denkbar ist, wenn die geschaffene Infrastruktur überwiegend für nichtwirtschaftliche Zwecke genutzt und der Wissenschaft bzw. kooperierenden Unternehmen zur Verfügung gestellt wird. Eine gewisse Kapazität des Supercomputers und des Personals kann gleichwohl auch für wirtschaftliche Tätigkeiten zur Verfügung gestellt werden, wobei diese wirtschaftlichen Tätigkeiten in der Regel keinen Verlust erwirtschaften und zu einem Marktpreis angeboten werden müssen. Eine Trennungsrechnung muss buchhalterisch und im Jahresabschluss klarstellen, dass keine Quersubventionierung wirtschaftlicher Tätigkeiten erfolgt.

Vergaberecht

Das Vergaberecht ist bei einer rein öffentlichen Finanzierung voll einschlägig.

Wird das LEAM-Projekt nach dem ersten in Betracht kommenden Modell vollständig öffentlich finanziert, dann handelt es sich bei dem für das LEAM-Projekt zu schaffenden Rechtsträger um einen öffentlichen Auftraggeber im Sinne von §§ 98, 99 GWB. In diesem Fall sind die in Kapitel 11.4 aufgeführten Voraussetzungen von § 99 Nr. 2 lit. a) GWB erfüllt, denn es handelt sich in diesem Fall um eine juristische Person, die zu dem besonderen Zweck gegründet wurden, im Allgemeininteresse liegende Aufgaben nichtgewerblicher Art zu erfüllen und die mindestens überwiegend – hier: vollständig – von Gebietskörperschaften finanziert wird.

Dabei ist das Merkmal der Aufgaben nichtgewerblicher Art selbst dann erfüllt, wenn neben der Kernaufgabe der KI-Forschung und -Entwicklung im Rahmen des LEAM-Projekts auch Consulting oder die Vermietung von Rechenzeit angeboten werden, da es für die Frage der Nichtgewerblichkeit auf eine Gesamtschau ankommt, die Faktoren wie

eine Entkoppelung von Wettbewerb und Marktmechanismen und das Vorliegen einer Gewinnerzielungsabsicht berücksichtigt. Insoweit treten Nebenbetätigungen wie Consulting und Rechenzeitvermietung hinter der weit überwiegenden, nicht auf Gewinnerzielung ausgerichteten Haupttätigkeit zurück.

Öffentliches Dienst- und Vergütungsrecht

Beim Modell der rein öffentlichen Finanzierung gelten Art. 33 Abs. 2 GG und die Einschränkungen des öffentlichen Vergütungsrechts uneingeschränkt.

Die Geltung des verfassungsrechtlichen Grundsatzes der Bestenauslese führt dazu, dass eine rechtssichere Gestaltung von Bewerbungsverfahren einen gewissen Mehraufwand bedeutet. So muss beispielsweise für zu besetzende Stellen ein Anforderungsprofil festgelegt und Fragen im Bewerbungsgespräch hieran ausgerichtet werden und es gelten Dokumentationspflichten. Stellenausschreibungen müssen zudem fristgebunden sein.

Je nach Struktur des Projekts kommen zudem TVöD bzw. TVL zur Anwendung mit der Folge, nur vergleichbar unattraktiv entlohnen zu können. Für Angestellte bei rein oder überwiegend staatlich finanzierten Unternehmen, auch wenn diese eine privatrechtliche Form haben, greift das Besserstellungsverbot. Dieses ist einzuhalten bei einer institutionellen Förderung (Zuwendungen zur Deckung der gesamten oder eines nicht abgegrenzten Teils der Ausgaben des Zuwendungsempfängers) sowie bei Projektförderung, wenn die Gesamtausgaben des Zuwendungsempfängers überwiegend aus Zuwendungen der öffentlichen Hand bestritten werden (vgl. § 8 Haushaltsgesetz 2021). Der Bundesfinanzminister kann bei Vorliegen zwingender Gründe zwar Ausnahmen erlassen, dies wird jedoch restriktiv gehandhabt.

Für das EU-Beihilferecht und das öffentliche Dienst- und Vergütungsrecht kann jedoch noch ein gemeinsamer Hinweis gegeben werden: Selbst bei einer rein oder überwiegend öffentlichen Finanzierung des LEAM-Projekts ist es möglich, dass neben der Trägerstruktur des LEAM-Projekts auch eine Service-GmbH errichtet wird, die sich auf wirtschaftliche Aktivitäten fokussiert und zu deren Erbringung Infrastruktur des LKS anmietet (unter Berücksichtigung der oben dargestellten Kapazitätsbeschränkungen). Die Ausgliederung von wirtschaftlichen Aktivitäten im Sinne des EU-Beihilferechts in eine wirtschaftliche Einheit ist beihilferechtlich bei Einhaltung der Regeln zum Verbot der Quersubventionierung und der Kapazitätsbeschränkungen für die wirtschaftliche Nutzung geförderter Infrastrukturen grundsätzlich zulässig.

Sie ist in der deutschen Wissenschaftslandschaft erprobt und bietet verschiedene Vorteile:

- Die Trennungsrechnung, die die geförderte Einrichtung führen muss, wird deutlich vereinfacht, wenn alle wirtschaftlichen Tätigkeiten in eine dafür eingerichtete GmbH ausgegliedert sind.
- Die rechtlich von der geförderten Struktur getrennte wirtschaftliche Einheit (also beispielsweise eine gesonderte Service-GmbH) genießt größere Handlungsspielräume in Fragen der Vergütung und Preisbestimmung. Insbesondere kann sie Beschäftigte der geförderten nichtwirtschaftlichen Einheit beispielsweise in Nebentätigkeit beschäftigen und vergüten, um auf diese Weise im Rahmen des rechtlich Zulässigen ein interessantes Vergütungspaket zu bieten. Da zudem die Beschäftigten dann mit der Nebentätigkeit noch eigene Anstrengungen unternehmen müssen, um diese Zusatzvergütung zu erreichen, ist in der Regel auch in der Außendarstellung und in der Politik eine Akzeptanz solcher Lösungen erreichbar. Diese Vorgehensweise ist insbesondere in der Wissenschaft erprobt. Rechtliche Parameter müssen im Einzelnen ausgearbeitet werden.
- In der Umsetzung würden Infrastruktur/Kapazitäten der wirtschaftlich orientierten Einheit/GmbH von der geförderten Einheit angemietet, was beihilferechtlich grundsätzlich zulässig ist und eine klare Trennung zwischen wirtschaftlichen und nichtwirtschaftlichen Bereichen ermöglicht.

Im Rahmen der konkreten Ausgestaltung kann es vor diesem Hintergrund sinnvoll sein, die Trennung von nichtwirtschaftlichen und wirtschaftlichen Aktivitäten des Projekts LEAM in zwei rechtlich getrennten Strukturen zu untersuchen.

SPOTLIGHT

Ubermetrics Technologies GmbH

An verschiedenen Stellen in der Studie stellen wir ausgewählte deutsche Startups und Unternehmen vor, die KI-Foundation-Modelle bereits nutzen. Sie berichten über ihre Anwendungsfälle, Herausforderungen und die Vorteile europäischer Modelle.



Patrick Bunk, Founder und CEO von Ubermetrics.

Die **Ubermetrics** Content Intelligence-Plattform basiert auf einer für die Bedürfnisse von Kommunikatoren speziell entwickelten und trainierten Künstlichen Intelligenz, welche öffentlich verfügbare Inhalte und Daten in aussagekräftige Insights für effektive Kommunikationsstrategien umwandelt. Mit der Verarbeitung von über 50.000 Artikeln pro Minute und Inhalte von mehr als 460 Millionen Quellen ist Ubermetrics die führende Content Intelligence Plattform für Marketing- und PR-Experten.

Wo setzt ihr Foundation-Modelle ein? Was ist euer Use-Case?

Ubermetrics extrahiert Texte aus Millionen von Internetquellen und wertet diese unter dem Einsatz menschlicher Experten und automatisiert durch KI-Verfahren für Kunden aus. Wir setzen Foundation-Modelle bislang begrenzt in einigen Teilen der Ergebnisdarstellung ein. Beispielsweise setzen wir ein RoBERTa Modell ein, um automatisiert Entitäten (bspw. Personen, Orte, Produkte etc.) in Texten zu erkennen, um diese für die weitere Analyse nutzbar zu machen (sog. Named Entity Recognition / -Linking). Allerdings schöpfen wir damit die Möglichkeiten der Modelle für unsere Produkte erst ansatzweise aus.

Welchen Einfluss haben KI-Foundation-Modelle auf euer Geschäftsmodell bzw. eure Projekte?

Foundation-Modelle sind von überragender Bedeutung für die Wettbewerbsfähigkeit und den Erfolg von Ubermetrics, bzw. UNICEPTA. Bislang mussten wir aufwändig für jeden einzelnen Aspekt in unserer Produktionskette und teilweise für jede Sprache ein neues KI-Verfahren implementieren oder selbst erforschen (bspw. Sentimentanalyse). Diese könnten zu einem großen Teil durch auf Foundation-Modellen basierten Verfahren ersetzt werden- bei gleichzeitiger Verbesserung der Treffsicherheit. Das betrifft sowohl die Textsammlung als auch die Aufbereitung (bspw. Spracherkennung) und die Auswertung. Foundation-Modelle erlauben uns nun eine umfassende Basis, auf der wir die Entwicklung besserer Verfahren und die Entwicklung neuer Produkte aufsetzen können. Für uns sinkt also einerseits der Aufwand in der Produktion und wir werden andererseits in die Lage versetzt, bessere bzw. für uns gänzlich neue Produkte und Features anbieten zu können. Bspw. könnten wir Summaries über mehrere Artikel hinweg und Abstracts von Artikeln teilautomatisieren, um unsere Wissensarbeiter:innen bei weniger anspruchsvollen Tätigkeiten maschinell zu unterstützen. Wir schätzen den wirtschaftlichen Wert durch die Implementation auf 60 % unseres Umsatzes ein.

Welche Schwierigkeiten und Probleme siehst du, dass nur USA und China derzeit KI-Foundation-Modelle umfassend bereitstellen?

Die größten Schwierigkeiten und Probleme liegen in dem Verlust digitaler Souveränität und in der schwächeren Unterstützung relevanter europäischer Sprachen. Sollten wir gezwungen sein, außereuropäische Modelle über APIs zu nutzen, müssten wir dazu relevante und zum Teil sensible Geschäftsdaten preisgeben. Dies ist nur schwer vereinbar mit europäischen Datenschutzstandards und ein Hauptgrund für unseren bisher nur sporadischen Einsatz solcher Modelle.

Wie würden euch europäische Modelle - Open Source, alle europäischen Sprachen abdeckend, mit hohen Datenschutzstandards und minimalem Bias - helfen?

Von europäischen Open-Source-Modellen erhoffen wir uns neben der besseren Sprachabdeckung u.a. durch die europäischen Datenschutzstandards eine größere Akzeptanz unserer Kunden. Gleichzeitig erwarten wir eine bessere Planbarkeit in Bezug auf Weiterentwicklungen.

Das Modell der privaten Finanzierung

Im Gegensatz zur rein öffentlichen Kapitalbeschaffung steht grundsätzlich das Modell der privaten Finanzierung bzw. der überwiegenden privaten Beherrschung der gewählten Struktur.

Angenommen, es werden keine staatlichen Mittel in Anspruch genommen, sondern das Modell wird rein privat finanziert und betrieben, so gelten die oben genannten Beschränkungen nicht. Das LEAM-Projekt muss zwar in diesem Fall letztlich „sich selbst tragen“ bzw. (wenn nicht die gemeinnützige Form gewählt wird) und zum wirtschaftlichen Überleben sogar Gewinne erwirtschaften, doch greifen in diesem Fall weder beihilferechtliche noch vergaberechtlich noch tarifliche Beschränkungen. Das LEAM-Projekt kann (und muss) dann „frei am Markt“ agieren. Aufträge können schnell vergeben werden, am Markt übliche Gehälter bezahlt werden. Das gilt selbst dann, wenn aus steuerrechtlichen Gründen die Wahl auf eine gemeinnützige Gesellschaft fällt (mit bestimmten Einschränkungen des Gemeinnützigkeitsrechts, die hier nicht abschließend dargestellt werden können).

EU-Beihilfenrecht

Einer solchen privat strukturierten bzw. überwiegend privat finanzierten Einrichtung ist es dann wiederum möglich, sich allgemein auf Fördermittel für wissenschaftliche Infrastruktur und Projekte, im Wege der Projektförderung zum Beispiel bei der DFG, bei dem Bundesministerium für Forschung und Bildung oder bei der EU zu bewerben – so wie jeder anderen öffentlichen oder privaten Einrichtung auch. Diese Projektförderung wird jedoch nicht nur kompetitiv für begrenzte Laufzeiten vergeben (das heißt die Mittel sind zeitlich beschränkt und ihr Erhalt aufgrund des Wettbewerbscharakters öffentlicher Ausschreibungen nicht gesichert), sondern ist dann auch gegebenenfalls wieder mit öffentlich-rechtlichen Pflichten verbunden. Das können beispielsweise wiederum Beschränkungen in der Vergütung sein (gegebenenfalls gilt das Besserstellungsverbot, s. oben) oder es sind bei der Vergabe der Mittel an Unternehmen durch die geförderte Infrastruktur (also hier dann das LEAM-Projekt, das beispielsweise Rechenkapazitäten einkaufen will) öffentliche Ausschreibungen durchzuführen. Das hängt von dem

jeweiligen Fördertopf, der in Anspruch genommen werden soll, und der Förderintensität ab.

Es ist jedoch nicht ausgeschlossen, dass auch der Bund oder ein Land dann wiederum Finanzierungshilfe leisten, wenn die zugrundeliegende Einrichtung ihrem Wesen nach privat finanziert und in der Gesellschafterstruktur auch betrieben werden soll. Deutschland kennt zahlreiche Förderinstrumente für Start-Ups und Unternehmen mit Technologiebezug, die in ihrer Bandbreite hier jedoch nicht alle dargestellt werden können. Auch der bekannte High-Tech-Gründerfonds gehört dazu, ebenso wie Kreditprogramme der staatlichen Förderbank KfW auf Bundesebene oder der Förderbanken auf Landesebene. Unabhängig davon, dass sich hieraus auch wiederum beihilferechtliche Prüfungen ergeben können (dazu sogleich), können solche Mittel oftmals nur eine „Anschubfinanzierung“ von beschränktem Umfang sein, neben die erhebliche Privatmittel treten müssen (siehe auch die Ausführungen zur AGVO und zum Public-Private-Partnership, dazu sogleich).

Solche Finanzierungen müssen sich dann, wenn sie einen erheblichen Umfang haben sollen, voraussichtlich als „Beihilfen“ für ein Unternehmen vor allem an den Regelungen der AGVO messen lassen, soll ein langwieriges Notifizierungsverfahren vermieden werden. Beihilfen, die auf Grundlage der AGVO gewährt werden, müssen bestimmte Transparenzvorgaben erfüllen und einen Anreizeffekt haben. Außerdem müssen Informationen zu den Beihilfen auf einer „Beihilfe-Website“ des Mitgliedstaates bzw. seiner handelnden Körperschaft veröffentlicht und der Kommission mitgeteilt werden. Die Mitgliedstaaten sind zum Monitoring verpflichtet. Für die Förderung des LKS könnten insbesondere folgende Kategorien der AGVO relevant sein:

Art. 17 AGVO: Investitionsbeihilfen für KMU

Bis zu 20 % der Investitionskosten kleiner bzw. 10 % der Investitionskosten mittlerer Unternehmen für immaterielle und materielle Investitionsgüter. Darunter fallen u. a. die Errichtung einer neuen Produktionsstätte sowie die Deckung von Lohnkosten der durch das Investitionsvorhaben geschaffenen Arbeitsplätze.

Art. 18 AGVO: Investitionsbeihilfen für KMU bei der Inanspruchnahme von Beratungsdiensten

Mit einer Höhe von bis zu 50 % der Kosten dürfen Beratungsleistungen externer Berater bezuschusst werden, insofern es sich dabei nicht um gewöhnliche Werbungskosten und fortlaufende Dienstleistungen handelt.

Art. 26 AGVO: Investitionsbeihilfen für Forschungsinfrastrukturen

Bau und Ausbau von Forschungsinfrastrukturen können mit bis zu 50 % der Kosten als bevorzugte Beihilfe nach den Regeln der AGVO gefördert werden. Der Preis für Betrieb und Nutzung der so geförderten Infrastruktur muss dem Marktpreis entsprechen. Die Infrastruktur muss außerdem mehreren Nutzer:innen offenstehen und der Zugang muss zu transparenten und diskriminierungsfreien Bedingungen gewährt werden. Unternehmen, die mindestens 10 % der Investitionskosten der Infrastruktur finanziert haben, können einen bevorzugten Zugang erhalten. Wenn eine Forschungsinfrastruktur sowohl wirtschaftliche als auch nichtwirtschaftliche Tätigkeiten ausübt, ist eine Trennungsrechnung zwingend.

Art. 28 AGVO: Innovationsbeihilfen für KMU

Bis zu 50 % der Kosten, die kleinen und mittleren Unternehmen u. a. für die Abordnung hochqualifizierten Personals für Forschungs- und Wissensverbreitung oder für Innovationsberatungsdienste oder innovationsunterstützende Dienstleistungen entstehen, dürfen als AGVO-Beihilfen finanziert werden. Bis zu einem Betrag von 200.000 EUR in drei Jahren dürfen auch bis zu 100 % der Kosten für Innovationsberatungsdienste und innovationsunterstützende Dienstleistungen finanziert werden.

Vergaberecht

Bei einer rein privaten Finanzierung ist das LEAM-Projekt kein öffentlicher Auftraggeber, sodass bei seiner Beschaffung von Leistungen das Vergaberecht nicht zur Anwendung kommt. Das führt zu deutlich freieren Beschaffungsvorgängen.

Gegebenenfalls kann es im Rahmen einer von einer privaten Einheit beantragten Projektfinanzierung dazu kommen, dass die Fördermittelbedingungen Vorschriften zur öffentlichen Vergabe anzuschaffender Infrastruktur enthalten. Das kann, muss aber nicht der Fall sein.

Öffentliches Dienstrecht

Das öffentliche Dienstrecht kommt in diesem Fall in Bezug auf die Personalgewinnung nicht zur Anwendung. Da das Besserstellungsverbot nur bei überwiegend öffentlicher Finanzierung greift, entfällt dieses bei einer (überwiegend) privaten Finanzierung. In Bezug auf die Arbeitnehmer:innengewinnung ist eine rein private Finanzierung insgesamt deutlich freier und eine höhere Vergütung für Arbeitnehmer:innen attraktiver.

Das Modell der Public-Private-Partnership

Die Mischform zu den vorhergehenden Modellen stellt das Konzept der Public-Private-Partnership, also der finanziellen und/oder (gesellschafts-)rechtlich statuierten Zusammenarbeit zwischen Privaten und dem Staat dar. In der Praxis sind hier verschiedenste Ausprägungen denkbar.

Bevor die Anwendbarkeit und Bedeutung von Beihilfen- Vergabe- und öffentlichem Dienstrecht auf das Modell der Public-Private-Partnership dargestellt werden kann, ist zunächst zu erklären, was eine solche ist, welche Formen möglich sind und inwieweit der Staat sich an einer solchen beteiligen darf.

Begriffserklärung

Eine sog. Public-Private-Partnership (nachfolgend auch „PPP“), zu Deutsch Öffentlich-private-Partnerschaft, kurz ÖPP, stellt eine Mischform zwischen einer rein öffentlichen Finanzierung und Struktur und einer ausschließlich privaten Ausgestaltung dar.

Es handelt sich bei einer PPP um eine partnerschaftliche Zusammenarbeit zwischen der öffentlichen Hand und privaten Unternehmen zur Realisierung eines öffentlichen Projekts oder Erbringung von Leistungen, die der Erfüllung einer öffentlich-rechtlichen Aufgabe dienen.

Für die Betrachtung der rechtlichen Möglichkeiten, Anforderungen und einschlägigen Vorschriften kommt es auf die konkrete Ausgestaltung der jeweiligen Kooperation an. Für

PPP existieren nur lückenhafte Regelungen in verschiedenen Gesetzen, aber keine umfassenden gesetzlichen Vorgaben.

Ausgestaltungsmodelle

Für die konkrete Ausgestaltung von PPP kommen zahlreiche verschiedene Organisationsmodelle in Frage. Wiederum sind diese nicht gesetzlich definiert, sondern haben sich in der Praxis herausgebildet. Dabei erfolgt die Kategorisierung von Organisationsmodellen häufig uneinheitlich.

Anzahl, Namen und genaue Beschreibung der Organisationsformen unterscheiden sich je nach Betrachtung. Im Groben kann aber folgende Kategorisierung erfolgen:

Konzessionsmodell

Beim Konzessionsmodell überträgt die öffentliche Hand einem privaten Unternehmen eine öffentliche Aufgabe. Der Private betreibt die betroffene Einrichtung dabei in eigenem Namen, auf eigene Rechnung und auf eigenes wirtschaftliches Risiko. Nur er tritt hierbei gegenüber Dritten auf. Er generiert Einnahmen dadurch, dass er von Dritten für die Nutzung einer Sache bzw. die Inanspruchnahme einer Leistung ein Entgelt verlangt. Im Falle von LEAM wäre insoweit die Vermietung von Rechenzeit sowie Consulting möglich. Zum Teil wird bei Anwendung des Konzessionsmodells zudem vereinbart, dass die öffentliche Hand als Konzessionär an den Privaten eine Konzessionsabgabe entrichtet. Bei diesem Modell ist allerdings anzumerken, dass für eine solche Konzessionsvergabe eine europaweite Ausschreibung nötig ist.

Die Problematik besteht bei einem Konzessionsmodell jedoch darin, dass es sich bei dem Betrieb des KI-Hochleistungsrechenzentrums denklogisch um eine (zwingende) öffentliche Aufgabe handeln muss, die dem Staat obliegt (im Gegensatz zu einer staatlichen Förderung von wünschenswerten Aktivitäten, zum Beispiel in der Wissenschaft). Das erscheint derzeit nur schwer darstellbar, bedarf jedoch auch einer Erörterung mit politisch Verantwortlichen, ob sie gegebenenfalls einen solchen Tatbestand durch Verordnung oder Gesetz schaffen wollen. Bis dahin sind Zweifel an der Anwendbarkeit des Konzessionsmodells gegeben.

Leasingmodell

Beim Leasingmodell beauftragt die öffentliche Hand eine Privatperson oder eine Leasinggesellschaft mit der Planung, Errichtung und Finanzierung eines Vorhabens gegen Zahlung einer vereinbarten Leasingrate. Die Privatperson wird Eigentümer:in bzw. Inhaber:in des Objekts, aber gewährt der öffentlichen Hand das Nutzungsrecht während der vereinbarten Leasingdauer. Hierzu erfolgt gegebenenfalls ein Widmungsakt, mit dem das Objekt zu einer öffentlichen Sache wird.

Betreibermodell

Beim Betreibermodell verwirklicht eine private Organisation vertraglich festgelegte Teilaufgaben in Bezug auf Planung, Entwicklung, Betrieb und Finanzierung und erhält hierfür ein vereinbartes Entgelt. Dabei ist sowohl möglich, dass der Private eher als technischer Erfüllungsgehilfe nur im Innenverhältnis zum Verwaltungsträger auftritt, als auch, dass er, näher am Konzessionsmodell, auch gegenüber Dritten agiert und von diesen ein Entgelt erhebt.

Betriebsführungsmodell

Beim Betriebsführungsmodell betreibt eine private Organisation eine öffentliche Einrichtung namens und im Auftrag der öffentlichen Hand auf deren Rechnung und deren Risiko. Er tritt gegenüber Dritten also nur in fremden Namen auf.

Die vorgenannten Modelle können unter dem Oberbegriff der „Vertragsmodelle“ zusammengefasst werden. Ihnen ist gemein, dass öffentliche Hand und Private hier auf vertraglicher Basis miteinander kooperieren, es jedoch nicht zur Schaffung eines gemeinsamen Rechtssubjektes kommt.

Beteiligungs- oder Kooperationsmodell

Anders liegt es bei dem sog. Beteiligungs- oder Kooperationsmodell. Bei diesem gründen die öffentliche Hand und private Akteure eine gemeinsame Gesellschaft. An dieser beteiligen sich sowohl öffentliche Stellen (z.B. Bund und/oder Bundesländer) als auch private Unternehmen als Gesellschafter:innen.

Aufgrund dieser Schaffung einer neuen Gesellschaft als von ihren Gesellschaftern zu unterscheidende eigenständige Rechtspersönlichkeit wird dieses Modell auch als „institutionelle PPP“ bezeichnet, aufgrund der gemeinsamen Inhaberschaft von öffentlichen und privaten Rechtsträgern auch als „gemischtwirtschaftliches Unternehmen“.

Es wird ein Gesellschaftsvertrag abgeschlossen. Neben der Rechtsform der Gesellschaft (s. Kapitel 12) ist in diesem zunächst festgelegt, wie groß der Anteil der Gesellschaftsanteile ist, den die jeweiligen Gesellschafter:innen tragen. Dabei kann der Anteil der öffentlichen Hand unterschiedlich groß sein. Sie kann, was in der Praxis häufig vorkommt, Mehrheitsanteileignerin sein; möglich ist aber auch eine Beteiligung zu einem Anteil von weniger als 50 %.

Der Gesellschaftsvertrag enthält weitere Regelungen zur Geschäftsführung und Vertretung der Gesellschaft, zu Rechten und Pflichten der Gesellschafter:innen, zur Verteilung von Gewinn und Verlust, zur Besetzung von Gesellschaftsorganen. Von den gesellschaftsrechtlichen Einflussnahmemöglichkeiten verspricht man sich auch eine verbesserte Steuerungs- und Kontrollmöglichkeiten durch die öffentliche Hand Zu Vor- und Nachteilen verschiedener Gesellschaftsformen (s. Kapitel 12).

Nach dem Verständnis von LEAM als auf Dauer angelegtes Projekt erscheint die Gründung einer gemeinsamen Gesellschaft als passendste unter den möglichen Ausgestaltungsformen von PPPs. Daher fokussieren sich die folgenden Ausführungen auf die rechtlichen Auswirkungen einer solchen Ausgestaltung.

Voraussetzungen für öffentlich-private Kooperationen

Bei der Erfüllung seiner Aufgaben hat der Staat grundsätzlich einen weiten Gestaltungsspielraum. Das betrifft auch die Frage, ob er in privatrechtlicher Form tätig wird und inwieweit er private Akteure zur Wahrnehmung staatlicher Aufgaben heranziehen kann. Abgesehen von für LEAM nicht einschlägigen Ausnahmebereichen ist daher auch die Beteiligung an Public-Private-Partnerships möglich.

Auch weitere haushaltsrechtliche Regeln für die Beteiligung der öffentlichen Hand an Unternehmen – etwa die Begrenzung der Einzahlungsverpflichtung auf einen bestimmten Betrag und die Sicherung eines angemessenen staatlichen Einflusses in den Überwachungsorganen der Gesellschaft können durch das LEAM-Projekt eingehalten werden.

EU-Beihilfenrecht

Im Hinblick auf das EU-Beihilfenrecht ergeben sich für Public-Private-Partnerships keine speziellen Rechtsfragen. Durch die Beteiligung des Staates und die Finanzierung des Projekts mit staatlichen Mitteln bleibt das Beihilferecht grundsätzlich anwendbar, wobei sich jedoch für die Finanzierung nichtwirtschaftlicher Aktivitäten die gleichen Maßstäbe ergeben wie bei der rein öffentlichen Finanzierung.

Für die Unterscheidung zwischen wirtschaftlichen und nichtwirtschaftlichen Tätigkeiten – die entscheidend für die Frage des Vorliegens einer staatlichen Beihilfe ist – gelten, da es insoweit nicht auf die Rechtsform und die Art der beteiligten Gesellschafter:innen ankommt, sondern die Abgrenzung tätigkeitsbezogen erfolgt, die in Kapitel 11.4 ausgeführten Grundsätze. Auch insoweit ist also der FuE-Rahmen maßgeblich. Dessen Anforderungen, insb. das Gebot der Einführung einer Trennungsrechnung und das Verbot einer Quersubventionierung wirtschaftlicher Tätigkeiten sind auf PPP zu übertragen. Möglich ist allerdings, die Struktur hier bei einer Finanzierung Privater so auszugestalten, dass im stärkeren Umfang wirtschaftliche Aktivitäten möglich sind; dies insbesondere dann, wenn die privaten Investoren sich bereit erklären, etwaige Verluste der wirtschaftlichen Tätigkeiten selbst auszugleichen (so dass aus staatlichen Mitteln keine Quersubventionierung erfolgt). Einzelheiten müssen dann bei der konkreten Ausgestaltung evaluiert werden.

Soweit es sich begrifflich um eine Beihilfe handelt, kann auch beim PPP-Finanzierungsmodell eine Ausnahmenvorschrift greifen. Insoweit kommen insb. die oben aufgeführten AGVO-Beihilfen in Betracht. Liegt eine nicht unter Ausnahmenvorschriften fallende staatliche Beihilfe vor, ist ein Notifizierungsverfahren durchzuführen.

Vergaberecht

Für die Frage, inwieweit das Vergaberecht auf PPPs Anwendung findet, sind verschiedene Themenkomplexe voneinander abzugrenzen: Zu unterscheiden sind einerseits eine Zusammenarbeit, die einem der oben unter dargestellten Vertragsmodelle (Konzessions-, Leasing-, Betreiber- oder Betriebsführungsmodell) zuzuordnen ist, andererseits PPPs nach dem Beteiligungsmodell, bei dem eine gemeinsame Gesellschaft gegründet wird. Bei diesen wiederum ist danach zu differenzieren, ob es um die Anwendbarkeit von Vergaberecht auf die Gesellschaftsgründung oder um die Beschaffung von Leistungen durch die bestehende Gesellschaft geht.

Für Kooperationen nach einem der Vertragsmodelle kann aufgrund der Vielfältigkeit möglicher Modelle im derzeitigen Stadium keine vollständige vergaberechtliche Bewertung erfolgen. Da ihnen aber gemein ist, dass der Staat hier alle oder bestimmte Teile der Projektführung auf Private überträgt, wird es sich bei dem Projektvertrag häufig um einen öffentlichen Auftrag im Sinne von § 103 Abs. 1 GWB oder eine Konzession nach § 105 Abs. 1 GWB handeln. Daher muss dem Abschluss eines Projektvertrags mit einem privaten Kooperationspartner in aller Regel ein formelles Vergabeverfahren vorausgehen, aufgrund des hohen Projektvolumens mit europaweiter Bekanntmachung.

Für die Gründung einer Joint Venture (Beteiligungsmodell) stellt sich die Rechtslage folgendermaßen dar: Die Gründung der gemischtwirtschaftlichen Gesellschaft selbst stellt grundsätzlich keinen beschaffungsrelevanten Vorgang dar, unterliegt also nicht

dem Vergaberecht. Anders liegt es jedoch, wenn mit dem Gründungsakt zugleich eine unmittelbare Übertragung von Aufgaben von der öffentlichen Hand an die Gesellschaft einhergeht. In diesem Fall wird das beteiligte private Unternehmen nicht nur Gesellschafter, sondern zugleich auch Leistungserbringer gegenüber dem Staat. Im Rahmen einer vorzunehmenden Gesamtbetrachtung stellt sich dieses Gesamtkonstrukt insgesamt als ausschreibungspflichtiger Vorgang in Form eines öffentlichen Auftrags im Sinne von § 103 Abs. 1 GWB bzw. als Konzession nach § 105 Abs. 1 GWB dar, da Gesellschaftsgründung und Aufgabenwahrnehmung ein unteilbares Ganzes sind. Dasselbe gilt, wenn ein gemischtwirtschaftliches Unternehmen durch die öffentliche Hand und private Unternehmen gegründet wird, um eigene Aufgaben in einem entsprechend definierten Bereich zu übernehmen und die Leistungen im Anschluss durch die Gesellschafter erbracht werden.

Werden später – von der Gesellschaftsgründung zu unterscheiden – durch die Gesellschaft für sie erforderliche Aufträge in einem nachgelagerten Schritt an Dritte vergeben, so ist die Gesellschaft in der Regel selbst als öffentlicher Auftraggeber nach § 99 Nr. 2 lit. a) GWB einzuordnen. Beschafft sie sich Leistungen am Markt, so unterliegt dieser Beschaffungsakt in aller Regel den vergaberechtlichen Vorgaben. Die oben in Kapitel 11.4 dargestellten Voraussetzungen dieser Vorschrift liegen vor. Insbesondere folgende sind gegeben: Wie auch im Rahmen der Ausführungen zum Vergaberecht bei rein öffentlicher Finanzierung dargestellt ist bei den im LEAM-Projekt geplanten Aktivitäten im Rahmen der vorzunehmenden Gesamtschau insgesamt von im Allgemeininteresse liegenden Aufgaben nichtgewerblicher Art auszugehen. Insoweit ergeben sich keine Unterschiede. Die weitere Voraussetzung des § 99 Nr. 2 lit. a) GWB, dass die Gesellschaft überwiegend von Gebietskörperschaften – hier voraussichtlich Bund und Standortland – finanziert wird, liegt vor, wenn – wie zu erwarten – die Finanzierung zu mehr als 50 % staatlich ist.

Öffentliches Dienst- und Vergütungsrecht

Auch bei von der öffentlichen Hand beherrschten Public-Private-Partnerships (also solchen mit staatlicher Beteiligung von mehr als 50 %) muss die Stellenbesetzung nach Art. 33 Abs. 2 GG erfolgen. Der bisherigen Rechtsprechung des Bundesarbeitsgerichts zu öffentlichen Betrieben in privater Rechtsform ist zu entnehmen, dass es auch für öffentlich beherrschte gemischtwirtschaftliche Unternehmen von einer Geltung der Vorschrift ausgehen wird.

Vergütungsrechtliche Einschränkungen gelten nur, wenn aufgrund der Beihilföhe das Besserstellungsverbot zur Anwendung kommt.

11.6 Abschließende Übersicht

Die nachfolgende Tabelle stellt überblicksmäßig die rechtlichen Vor- und Nachteile der einzelnen Modelle in Bezug auf die drei grundsätzlichen Finanzierungsarten dar:

	Öffentlich finanziert	Privat finanziert	Public-Private-Partnership
EU-Beihilferecht	<ul style="list-style-type: none"> - voll anwendbar - Trennungsrechnung - Beschränkung wirtschaftlicher Tätigkeiten - FuE-Rahmen 2022 mit vielen Privilegien 	<ul style="list-style-type: none"> - nicht anwendbar oder nur eingeschränkt - Beihilfen nach AGVO denkbar 	<ul style="list-style-type: none"> - mögliche Beschränkung wirtschaftlicher Tätigkeit und Trennungsrechnung; FuE-Rahmen - im Übrigen Beihilfen nach AGVO denkbar
Vergaberecht	<ul style="list-style-type: none"> - voll anwendbar - öffentliche Ausschreibungen notwendig 	<ul style="list-style-type: none"> - grundsätzlich nicht anwendbar 	<ul style="list-style-type: none"> - anwendbar auf den Gründungsakt, wenn damit zugleich Aufgabenübertragung auf den privaten Partner und bei Konzessionierung - in aller Regel anwendbar auf Leistungsbeschaffung durch die gemischtwirtschaftliche Gesellschaft
Öffentliches Dienst- und Vergütungsrecht	<ul style="list-style-type: none"> - voll anwendbar - Besserstellungsverbot greift ein 	<ul style="list-style-type: none"> - in der Mitarbeiterauswahl und Vergütung frei (ggf. steuerliche Einschränkungen bei Gemeinnützigkeit) - Marktvergütungen möglich 	<ul style="list-style-type: none"> - Mitarbeitergewinnung unterliegt Vorgaben der Bestenauslese, wenn staatliche Beteiligung über 50 % - Besserstellungsverbot je nach Beihilfenhöhe möglich

Tabelle 21: Übersicht der Vor- und Nachteile der drei Finanzierungsmodelle mit Hinblick auf die rechtlichen Rahmenbedingungen

Gesellschaftsstruktur von LEAM



12. Gesellschaftsstruktur von LEAM

Die Gesellschaftsstruktur von LEAM ist stark abhängig von der Finanzierungsart des Projektes. Hier bieten sich je nach Finanzierungsstruktur verschiedene Gesellschaftsformen an. Im Rahmen der Finanzierung sollten diese genauer evaluiert und bewertet werden. Generell ist aufgrund der Struktur des Projektes die Gründung einer Projektentwicklungsgesellschaft denkbar.

Grundsätzlich können verschiedene Gesellschaftsformen unterschieden werden. Diese sollen anhand der Finanzierungsform kurz beleuchtet werden.

12.1 Öffentliche Finanzierung

Sofern eine rein öffentliche Finanzierung gewählt wird, kommen neben den Rechtsformen des öffentlichen Rechts (vor allem der Anstalt des öffentlichen Rechts) grundsätzlich mehrere privatrechtliche Formen in Betracht. Auch in der deutschen Förder- und Wissenschaftspraxis gibt es Einrichtungen, die als e.V. organisiert sind (zum Beispiel Max-Planck-Gesellschaft oder Fraunhofer) oder auch als GmbHs. Insbesondere die GmbH – gegebenenfalls auch in Form der gemeinnützigen GmbH mit dem besonderen steuerrechtlichen Einschlag – kann in der Praxis heute eine Rolle für die Aufgabenerfüllung der öffentlichen Hand spielen. Dem Bund oder einem Land ist hierbei eine Gesellschaftsbeteiligung nach dem deutschen Haushaltsrecht im Umfang der Investition möglich (wenn also beispielsweise der Bund 90 % der Investition trägt, würde er auch 90 % der Geschäftsanteile an der GmbH enthalten). Dies folgt auch aus der Logik, dass das EU-Beihilferecht und das Vergaberecht sich mit ihren Möglichkeiten und Beschränkungen nicht an der Rechtsform, sondern vielmehr an der tatsächlich ausgeübten Tätigkeit orientieren. Das GmbH-Recht ist zudem hinreichend flexibel, um auch besondere Organstrukturen abzubilden und gleichzeitig eine gewisse Flexibilität der Handelnden, insbesondere der Geschäftsführung, zu erlauben. Bei einer GmbH-Konstruktion kann es sich zudem anbieten, eine eher nicht-wirtschaftlich orientierte Träger-GmbH für die Infrastruktur (ggf. als gGmbH) mit einer wirtschaftlich orientierten GmbH zu verbinden.

Die Aktiengesellschaft und die Genossenschaft erscheinen dagegen in diesem Modell nicht sinnvoll, da sie haushaltsrechtlich für den Bund und ein beteiligtes Bundesland nicht darstellbar erscheinen und zudem auch in der rechtlichen Flexibilität Probleme auslösen können.

12.2 Private Finanzierung

Im Rahmen einer privaten Finanzierung sind grundsätzlich alle Gesellschaftsformen, die oben dargestellt sind, denkbar.

Die Rechtsform der AG ist im Hinblick auf die Übertragbarkeit von Anteilen und die – auch kurzfristige – Kapitalbeschaffung flexibel. Durch strenge gesetzliche Vorgaben wird indes der unternehmerische Gestaltungsspielraum reduziert. Zudem ist eine AG eher für ein kapitalmarktorientiertes Unternehmen sinnvoll. Übertragbarkeit und kurzfristige Kapitalbeschaffung sind bei einer GmbH im Vergleich dazu erschwert. Im Gegensatz zum Vorstand einer AG ist die Geschäftsführung einer GmbH an Weisungen der Gesellschafter:innen gebunden. Formale Vorgaben sind weniger streng. Die demokratischen Strukturen und Flexibilität der Genossenschaft können je nach Sichtweise als Vor- oder Nachteil begriffen werden. Nachteil einer Stiftung ist der Wegfall des Zugriffs auf das eingeflossene Vermögen. Der Betrieb eines wirtschaftlichen Unternehmens mit Gewinnerzielungsabsicht als eingetragener Verein ist so nicht möglich.

12.3 Public-Private-Partnership

Auch hier sind grundsätzlich verschiedene Gesellschaftsformen denkbar.

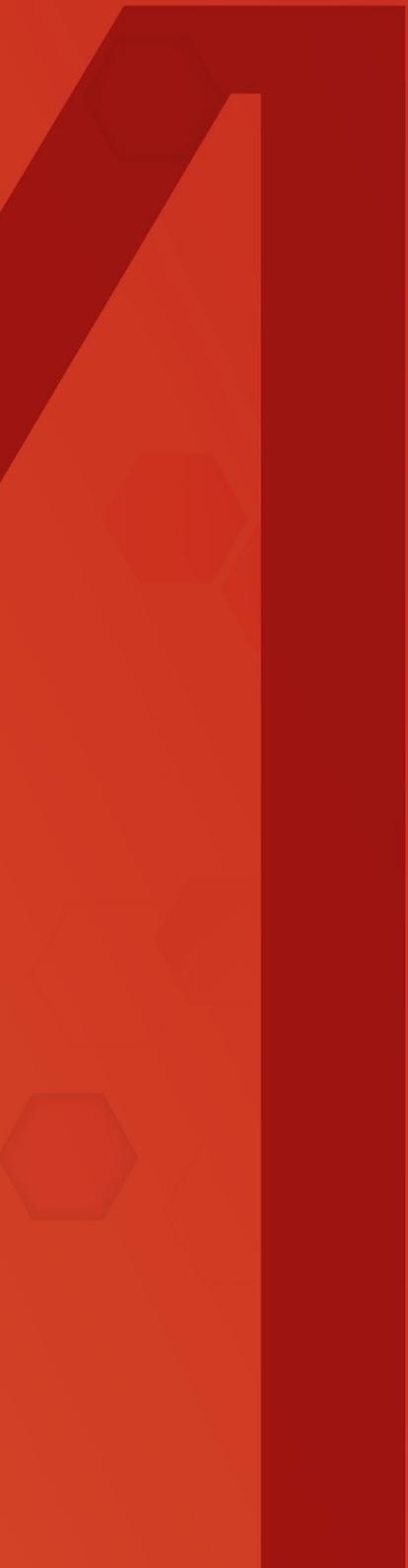
Für PPPs wird häufig die Rechtsform der GmbH gewählt. Für diese spricht, dass sie aufgrund ihrer Satzungsfreiheit und der Weisungsbindung der Geschäftsführung gute Voraussetzungen bietet, um die haushaltsrechtlichen Pflichten der öffentlichen Hand zur Einwirkung und Kontrolle gegenüber der Geschäftsführung hinreichend Rechnung zu tragen. In Form der GmbH schützt die öffentliche Hand zudem vor untragbaren Haftungsrisiken.

Der satzungsmäßigen Festlegung der öffentlichen Aufgabe kommt im gemischtwirtschaftlichen Unternehmen auch die wichtige Bedeutung zu, der Gefahr einer Anfechtung von am Gemeinwohl orientierten Gesellschafterbeschlüssen durch überstimmte Dritte wegen Treuepflichtverstoßes oder der Verfolgung von Sondervorteilen zu begegnen.

Vorteil der GmbH gegenüber der AG ist im Rahmen von Public-Private-Partnerships, dass sie mehr Steuerungsmöglichkeiten der Gesellschafter:innen zulässt. Das GmbH-Recht ist insoweit gestaltungsoffener, weil es die Regelung der gesellschaftsvertraglichen Rechtsverhältnisse weitgehend den Gesellschafter:innen überlässt.

Eine GmbH ermöglicht es bei einer PPP zudem, die Geschäftsanteile zwischen Bund, Land und Privatwirtschaft angemessen aufzuteilen und gleichzeitig der Geschäftsführung der GmbH gewisse Freiheiten und Flexibilität bei der Umsetzung der Ziele des Projekts LEAM zu gewährleisten.

Nicht ausgeschlossen ist es andererseits, gemischtwirtschaftliche Unternehmen in öffentlicher Rechtsform zu bilden. Hierbei sind allerdings erhöhte Anforderungen geboten, weil für öffentliche Rechtsformen das verfassungsrechtliche Legitimationsgebot umfassend gilt.



Szenario für ein LEAM KI-Servicezentrum



13. Szenario für ein LEAM KI-Servicezentrum

Auf Basis der vorangegangenen Ausführungen lassen sich verschiedene Szenarien für die Gestaltung und Implementierung des LKS ableiten. Im Folgenden wird eine mögliche Konzeption entwickelt und erläutert. Hierbei ist zu berücksichtigen, dass es in allen Dimensionen noch Gestaltungsoptionen und Alternativen geben kann. Diese müssen dann im weiteren Verlauf der Konzeption und Verhandlungen detaillierter spezifiziert und zur Entscheidung gebracht werden.

Die Governance-Struktur des LKS ist in Abbildung 27 schematisch dargestellt und besteht aus folgenden Komponenten:

LEAM-Institut

Das LEAM-Institut übernimmt folgende Aufgaben:

- Gesamtkoordination der LEAM-Aktivitäten und des LKS
- Steuerung der Entwicklung von KI-Foundation-Modellen
- Bereitstellung von Foundation-Modellen als Open Source
- Entwicklung und Umsetzung des Dienstes "Training-as-a-Service"

Die Finanzierung erfolgt über öffentliche Mittel unter Berücksichtigung der Beihilfe-Konformität mit Gehaltsstrukturen des öffentlichen Dienstes (TVÖD). Die Anmietung von Infrastrukturen (Housing, Office, evtl. Rechenkapazitäten) bzw. die Anschaffung eines KI-Supercomputer muss dafür öffentlich ausgeschrieben werden. Um dem EU-Beihilferecht gerecht zu werden, erfolgt die Nutzung der Kapazitäten des Supercomputers zu 80% für nicht-wirtschaftliche Zwecke - das heißt Forschung und Entwicklung durch Wirtschaft und Wissenschaft im Open-Source-Verfahren - und zu 20% für wirtschaftliche Zwecke.

LEAM-Servicegesellschaft

Die LEAM-Servicegesellschaft übernimmt folgende Aufgaben:

- Angebot eines Services für Model-Tuning mit Fokus auf Kunden aus der Wirtschaft
- Bereitstellung von Kapazitäten für den Betrieb von KI-Anwendungen (Inference)
- Bereitstellung von GPU Rechenkapazitäten
- Beratungsdienstleistungen (Consulting)

Diese Services werden in der Regel kommerziell angeboten.

Mögliche Finanzierungsquellen für die Servicegesellschaft sind:

- Finanzinvestoren
- Unternehmen bzw. Joint Ventures
- Eine Public-Private Partnership (PPP) mit Beteiligung des Bundes / eines Landes

Die Möglichkeit für eine weitere Unterstützung durch öffentliche Mittel (z.B. Anschubfinanzierung) ist im weiteren Verlauf zu prüfen und die Rechtsform sowie die kommerzielle Ausrichtung (gewinnorientiert / gemeinnützig) zu definieren.

Da die Gesellschaft nicht dem öffentlichen Dienst- und Vergütungsrecht unterliegt, kann sie marktübliche Gehälter bezahlen. Dadurch steigt die Attraktivität für hochqualifizierte Talente.

Infrastruktur

Die Bereitstellung der notwendigen Supercomputing-Infrastruktur lässt sich mit folgenden zwei alternativen Ansätzen gestalten (s. Kapitel 9.2):

- Anschaffung eines KI-Supercomputers
- Einkauf von GPU RZ-Leistungen bei einem externen Provider, der einen KI-Supercomputer bereitstellt

Housing

Die notwendige Housing Infrastruktur wird von einem externen Dienstleister angemietet (s. Kapitel 9.2).

Förderprojekt

Ein erstes Foundation-Modell wird im Rahmen eines öffentlich geförderten Ankerprojekts erstellt, an dem Wirtschaft und Wissenschaft gemeinsam arbeiten. Die Gestaltung dieses Projektes kann sich an dem bestehenden Projekt OpenGPT-X orientieren und die Ergebnisse von OpenGPT-X übernehmen und weiterentwickeln.

Im Rahmen dieses Ankerprojekts werden folgende Zielsetzungen erreicht:

- Aufbau und Pflege eines Daten-Korpus für Sprachmodelle
- Implementierung und Test der notwendigen Training-as-a-Service Prozesse
- Bereitstellung des entwickelten Foundation-Modells als Open Source für die Wirtschaft

KI-Compute-Voucher

Start-ups sollen im Rahmen der LEAM-Initiative die Möglichkeit erhalten, sich an der Entwicklung von Foundation-Modellen zu beteiligen sowie - vor allem - bereitgestellte Foundation-Modelle zu tunen und auf dieser Basis eigene KI-Anwendungen und Geschäftsmodelle zu entwickeln.

Um diese Zielsetzung zu unterstützen, stellt die öffentliche Hand sog. KI-Compute-Voucher bereit. Hiermit können (auch kleinere) Entwicklungsprojekte für Start-ups gefordert werden. Die genaue Gestaltung und der Umfang dieses Programms sind im weiteren Verlauf zu definieren.

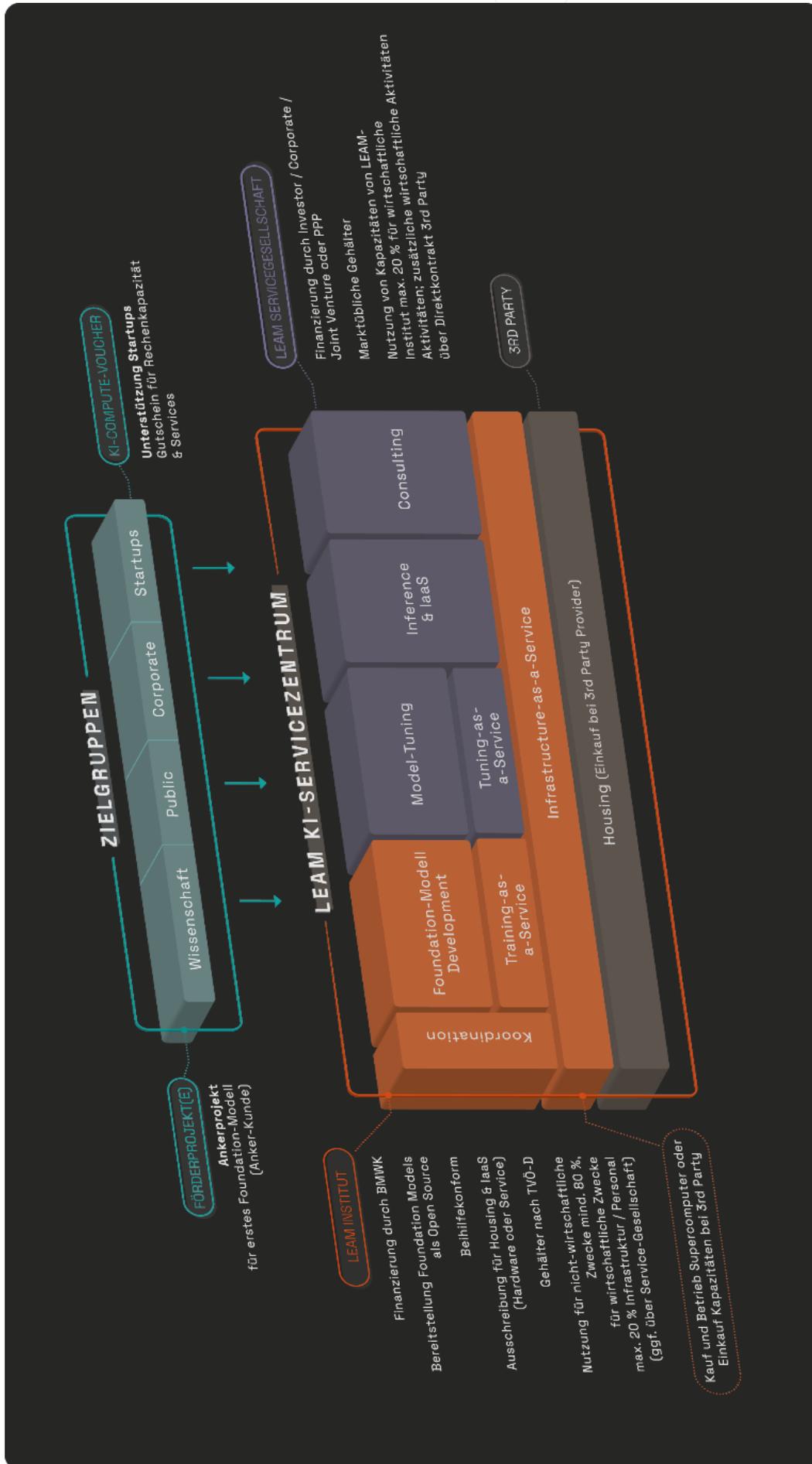


Abb. 27: Szenario für die Struktur des LEAM-KI-Servicezentrums

Fazit

LEAM Machbarkeitsstudie



14. Fazit

Die Umfragen und Interviews mit Expert:innen aus Wirtschaft und Wissenschaft zeigen ein eindeutiges Bild: KI-Foundation-Modelle stellen die nächste Entwicklung in der Erfolgsgeschichte der Künstlichen Intelligenz dar. Dabei sind die aktuell populären Sprachmodelle nur der erste Schritt. In den nächsten Jahren werden noch weitaus performantere und auf noch vielfältigere Daten trainierte Modelle den Markt weiter revolutionieren.

Die deutsche Wirtschaft hat diesen Trend erkannt und evaluiert bereits Möglichkeiten, die KI-Foundation-Modellen effektiv in internen Prozessen und als Produkte zu nutzen. Dabei ist sie aktuell aber weitgehend von proprietären, amerikanischen Foundation-Modellen abhängig. Dies stellt die Unternehmen vor große Herausforderungen in den Bereichen Datenschutz, Qualität und Zugriff auf die Modelle. Ein entscheidender Wettbewerbsnachteil gegenüber der ausländischen Konkurrenz droht. Europäischen Standards entsprechende, mit hochwertigen und vielfältigen Daten trainierte und Open Source verfügbare Foundation-Modelle würden diese Herausforderungen bewältigen und dazu beitragen, dass die deutsche Wirtschaft umfangreich von KI-Foundation-Modellen profitiert.

Dabei wurden sowohl von Wirtschaft als auch der Wissenschaft drei Kernherausforderungen identifiziert:

- (1) Es bedarf einer Vielzahl an Expert:innen für das Thema KI-Foundation-Modelle.
- (2) Diese Expert:innen benötigen Zugriff auf qualitative hochwertige Daten verschiedener Arten sowie
- (3) Zugriff auf eine hinreichend mächtige Infrastruktur, die für die aktuelle KI-Technologie und die Prozesse der KI-Entwicklung optimiert ist.

Es gilt, diese Herausforderungen in einer gemeinsamen Anstrengung der Wirtschaft, der Wissenschaft und des Staates zu lösen.

14.1 Beurteilung der Machbarkeit

Diese Studie hat die Notwendigkeit, die Chancen und den Bedarf von KI-Foundation-Modellen in Deutschland untersucht. Dabei wurde vor allem die Machbarkeit für den Aufbau und den Betrieb eines dedizierten KI-Rechenzentrums beleuchtet. Dafür wurden in dieser Studie die Bereiche Software, Hardware, bauliche Infrastruktur, Personal, Organisationsstruktur sowie Finanzierung betrachtet. Die Machbarkeit in diesen Bereichen soll hier noch einmal abschließend beurteilt werden.

Software

Die für das Training und Entwicklung der KI-Foundation-Modelle benötigten Softwareframeworks und -tools stellen keine zentrale Herausforderung dar. Die notwendigen Technologien sind bereits vorhanden und überwiegend als Open-Source-Software verfügbar.

Hardware

Es gibt aktuell in Europa kein dediziertes KI-Rechenzentrum, das für die Entwicklung international kompetitiver Foundation-Modelle ausreicht. Um zum aktuellen Stand der amerikanischen Hyperscaler aufzuschließen, müssen rund 4.500 GPU im Rechenzentrum verbaut werden. Hersteller und Collocation-Anbieter haben signalisiert, dass der Aufbau eines solchen Clusters hardwareseitig zeitnah möglich ist.

Bauliche Infrastruktur

Von dem Aufbau einer eigenen baulichen Infrastruktur sollte aufgrund der Kosten und der zu erwartenden Bauzeit abgesehen werden. Verschiedene Collocation-Anbieter haben aber bereits signalisiert, dass sie in der Lage wären, in ihrer Infrastruktur zu betreiben. Es sollte daher auf diese Möglichkeit zurückgegriffen werden.

Personal

Für den Betrieb eines KI-Rechenzentrums wird ein Team aus hochspezialisierten Expert:innen benötigt. Der Aufbau des Teams für den Aufbau und den Betrieb der Services ist kurzfristig eine Herausforderung. Mittel- und langfristig eine große Chance, um talentierte Wissenschaftler zu halten.

Organisationsstruktur

Für den Betrieb eines KI-Rechenzentrums schlagen die Autor:innen die Einrichtung einer eigenen Organisation, dem LEAM KI-Servicezentrum, vor. Dieses LKS wird den Aufbau der Infrastruktur begleiten und den identifizierten Zielgruppen spezialisierte Services bereitstellen. Der Fokus soll darauf liegen, der Wirtschaft KI-Foundation-Modelle für die Entwicklung von Anwendungen bereitzustellen.

Finanzierung

Die Autor:innen kalkulieren für den Aufbau und Betrieb eines KI-Rechenzentrums über vier Jahre einen Bedarf von rund 380 Millionen Euro. Für die Finanzierung wurde ein Modell entwickelt, das öffentliche und privatwirtschaftliche Mittel berücksichtigt.

Zusammenfassend lässt sich festhalten, dass der Aufbau von LEAM-Infrastruktur und -Services für die deutsche Wirtschaft ein entscheidender Wirtschaftsfaktor ist. Die Umsetzung ist mit der Beteiligung der öffentlichen Hand, der Forschung und Wissenschaft sowie der Wirtschaft realisierbar.

14.2 Ausblick

In engem Austausch mit der Politik, der Wirtschaft, potentiellen Standorten, Anbieter:innen und Anwender:innen sollte das Thema nun weiter vorangetrieben werden. Dabei gilt es vor allem, die Finanzierungsmöglichkeiten weiter zu präzisieren.

Aktuelle Entwicklungen wie der Hype um ChatGPT zeigen, wie dynamisch sich die Forschung und Anwendung von Foundation-Modellen entwickelt. Die Gefahr, dass sich in den USA monopolartige Strukturen - ähnlich Google bei Suchmaschinen - bilden, wird fortlaufend größer. Damit Deutschland nicht weiter zurückfällt, ist eine zeitnahe Umsetzung des Konzeptes für die LEAM-Infrastruktur nötig. Das Momentum, das im Rahmen der Erstellung dieser Machbarkeitsstudie gewonnen wurde - durch eine enge Zusammenarbeit zwischen Forschung, Wissenschaft, Wirtschaft und Start-ups - sollte genutzt werden, um die Zielsetzung weiter voranzutreiben.

Eine Möglichkeit ist die zeitnahe Gründung einer Projektentwicklungsgesellschaft, welche die Grundlagen für den Aufbau dieses strategischen KI-Leuchtturmprojekts erarbeitet.

Verzeichnisse & Methodik der LEAM Machbarkeitsstudie



I. Quellenverzeichnis

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, 265–283. <https://dl.acm.org/doi/10.5555/3026877.3026899>
- Abseil Python Common Libraries. (2022). [Python]. Abseil. <https://github.com/abseil/abseil-py> (Original work published 2017)
- Agarwal, O., Ge, H., Shakeri, S., & Al-Rfou, R. (2021). *Knowledge Graph Based Synthetic Corpus Generation for Knowledge-Enhanced Language Model Pre-training* (arXiv:2010.12688). arXiv. <https://doi.org/10.48550/arXiv.2010.12688>
- AI accelerator. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=AI_accelerator&oldid=1123373022
- Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., ... Simonyan, K. (2022). Flamingo: A Visual Language Model for Few-Shot Learning. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in Neural Information Processing Systems*. <https://openreview.net/forum?id=EbMuimAbPbs>
- AMD. (o.D.). *AMD ROCm™ Open Ecosystem*. Retrieved 28 November 2022, from <https://www.amd.com/en/graphics/servers-solutions-rocm>
- AMD. (2021, August 11). *New AMD Instinct™ MI200 Series Accelerators Bring Leadership HPC and AI Performance to Power Exascale Systems and More*. <https://www.amd.com/en/press-releases/2021-11-08-new-amd-instinct-mi200-series-accelerators-bring-leadership-hpc-and-ai>
- An updated set of basic linear algebra subprograms (BLAS). (2002). *ACM Transactions on Mathematical Software*, 28(2), 135–151. <https://doi.org/10.1145/567806.567807>
- An, W., Guo, Y., Bian, Y., Ma, H., Yang, J., Li, C., & Huang, J. (2022). MoDNA: Motif-oriented pre-training for DNA language model. *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 1–5. <https://doi.org/10.1145/3535508.3545512>
- Apache Flink. (2022). [Java]. The Apache Software Foundation. <https://github.com/apache/flink> (Original work published 2014)
- Apache Hadoop. (2022). [Java]. The Apache Software Foundation. <https://github.com/apache/hadoop> (Original work published 2014)
- Apache Spark. (2022). [Scala]. The Apache Software Foundation. <https://github.com/apache/spark> (Original work published 2014)
- Arakelyan, G., Soghomonyan, G., & The Aim team. (2020). *Aim* (3.9.3) [TypeScript]. <https://doi.org/10.5281/zenodo.6536395>
- ASHRAE. (2022). *Data Center Power Equipment Thermal Guidelines and Best Practices*. ASHRAE Technical Committee (TC) 9.9 Mission Critical Facilities, Data Centers, Technology Spaces, and Electronic Equipment.
- Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., ... Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557), 871–876. <https://doi.org/10.1126/science.abj8754>
- Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 12449–12460.
- Bannour, N., Ghannay, S., Névél, A., & Ligozat, A.-L. (2021). *Evaluating the carbon footprint of NLP methods: A survey and analysis of existing tools*. 11–21. <https://doi.org/10.18653/v1/2021.sustainlp-1.2>
- Basic Linear Algebra Subprograms. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Basic_Linear_Algebra_Subprograms&oldid=1120747813
- Benaïc, N., & Hogarth, I. (2022). *State of AI 2022*. <https://www.stateof.ai>

- Bitkom e.V. (2022a, January 3). *IT-Fachkräftelücke wird größer: 96.000 offene Jobs* | Bitkom e.V. <https://www.bitkom.org/Presse/Presseinformation/IT-Fachkraefteluecke-wird-groesser>
- Bitkom e.V. (2022b). *Bitkom Positionspapier: Perspektiven für eine nachhaltige Rechenzentren-Wirtschaft bis 2030*. <https://www.bitkom.org/sites/main/files/2022-10/2210-Positionspapier-Nachhaltige-Rechenzentren.pdf>
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). *On the Opportunities and Risks of Foundation Models* (arXiv:2108.07258). arXiv. <https://doi.org/10.48550/arXiv.2108.07258>
- Boroditsky, L. (2012, March 15). *Linguistik: Wie die Sprache das Denken formt*. Spektrum.de. <https://www.spektrum.de/news/linguistik-wie-die-sprache-das-denken-formt/1145804>
- Branch, H. J., Cefalu, J. R., McHugh, J., Hujer, L., Bahl, A., Iglesias, D. del C., Heichman, R., & Darwishi, R. (2022). *Evaluating the Susceptibility of Pre-Trained Language Models via Handcrafted Adversarial Examples* (arXiv:2209.02128). arXiv. <https://doi.org/10.48550/arXiv.2209.02128>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901. <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>
- Brundage, M., Mayer, K., Eloundou, T., Agarwal, S., Adler, S., Krueger, G., Leike, J., & Mishkin, P. (2022). Lessons learned on language model safety and misuse. *OpenAI*. <https://openai.com/blog/language-model-safety-and-misuse/>
- Bundesministerium für Digitales und Verkehr. (2022). *Digitalstrategie Deutschland*. <https://digitalstrategie-deutschland.de/medien/>
- Campa, C., Kawalek, C., Vo, H., & Bessoudo, J. (2020, May 14). *Defining AI Innovation with NVIDIA DGX A100*. NVIDIA Technical Blog. <https://developer.nvidia.com/blog/defining-ai-innovation-with-dgx-a100/>
- Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., & Erlingsson, U. (2021). Extracting training data from large language models. *30th USENIX Security Symposium (USENIX Security 21)*, 2633–2650. <https://www.usenix.org/system/files/sec21-carlini-extracting.pdf>
- Carpintero, A. G. (2021, November 21). *MLOps with Docker and Jenkins: Automating Machine Learning Pipelines*. Medium. <https://towardsdatascience.com/mlops-with-docker-and-jenkins-automating-machine-learning-pipelines-a3a4026c4487>
- Cen, S., & Shah, D. (2021). Regulating algorithmic filtering on social media. *Advances in Neural Information Processing Systems*, 34, 6997–7011. <https://proceedings.neurips.cc/paper/2021/hash/38b4f06e27fd4f6fdcceabc6f5c068ea-Abstract.html>
- Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., & Mordatch, I. (2021). Decision Transformer: Reinforcement Learning via Sequence Modeling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, & J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems* (Vol. 34, pp. 15084–15097). Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2021/file/7f489f642a0ddb10272b5c31057f0663-Paper.pdf>
- Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., & others. (2021). Evaluating large language models trained on code. *ArXiv Preprint ArXiv:2107.03374*.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., ... Fiedel, N. (2022). *PaLM: Scaling Language Modeling with Pathways* (arXiv:2204.02311). arXiv. <https://doi.org/10.48550/arXiv.2204.02311>
- Chowdhury, R., Bouatta, N., Biswas, S., Floristean, C., Kharkar, A., Roy, K., Rochereau, C., Ahdritz, G., Zhang, J., Church, G. M., Sorger, P. K., & AlQuraishi, M. (2022). Single-sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology*, 40(11), Article 11. <https://doi.org/10.1038/s41587-022-01432-w>
- Climate Neutral Data Centre Pact – The Green Deal need Green Infrastructure*. (o.D.). Retrieved 19 December 2022, from <https://www.climateneutraldatacentre.net/>
- CodeCarbon. (2020). *CodeCarbon—CodeCarbon 2.0.0 documentation*. <https://mlco2.github.io/codecarbon/>

- Creemers, A. B., Englander, A., Gabriel, M., Hecker, D., Mock, M., Poretschkin, M., Julia Rosenzweig, J., Rostalski, F., Volmer, J., & Voosholz, J. (2019). Vertrauenswürdiger Einsatz von Künstlicher Intelligenz. Handlungsfelder aus philosophischer, ethischer, rechtlicher und technologischer Sicht als Grundlage für eine Zertifizierung von Künstlicher Intelligenz. *Fraunhofer-Institut Für Intelligente Analyse-Und Informationssysteme (IAIS)*. https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper_KI-Zertifizierung.pdf
- CSTB Releases Report *Fostering Responsible Computing Research: Foundations and Practices* » *CCC Blog*. (2022, May 16). <https://cccblog.org/2022/05/16/cstb-releases-report-fostering-computing-research-foundations-and-practices/>
- Dask. (2022). [Python]. dask. <https://github.com/dask/dask> (Original work published 2015)
- Data protection in the EU. (o.D.). [Text]. European Commission - European Commission. Retrieved 28 November 2022, from https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en
- Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Large Clusters. *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, 137–150.
- Deep Lake. (2022). [Python]. Activeloop. <https://github.com/activeloopai/deeplake> (Original work published 2019)
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. <https://doi.org/10.48550/arXiv.1810.04805>
- Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., & Gardner, M. (2021). Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1286–1305. <https://doi.org/10.18653/v1/2021.emnlp-main.98>
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2022, March 23). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. International Conference on Learning Representations. <https://openreview.net/forum?id=YicbFdNTTy>
- Dürr, B. (2018). *IT-Räume und Rechenzentren planen und betreiben: Handbuch der baulichen Maßnahmen und Technischen Gebäudeausrüstung* (Vol. 2). Verlag Bau+Technik.
- ELE Consortium. (2022). *Digital Language Equality in Europe by 2030: Strategic Agenda and Roadmap*. <https://european-language-equality.eu/agenda/>
- Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., & Joulin, A. (2022). Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1), 4839–4886.
- FFCV. (2022). [Python]. FFCV. <https://github.com/libffcv/ffcv> (Original work published 2021)
- Frostig, R., Johnson, M., & Leary, C. (2018). *Compiling machine learning programs via high-level tracing*. <https://mlsys.org/Conferences/doc/2018/146.pdf>
- Gehlhaus, D., & Koslosky, L. (2022). Training Tomorrow's AI Workforce. *Center for Security and Emerging Technology*. <https://cset.georgetown.edu/publication/training-tomorrows-ai-workforce/>
- Gehlhaus, D., Koslosky, L., Goode, K., & Perkins, C. (2021). U.S. AI Workforce: Policy Recommendations. *Center for Security and Emerging Technology*. <https://cset.georgetown.edu/publication/u-s-ai-workforce-policy-recommendations/>
- Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). *RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models* (arXiv:2009.11462). arXiv. <https://doi.org/10.48550/arXiv.2009.11462>
- Glaese, A., McAleese, N., Trębacz, M., Aslanides, J., Firoiu, V., Ewalds, T., Rauh, M., Weidinger, L., Chadwick, M., Thacker, P., Campbell-Gillingham, L., Uesato, J., Huang, P.-S., Comanescu, R., Yang, F., See, A., Dathathri, S., Greig, R., Chen, C., ... Irving, G. (2022). *Improving alignment of dialogue agents via targeted human judgements* (arXiv:2209.14375). arXiv. <https://doi.org/10.48550/arXiv.2209.14375>
- GlusterFS. (2022). [C]. Gluster.org. <https://github.com/gluster/glusterfs> (Original work published 2011)
- Gopani, A. (2021, July 16). *JAX Vs TensorFlow Vs PyTorch: A Comparative Analysis*. Analytics India Magazine. <https://analyticsindiamag.com/jax-vs-tensorflow-vs-pytorch-a-comparative-analysis/>
- Hansell, S. (2002, April 8). Google's Toughest Search Is for a Business Model. *The New York Times*. <https://www.nytimes.com/2002/04/08/business/google-s-toughest-search-is-for-a-business-model.html>

- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), Article 7825. <https://doi.org/10.1038/s41586-020-2649-2>
- Hensel, M., & Ostler, U. (2020, November 22). Die beliebtesten Anbieter von Technik für das High-Performance Computing. *Datacenter Insider*. <https://www.datacenter-insider.de/die-beliebtesten-anbieter-vontechnik-fuer-das-high-performance-computing-a-980532/>
- Hickmann, H., & Koneberg, F. (2022). *Die Berufe mit den aktuell größten Fachkräftelücken*. 67. <https://www.iwkoeln.de/studien/helen-hickmann-filiz-koneberg-die-berufe-mit-den-aktuell-groessten-fachkraefteluecken.html>
- Hintemann, Dr. R., Hinterholzer, S., Graß, M., & Grothey, T. (2022). *Bitkom-Studie: Rechenzentren in Deutschland 2021 – Aktuelle Marktentwicklungen*. Borderstep Institut. <https://www.bitkom.org/sites/main/files/2022-02/10.02.22-studie-rechenzentren.pdf>
- Hintemann, Dr. R., Hinterholzer, S., & Grothey, T. (2021). *Herausforderungen und Chancen durch den Boom beim Neubau von Rechenzentren*. Hessische Staatskanzlei, Ministerin für Digitale Strategie und Entwicklung. https://digitales.hessen.de/sites/digitales.hessen.de/files/2022-05/rechenzentrumsmarkt_hessen.pdf
- Hintemann, R. (2020). *Data centers 2018. Efficiency gains are not enough: Data center energy consumption continues to rise significantly - Cloud computing boosts growth*. <https://doi.org/10.13140/RG.2.2.26033.40800>
- Hintemann, R., & Clausen, J. (2018). *Bedeutung digitaler Infrastrukturen in Deutschland*.
- Hinton, G., Vinyals, O., & Dean, J. (2015). *Distilling the Knowledge in a Neural Network* (arXiv:1503.02531). arXiv. <https://doi.org/10.48550/arXiv.1503.02531>
- Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., & Salimans, T. (2022). *Imagen Video: High Definition Video Generation with Diffusion Models* (arXiv:2210.02303). arXiv. <https://doi.org/10.48550/arXiv.2210.02303>
- Hugging Face. (o.D.). *Model Parallelism*. Retrieved 28 November 2022, from <https://huggingface.co/docs/transformers/v4.15.0/parallelism>
- Hydra. (2022). [Python]. Meta Research. <https://github.com/facebookresearch/hydra> (Original work published 2019)
- Intel. (2022, June 29). *Second-Gen Habana Gaudi2 Outperforms Nvidia A100*. Intel. <https://www.intel.com/content/www/us/en/newsroom/news/second-gen-habana-gaudi2-outperforms-nvidia-a100.html>
- ISO - ISO/IEC 27001 and related standards—Information security management. (o.D.). ISO. Retrieved 28 November 2022, from <https://www.iso.org/isoiec-27001-information-security.html>
- ISO 27017 and ISO 27018 Certification | DEKRA. (o.D.). Retrieved 6 December 2022, from <https://www.dekra.com/en/iso-27017-and-iso-27018-certification/>
- Izcard, G., & Grave, E. (2021). *Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering* (arXiv:2007.01282). arXiv. <https://doi.org/10.48550/arXiv.2007.01282>
- Jawahar, R. (2021, October 14). *Teaching AI to perceive the world through your eyes*. Meta AI. <https://ai.facebook.com/blog/teaching-ai-to-perceive-the-world-through-your-eyes/>
- Joseph, E., Riddle, M., Sorensen, T., & Conway, S. (2022). *The Economic and Societal Benefits of Linux Supercomputers*. <https://davidbader.net/publication/2022-hyperionresearch/>
- Jülich Forschungszentrum. (2022, June 15). *Erster europäischer Exascale-Superrechner kommt nach Jülich*. <https://www.fz-juelich.de/de/aktuelles/news/pressemitteilungen/2022/exascale-standort-entscheidung>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), Article 7873. <https://doi.org/10.1038/s41586-021-03819-2>
- Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
- Kale, M., Siddhant, A., Al-Rfou, R., Xue, L., Constant, N., & Johnson, M. (2021). nmT5—Is parallel data still relevant for pre-training massively multilingual language models? *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on*

- Natural Language Processing (Volume 2: Short Papers)*, 683–691. <https://doi.org/10.18653/v1/2021.acl-short.87>
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling Laws for Neural Language Models* (arXiv:2001.08361). arXiv. <https://doi.org/10.48550/arXiv.2001.08361>
- Keras. (2022). [Python]. Keras. <https://github.com/keras-team/keras> (Original work published 2015)
- Khan, S. M., & Mann, A. (2020, April). AI Chips: What They Are and Why They Matter. *Center for Security and Emerging Technology*. <https://cset.georgetown.edu/publication/ai-chips-what-they-are-and-why-they-matter/>
- Kosson, A., Chiley, V., Venigalla, A., Hestness, J., & Köster, U. (2021). *Pipelined Backpropagation at Scale: Training Large Models without Batches* (arXiv:2003.11666). arXiv. <https://doi.org/10.48550/arXiv.2003.11666>
- Kubeflow. (2022). [Jsonnet]. Kubeflow. <https://github.com/kubeflow/kubeflow> (Original work published 2017)
- Kubernetes (K8s). (2022). [Go]. Kubernetes. <https://github.com/kubernetes/kubernetes> (Original work published 2014)
- Lamonica, M. (2014, June 11). *HP's Water-Cooled Supercomputer is Designed for the Hydrophobic*. IEEE Spectrum. <https://spectrum.ieee.org/a-watercooled-supercomputer-for-the-hydrophobic->
- Li, C. (2020, June 3). *OpenAI's GPT-3 Language Model: A Technical Overview*. <https://lambdalabs.com/blog/demystifying-gpt-3>
- Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., ... Koreeda, Y. (2022). *Holistic Evaluation of Language Models* (arXiv:2211.09110). arXiv. <https://doi.org/10.48550/arXiv.2211.09110>
- Lin, S., Hilton, J., & Evans, O. (2021). *TruthfulQA: Measuring How Models Mimic Human Falsehoods* (arXiv:2109.07958). arXiv. <https://doi.org/10.48550/arXiv.2109.07958>
- Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Costa, A. dos S., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2022). *Evolutionary-scale prediction of atomic level protein structure with a language model* (p. 2022.07.20.500902). bioRxiv. <https://doi.org/10.1101/2022.07.20.500902>
- Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8, 726–742. https://doi.org/10.1162/tacl_a_00343
- Liu, Y., Liu, P., Radev, D., & Neubig, G. (2022). BRIO: Bringing Order to Abstractive Summarization. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2890–2903. <https://doi.org/10.18653/v1/2022.acl-long.207>
- Maszke, S. (2022). *Torchdatasets* [Python]. <https://github.com/szymonmaszke/torchdatasets> (Original work published 2019)
- Merkel, D. (2014). Docker: Lightweight Linux containers for consistent development and deployment. *Linux Journal*, 2014(239), 2:2.
- Message Passing Interface. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Message_Passing_Interface&oldid=1112449606
- Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., & Sivic, J. (2019). HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2630–2640. <https://doi.org/10.1109/ICCV.2019.00272>
- Mishkin, P., Ahmad, L., Brundage, M., Krueger, G., & Sastry, G. (2022). *DALL·E 2 Preview—Risks and Limitations*. <https://github.com/openai/dalle-2-preview/blob/main/system-card.md> (<https://github.com/openai/dalle-2-preview/blob/main/system-card.md>)
- MLflow: A Machine Learning Lifecycle Platform. (2022). [Python]. MLflow. <https://github.com/mlflow/mlflow> (Original work published 2018)
- Mo, S., Fu, X., Hong, C., Chen, Y., Zheng, Y., Tang, X., Lan, Y., Shen, Z., & Xing, E. (2021, September 24). *Multi-modal Self-supervised Pre-training for Large-scale Genome Data*. NeurIPS 2021 AI for Science Workshop. <https://openreview.net/forum?id=fdV-GZ4LPfn>
- Moritz, P., Nishihara, R., Wang, S., Tumanov, A., Liaw, R., Liang, E., Elibol, M., Yang, Z., Paul, W., Jordan, M. I., & Stoica, I. (2018). *Ray: A Distributed Framework for Emerging AI Applications* (arXiv:1712.05889). arXiv. <https://doi.org/10.48550/arXiv.1712.05889>

- Mudigere, D., Hao, Y., Huang, J., Jia, Z., Tulloch, A., Sridharan, S., Liu, X., Ozdal, M., Nie, J., Park, J., Luo, L., Yang, J. A., Gao, L., Ivchenko, D., Basant, A., Hu, Y., Yang, J., Ardestani, E. K., Wang, X., ... Rao, V. (2022). *Software-Hardware Co-design for Fast and Scalable Training of Deep Learning Recommendation Models* (arXiv:2104.05158). arXiv. <https://doi.org/10.48550/arXiv.2104.05158>
- Mujkanovic, N., Sivalingam, K., & Lazzaro, A. (2020). *Optimising AI Training Deployments using Graph Compilers and Containers* (arXiv:2008.11675). arXiv. <https://doi.org/10.48550/arXiv.2008.11675>
- Nagrani, A., Seo, P. H., Seybold, B., Hauth, A., Manen, S., Sun, C., & Schmid, C. (2022). Learning Audio-Video Modalities from Image Captions. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), *Computer Vision – ECCV 2022* (pp. 407–426). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19781-9_24
- Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., & Schulman, J. (2021). *WebGPT: Browser-assisted question-answering with human feedback* (arXiv:2112.09332). arXiv. <https://doi.org/10.48550/arXiv.2112.09332>
- Narayanan, D., Shoeybi, M., Casper, J., LeGresley, P., Patwary, M., Korthikanti, V. A., Vainbrand, D., Kashinkunti, P., Bernauer, J., Catanzaro, B., Phanishayee, A., & Zaharia, M. (2021). *Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM* (arXiv:2104.04473). arXiv. <https://doi.org/10.48550/arXiv.2104.04473>
- Nguyen, T. T., Trahay, F., Domke, J., Drozd, A., Vatai, E., Liao, J., Wahib, M., & Gerofi, B. (2022). Why Globally Re-shuffle? Revisiting Data Shuffling in Large Scale Deep Learning. *2022 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 1085–1096. <https://doi.org/10.1109/IPDPS53621.2022.00109>
- Nichol, A. (2022, June 28). DALL·E 2 Pre-Training Mitigations. *OpenAI*. <https://openai.com/blog/dall-e-2-pre-training-mitigations/>
- NVIDIA. (o.D.). *NCCL and MPI — NCCL 2.15.5 documentation*. Retrieved 28 November 2022, from <https://docs.nvidia.com/deeplearning/nccl/user-guide/docs/mpi.html>
- NVIDIA Developer. (2013, July 2). *CUDA Toolkit—Free Tools and Training*. NVIDIA Developer. <https://developer.nvidia.com/cuda-toolkit>
- Ofeidis, I., Kiedanski, D., & Tassioulas, L. (2022). *An Overview of the Data-Loader Landscape: Comparative Performance Analysis* (arXiv:2209.13705). arXiv. <https://doi.org/10.48550/arXiv.2209.13705>
- Ofer, D., Brandes, N., & Linial, M. (2021). The language of proteins: NLP, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, 19, 1750–1758. <https://doi.org/10.1016/j.csbj.2021.03.022>
- Open MPI. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Open_MPI&oldid=1120683830
- OpenAI. (2022a, July 18). Reducing Bias and Improving Safety in DALL·E 2. *OpenAI*. <https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2/>
- OpenAI. (2022b, November 30). *ChatGPT: Optimizing Language Models for Dialogue*. OpenAI. <https://openai.com/blog/chatgpt/>
- Paaß, G., & Giesselbach, S. (2023). *Foundation Models for Natural Language Processing*. Springer Cham. <https://link.springer.com/book/9783031231896>
- Papers with Code—Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. (o.D.). Retrieved 10 January 2023, from <https://paperswithcode.com/paper/exploring-the-limits-of-transfer-learning>
- Parliament, E. (2018). *European Parliament resolution of 11 September 2018 on language equality in the digital age (2018/2028(INI))*. https://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_DE.html
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library* (arXiv:1912.01703). arXiv. <https://doi.org/10.48550/arXiv.1912.01703>
- Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2022). The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. *Computer*, 55(7), 18–28. <https://doi.org/10.1109/MC.2022.3148714>
- Philippe Lorenz & Kate Saslow. (2019). *Demystifying AI & AI Companies. What foreign policy makers need to know about the global AI industry*. <https://www.stiftung-nv.de/de/publikation/demystifying-ai-ai-companies-what-foreign-policy-makers-need-know-about-global-ai>

- Piloto, L. S., Weinstein, A., Battaglia, P., & Botvinick, M. (2022). Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, 6(9), Article 9. <https://doi.org/10.1038/s41562-022-01394-8>
- Poretschkin, M. (2022). ZERTIFIZIERTE KI | Qualität sichern. Fortschritt gestalten. ZERTIFIZIERTE KI. <https://www.zertifizierte-ki.de/>
- Poretschkin, M., Mock, M., & Wrobel, S. (2021). Zur Systematischen Bewertung der Vertrauenswürdigkeit von KI-Systemen. *Regulierung Für Algorithmen Und Künstliche Intelligenz*, 175–202. <https://doi.org/10.5771/9783748927990>
- Pushkarna, M., Zaldivar, A., & Kjartansson, O. (2022). *Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI* (arXiv:2204.01075). arXiv. <https://doi.org/10.48550/arXiv.2204.01075>
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763. <https://proceedings.mlr.press/v139/radford21a.html>
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1–67.
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical Text-Conditional Image Generation with CLIP Latents* (arXiv:2204.06125). arXiv. <https://doi.org/10.48550/arXiv.2204.06125>
- Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J. T., Eccles, T., Bruce, J., Razavi, A., Edwards, A., Heess, N., Chen, Y., Hadsell, R., Vinyals, O., Bordbar, M., & de Freitas, N. (2022). *A Generalist Agent* (arXiv:2205.06175). arXiv. <https://doi.org/10.48550/arXiv.2205.06175>
- Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2022). *FastSpeech 2: Fast and High-Quality End-to-End Text to Speech* (arXiv:2006.04558). arXiv. <https://doi.org/10.48550/arXiv.2006.04558>
- Research and Markets Ltd. (2021, June). *Europe Data Center Colocation Market: Market Size, Forecast, Insights, and Competitive Landscape*. <https://www.researchandmarkets.com/reports/5511065/europe-data-center-colocation-market-market>
- Reuther, A., Michaleas, P., Jones, M., Gadepally, V., Samsi, S., & Kepner, J. (2022). *AI and ML Accelerator Survey and Trends* (arXiv:2210.04055). arXiv. <https://doi.org/10.48550/arXiv.2210.04055>
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). *High-Resolution Image Synthesis with Latent Diffusion Models* (arXiv:2112.10752). arXiv. <https://doi.org/10.48550/arXiv.2112.10752>
- Schödwell, B., Zarnekow, D. R., Liu, R., Gröger, J., & Wilkens, M. (2018). *Kennzahlen und Indikatoren für die Beurteilung der Ressourceneffizienz von Rechenzentren und Prüfung der praktischen Anwendbarkeit*.
- Schreiner, M. (2022, January 29). *Meta's AI chief: Three major challenges of artificial intelligence*. THE DECODER. <https://the-decoder.com/metas-ai-chief-three-major-challenges-of-artificial-intelligence/>
- Schuhmann, C. (2021, August 8). *LAION-400-MILLION OPEN DATASET | LAION*. LAION. <https://laion.ai/blog/laion-400-open-dataset>
- Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). *Compute Trends Across Three Eras of Machine Learning* (arXiv:2202.05924). arXiv. <https://doi.org/10.48550/arXiv.2202.05924>
- Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). *Retrieval Augmentation Reduces Hallucination in Conversation* (arXiv:2104.07567). arXiv. <https://doi.org/10.48550/arXiv.2104.07567>
- Simons, G. J., & Frese, A. (2021). *Zukunft regional – digital: Das Rheinische Revier; Machbarkeitsstudie Dateninfrastrukturen im Rheinischen Revier*. Ministerium für Wirtschaft, Innovation, Digitalisierung und Energie des Landes Nordrhein-Westfalen. https://www.wirtschaft.nrw/sites/default/files/documents/machbarkeitsstudie_dateninfrastrukturen_lang_de.pdf
- Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., & Taigman, Y. (2022). *Make-A-Video: Text-to-Video Generation without Text-Video Data* (arXiv:2209.14792). arXiv. <https://doi.org/10.48550/arXiv.2209.14792>
- Sohofi, A., Yu, T., Aribal, A., Loetsch, W., Team, S. D., & Wollmann, T. (2022). *Squirrel* [Python]. <https://github.com/merantix-momentum/squirrel-core> (Original work published 2022)
- Sonnenburg, S., Braun, M. L., Ong, C. S., Bengio, S., Bottou, L., Holmes, G., LeCun, Y., Müller, K.-R., Pereira, F., Rasmussen, C. E., & Schölkopf, B. (2007). The Need for Open Source Software in Machine Learning. *Journal of Machine Learning Research*, 8(81), 2443–2466.

- Stobbe, Dr. L., Proske, M., Zedel, H., Hintemann, Dr. R., Clausen, Dr. J., & Beucker, Dr. S. (2015). *Entwicklung des IKT-bedingten Strombedarfs in Deutschland* [Abschlussbericht]. Fraunhofer-Institut für Zuverlässigkeit und Mikrointegration.
https://www.bmwk.de/Redaktion/DE/Downloads/E/entwicklung-des-ikt-bedingten-strombedarfs-in-deutschland-abschlussbericht.pdf?__blob=publicationFile&v=3
- Stöcker, C., & Dambeck, H. (2006, December 19). Deutsch-französische Suchmaschine: Quaero ist geplatzt. *Der Spiegel*. <https://www.spiegel.de/netzwelt/web/deutsch-franzoesische-suchmaschine-quaero-ist-geplatzt-a-455558.html>
- Streim, A. (2022, November 16). *Trotz Krieg und Krisen: In Deutschland fehlen 137.000 IT-Fachkräfte* | *Presseinformation* | Bitkom e.V. <https://www.bitkom.org/Presse/Presseinformation/Deutschland-fehlen-137000-IT-Fachkraefte>
- Suzgun, M., Scales, N., Schärli, N., Gehrman, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., & Wei, J. (2022). *Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them* (arXiv:2210.09261). arXiv. <https://doi.org/10.48550/arXiv.2210.09261>
- System and Organization Controls: SOC Suite of Services*. (o.D.). AICPA. Retrieved 28 November 2022, from <https://www.aicpa.org/interestareas/frc/assuranceadvisoryservices/sorhome.html>
- Tachyum. (2022, October 4). *Tachyum Unveils Details of Architecture and Prodigy Design in Overview White Paper* | Tachyum. <https://www.tachyum.com/media/press-releases/2022/10/04/tachyum-unveils-details-of-architecture-and-prodigy-design-in-overview-white-paper/>
- Tamkin, A., & Ganguli, D. (2021, February 5). *How Large Language Models Will Transform Science, Society, and AI*. Stanford HAI. <https://hai.stanford.edu/news/how-large-language-models-will-transform-science-society-and-ai>
- Terraform*. (2022). [Go]. HashiCorp. <https://github.com/hashicorp/terraform> (Original work published 2014)
- THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH RESOURCE TASK FORCE (NAIRRTF)*. (o.D.). National Artificial Intelligence Initiative. Retrieved 10 January 2023, from <https://www.ai.gov/nairrtf/>
- The role of data centers in an interconnected world—DECIX – Without You. (o.D.). *DECIX – Without You*. Retrieved 19 December 2022, from <https://withoutyou.de-cix.net/the-role-of-data-centers/>
- Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., ... Le, Q. (2022). *LaMDA: Language Models for Dialog Applications* (arXiv:2201.08239). arXiv. <https://doi.org/10.48550/arXiv.2201.08239>
- Ung, M., Xu, J., & Boureau, Y.-L. (2022). SaFeRDialogues: Taking Feedback Gracefully after Conversational Safety Failures. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6462–6481. <https://doi.org/10.18653/v1/2022.acl-long.447>
- van Rossum, G. (1995). *Python reference manual* (R 9525). Article R 9525. <https://ir.cwi.nl/pub/5008>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). Curran Associates, Inc.
<https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>
- Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., & Yang, H. (2022). *OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework* (arXiv:2202.03052). arXiv. <https://doi.org/10.48550/arXiv.2202.03052>
- Wannemacher, K., & Bodmann, L. (2021). *Künstliche Intelligenz an den Hochschulen. Potenziale und Herausforderungen in Forschung, Studium und Lehre sowie Curriculumentwicklung: Arbeitspapier Nr. 59*. Berlin: Hochschulforum Digitalisierung.
https://hochschulforumdigitalisierung.de/sites/default/files/dateien/HFD_AP_59_Kuenstliche_Intelligenz_Hochschulen_HIS-HE.pdf
- Weidmann, Dr. R. E., & Krüger, Dr. T. (2020, November 30). *Dr. Béla Waldhauser, Telehouse Deutschland: Digitale Effizienz in Rechenzentren*. <https://detecon.com/de/journal/dr-bela-waldhauser-telehouse-deutschland-digitale-effizienz-rechenzentren>
- Wikipedia contributors. (2022). *Selene (supercomputer)*—Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/w/index.php?title=Selene_\(supercomputer\)&oldid=1109224992](https://en.wikipedia.org/w/index.php?title=Selene_(supercomputer)&oldid=1109224992)
- Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., & Duan, N. (2022). NÜWA: Visual Synthesis Pre-training for Neural visUal World creAtion. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.),

- Computer Vision – ECCV 2022* (pp. 720–736). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19787-1_41
- Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). *mT5: A massively multilingual pre-trained text-to-text transformer* (arXiv:2010.11934). arXiv. <https://doi.org/10.48550/arXiv.2010.11934>
- Yazdani-Jahromi, M., Yousefi, N., Tayebi, A., Kolanthai, E., Neal, C. J., Seal, S., & Garibay, O. O. (2022). AttentionSiteDTI: an interpretable graph-based model for drug-target interaction prediction using NLP sentence-level relation classification. *Briefings in Bioinformatics*, 23(4). <https://doi.org/10.1093/bib/bbac272>
- Yin, P., Neubig, G., Yih, W., & Riedel, S. (2020). TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8413–8426. <https://doi.org/10.18653/v1/2020.acl-main.745>
- Yoo, A. B., Jette, M. A., & Grondona, M. (2003). SLURM: Simple Linux Utility for Resource Management. In D. Feitelson, L. Rudolph, & U. Schwiegelshohn (Eds.), *Job Scheduling Strategies for Parallel Processing* (pp. 44–60). Springer. https://doi.org/10.1007/10968987_3
- Yuan, A., Coenen, A., Reif, E., & Ippolito, D. (2022). Wordcraft: Story Writing With Large Language Models. *27th International Conference on Intelligent User Interfaces*, 841–852. <https://doi.org/10.1145/3490099.3511105>
- Yuan, B., He, Y., Davis, J. Q., Zhang, T., Dao, T., Chen, B., Liang, P., Re, C., & Zhang, C. (2022). *Decentralized Training of Foundation Models in Heterogeneous Environments* (arXiv:2206.01288). arXiv. <https://doi.org/10.48550/arXiv.2206.01288>
- Zellers, R., Lu, J., Lu, X., Yu, Y., Zhao, Y., Salehi, M., Kusupati, A., Hessel, J., Farhadi, A., & Choi, Y. (2022). *MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound* (arXiv:2201.02639). arXiv. <https://doi.org/10.48550/arXiv.2201.02639>
- Zhang, D., Maslej, N., Brynjolfsson, E., Etchemendy, J., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Sellitto, M., Sakhaee, E., Shoham, Y., Clark, J., & Perrault, R. (2022). *Artificial Intelligence Index Report 2022*. Stanford Institute for Human-Centered AI, Stanford University. https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf
- Zhang, H., Gong, Y., Shen, Y., Li, W., Lv, J., Duan, N., & Chen, W. (2021). Poolingformer: Long document modeling with pooling attention. *International Conference on Machine Learning*, 12437–12446. <https://arxiv.org/abs/2105.04371>
- Zhang, Y., Qin, J., Park, D. S., Han, W., Chiu, C.-C., Pang, R., Le, Q. V., & Wu, Y. (2020). *Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition* (arXiv:2010.10504). arXiv. <https://doi.org/10.48550/arXiv.2010.10504>
- Zhang, Z., Zhang, A., Li, M., & Smola, A. (2022). *Automatic Chain of Thought Prompting in Large Language Models* (arXiv:2210.03493). arXiv. <https://doi.org/10.48550/arXiv.2210.03493>

II. Abbildungsverzeichnis

Abbildungsnummer	Name	Seite
Abb. 1	Anzahl der Parameter großer KI-Sprachmodelle seit GPT-3 (Open Source Modelle rot markiert). Quelle: state of ai Report 2022 (Benaic & Hogarth, 2022)	19
Abb. 2	Verdeutlichung der Self-Attention am Satz "Die Bank verleiht Geld". Das Token Bank (unten) hat eine hohe Korrelation mit dem Token Geld (oben), wobei die Korrelation zu den anderen Token geringer ausfällt.	24
Abb. 3	Zentraler Bestandteil der Foundation-Modelle sind Schichten mit Self-Attention Blöcken (blau), die kontextsensitive Einbettungsvektoren (violett) von Eingabetoken (grün) berechnen. Die logistische Schicht L prognostiziert die Wahrscheinlichkeit der Ausgabtoken. Beim Training werden die Parameter so optimiert, dass die Wahrscheinlichkeiten der korrekten fehlenden bzw. nächsten Token (gelb) möglichst hoch sind.	27
Abb. 4	Ein Foundation-Modell kann Informationen aus verschiedenen Datenquellen verschiedener Modalitäten berücksichtigen. Dieses eine Modell kann dann eine Vielzahl von nachgelagerten Aufgaben lösen (Bommasani et al., 2021).	29
Abb. 5	Die Genauigkeit des „few-shot“-Lernens von GPT-3 wird durch die Erweiterung der Modellgröße und der Anzahl der präsentierten Beispiele erhöht (Brown et al., 2020).	31
Abb. 6	Von einem Foundation-Modell mit Hilfe von Retrieval gefundene Antwort auf eine Frage im Natural Question Benchmark. Aktuelle Modelle erreichen eine Genauigkeit (F1) von 80% (Zhanag et al., 2021).	35
Abb. 7	Zusammenfassung eines Textes von 800 Wörtern durch das Modell BRIO im Vergleich zu der von Experten erstellten Zusammenfassung (Liu et al., 2022).	36
Abb. 8	Zu unterschiedlichen Texten von DALL-E 2 erzeugte Bilder (Ramesh et al., 2022)	40
Abb. 9	Zu unterschiedlichen Texten von CogVideo erzeugte Videos (Hong et al., 2022)	41
Abb. 10	Das Gato-Modell generiert aus den aktuellen Zuständen (hellblau) neue Aktionen (dunkelblau). Die Umgebung produziert daraus neue Zustände, usw. Das Modell kann Texte, Messwerte, Bilder, usw. verarbeiten.	44
Abb. 11	Um 2016 tauchte ein neuer Trend zu sehr großen Modellen auf, die von großen Internetfirmen trainiert wurden (rot). Diese waren in der Lage waren, die notwendigen Investitionen zu finanzieren. Die untere blaue Linie	55

	veranschaulicht den Berechnungsaufwand der anderen Modelle, z.B. von Universitäten (Sevilla et al., 2022).	
Abb. 12	Trainingsleistung (1 ExaFLOPs = 10^{18} FLOPs) unterschiedlicher Foundation-Modelle gegenüber Veröffentlichungsjahr nach Ländern	56
Abb. 13	Trainingsleistung (1018 FLOPs = 1 ExaFLOPs) gegenüber Veröffentlichungsjahr nach Ländern. Der Abbildung (intern)1.12: Großteil der Foundation-Modelle wurden in den USA entwickelt. Aus Deutschland und anderen EU Ländern wurden tendenziell kleinere Foundation-Modelle mit niedrigerem Trainingsaufwand veröffentlicht. Das BLOOM Modell stellt eine Ausnahme dar und wurde von einem Wissenschaftskollektiv bestehend aus über 250 Institutionen auf dem Jean Zay Supercomputer in Frankreich trainiert.	57
Abb. 14	Ergebnisse der Umfrage mit KMUs zu deren Einsatz von Foundation-Modellen	69
Abb. 15	Ergebnisse der Umfrage mit KMUs zu Hindernissen beim Einsatz von Foundation-Modellen	70
Abb. 16	Ergebnisse der Umfrage mit KMUs zur Bedeutung von unterschiedlichen Aspekten der Foundation-Modell-Entwicklung	72
Abb. 17	Ergebnisse der Umfrage mit KMUs zur Relevanz von unterschiedlichen Arten von KI-Modellen bei der Foundation-Modell-Entwicklung (Antworten mit einer Antwortrate von weniger als 20 % wurden ausgelassen. Die vollständigen Antworten befinden sich in Anhang A.2.)	73
Abb. 18	Simplifizierte Darstellung der Hard- und Software-Infrastruktur von HPCs	110
Abb. 19	Die Architektur des Trainings- & Applikations-Layers im Detail	113
Abb. 20	Die Architektur des System- und Data-Storage & Loading-Layers im Detail	120
Abb. 21	Die Architektur des Framework- & Service-Layers im Detail	122
Abb. 22	MLPerf hardware: accelerators (Zhang et al., 2022, S.18)	138
Abb. 23	Beispielrechnung Bau und Betrieb eines eigenen HPC-Rechenzentrums	160
Abb. 24	Delphi-Befragung: Wie beurteilen Sie folgende Risiken für die Entwicklung des Rechenzentrumsmarktes in Deutschland? (Hintemann et al., 2022, S. 37)	162
Abb. 25	Organisationseinheiten des LEAM-KI-Servicezentrums	169
Abb. 26	Das LEAM-Board als zentrale Governance-Einheit des LKS	181
Abb. 27	Szenario für die Struktur des LEAM-KI-Servicezentrums	225

III. Tabellenverzeichnis

Tabellennummer	Name	Seite
Tabelle 1	Eine Auswahl möglicher Anwendungen auf Basis von Sprachmodellen	33
Tabelle 2	Eine Auswahl möglicher Anwendungen auf Basis von multimodalen Modellen	39
Tabelle 3	Übersicht der wichtigsten Information zu der Umfrage für die Wirtschaft	67
Tabelle 4	Befragte Experten aus der Wirtschaft	68
Tabelle 5	Im Betrieb von LEAM werden für Training, Tuning und Inference Tausende GPUs benötigt	111
Tabelle 6	Beispiele für Rechenzentren	128
Tabelle 7	Größen von Rechenzentren	129
Tabelle 8	Übersicht über die Kühlmöglichkeiten in Rechenzentren	132
Tabelle 9	Compute Anforderungen für die Berechnung eines Foundationmodells	137
Tabelle 10	HPC-Standorte in Deutschland	142
Tabelle 11	Verfügbarkeitsklassen (VK1 - VK4)	152
Tabelle 12	Schutzklassen nach DIN EN 50600-1	154
Tabelle 13	Übersicht über die OE Housing	170
Tabelle 14	Übersicht über die Training-as-a-Service	174
Tabelle 15	Übersicht der Kosten des LEAM-KI-Servicezentrums	185
Tabelle 16	Gesamtkostenstruktur des LEAM-KI-Servicezentrums bei einer Abschreibungsdauer von vier Jahren	186
Tabelle 17	Kosten des LEAM-KI-Servicezentrums bei einem Einkauf der GPU-RZ-Leistung	188
Tabelle 18	Übersicht der Einnahmen durch das Model-Training	190
Tabelle 19	Übersicht der Einnahmen durch die Beratung	191
Tabelle 20	Gegenüberstellung der drei Finanzierungsszenarien für das LKS	193
Tabelle 21	Übersicht der Vor- und Nachteile der drei Finanzierungsmodelle mit Hinblick auf die rechtlichen Rahmenbedingungen	218
Tabelle 22	Übersicht über die drei Online-Umfragen	320
Tabelle 23	Übersicht der Interviewpartner:innen im Bereich Wirtschaft	321
Tabelle 24	Übersicht der Interviewpartner:innen im Bereich Rechenzentrum und Hardware	322
Tabelle 25	Übersicht der Interviewpartner:innen im Bereich Wissenschaft	323
Tabelle 26	Übersicht der sonstigen Interviewpartner:innen	324
Tabelle 27	Kosten für eine GPU-Stunde auf einer NVIDIA A100 Tensor Core GPU 80 GB nach Anbieter. Preise in US-Dollar wurden in Euro umgerechnet zu einem Kurs von \$1 = 0,948768EUR (Dollarkurs am 06.12.2022)	369

IV. Abkürzungsverzeichnis

Abb	Abbildung
Abs	Absatz
AEUV	Vertrag über die Arbeitsweise der Europäischen Union
AI	Artificial Intelligence - Künstliche Intelligenz
API	Application Programming Interface
Art	Artikel
B2B	Business to Business
BERT	Bidirectional Encoder Representations from Transformers
BMWK	Bundesministerium für Wirtschaft und Klimaschutz
bspw	beispielsweise
bzgl	bezüglich
bzw	beziehungsweise
ca	circa
DMZ	Demilitarized Zone
DNA	Desoxyribonukleinsäure
EU	Europäische Union
F&E	Forschung und Entwicklung
FLOP	oder FLOPs - Floating Point Operations
FTE	Full-Time-Equivalents
GG	Grundgesetz
ggf	Gegebenenfalls
GPT	Generative Pre-trained Transformer
GWB	Gesetz gegen Wettbewerbsbeschränkungen
HPC	High Performance Computing
IaaS	Infrastruktur-as-a-Service
IT	Information Technologies

KI	Künstliche Intelligenz
KMU	kleine und mittlere Unternehmen
LEAM	Large European Artificial Intelligence Models
LKS	LEAM KI-Servicezentrum
Mio	Millionen
ML	Machine Learning
Mrd	Milliarden
NLP	Natural Language Processing
NLU	Natural Language Understanding
o.D.	ohne Datum
OE	Organisationseinheit
OSS	Open Source Software
Pers	Person
PPP	Public-Private-Partnership
PUE	Power usage effectiveness
RZ	Rechenzentrum
S	Seite
Seq2Seq	Sequence to Sequence
sog	sogenannte
TVöD	Tarifvertrag für den öffentlichen Dienst
u.a.	unter anderem
USV	unterbrechungsfreie Stromversorgung
usw	Und so weiter
vgl	vergleiche
VPN	Virtual Private Network
z.B.	zum Beispiel
Ziff	Ziffer

V. Methodik der Machbarkeitsstudie

Im Folgenden werden die bei der Durchführung der Machbarkeitsstudie angewandten Methoden beschrieben. Sie lassen sich in drei übergeordnete Bereiche einteilen: Literatur- und Internetrecherche, die Erhebung von Primärdaten in qualitativer und quantitativer Form sowie deren Analyse.

Zur besseren Nachvollziehbarkeit für den/die Leser:in können die einzelnen Schritte der methodischen Vorgehensweise wie folgt umrissen werden:

1. Literatur- und Internetrecherche zu KI-Foundation-Modellen, Softwareanforderungen, Hochleistungsrechenzentren sowie organisatorische und finanzielle Rahmenbedingungen
2. Identifikation relevanter Zielgruppen für die quantitative Online-Umfrage:
 - a. Kleine und mittlere KI-Unternehmen sowie KI-Initiativen
 - b. Großunternehmen mit KI-Abteilungen
 - c. KI-Forscher:innen
3. Erstellung, Versendung und Auswertung einer Online-Umfrage pro Zielgruppe
4. Identifikation relevanter Akteursgruppen für qualitative Interviews:
 - a. Großunternehmen mit KI-Abteilungen
 - b. Betreiber von Rechenzentren und Expert:innen von Hardware
 - c. KI-Forscher:innen
 - d. Sonstige (bspw. KI-Initiativen, Vertreter der Bundesländer)
5. Erstellung mehrerer Leitfäden/ Fragenkataloge basierend auf den Zielgruppen
6. Durchführung und Auswertung der leitfadengestützten Interviews
7. Beurteilung der Machbarkeit anhand aller vorliegenden Informationen

Internet- und Literaturrecherche

Die Autor:innen der einzelnen Kapitel nutzten Internet- und Literaturrecherche als Ausgangspunkt für die Studie. Sie bietet in vielen Fällen einen geeigneten Einstieg in die Thematik. Die Erkenntnisse flossen außerdem in die Online-Umfrage sowie die Interviewleitfäden ein.

Im Rahmen der Recherchen wurden relevante Quellen aus der Literatur genutzt, um grundlegendes Wissen zu KI-Foundation-Modellen, ihren Eigenschaften und Besonderheiten sowie momentan existierenden Modellarten zu gewinnen. Außerdem wurde der aktuelle Entwicklungsstand und Einsatz von KI-Foundation-Modellen im internationalen Vergleich recherchiert.

Im Bereich der Softwareanforderungen wurden aktuelle Studien und das Expertenwissen der Autor:innen im Bereich der Foundation-Modellentwicklung herangezogen, um eine ausführliche Übersicht über die benötigte Software und das Personal zu erhalten. Darüber hinaus wurde die bestehende Literatur verwendet, um einen Überblick über bestehende Rechenzentren und die Herausforderungen für das KI-Training zu gewinnen. Schließlich bot die Recherche auch einen Einstieg in eine mögliche Governance, Gesellschaftsform sowie Finanzierungsmöglichkeiten.

Datenerhebung

Die Datenerhebung bestand aus zwei Aspekten: quantitativen Online-Umfragen und qualitativen Experteninterviews. Die quantitativen Daten wurden in Form einer webbasierten Online-Umfrage erhoben und an drei Akteursgruppen versendet:

- (1) Kleine und mittlere KI-Unternehmen & KI-Start-Ups,
- (2) Großunternehmen und
- (3) KI-Wissenschaftler:innen.

Das Ziel der qualitativen Datenerhebung war der Gewinn eines allgemeinen Überblicks über den wirtschaftlichen Bedarf von KI- Foundation-Modellen und deren wissenschaftlichen und technischen Voraussetzungen. Die qualitativen, durch leitfadengestützten Interviews wurden mit ausgewählten Expert:innen aus den Bereichen Wissenschaft, Wirtschaft, Politik und Recht durchgeführt. In den Interviews konnte tiefer auf Erkenntnisse aus der quantitativen Umfrage sowie neu aufkommende Themen eingegangen werden.

Online-Umfragen

Die Online-Umfrage wurde maßgeblich aus der Internet- und Literaturrecherche abgeleitet und dient dazu, wesentliche Faktoren wie beispielsweise die Relevanz, Nutzung und Entwicklung verschiedener KI-Foundation-Modelle quantitativ messbar zu machen. Die Befragung fand vom 31. Oktober bis zum 23. November 2022 statt.

Die drei Fragenkataloge wurden von den Projektpartnern der Machbarkeitsstudie unter Koordination des KI Bundesverbands ausgearbeitet. Sie bestehen aus 20 - 46 Fragen (s. Tabelle 22), die überwiegend in Multiple-Choice-Form aufgebaut sind. Bei der Auswertung ihrer Ergebnisse ist somit zu berücksichtigen, dass pro Frage meist mehrere Antwortmöglichkeiten gegeben waren. Zusätzlich zu den Multiple-Choice Fragen wurden demographische Angaben erfragt und zum Schluss durch ein offenes Kommentarfeld die Möglichkeit gegeben, weitere Anmerkungen zu den Anforderungen von KI-Foundation-Modellen oder zu der Machbarkeitsstudie zu machen. Im Anhang A.1 sind die Umfragen im Detail zu finden.

Um eine hohe quantitative Fallzahl zu erlangen, wurden die drei Umfragen an ein breites Spektrum an Kontakten gesendet. Insgesamt gab es 71 Rückmeldungen von den kleinen und mittleren KI-Unternehmen und 21 Rückmeldungen von KI-Wissenschaftler:innen. Aufgrund der geringen Rücklaufquote bei den Großunternehmen wurden ausgewählte Antwortpartner:innen der Online-Umfrage zusätzlich um Interviews gebeten. Somit liegt bei dieser Auswertung der Fokus auf den Interviews.

Umfrage	Anzahl Fragen	Anzahl Empfänger	Anzahl Antworten	Antwortrate
Kleine und mittlere KI-Unternehmen	44*	398	71	17,8 %
große Unternehmen	20	94	10	10,6 %
Wissenschaft	23	74	21	28,4 %

* Fragen wurden je nach Antwort nur teilweise angezeigt

Tabelle 22: Übersicht über die drei Online-Umfragen

Experteninterviews

Den Interviews lagen Leitfäden zu Grunde, die eine offene Beantwortung der Fragen vorsahen. Dabei gab es für jede Zielgruppe eigene Leitfäden mit Fokus auf deren Expertise.

Die Interviews wurden per Videokonferenz durchgeführt. Der Zeitraum der Befragung war vom 25. Oktober 2022 bis zum 12. Januar 2023.

Bei der Auswahl der Interviewpartner:innen wurde zunächst eine Identifikation und Einstellung relevanter Akteursgruppen vorgenommen. Daraus hervorgehend wurden die folgenden Gruppen bzw. Institutionen herausgestellt:

- Großunternehmen
- KI-Forscher:innen
- Rechenzentrums- und Hardwareexpert:innen
- Sonstige Expert:innen (bspw. regionale Vertreter, KI-Initiativen, etc.)

Es wurden insgesamt Interviews mit 71 Gesprächspartner:innen durchgeführt. Die vollständige Liste der Interviewpartner:innen befindet sich im Anhang B.1.

Die Auswertung der unterschiedlichen Daten wurde von den verantwortlichen Projektpartner:innen unternommen und mit den Ergebnissen der Literatur- und Internetrecherche verbunden.

Auswertung der Online-Umfragen

Zur Auswertung der quantitativen Datenerhebung wurden die Ergebnisse der Online-Umfrage von KI-Unternehmen/KI-Initiativen, auf der einen Seite, und KI-Forscher:innen, auf der anderen, jeweils gesammelt, ausgewertet und grafisch aufbereitet. Die offenen Fragen wurden zur Quantifizierbarkeit auf spezifische Schlüsselwörter durchsucht. Eine grafische Aufarbeitung der Multiple-Choice Fragen befindet sich in Anhang A.2. Aufgrund der geringen Teilnehmerzahl fand keine Auswertung der Umfrage an Großunternehmen statt.

Auswertung der Experteninterviews

Um die qualitativen Daten in Form der leitfadengestützten Interviews auszuwerten, wurden die Interviews transkribiert und in Ergebnisprotokollen zusammengefasst. Im Falle der KI-Forscher:innen wurden die Ergebnisse in einer Interviewmatrix für eine interne Auswertung zusammengefasst. Die Interviewten aus der Wirtschaft wurden gebeten, die Ergebnisprotokolle für den Druck freizugeben. Die bis zum Redaktionsschluss freigegebenen Ergebnisprotokolle befinden sich im Anhang B.3.

DOWNLOAD DER KOMPLETTEN STUDIE MIT ALLEN ANHÄNGEN UNTER:

<https://leam.ai/feasibility-study-leam-2023/>



Anhang der LEAM Machbarkeitsstudie



Anhang

Anhang A - Zusätzliche Information zu den Umfragen

Anhang A.1 - Die Umfragen

Umfrage 1: Kleine und mittlere Unternehmen

Foundation-Modelle

1. Arbeiten Sie bereits an oder wollen Sie mit normalen KI-Modellen (nicht Foundation-Modellen) arbeiten?
2. Inwiefern haben Sie sich in Ihrem Unternehmen bereits mit Foundation-Modellen oder darauf basierenden Applikationen auseinandergesetzt
 - a. Welche Voraussetzungen müssen erfüllt werden, damit Sie Foundation-Modelle nutzen würden?
 - b. Inwiefern entwickeln Sie eigene Anwendungen auf Basis von Foundation-Modellen?
 - c. Inwiefern passen Sie existierende Foundation-Modelle an (Tuning)?
 - d. Inwiefern entwickeln Sie eigene Foundation-Modelle?
3. In welchen Bereichen sind für Sie Foundation-Modelle in Nutzung und Entwicklung relevant?

Gesamtwirtschaftliche Bedeutung

4. Welche Bedeutung messen Sie der Datenanalyse und KI für die gesamtwirtschaftliche Entwicklung in Deutschland zu?
5. Welche Bedeutung messen Sie den Foundation-Modellen für die KI und damit für gesamtwirtschaftliche Entwicklung in Deutschland zu?
6. Wie wichtig ist aus Ihrer Sicht der Aufbau eines deutschen und europäischen Ökosystems rund um Foundation Models für die Digitale Souveränität und Wettbewerbsfähigkeit?

Foundation-Modell Entwicklung

7. Welche Bedeutung hat für Sie die Verfügbarkeit von Foundation-Modellen, die in Europa entwickelt wurden und Werte wie Transparenz, Reduktion von Bias und Nachhaltigkeit berücksichtigen?
8. Welche Bedeutung hat für Sie, bzgl. der Entwicklung/Anwendung von Foundation-Modellen, die Zusammenarbeit mit wissenschaftlichen Institutionen?
9. Welche Bedeutung hat für Sie, bzgl. der Entwicklung/Anwendung von Foundation-Modellen, die Zusammenarbeit mit spezialisierten KMUs und Start-ups?
10. Haben Sie weitere Anmerkungen zu den Themen Bedeutung von, Gefahren durch und wissenschaftliche Fragestellungen zu Foundation-Modellen?

Technologische Anforderungen

11. Rechenleistung: Wie viele GPUs und andere Beschleuniger werden z.B. für Training und Inferenz von state-of-the-art Foundation-Modellen benötigt?

12. Verfügbarkeit: Welche Kriterien müssen Service Level Agreements (SLA) erfüllen, um die Entwicklung von Foundation-Modellen möglich zu machen?
13. Anforderungen der Entwicklung von Foundation-Modellen an die Software-Infrastruktur
14. Scheduling Infrastructure: Welche Scheduling Infrastructures sollen bei der Entwicklung eingesetzt werden?
15. Deployment Infrastructure: Welche Deployment Infrastructures müssen bei der Entwicklung unterstützt werden?
16. Access Control Levels & Datenhoheit: Wie soll das Laden der Daten in der Entwicklung angesteuert werden?
17. Anforderungen der Entwicklung von Foundation-Modellen an die Trainings-Software
18. Frameworks: Welche Frameworks sollen bei der Entwicklung von Foundation-Modellen eingesetzt werden?
19. Inference: Welche Voraussetzungen sollen Service Level Agreements (SLAs) mindestens erfüllen?
 - a. Inference: Welche Voraussetzungen sollen Service Level Agreements (SLAs) **bei der Verfügbarkeit** mindestens erfüllen?
 - b. Inference: Welche Voraussetzungen sollen Service Level Agreements (SLAs) **bei der erlaubten Latenz** mindestens erfüllen?
20. Anforderungen der Entwicklung von Foundation-Modellen an den Data-Storage-Layer
21. Internet- & Bandbreite: Wie schnell muss die Internetverbindung für die Entwicklung von Foundation-Modellen mindestens sein?
22. Speicherplatz für Daten: Wie hoch ist der Speicherbedarf für die Entwicklung eines Foundation-Modells?
23. Compliance: Welche regulatorischen Voraussetzungen müssen erfüllt werden, um Foundation-Modelle zu entwickeln?
 - a. Können sie die ISO/Normen benennen, die beachtet werden müssen?

Anforderungen der Entwicklung von Foundation-Modellen an Daten

24. Beschäftigen Sie sich mit möglichen Bias, Diskriminierung und Misrepresentation in Daten?
25. Wissen Sie, wie Sie Bias, Diskriminierung und Misrepresentation hinreichend quantifizieren können, um diese zu adressieren?
26. Ist Ihnen das Konzept der Model Cards / Data Set Cards im Bezug zur Erstellung neuer Daten geläufig?
 - a. Ist das Konzept für Ihre Domäne relevant oder hilfreich?

Anforderungen der Entwicklung von Foundation-Modellen an Mitarbeiter:innen

27. Welche Personalrollen (oder äquivalent) sind für die Entwicklung von Foundation-Modellen erforderlich?
28. In welchen dieser Personalrollen beobachten oder antizipieren Sie einen Mangel, um Foundation-Modelle zu entwickeln?

Wirtschaftliche Anforderungen

29. Welche Kosten halten Sie für die Entwicklung eines Foundation-Modells für realistisch?
30. Welche Kosten halten Sie für das Training eines Foundation-Modells für realistisch?
31. In welchem Zeitraum erwarten Sie, diese Investitionskosten zu amortisieren?
32. Wie erwarten Sie, dass sich diese Investitionskosten amortisieren könnten?

Zivilgesellschaftliche Anforderungen

33. Wie sind Sie über die Regulation der Entwicklung von Foundation-Modellen informiert?
34. Empfinden Sie Ihr Wissen über die Regulation als Enabler oder Disabler für Ihren möglichen Einsatz von Foundation-Modellen?
35. Wie sollten Ihrer Meinung nach Datensätze für die Entwicklung von Foundation-Modellen erhoben werden?
36. Wie hoch empfinden Sie Awareness, Relevanz und Akzeptanz bzgl. Foundation-Modellen innerhalb Ihrer Organisation?
37. Haben Sie weitere Anmerkungen zu den technologischen, wirtschaftlichen und zivilgesellschaftlichen Anforderungen der Entwicklung von KI-Foundation-Modellen?

Metadaten

38. Was ist Ihr Name?
39. Was ist Ihre E-Mail Adresse?
40. Wie groß ist Ihr Unternehmen?
41. Welcher Branche gehört Ihr Unternehmen an?
42. Was ist Ihre Position im Unternehmen?
43. Stehen Sie für einen weiteren Austausch zu den Themen Foundation-Modelle, digitale Souveränität sowie KI im allgemeinen zur Verfügung?

Ende

44. Abschließend möchten wir Ihnen die Gelegenheit bieten, uns Reaktionen, Ideen, Wünsche, Prioritäten oder Warnungen mitzuteilen, die für unsere weitere Arbeit wichtig sein könnten.

Umfrage 2: Große Unternehmen

Foundation-Modelle

1. Inwiefern haben Sie sich in Ihrem Unternehmen bereits mit Foundation-Modellen oder darauf basierenden Applikationen auseinandergesetzt

Nutzung

2. In welchen Bereichen sind für Sie Foundation-Modelle in Nutzung und Entwicklung relevant?
3. Inwiefern entwickeln Sie eigene Anwendungen auf Basis von Foundation-Modellen?
4. Inwiefern passen Sie existierende Foundation-Modelle an (Tuning)?

5. Inwiefern entwickeln Sie eigene Foundation-Modelle?

Gesamtwirtschaftliche Bedeutung

6. Welche Bedeutung messen Sie der Datenanalyse und KI für die gesamtwirtschaftliche Entwicklung in Deutschland zu?
7. Welche Bedeutung messen Sie den Foundation-Modellen für die KI und damit für gesamtwirtschaftliche Entwicklung in Deutschland zu?
8. Wie wichtig ist aus Ihrer Sicht der Aufbau eines deutschen und europäischen Ökosystems rund um Foundation Models für die Digitale Souveränität und Wettbewerbsfähigkeit?

Fragen der Entwicklung

9. Welche Bedeutung hat für Sie die Verfügbarkeit von Foundation-Modellen, die in Europa entwickelt wurden und Werte wie Transparenz, Reduktion von Bias und Nachhaltigkeit berücksichtigen?
10. Welche Bedeutung hat für Sie, bzgl. der Entwicklung/Anwendung von Foundation-Modellen, die Zusammenarbeit mit wissenschaftlichen Institutionen?
11. Welche Bedeutung hat für Sie, bzgl. der Entwicklung/Anwendung von Foundation-Modellen, die Zusammenarbeit mit spezialisierten KMUs und Start-ups?

LEAM

Ziel der LEAM Initiative ist der Aufbau eines dedizierten KI-Servicezentrums für die Wissenschaft und Industrie zur Erstellung von Foundation-Modellen und der Entwicklung darauf aufbauender Anwendungen.

12. Inwiefern würden Sie die Services von LEAM nutzen?
13. Inwiefern würden Sie bzw. Ihr Unternehmen in den Aufbau eines LEAM KI-Servicezentrums investieren (unter noch zu klärenden Voraussetzungen)?

Metadaten

14. Wie groß ist Ihr Unternehmen?
15. Welcher Branche gehört Ihr Unternehmen an?
16. Was ist Ihre Position im Unternehmen?
17. Was ist Ihr Name?
18. Was ist Ihre E-Mail Adresse?
19. Stehen Sie für einen weiteren Austausch zu den Themen Foundation-Modelle, digitale Souveränität sowie KI im allgemeinen zur Verfügung?
20. Welche Anregungen, Ideen oder auch Kritikpunkte möchten Sie uns mitgeben?

Umfrage 3: Wissenschaft

Persönliche Daten

1. Bitte füllen Sie Ihre persönlichen Daten aus

Foundation-Modelle

2. Wie schätzen Sie die Bedeutung von KI-Foundation-Modellen für die Wissenschaft ein?

3. Wie schätzen Sie die Bedeutung von KI-Foundation-Modellen für die Wirtschaft ein?
4. Arbeiten Sie bereits an oder wollen Sie mit normalen KI-Modellen (nicht Foundation-Modellen) arbeiten?
5. Arbeiten Sie bereits an oder wollen Sie mit Foundation-Modellen arbeiten?
 - Können Sie uns darüber etwas erzählen? Welche Daten werden in Ihrer Organisation aktuell für das Training von Foundation-Modellen verwendet?
 - Was hindert Sie daran, bereits heute Foundation-Modelle einzusetzen?
 - Wie hoch schätzen Sie die Wahrscheinlichkeit ein, Foundation-Modelle für Ihre Arbeit zu nutzen?
6. Wie schätzen Sie die Bedeutung verschiedener Arten von Foundation-Modellen in Gegenwart und Zukunft ein?
7. Gibt es weitere Arten von Foundation-Modellen, die Sie als bedeutend einschätzen?
8. Wie schätzen Sie die Gefahren ein, die mitunter im Zusammenhang mit Foundation-Modellen genannt wurden?
9. Sehen Sie zusätzliche Arten von potentiellen Gefahren?
10. Fast alle existierenden Foundation-Modelle, deren Architekturen und alle anderen wichtigen Neuerungen auf dem Gebiet kommen aus den USA und aus China. Warum? Was fehlt uns? Wo stehen wir in Deutschland im Vergleich in Bezug auf notwendige Voraussetzungen?
11. Sehen Sie weitere Voraussetzungen, die uns fehlen?
12. Welche Wichtigkeit haben Ihrer Meinung nach die folgenden wissenschaftlichen Fragestellungen?
 - Welche Wichtigkeit haben Ihrer Meinung nach die folgenden wissenschaftlichen Fragestellungen zur **Architektur**?
 - Welche Wichtigkeit haben Ihrer Meinung nach die folgenden wissenschaftlichen Fragestellungen zu **Daten und Datenverarbeitung**?
 - Welche Wichtigkeit haben Ihrer Meinung nach die folgenden wissenschaftlichen Fragestellungen zum **Pre-Training**?
 - Welche Wichtigkeit haben Ihrer Meinung nach die folgenden wissenschaftlichen Fragestellungen zum **Fine-Tuning (Nachtraining)**?
 - Welche Wichtigkeit haben Ihrer Meinung nach die folgenden wissenschaftlichen Fragestellungen zu **Multimedia**?
 - Welche Wichtigkeit haben Ihrer Meinung nach die folgenden wissenschaftlichen Fragestellungen zu **Sonstigen Themen**?
13. Welche weiteren relevanten Fragestellungen sehen Sie für die Forschung - auch besonders in Ihrem eigenen Fachgebiet? (Siehe Beispiele)
14. Wie schätzen Sie die Bedeutung von Foundation-Modellen für die folgenden Anwendungen ein?
15. Sehen Sie weitere Anwendungen, für die Foundation-Modelle bedeutend sein werden?
16. Haben Sie weitere Anmerkungen zu den Themen Bedeutung von, Gefahren durch und wissenschaftliche Fragestellungen zu Foundation-Modellen?

Bias und Diskriminierung

17. Beschäftigen Sie sich mit möglichen Bias, Diskriminierung und Misrepresentation in Daten?
18. Wissen Sie, wie Sie Bias, Diskriminierung und Misrepresentation hinreichend quantifizieren können, um diese zu adressieren?

Wirtschaftliche Anforderungen

19. Welche Kosten halten Sie für die Entwicklung eines Foundation-Modells für realistisch?
20. Welche Kosten halten Sie für das Training eines Foundation-Modells für realistisch?

Zivilgesellschaftliche Anforderungen

21. Wie sind Sie über die Regulation der Entwicklung von Foundation-Modellen informiert?
22. Empfinden Sie Ihr Wissen über die Regulation als Enabler oder Disabler für Ihren möglichen Einsatz von Foundation-Modellen?

Ende

23. Abschließend möchten wir Ihnen die Gelegenheit bieten, uns Reaktionen, Ideen, Wünsche, Prioritäten oder Warnungen mitzuteilen, die für unsere weitere Arbeit wichtig sein könnten.

Anhang A.2 - Graphische Auswertung der Umfragen

Die detaillierte Auswertung der Umfragen sind zu finden unter:

Umfrage 1: Kleine und Mittlere Unternehmen

- <https://form.typeform.com/report/VoH1BI5d/WHsdAKz0nUHNr76M>

Umfrage 2: Große Unternehmen

- <https://form.typeform.com/report/eDjpXmDx/RVEnkGRn1K7fhNja>

Umfrage 3: Wissenschaft

- <https://form.typeform.com/report/G3MBzf8d/iG33eUmp26emWkiC>

LEAM Umfrage KMU

71 Antworten

Arbeiten Sie bereits an oder wollen Sie mit normalen KI-Modellen (nicht Foundation-Modellen) arbeiten?

71 von 71 Personen haben diese Frage beantwortet

Wir setzen bereits KI bzw. darauf basierende Applikationen ein 64 Antw. 90.1%



Wir planen aktuell, KI bzw. darauf basierende Applikationen einzusetzen 4 Antw. 5.6%



Wir beabsichtigen aktuell nicht, KI bzw. darauf basierende Applikationen einzusetzen 2 Antw. 2.8%



Wir haben angefangen, uns mit KI bzw. darauf basierenden Applikationen auseinanderzusetzen 1 Antw. 1.4%

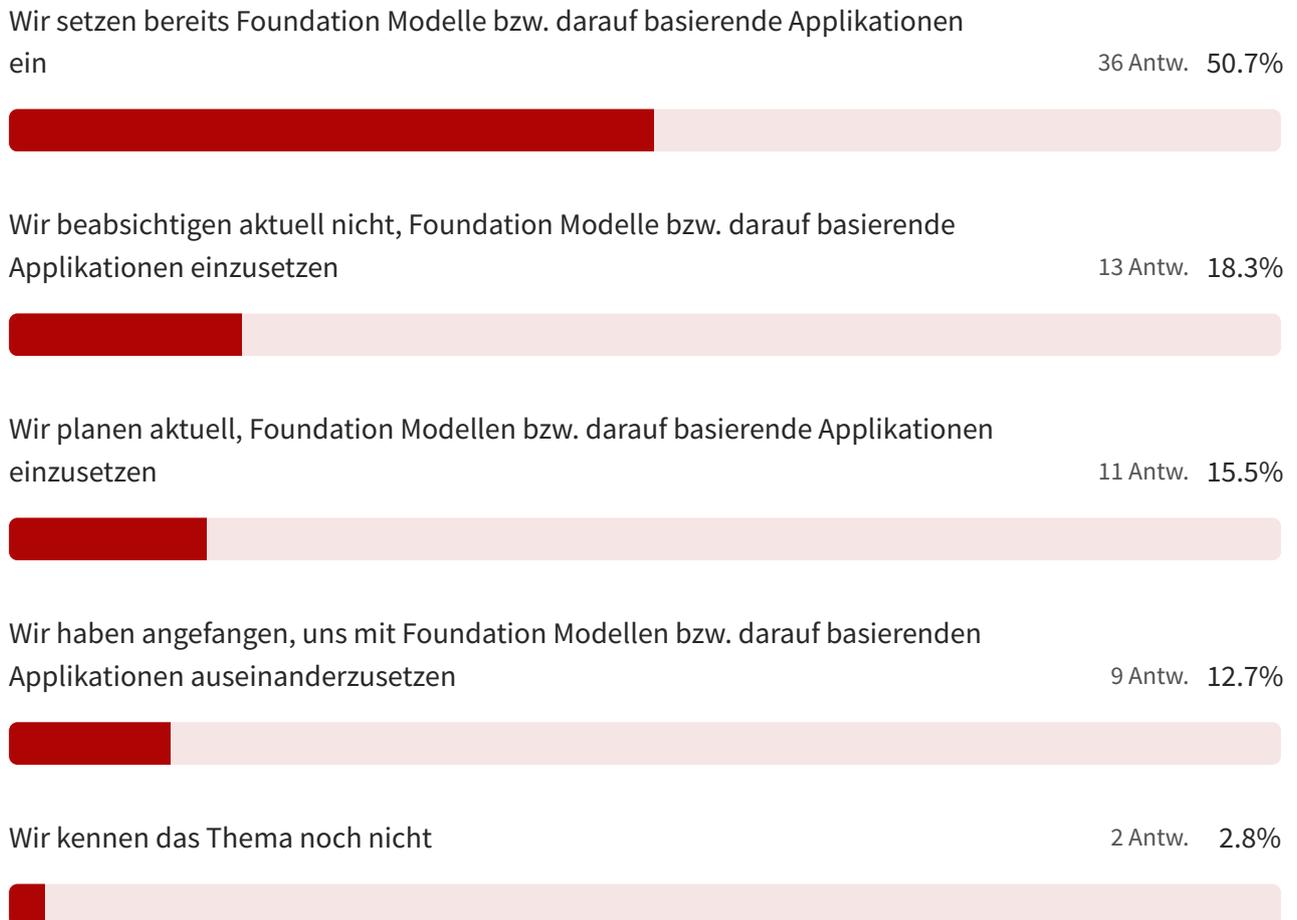


Wir kennen das Thema noch nicht 0 Antw. 0%



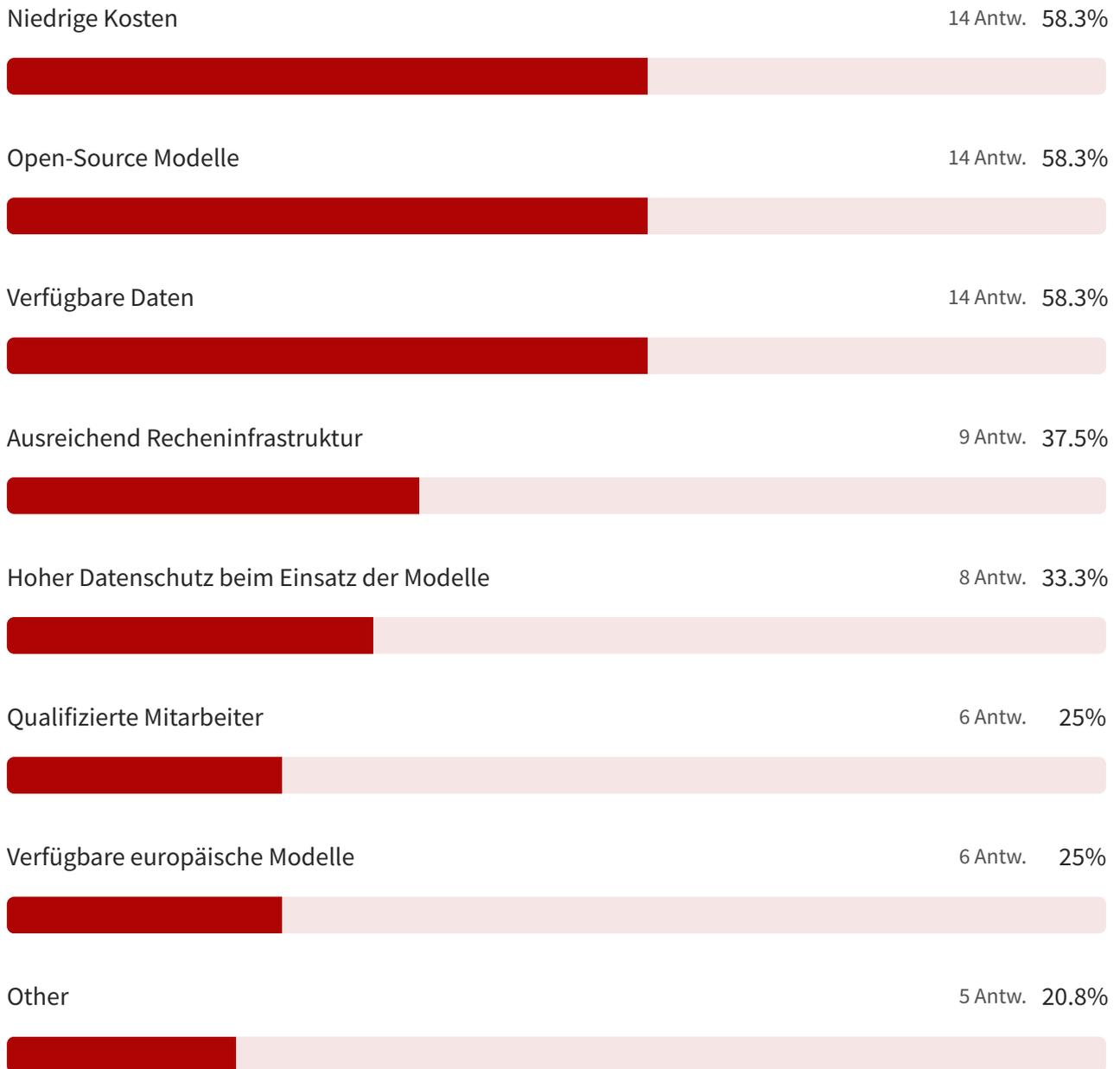
Inwiefern haben Sie sich in Ihrem Unternehmen bereits mit Foundation Modellen oder darauf basierenden Applikationen auseinandergesetzt

71 von 71 Personen haben diese Frage beantwortet



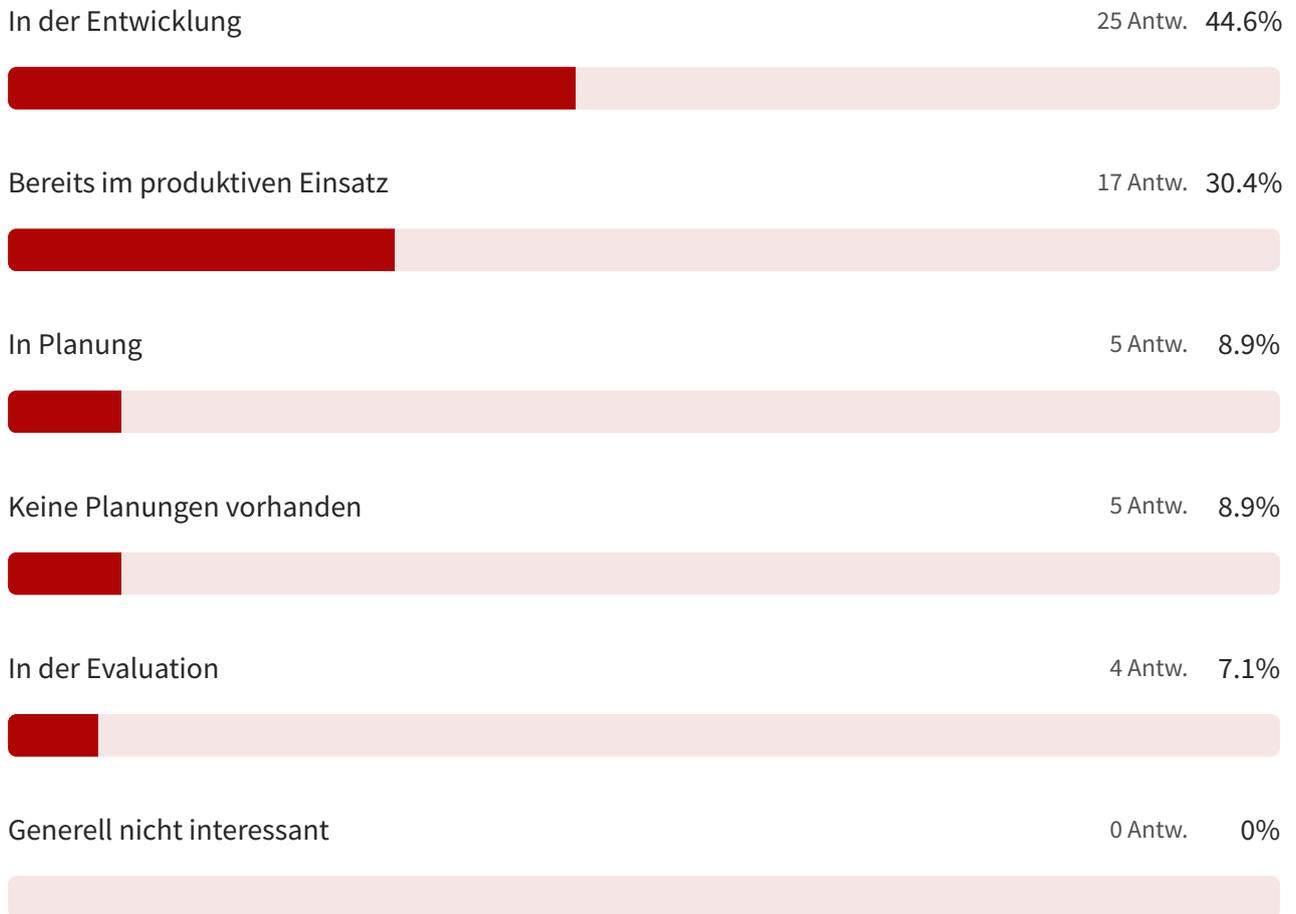
Welche Voraussetzungen müssen erfüllt werden, damit Sie Foundation-Modelle nutzen würden?

24 von 71 Personen haben diese Frage beantwortet



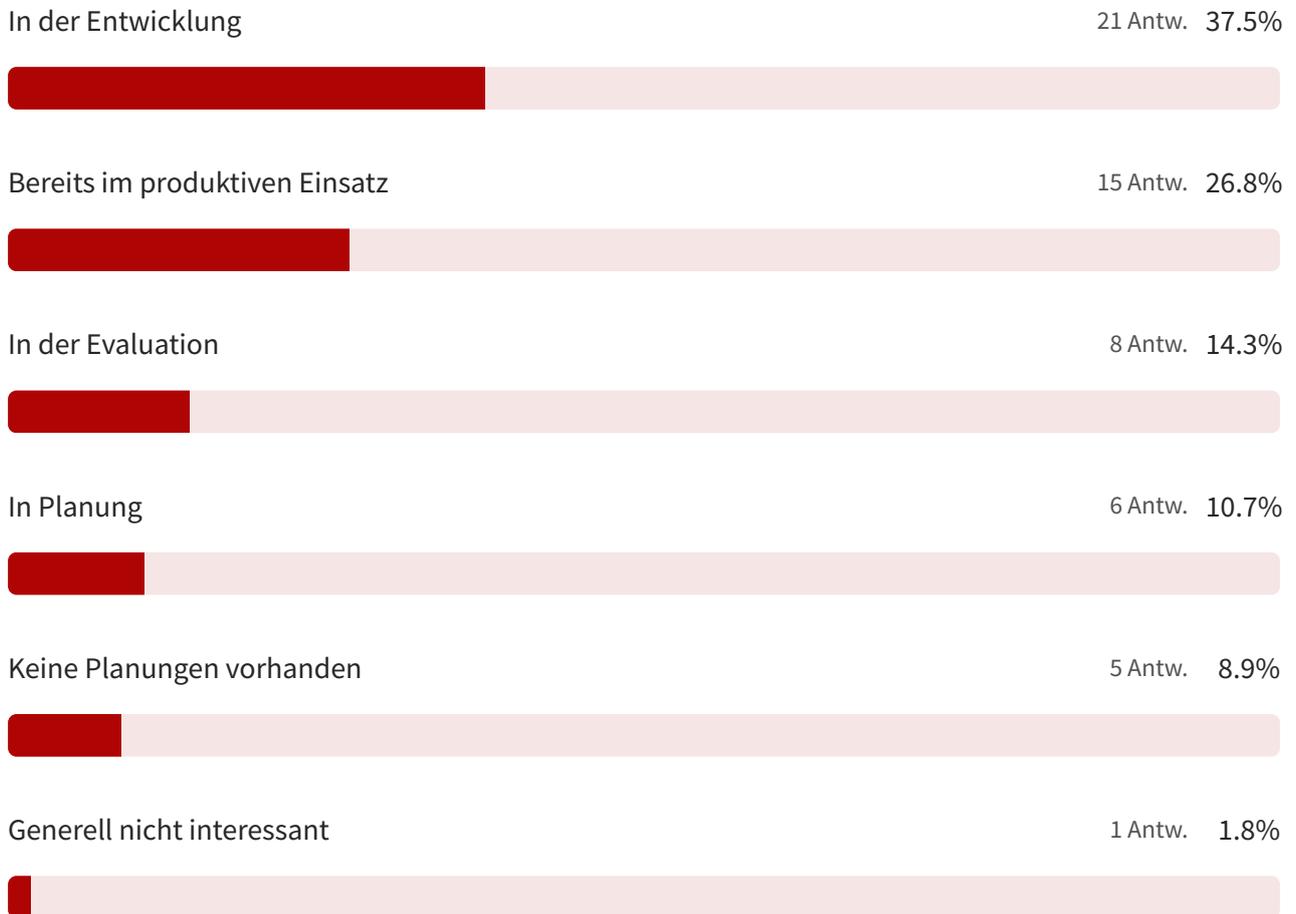
Inwiefern entwickeln Sie eigene Anwendungen auf Basis von Foundation Modellen?

56 von 71 Personen haben diese Frage beantwortet



Inwiefern passen Sie existierende Foundation Modelle an (Tuning)?

56 von 71 Personen haben diese Frage beantwortet



Inwiefern entwickeln Sie eigene Foundation Modelle?

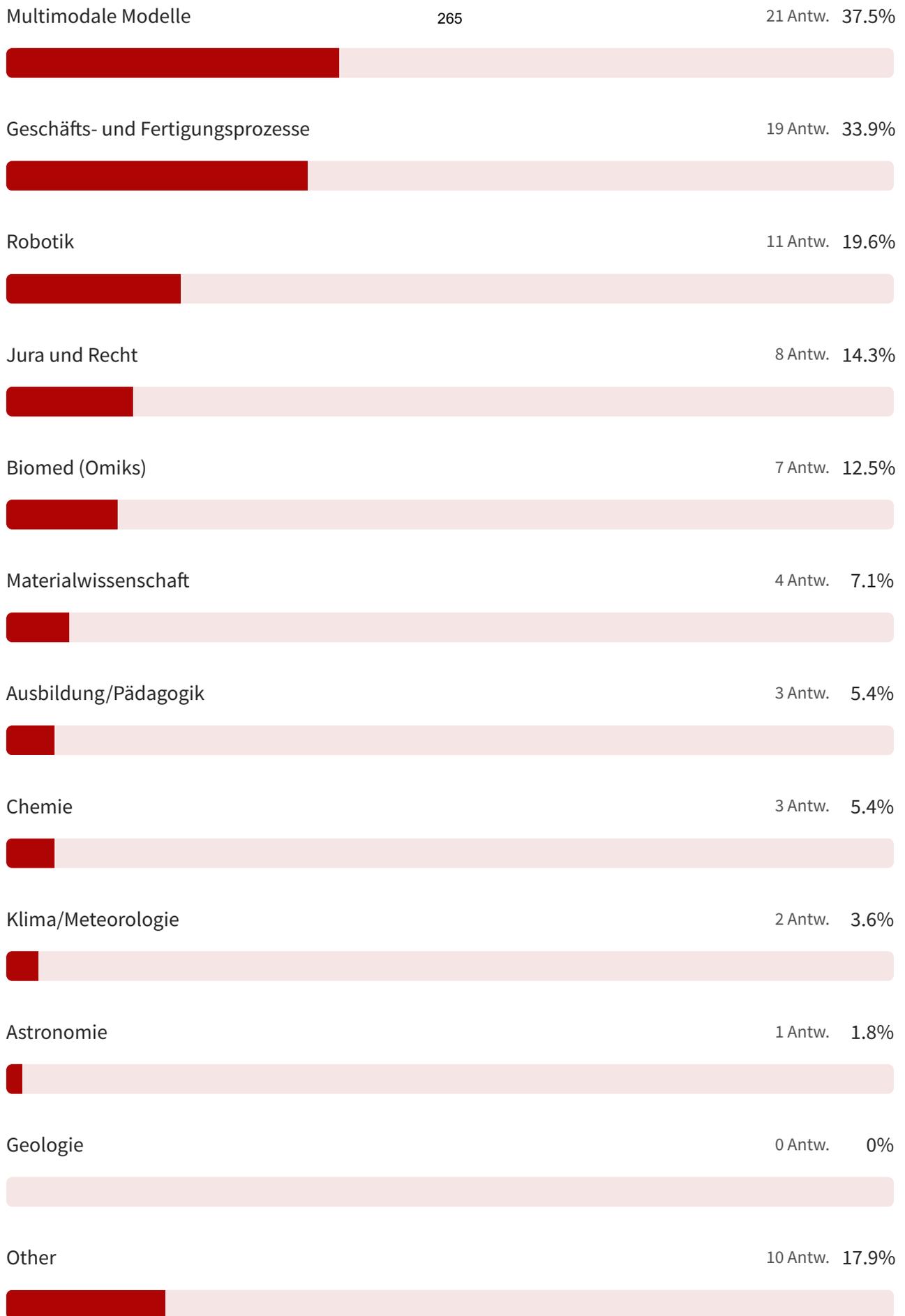
56 von 71 Personen haben diese Frage beantwortet



In welchen Bereichen sind für Sie Foundation Modelle in Nutzung und Entwicklung relevant?

56 von 71 Personen haben diese Frage beantwortet





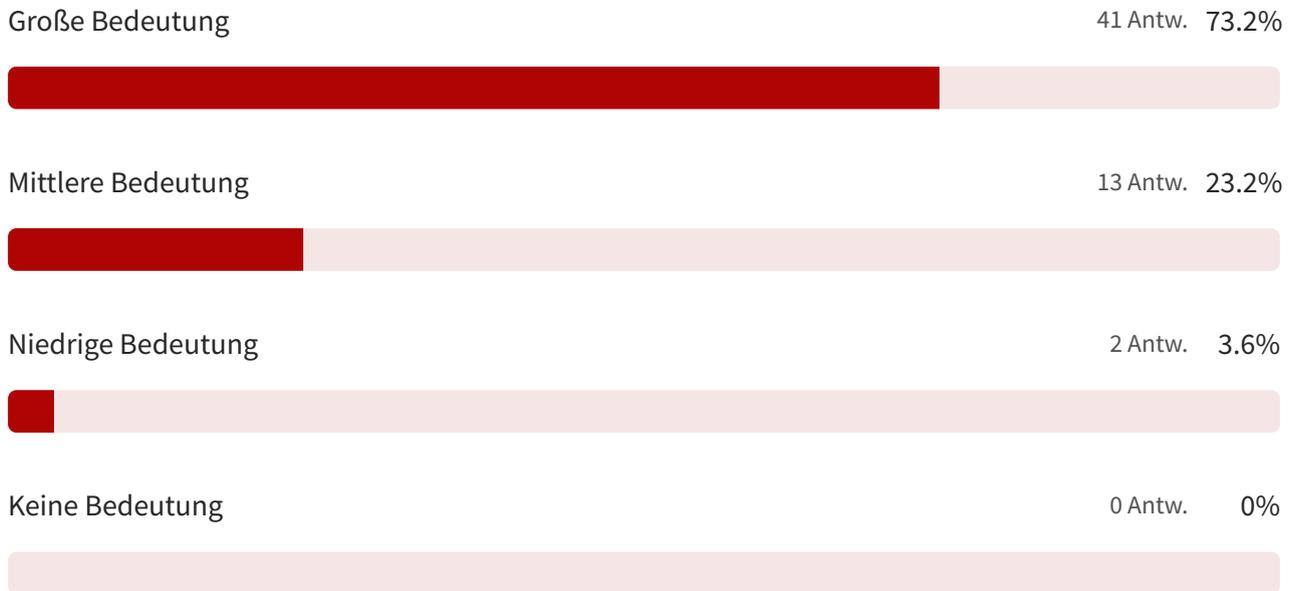
Welche Bedeutung messen Sie der Datenanalyse und KI für die gesamtwirtschaftliche Entwicklung in Deutschland zu?

71 von 71 Personen haben diese Frage beantwortet



Welche Bedeutung messen Sie den Foundation Modellen für die KI und damit für gesamtwirtschaftliche Entwicklung in Deutschland zu?

56 von 71 Personen haben diese Frage beantwortet



Wie wichtig ist aus Ihrer Sicht der Aufbau eines deutschen und europäischen Ökosystems rund um Foundation Models für die Digitale Souveränität und Wettbewerbsfähigkeit?

56 von 71 Personen haben diese Frage beantwortet



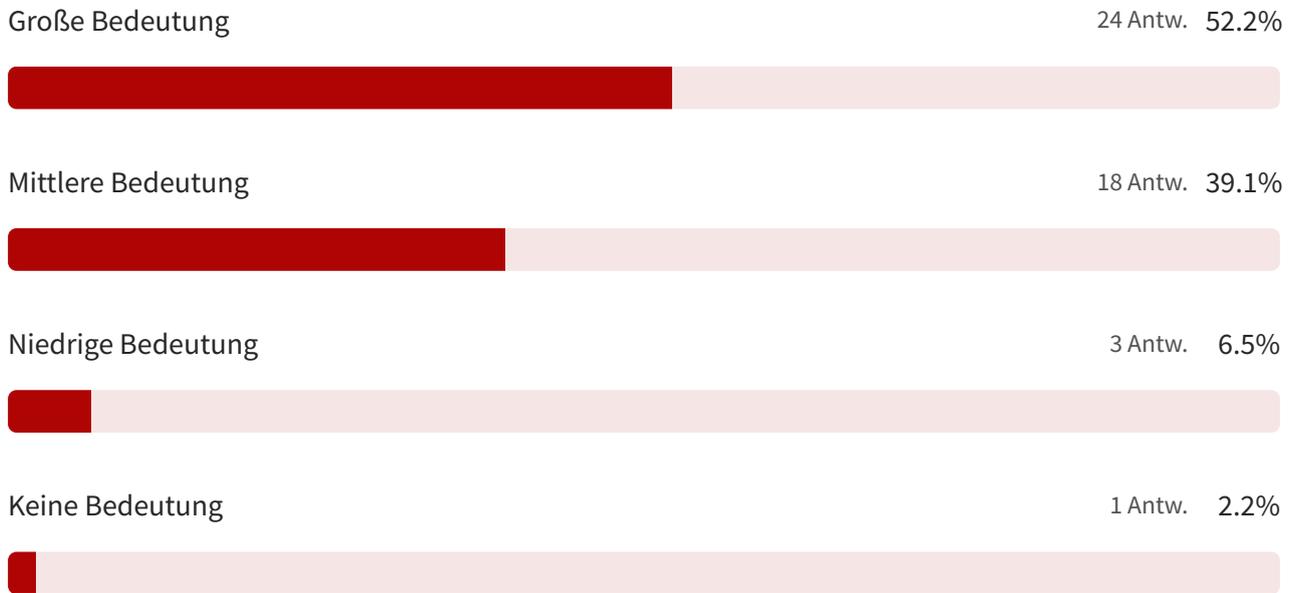
Welche Bedeutung hat für Sie die Verfügbarkeit von Foundation Modellen, die in Europa entwickelt wurden und Werte wie Transparenz, Reduktion von Bias und Nachhaltigkeit berücksichtigen?

46 von 71 Personen haben diese Frage beantwortet



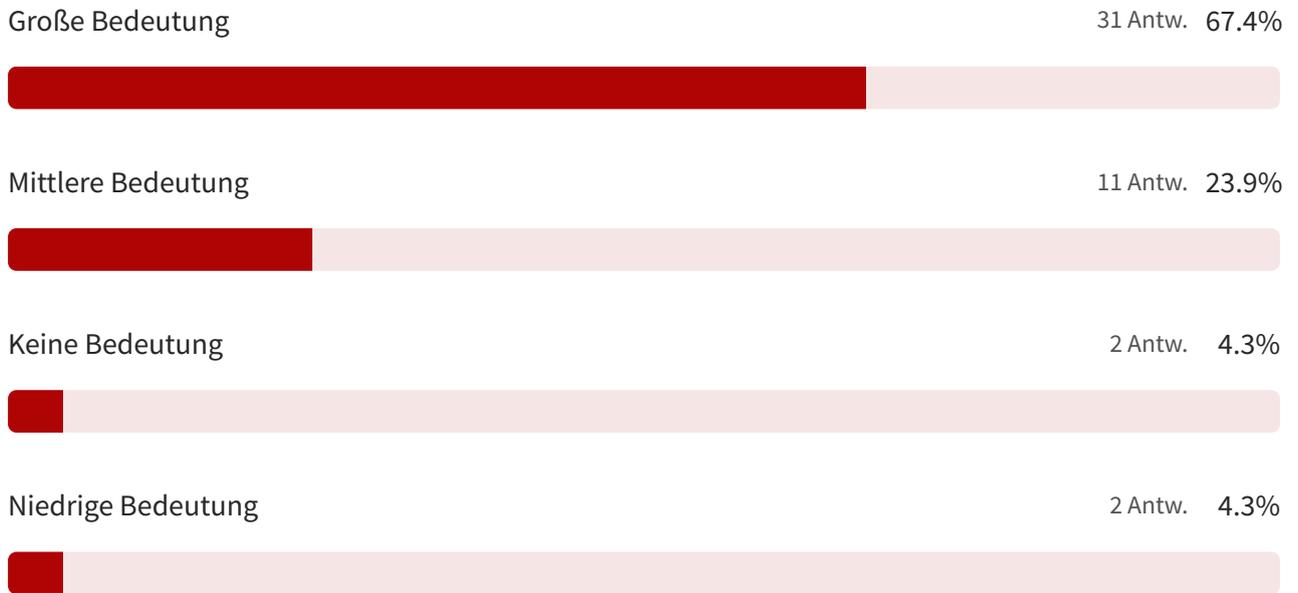
Welche Bedeutung hat für Sie, bzgl. der Entwicklung/Anwendung von Foundation-Modellen, die Zusammenarbeit mit wissenschaftlichen Institutionen?

46 von 71 Personen haben diese Frage beantwortet



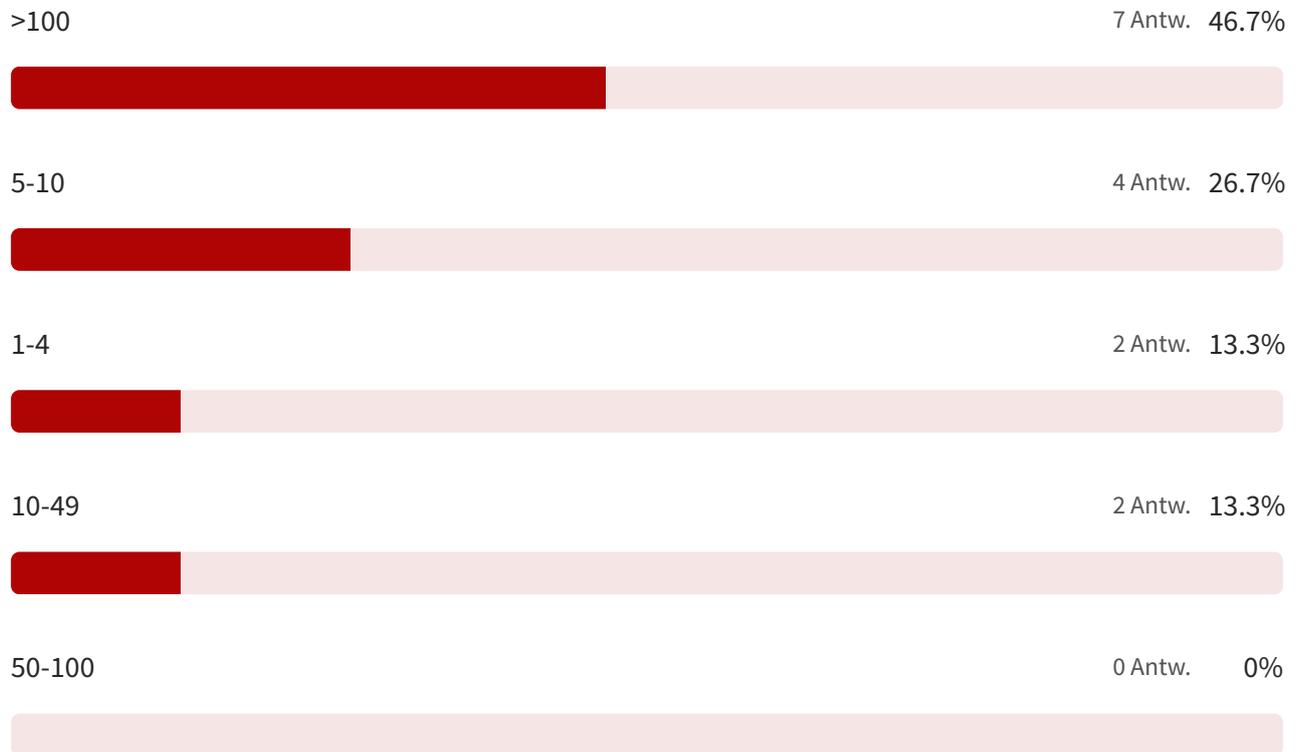
Welche Bedeutung hat für Sie, bzgl. der Entwicklung/Anwendung von Foundation-Modellen, die Zusammenarbeit mit spezialisierten KMUs und Startups?

46 von 71 Personen haben diese Frage beantwortet



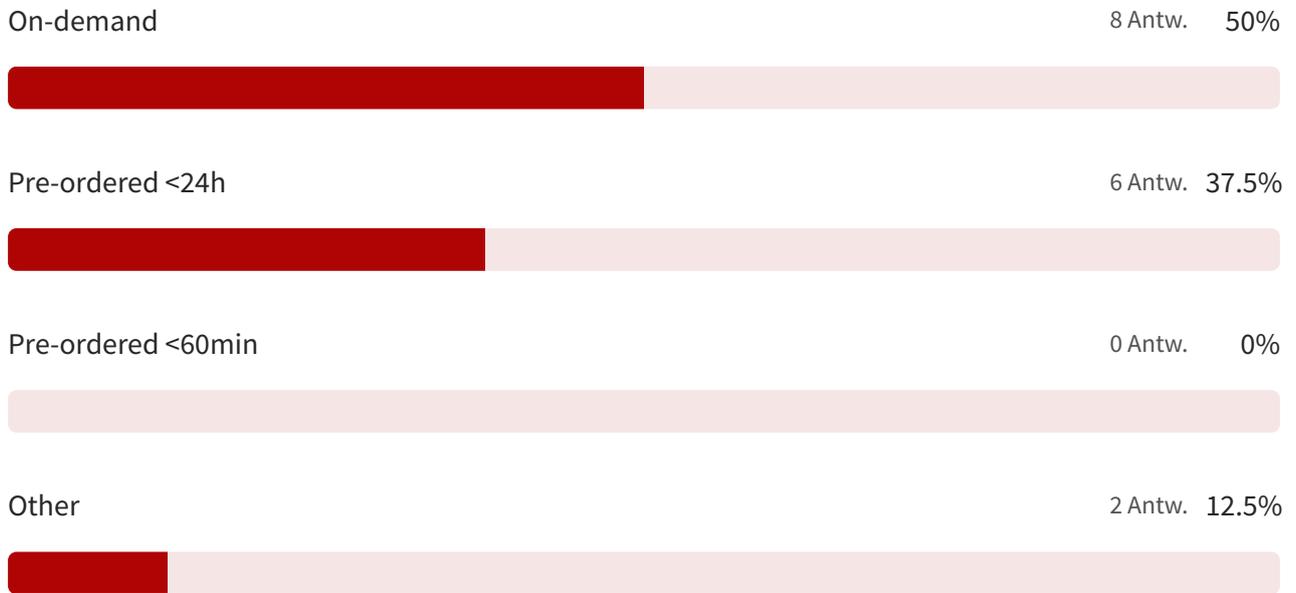
Rechenleistung: Wie viele GPUs und andere Beschleuniger werden z.B. für Training und Inferenz von state-of-the-art Foundation-Modellen benötigt?

15 von 71 Personen haben diese Frage beantwortet



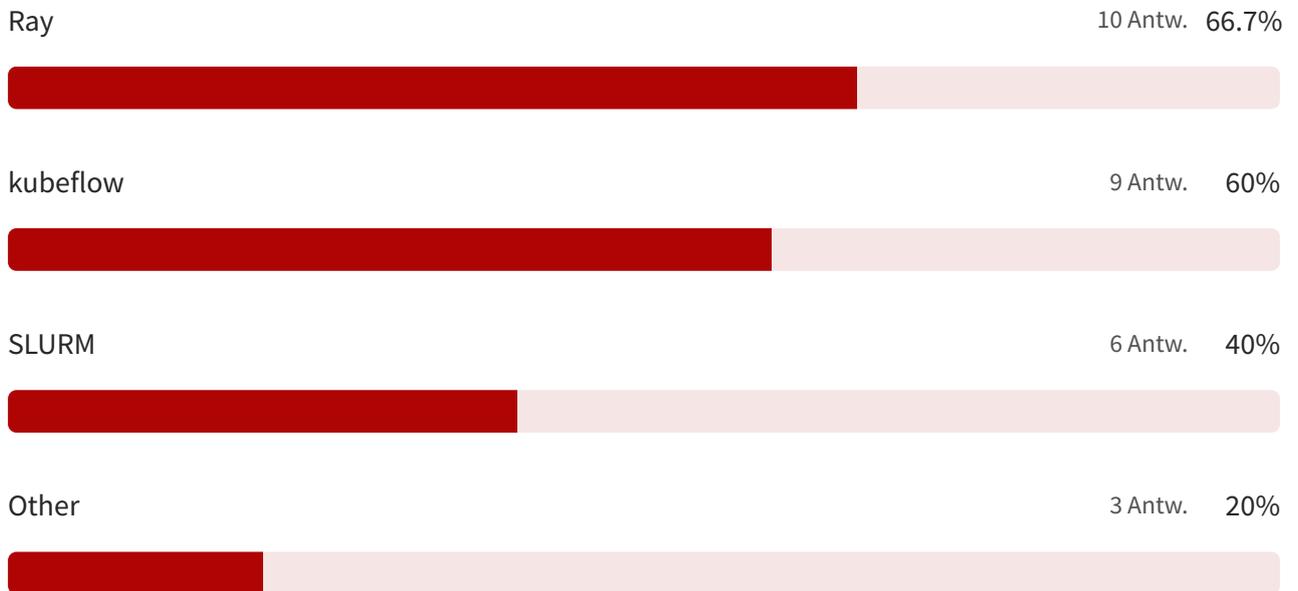
Verfügbarkeit: Welche Kriterien müssen Service Level Agreements (SLA) erfüllen, um die Entwicklung von Foundation-Modellen möglich zu machen?

16 von 71 Personen haben diese Frage beantwortet



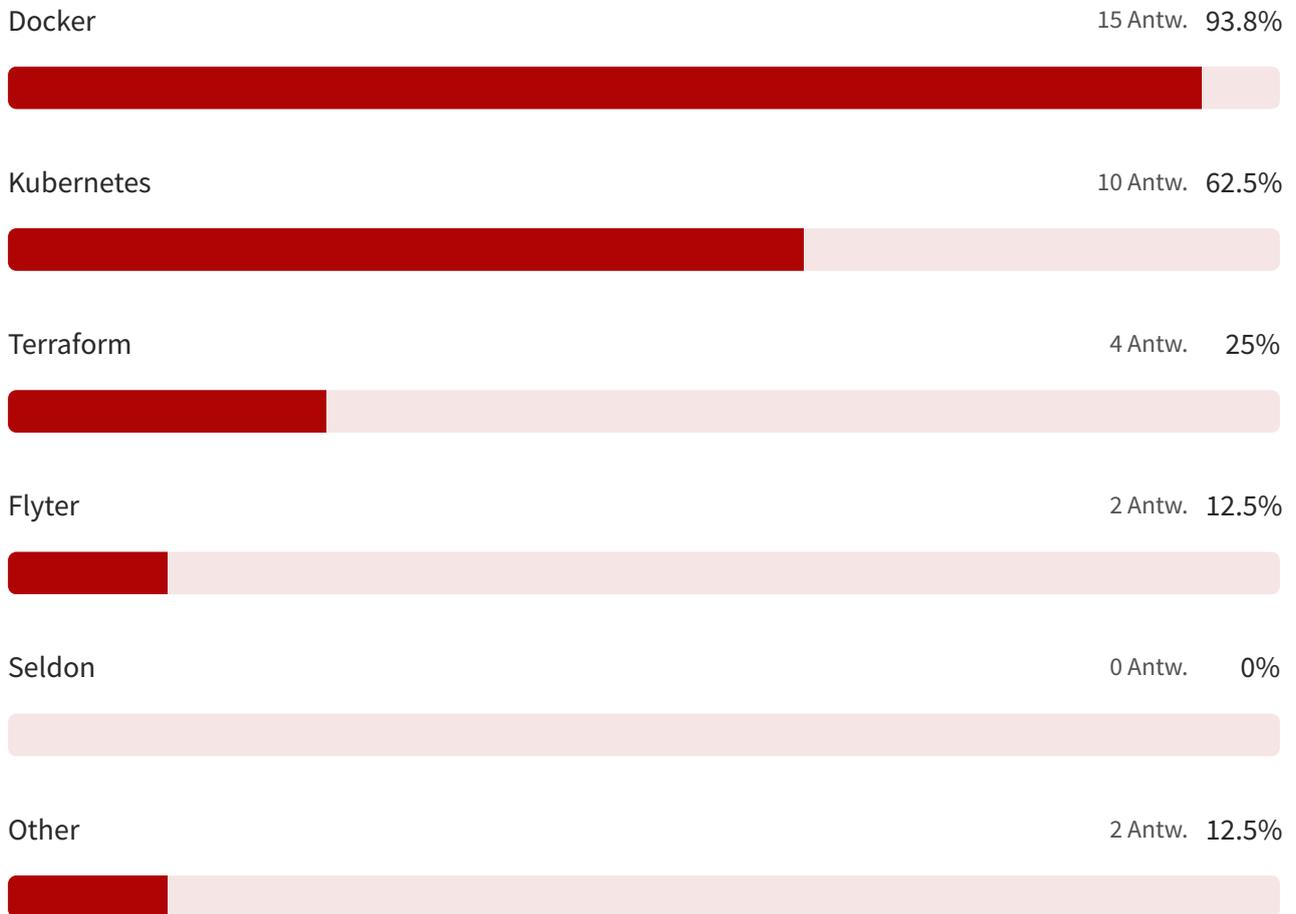
Scheduling Infrastructure: Welche Scheduling Infrastructures sollen bei der Entwicklung eingesetzt werden?

15 von 71 Personen haben diese Frage beantwortet



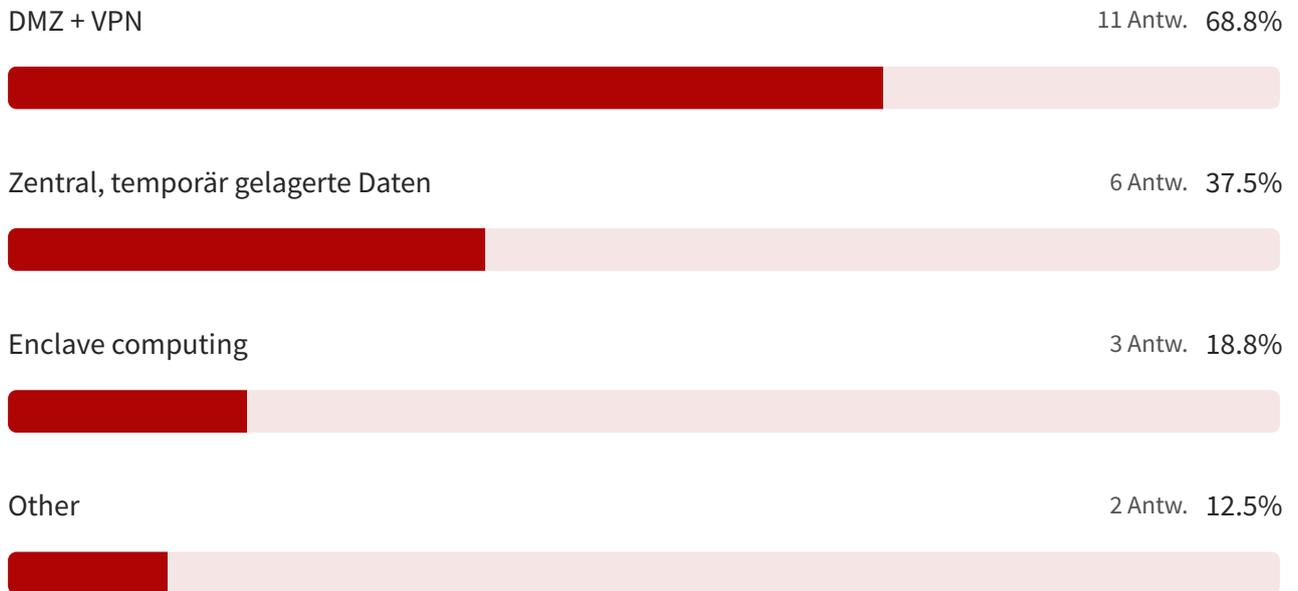
Deployment Infrastructure: Welche Deployment Infrastructures müssen bei der Entwicklung unterstützt werden?

16 von 71 Personen haben diese Frage beantwortet



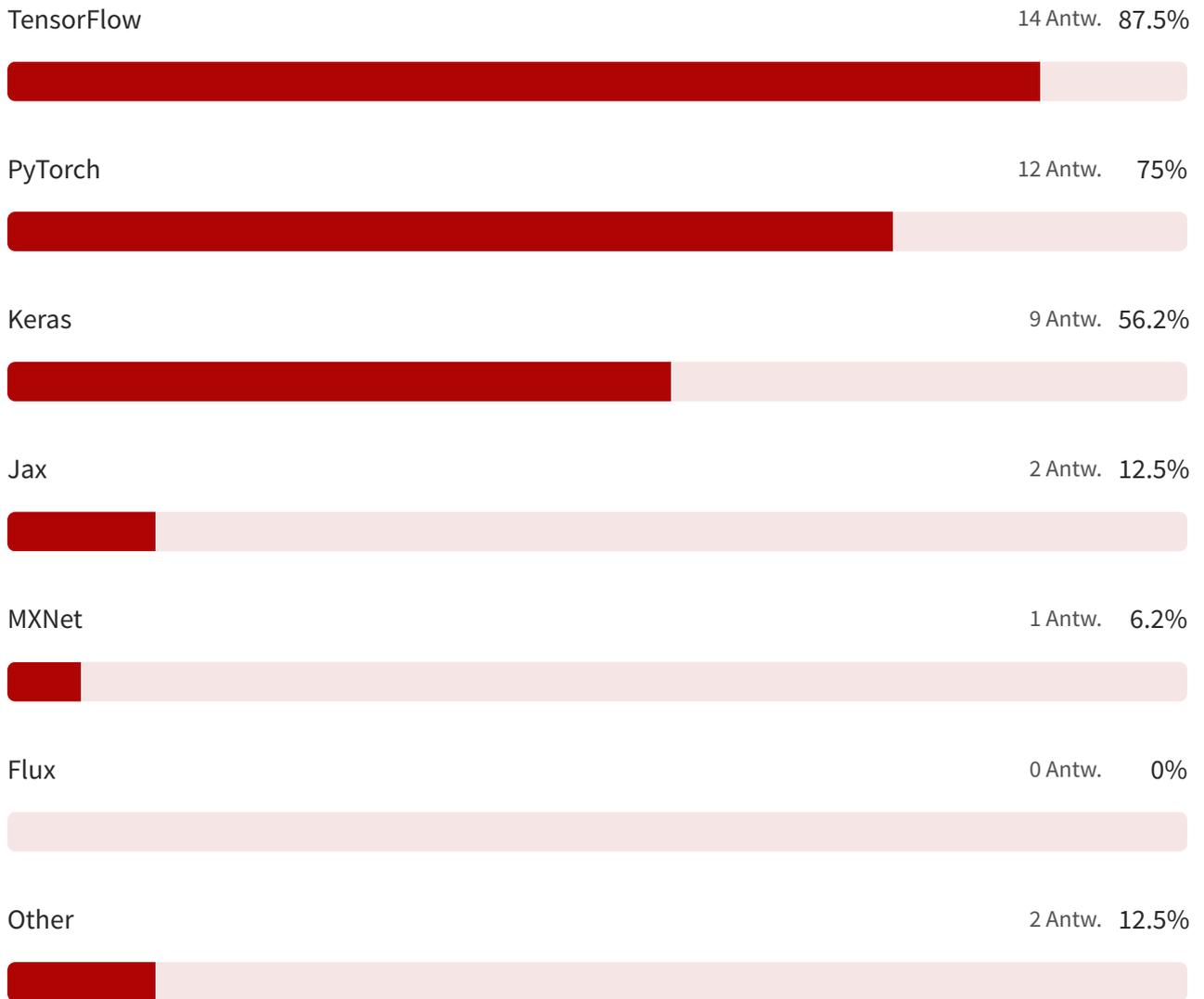
Access Control Levels & Datenhoheit: Wie soll das Laden der Daten in der Entwicklung angesteuert werden?

16 von 71 Personen haben diese Frage beantwortet



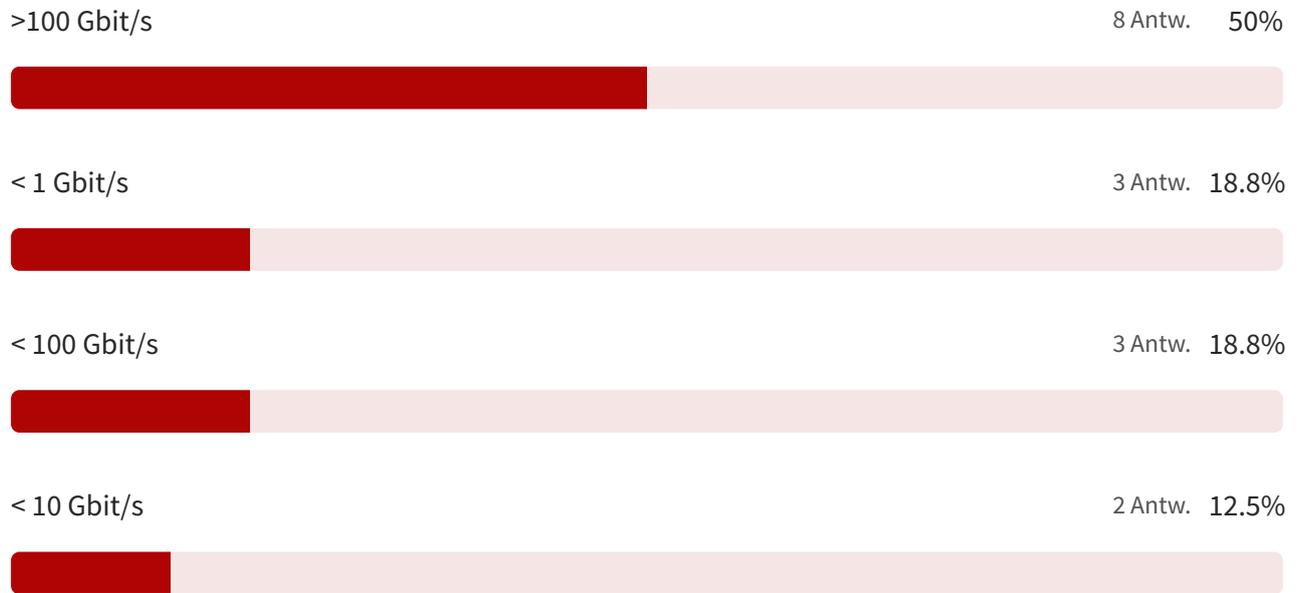
Frameworks: Welche Frameworks sollen bei der Entwicklung von Foundation-Modellen eingesetzt werden?

16 von 71 Personen haben diese Frage beantwortet



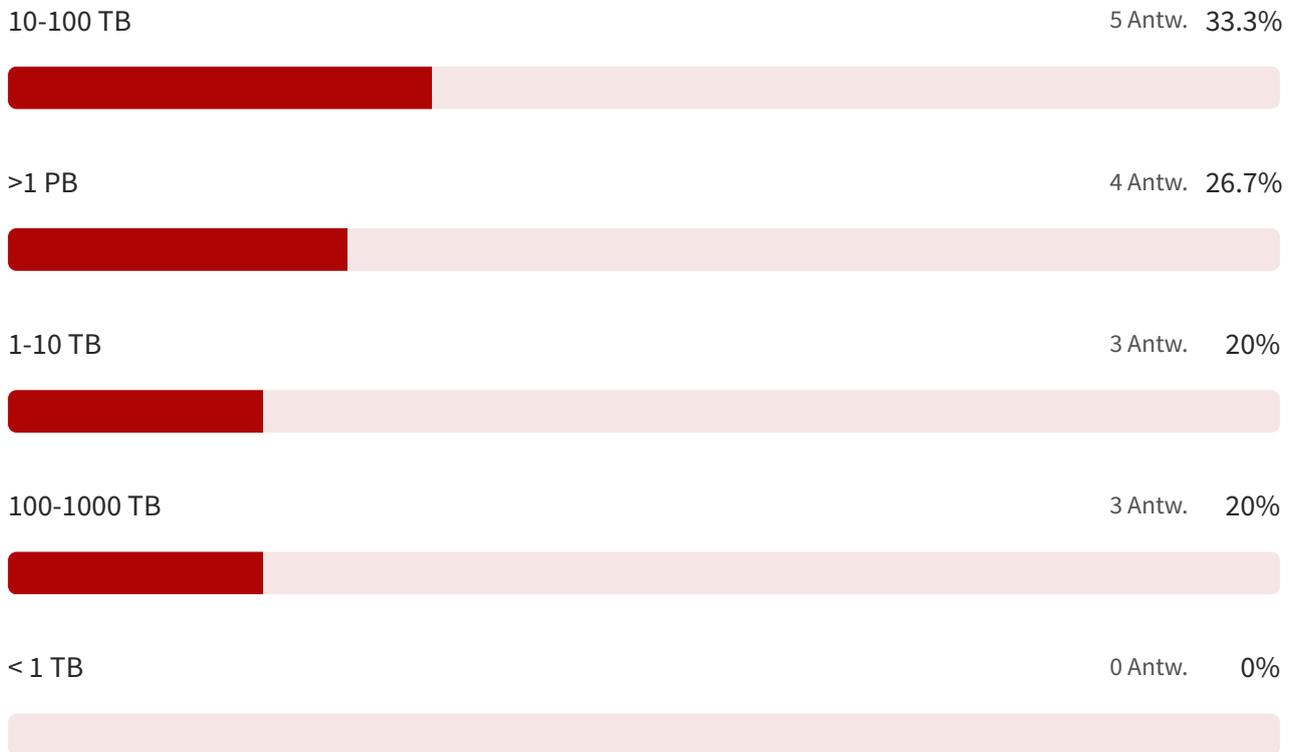
Internet- & Bandbreite: Wie schnell muss die Internetverbindung für die Entwicklung von Foundation-Modellen mindestens sein?

16 von 71 Personen haben diese Frage beantwortet



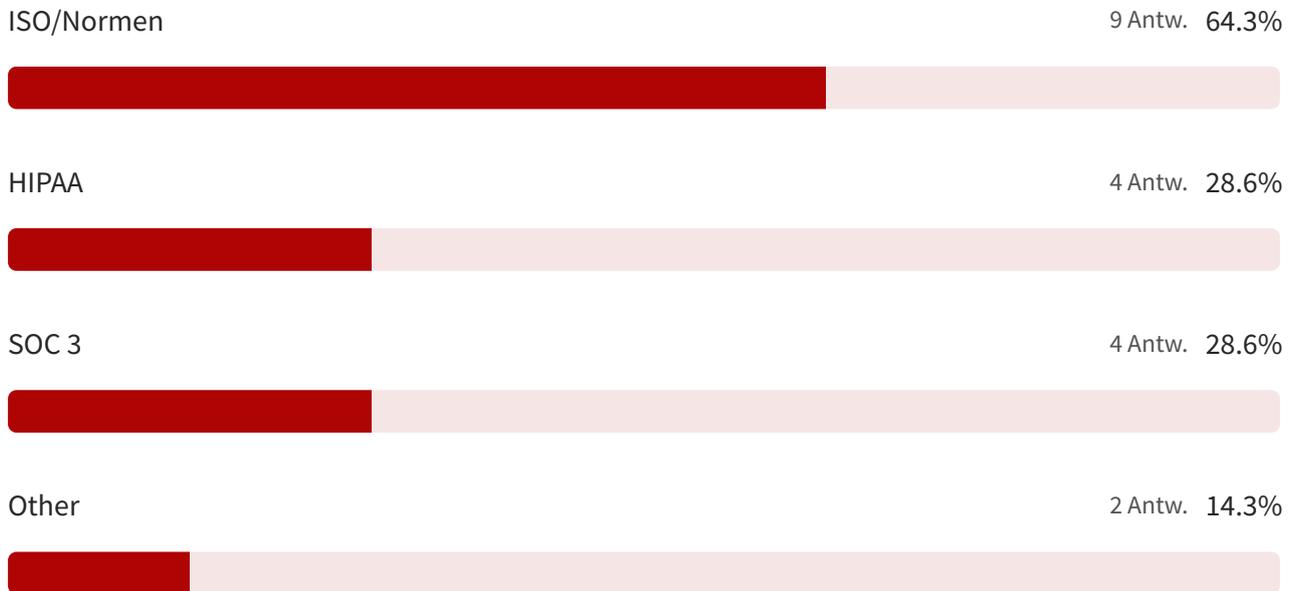
Speicherplatz für Daten: Wie hoch ist der Speicherbedarf für die Entwicklung eines Foundation-Modells?

15 von 71 Personen haben diese Frage beantwortet



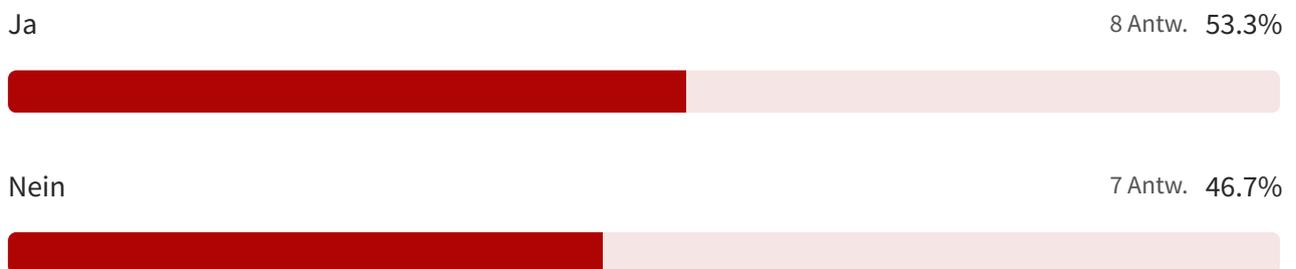
Compliance: Welche regulatorischen Voraussetzungen müssen erfüllt werden, um Foundation-Modelle zu entwickeln?

14 von 71 Personen haben diese Frage beantwortet



Beschäftigen Sie sich mit möglichen Bias, Diskriminierung und Misrepresentation in Daten?

15 von 71 Personen haben diese Frage beantwortet



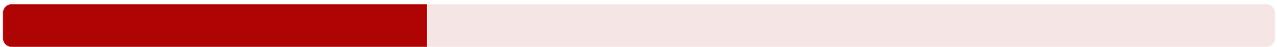
Wissen Sie, wie Sie Bias, Diskriminierung und Misrepresentation hinreichend quantifizieren können, um diese zu adressieren?

9 von 71 Personen haben diese Frage beantwortet

Ja 6 Antw. 66.7%



Nein 3 Antw. 33.3%



Ist Ihnen das Konzept der Model Cards / Data Set Cards im Bezug zur Erstellung neuer Daten geläufig?

16 von 71 Personen haben diese Frage beantwortet

Ja 9 Antw. 56.2%



Nein 7 Antw. 43.8%



Ist das Konzept für Ihre Domäne relevant oder hilfreich?

9 von 71 Personen haben diese Frage beantwortet

Ja

7 Antw. 77.8%



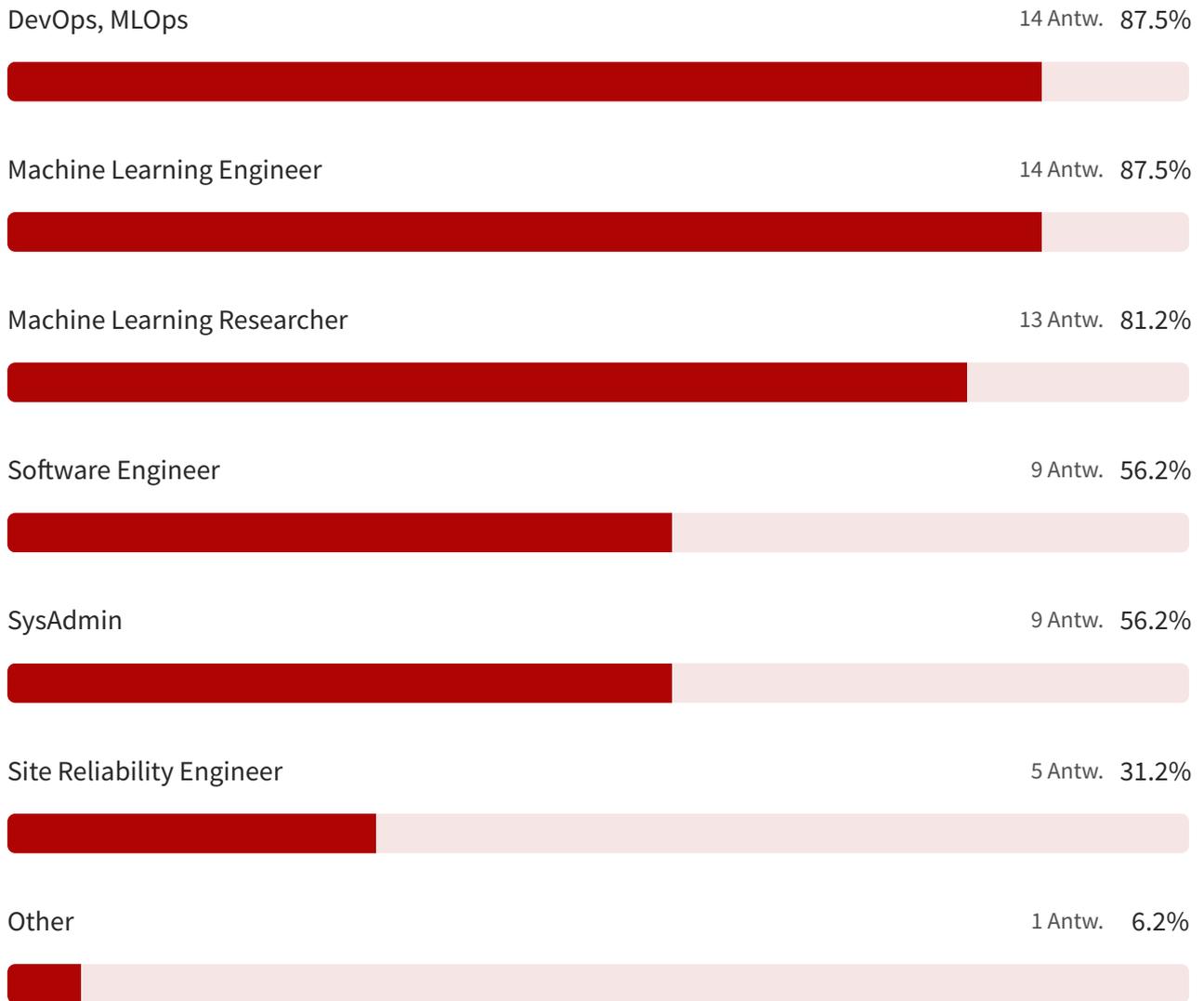
Nein

2 Antw. 22.2%



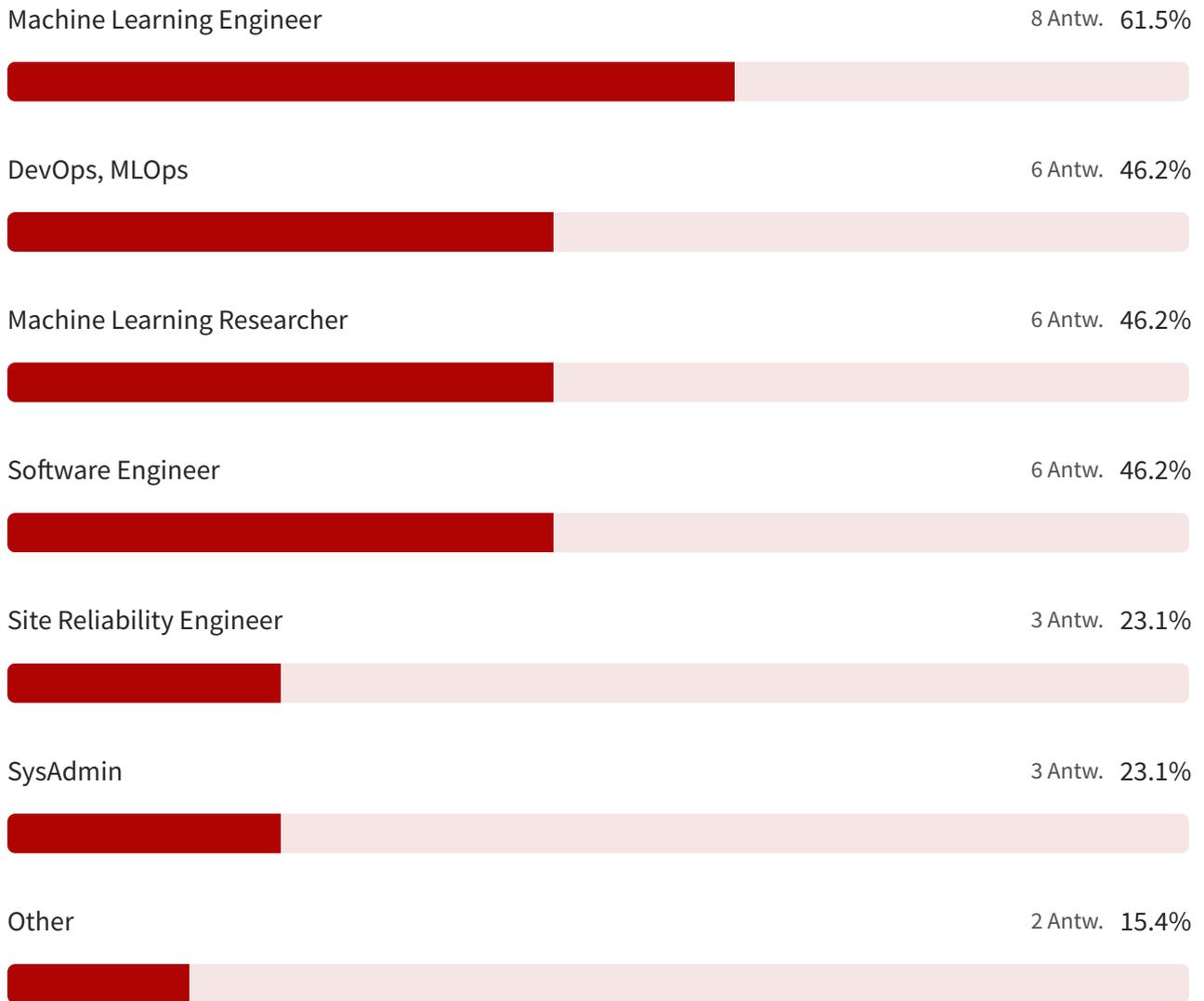
Welche Personalrollen (oder äquivalent) sind für die Entwicklung von Foundation-Modellen erforderlich?

16 von 71 Personen haben diese Frage beantwortet



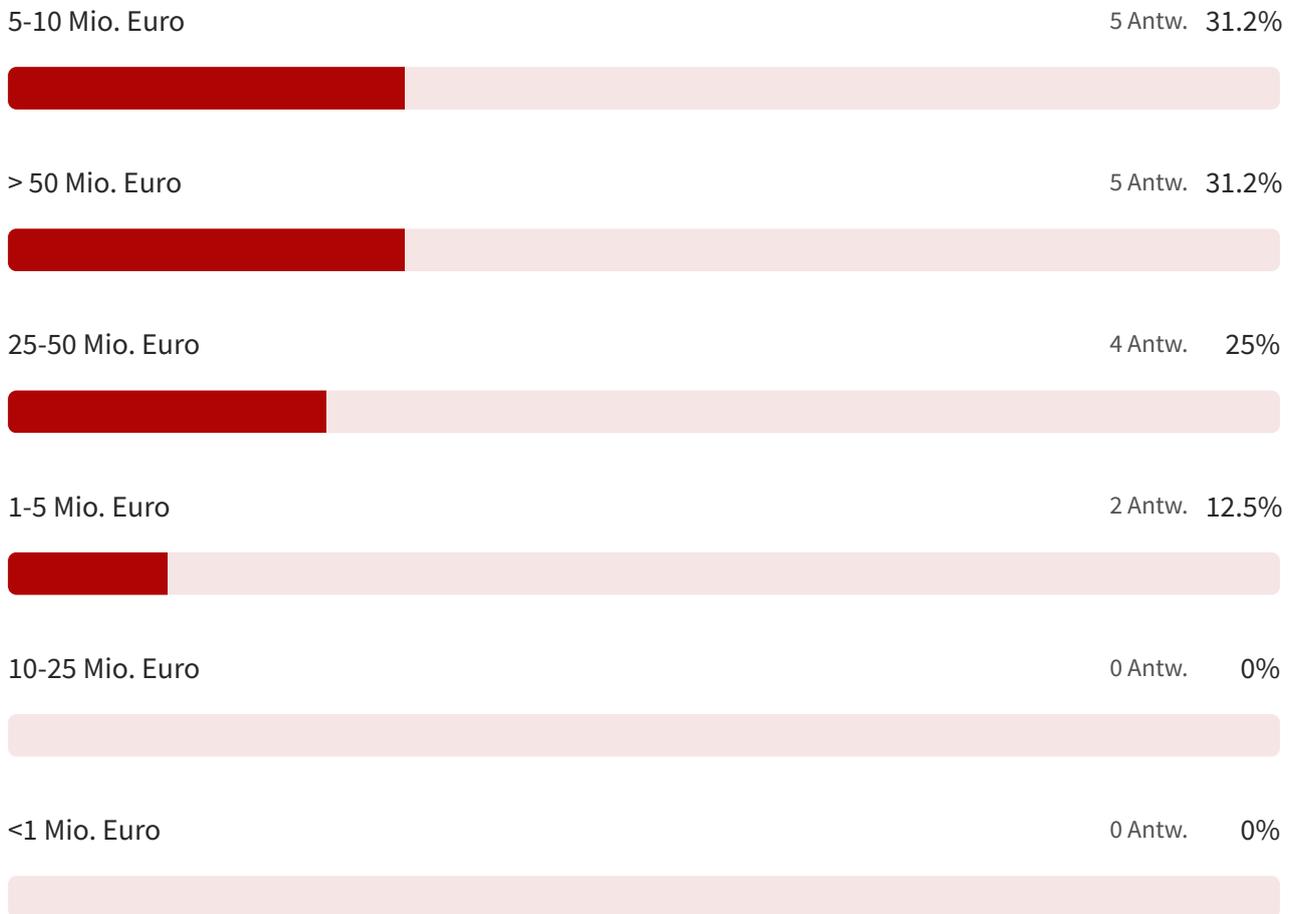
In welchen dieser Personalrollen beobachten oder antizipieren Sie einen Mangel, um Foundation-Modelle zu entwickeln?

13 von 71 Personen haben diese Frage beantwortet



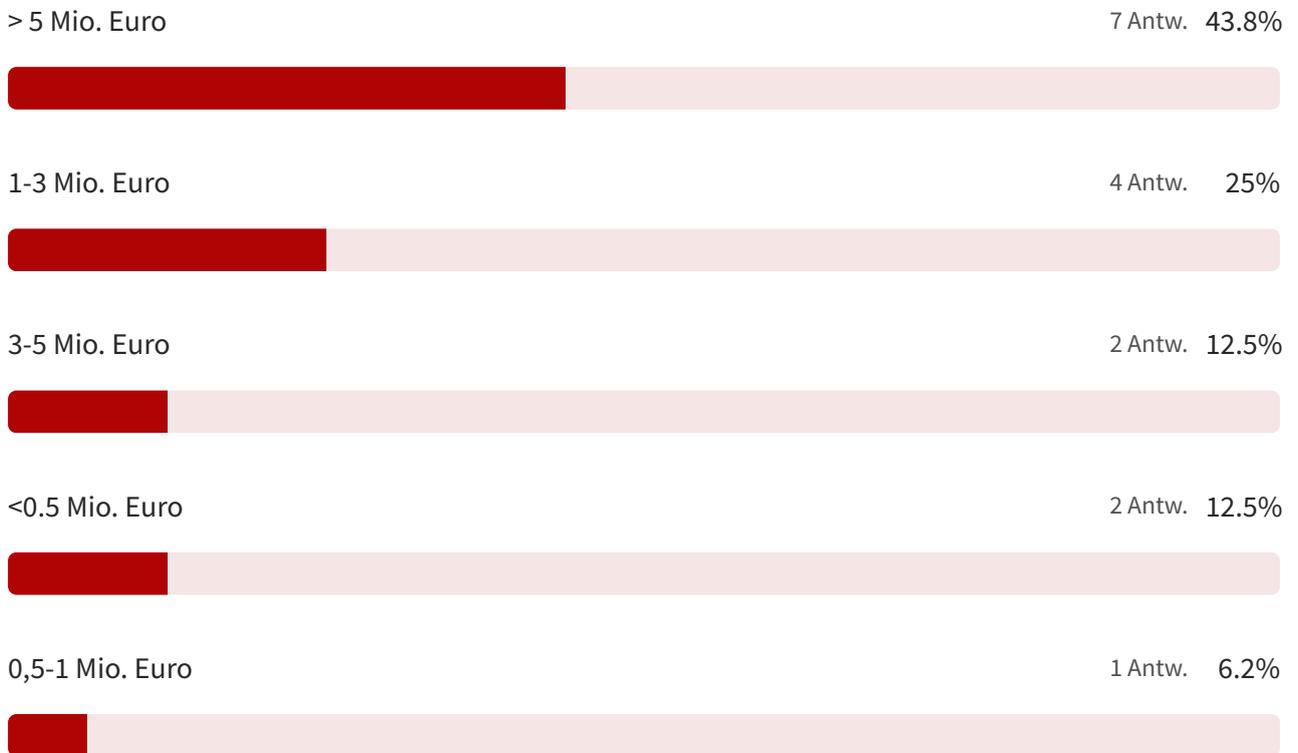
Welche Kosten halten Sie für die Entwicklung eines Foundation-Modells für realistisch?

16 von 71 Personen haben diese Frage beantwortet



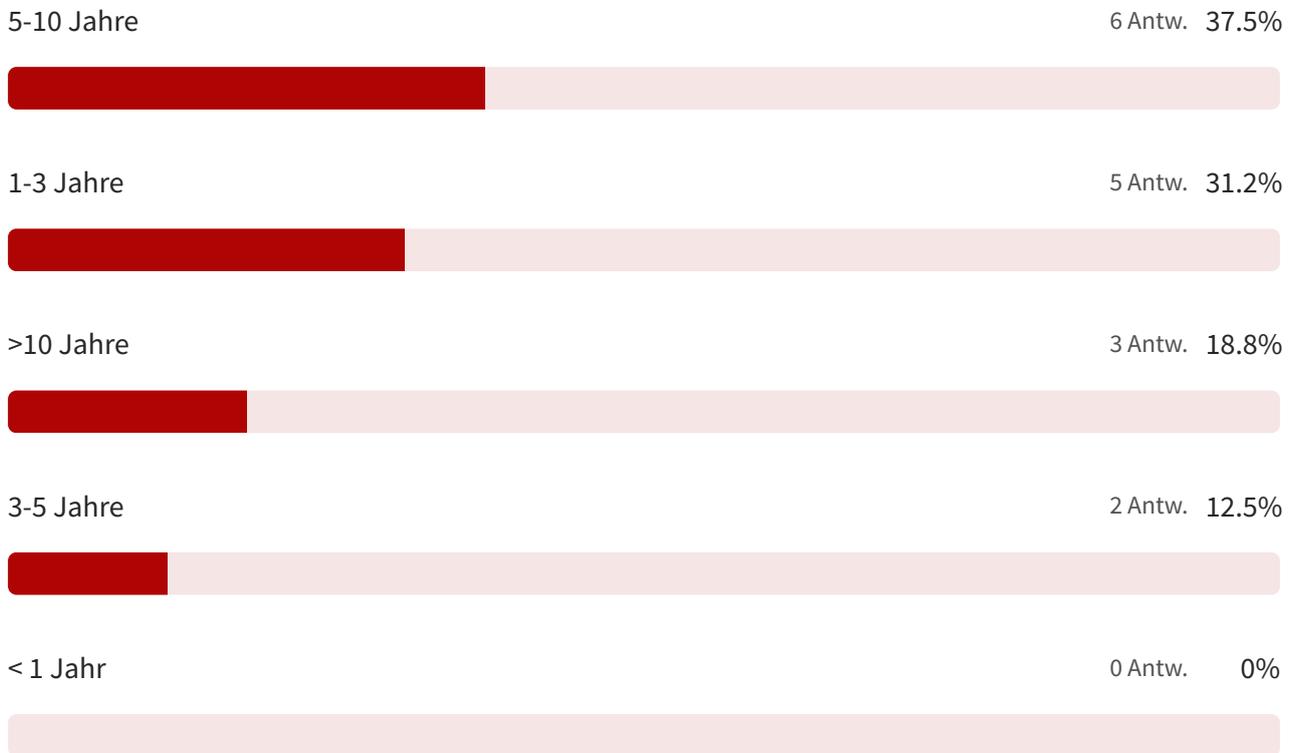
Welche Kosten halten Sie für das Training eines Foundation-Modells für realistisch?

16 von 71 Personen haben diese Frage beantwortet



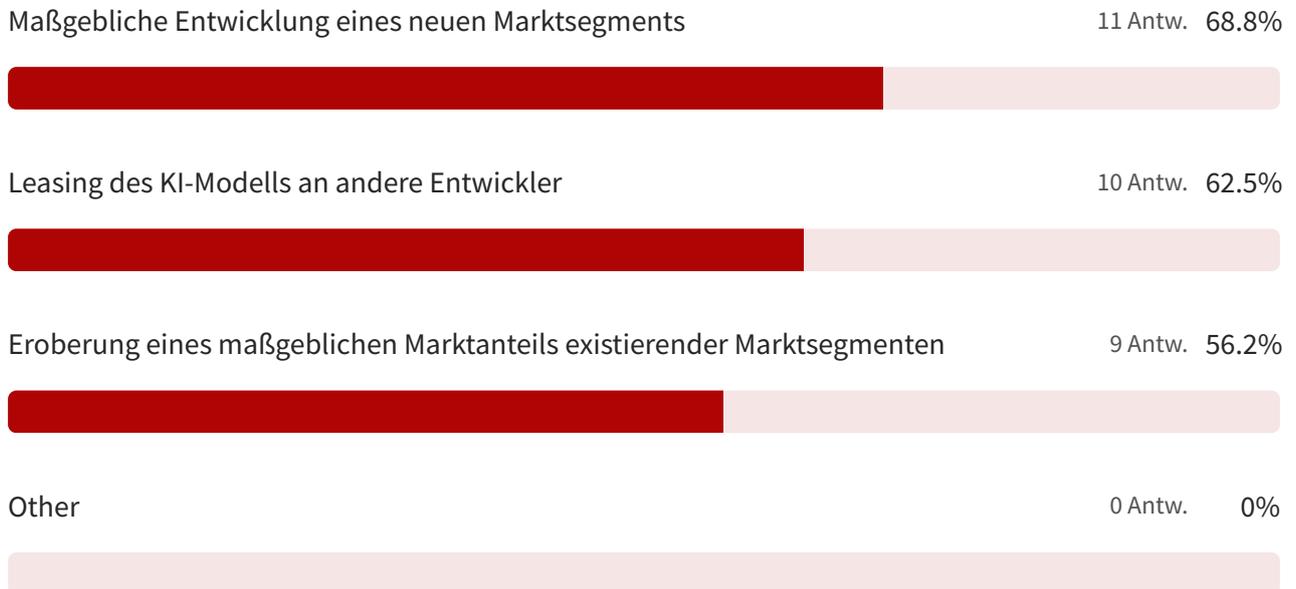
In welchem Zeitraum erwarten Sie, diese Investitionskosten zu amortisieren?

16 von 71 Personen haben diese Frage beantwortet



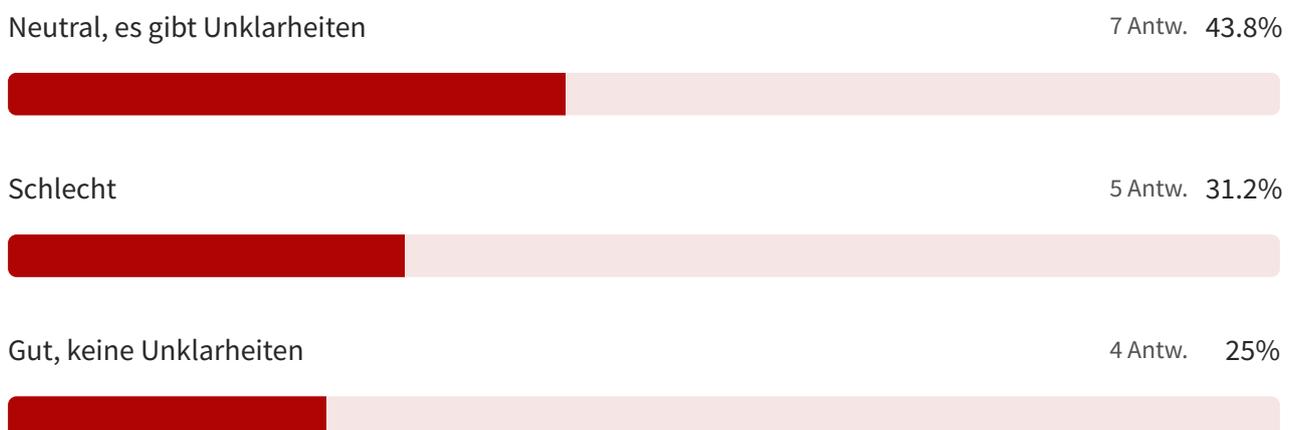
Wie erwarten Sie, dass sich diese Investitionskosten amortisieren könnten?

16 von 71 Personen haben diese Frage beantwortet



Wie sind Sie über die Regulation der Entwicklung von Foundation-Modellen informiert?

16 von 71 Personen haben diese Frage beantwortet



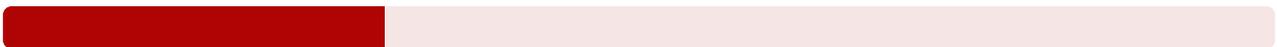
Empfinden Sie Ihr Wissen über die Regulation als Enabler oder Disabler für Ihren möglichen Einsatz von Foundation-Modellen?

10 von 71 Personen haben diese Frage beantwortet

Enabler 7 Antw. 70%



Disabler 3 Antw. 30%



Wie sollten Ihrer Meinung nach Datensätze für die Entwicklung von Foundation-Modellen erhoben werden?

16 von 71 Personen haben diese Frage beantwortet

Donation 13 Antw. 81.2%



Scraping 12 Antw. 75%



Centralized Third-Party collection 8 Antw. 50%

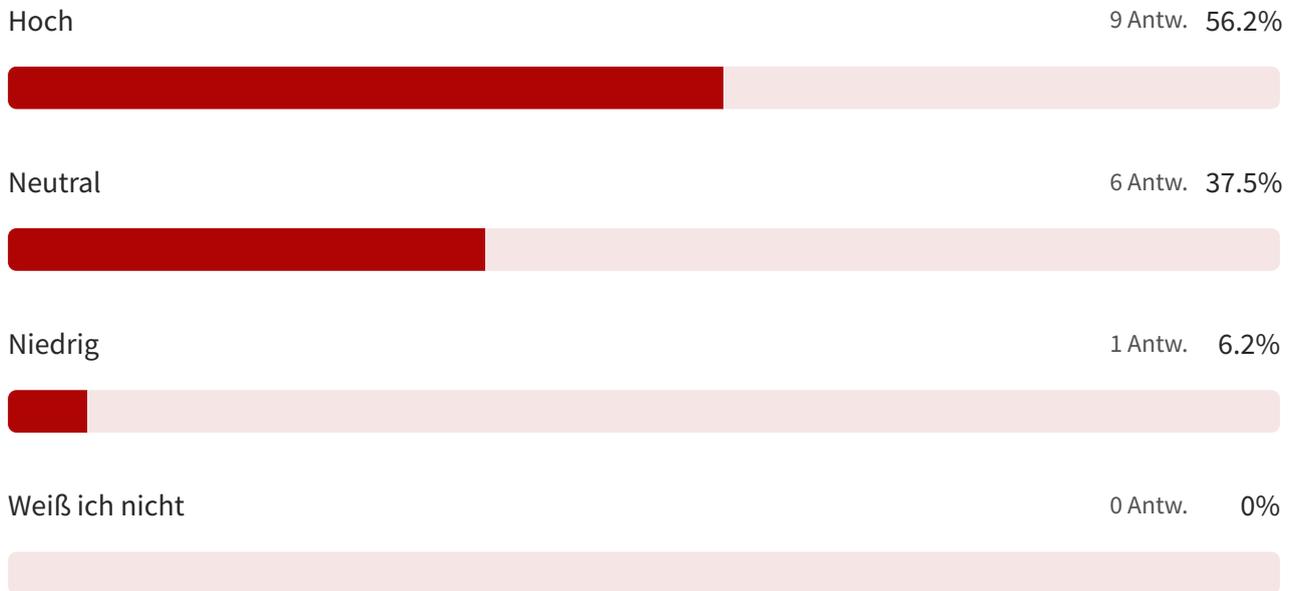


Other 3 Antw. 18.8%



Wie hoch empfinden Sie Awareness, Relevanz und Akzeptanz bzgl. Foundation-Modellen innerhalb Ihrer Organisation?

16 von 71 Personen haben diese Frage beantwortet



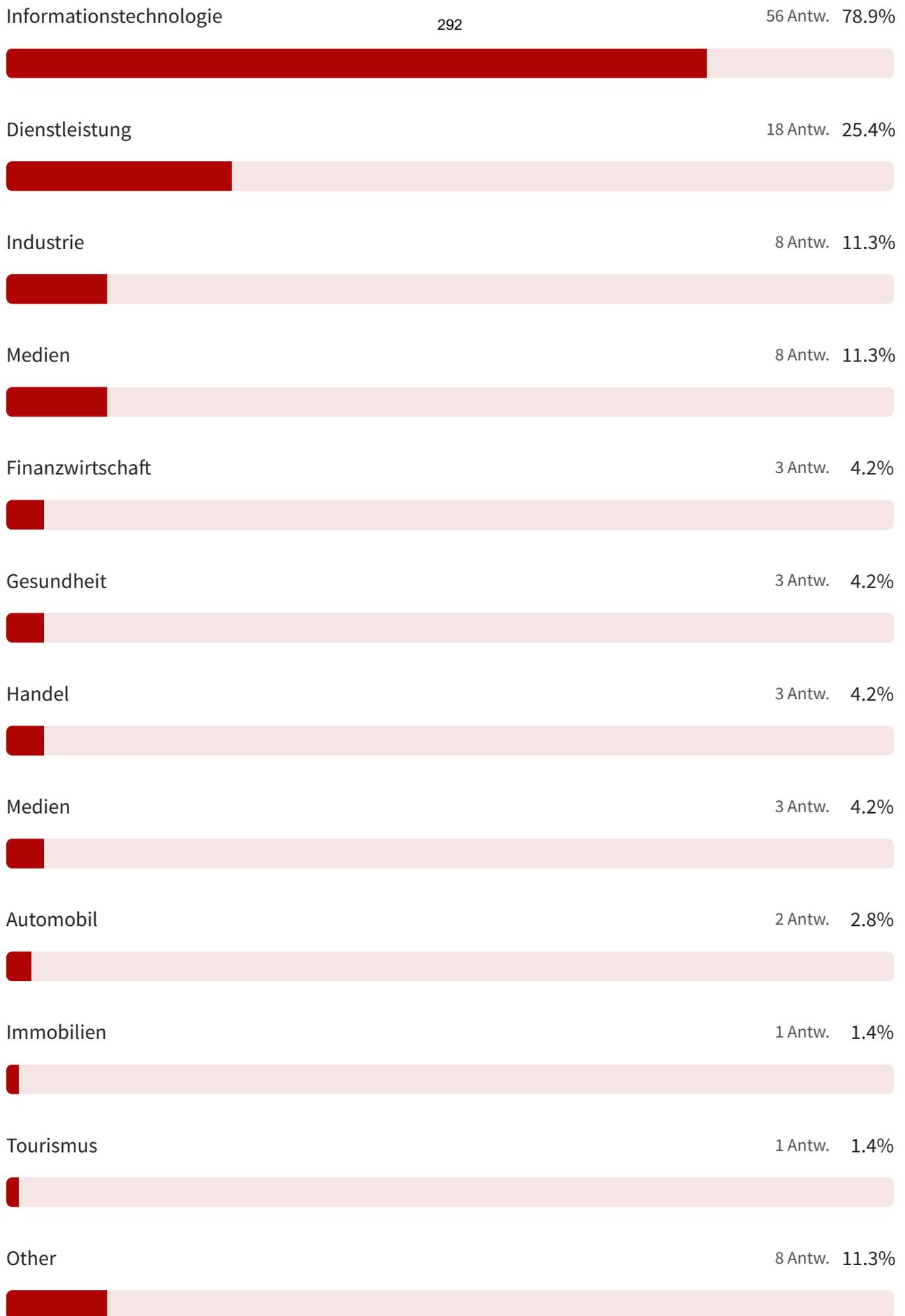
Wie groß ist Ihr Unternehmen?

71 von 71 Personen haben diese Frage beantwortet



Welcher Branche gehört Ihr Unternehmen an?

71 von 71 Personen haben diese Frage beantwortet



Stehen Sie für einen weiteren Austausch zu den Themen Foundation Modelle, digitale Souveränität sowie KI im allgemeinen zur Verfügung?

71 von 71 Personen haben diese Frage beantwortet

Ja 61 Antw. 85.9%



Nein 10 Antw. 14.1%



LEAM Umfrage Große Unternehmen

10 Antworten

Inwiefern haben Sie sich in Ihrem Unternehmen bereits mit Foundation Modellen oder darauf basierenden Applikationen auseinandergesetzt?

10 von 10 Personen haben diese Frage beantwortet

Wir setzen bereits Foundation Modelle bzw. darauf basierende Applikationen ein

5 Antw. 50%



Wir haben angefangen, uns mit Foundation Modellen bzw. darauf basierenden Applikationen auseinanderzusetzen

3 Antw. 30%



Wir beabsichtigen aktuell nicht, Foundation Modelle bzw. darauf basierende Applikationen einzusetzen

1 Antw. 10%



Wir planen aktuell, Foundation Modellen bzw. darauf basierende Applikationen einzusetzen

1 Antw. 10%



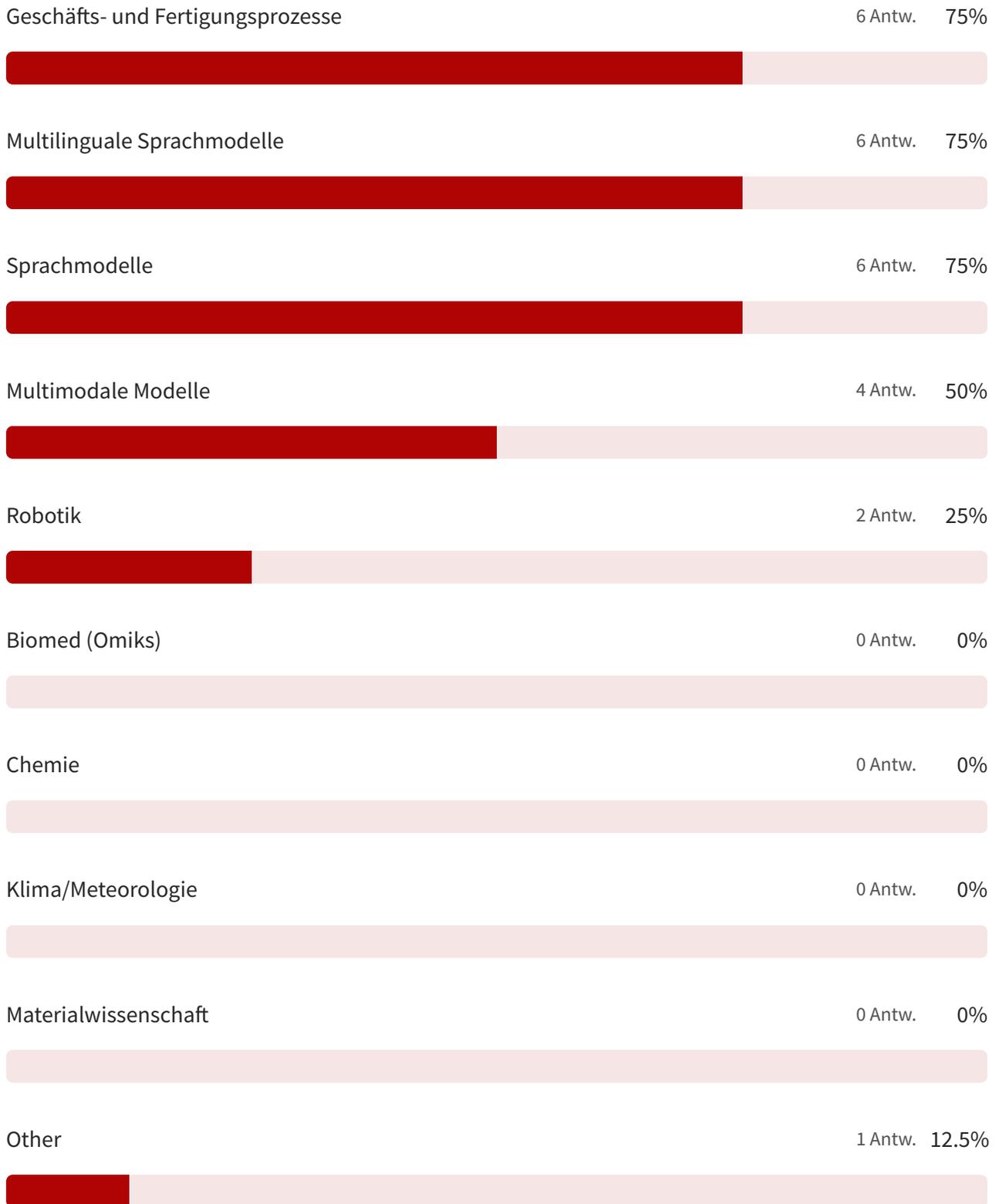
Wir kennen das Thema noch nicht

0 Antw. 0%



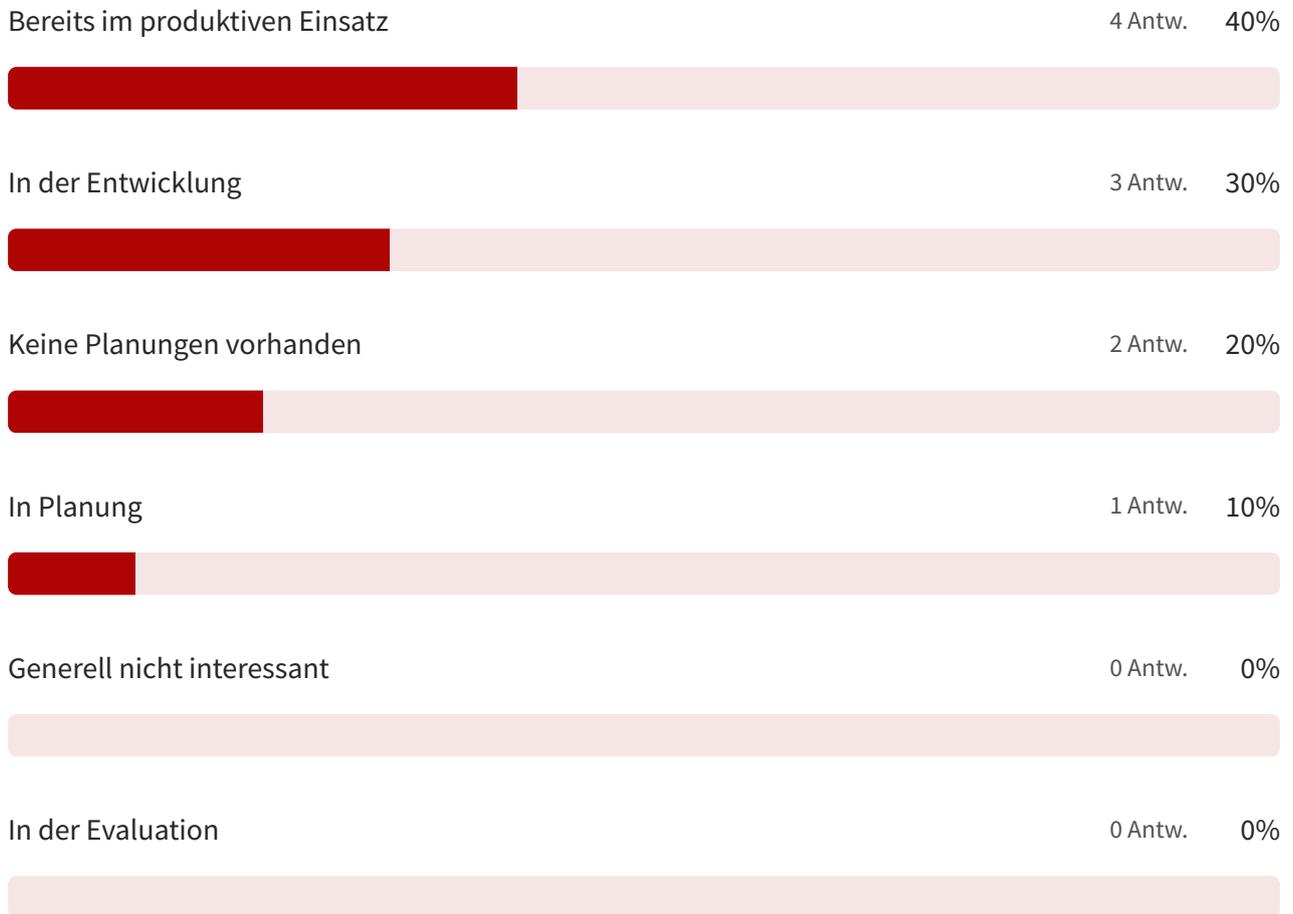
In welchen Bereichen sind für Sie Foundation Modelle in Nutzung und Entwicklung relevant?

8 von 10 Personen haben diese Frage beantwortet



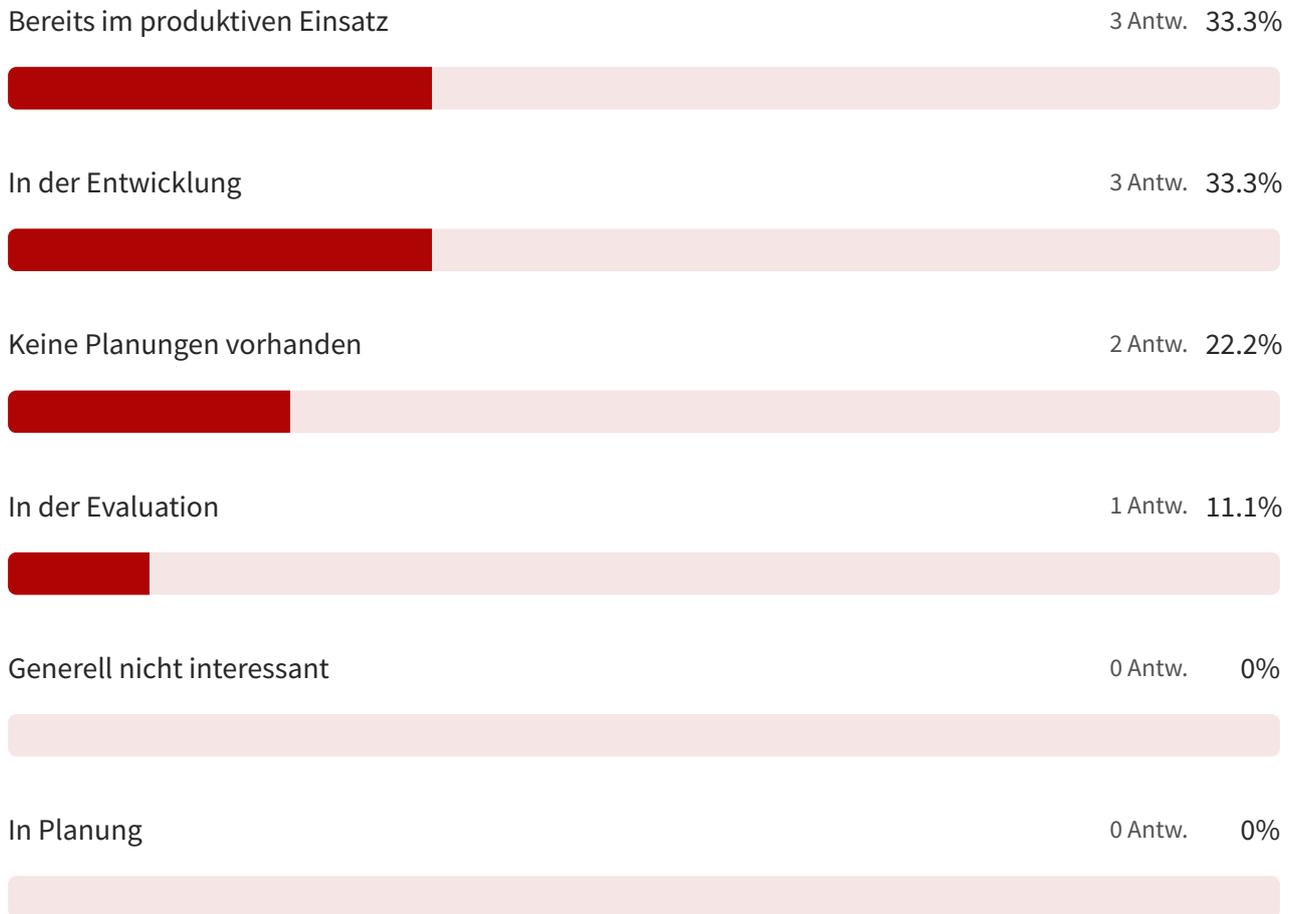
Inwiefern entwickeln Sie eigene Anwendungen auf Basis von Foundation Modellen?

10 von 10 Personen haben diese Frage beantwortet



Inwiefern passen Sie existierende Foundation Modelle an (Tuning)?

9 von 10 Personen haben diese Frage beantwortet



Inwiefern entwickeln Sie eigene Foundation Modelle?

10 von 10 Personen haben diese Frage beantwortet



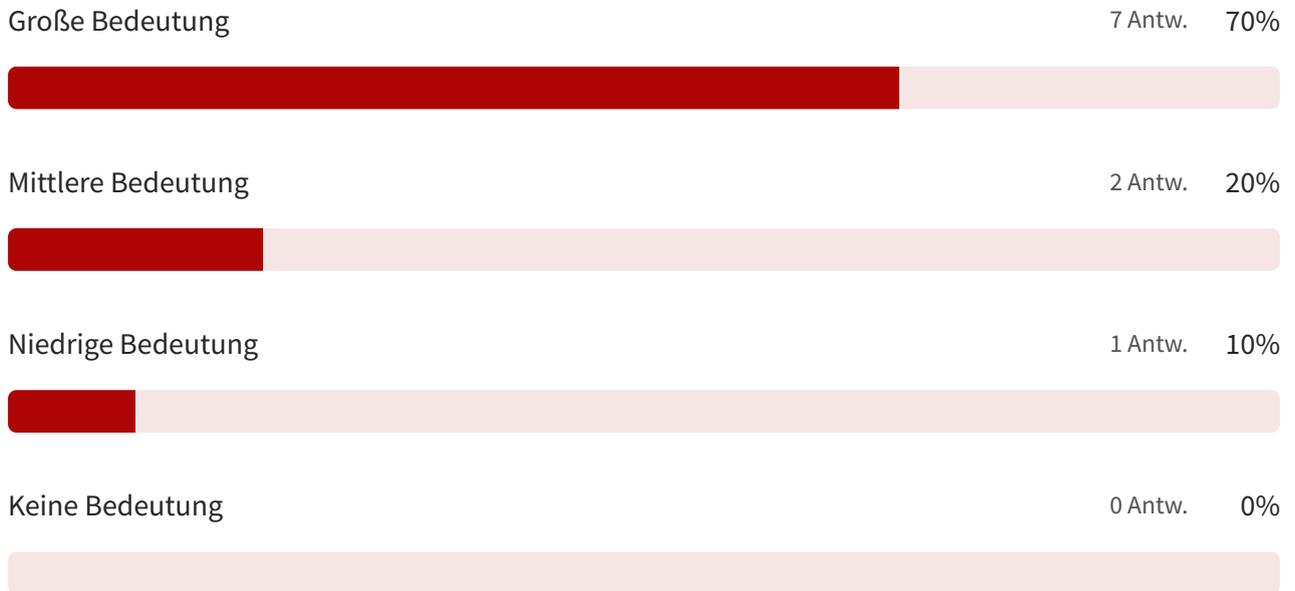
Welche Bedeutung messen Sie der Datenanalyse und KI für die gesamtwirtschaftliche Entwicklung in Deutschland zu?

10 von 10 Personen haben diese Frage beantwortet



Welche Bedeutung messen Sie den Foundation Modellen für die KI und damit für gesamtwirtschaftliche Entwicklung in Deutschland zu?

10 von 10 Personen haben diese Frage beantwortet



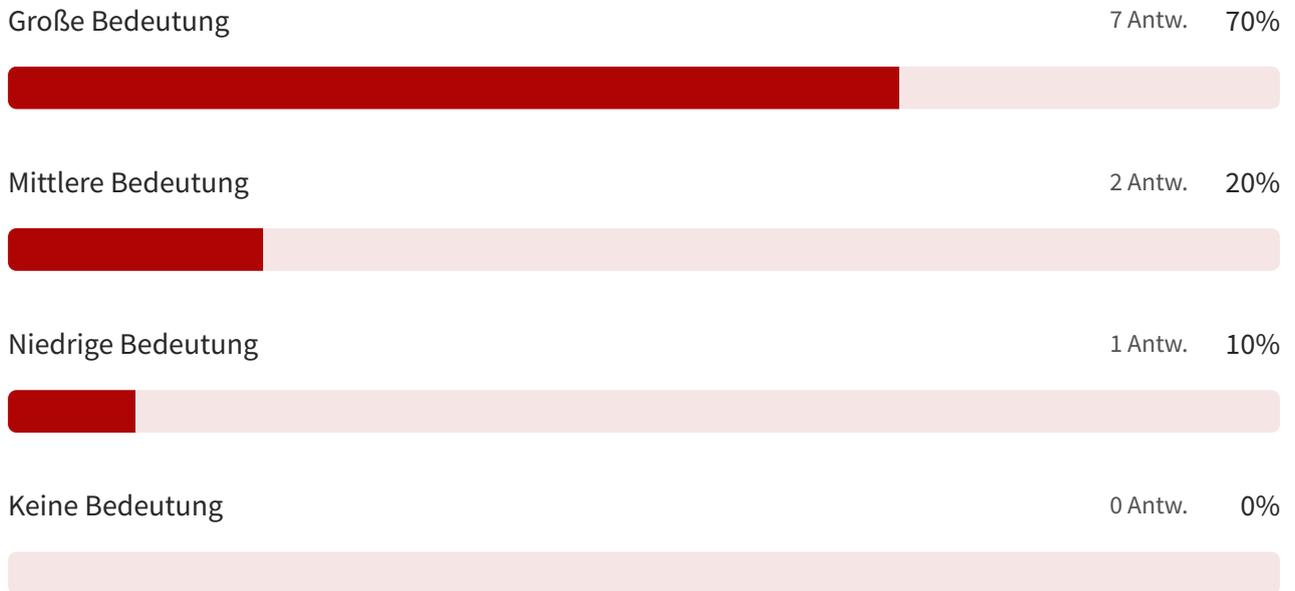
Wie wichtig ist aus Ihrer Sicht der Aufbau eines deutschen und europäischen Ökosystems rund um Foundation Models für die Digitale Souveränität und Wettbewerbsfähigkeit?

9 von 10 Personen haben diese Frage beantwortet



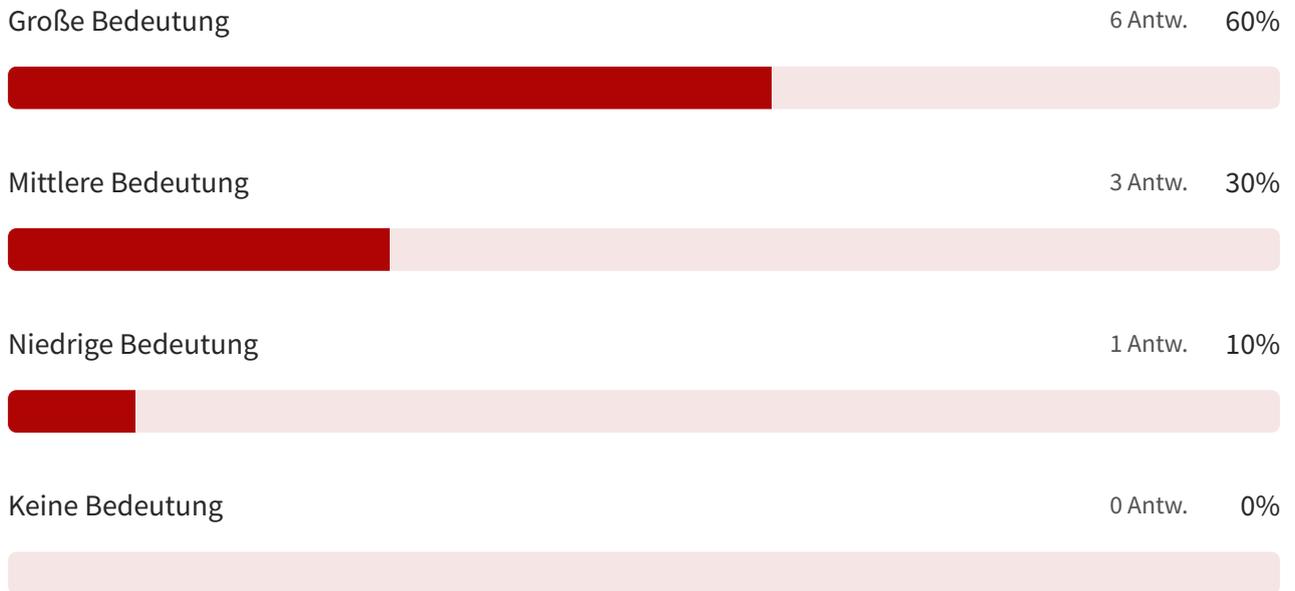
Welche Bedeutung hat für Sie die Verfügbarkeit von Foundation Modellen, die in Europa entwickelt wurden und Werte wie Transparenz, Reduktion von Bias und Nachhaltigkeit berücksichtigen?

10 von 10 Personen haben diese Frage beantwortet



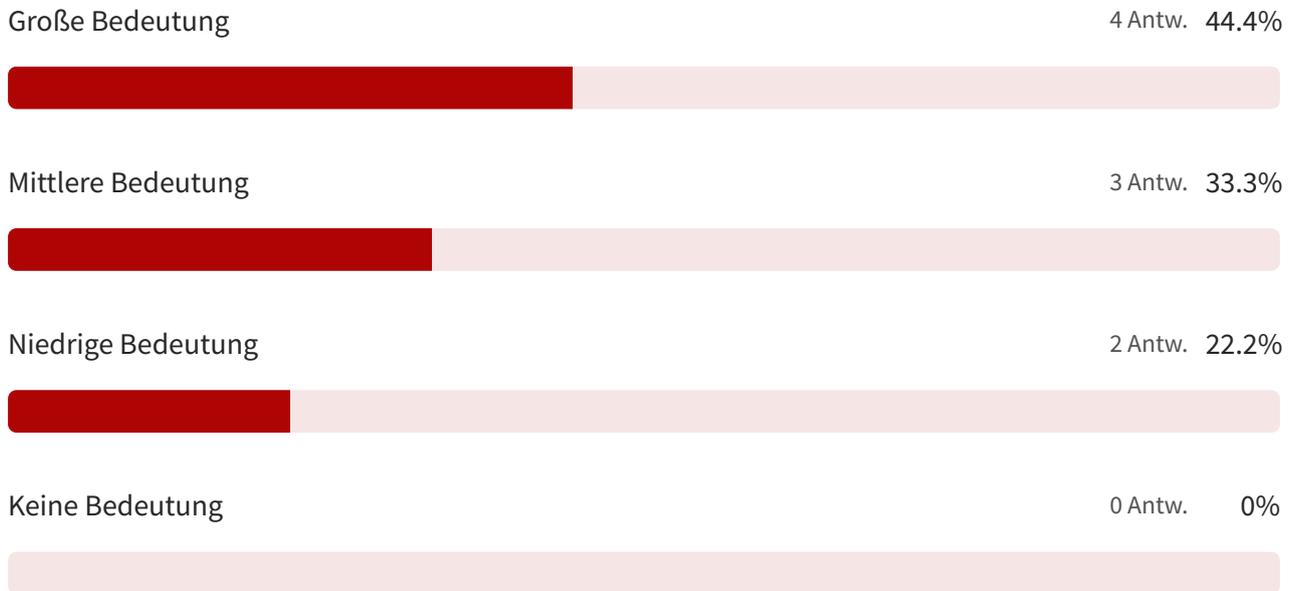
Welche Bedeutung hat für Sie, bzgl. der Entwicklung/Anwendung von Foundation-Modellen, die Zusammenarbeit mit wissenschaftlichen Institutionen?

10 von 10 Personen haben diese Frage beantwortet



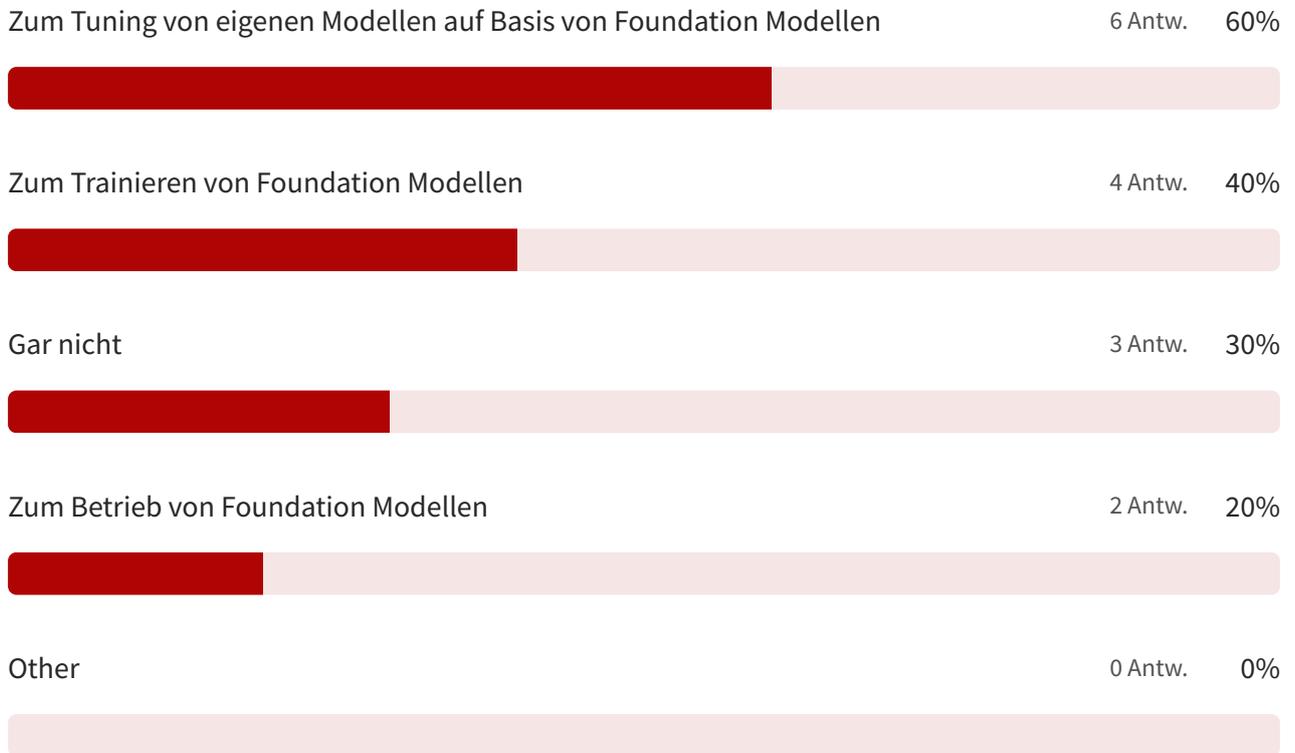
Welche Bedeutung hat für Sie, bzgl. der Entwicklung/Anwendung von Foundation-Modellen, die Zusammenarbeit mit spezialisierten KMUs und Startups?

9 von 10 Personen haben diese Frage beantwortet



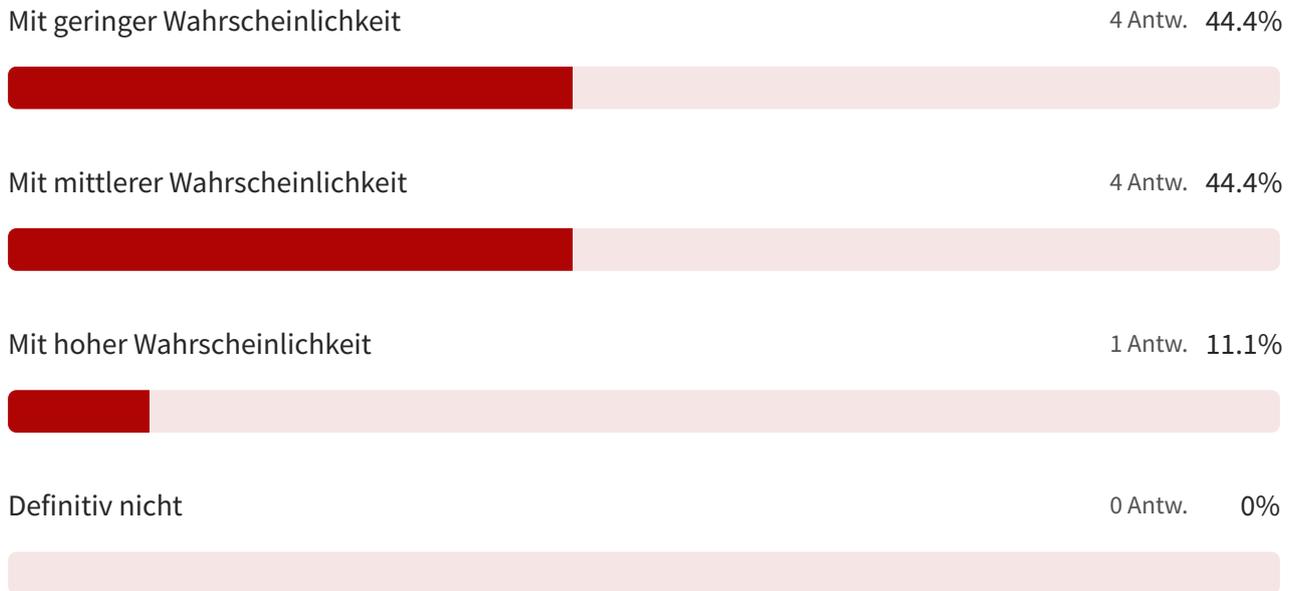
Inwiefern würden Sie die Services von LEAM nutzen?

10 von 10 Personen haben diese Frage beantwortet



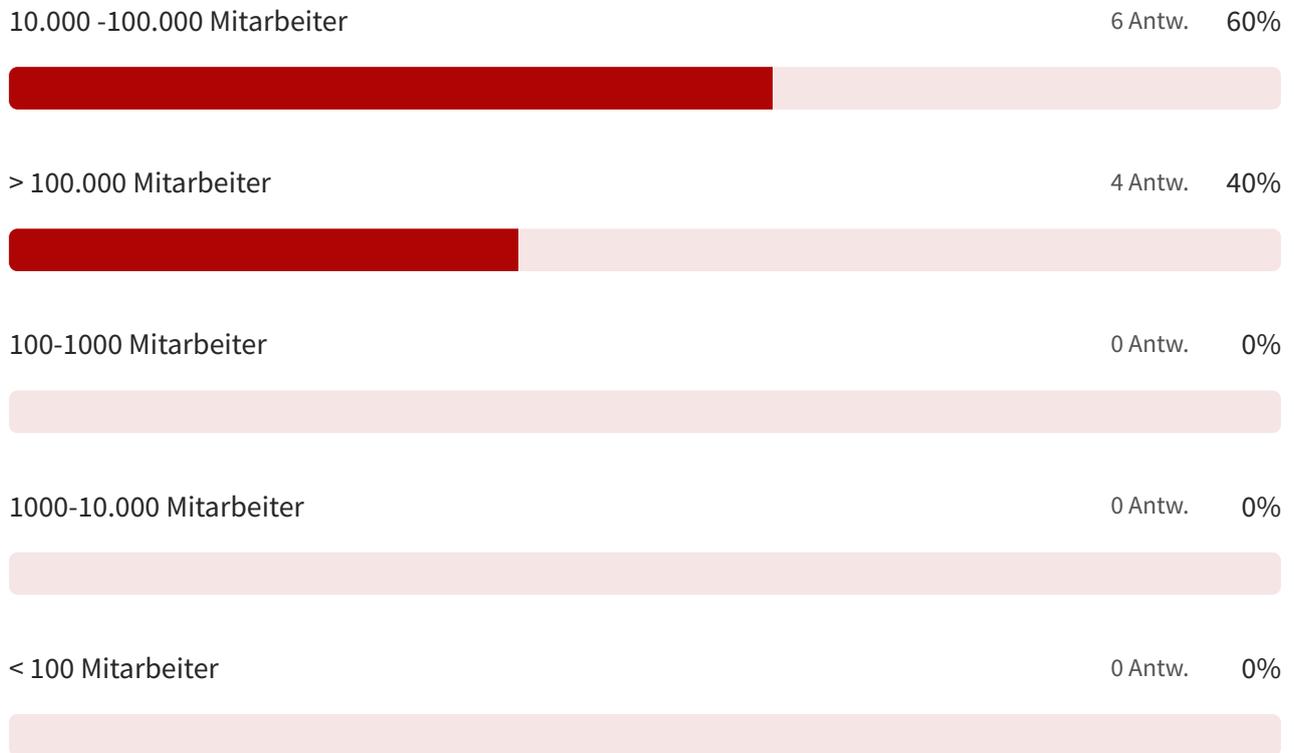
Inwiefern würden Sie bzw. Ihr Unternehmen in den Aufbau eines LEAM KI-Servicezentrums investieren (unter noch zu klärenden Voraussetzungen)?

9 von 10 Personen haben diese Frage beantwortet



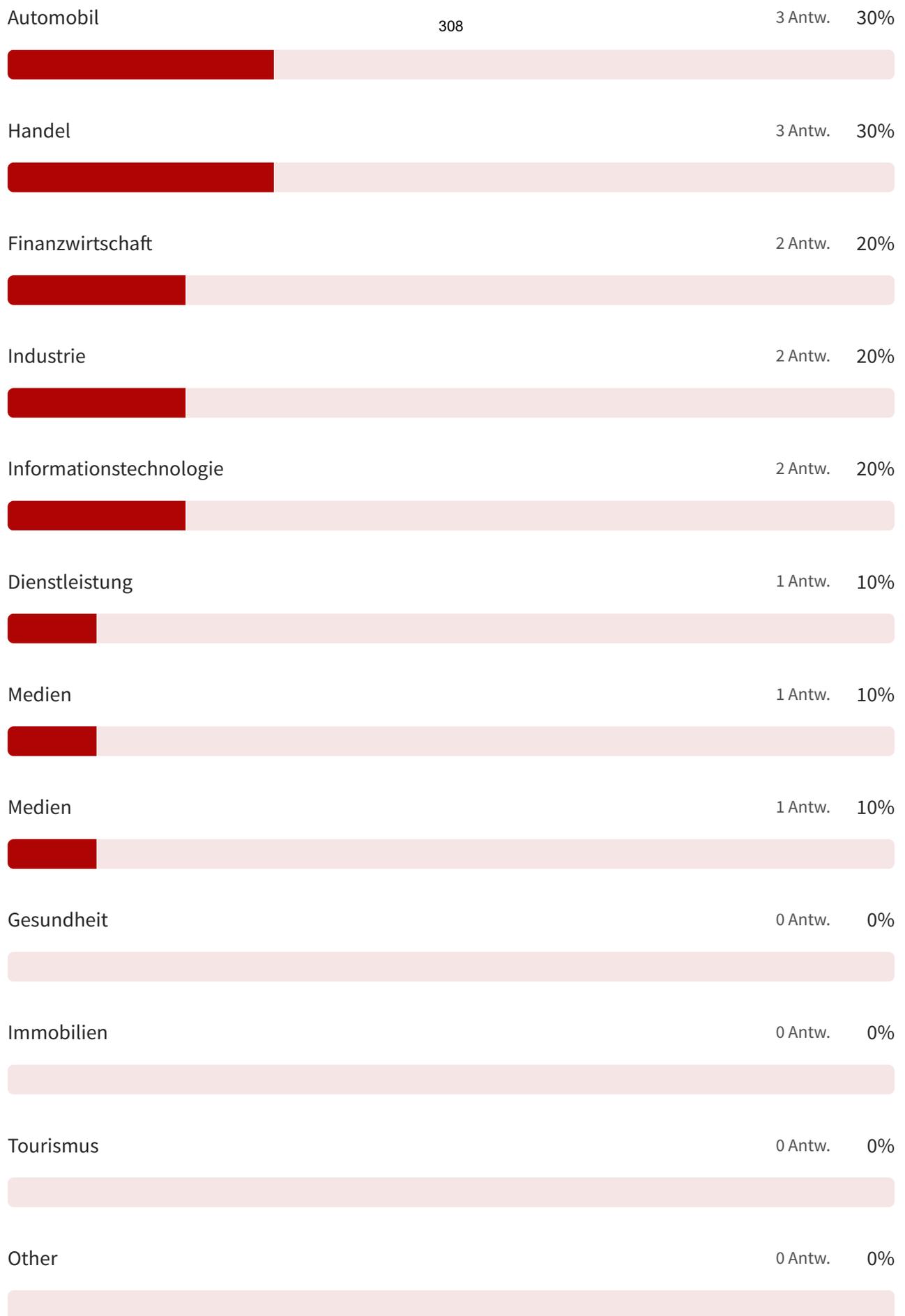
Wie groß ist Ihr Unternehmen?

10 von 10 Personen haben diese Frage beantwortet



Welcher Branche gehört Ihr Unternehmen an?

10 von 10 Personen haben diese Frage beantwortet



Stehen Sie für einen weiteren Austausch zu den Themen Foundation Modelle, digitale Souveränität sowie KI im allgemeinen zur Verfügung?

10 von 10 Personen haben diese Frage beantwortet

Ja

6 Antw. 60%



Nein

4 Antw. 40%

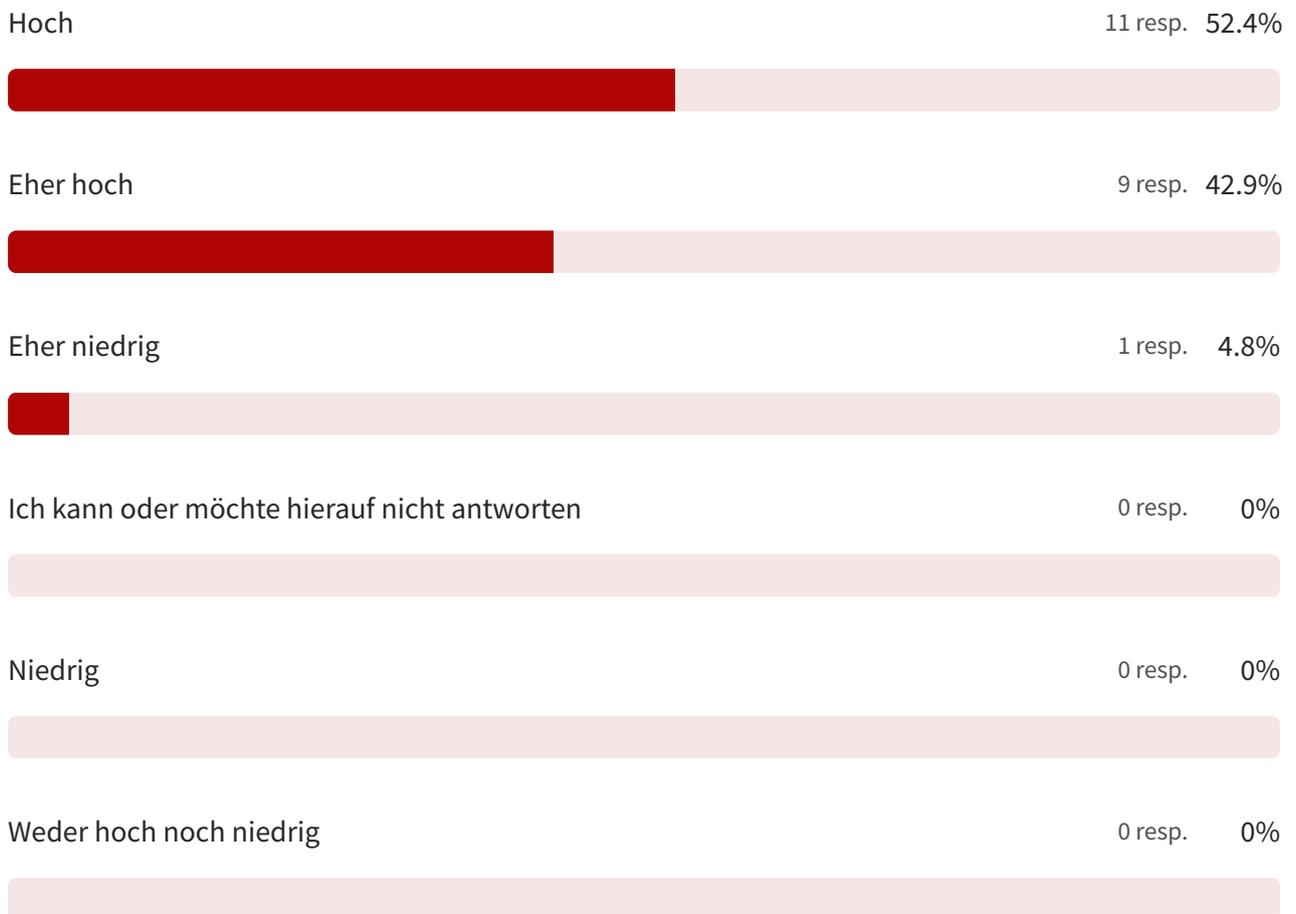


LEAM Umfrage - KI-Forscher

21 responses

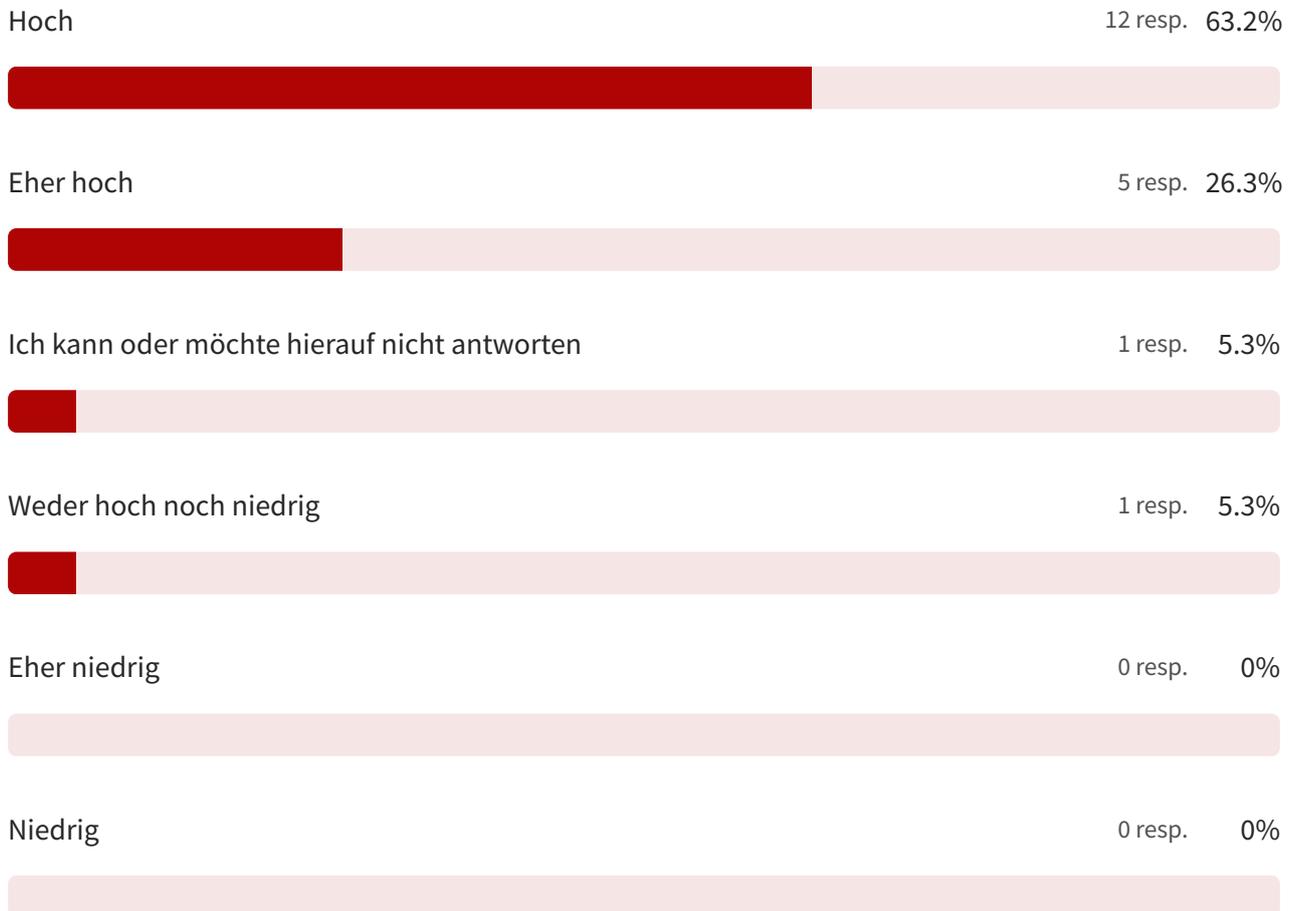
Wie schätzen Sie die Bedeutung von KI-Foundation-Modellen für die Wissenschaft ein?

21 out of 21 answered



Wie schätzen Sie die Bedeutung von KI-Foundation-Modellen für die Wirtschaft und andere Bereiche der Gesellschaft ein?

19 out of 21 answered



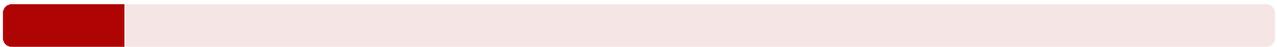
Arbeiten Sie bereits an oder wollen Sie mit normalen KI-Modellen (nicht Foundation-Modellen) arbeiten?

21 out of 21 answered

Bereits recht viel 17 resp. 81%



Schon etwas 2 resp. 9.5%



Ich kann oder möchte hierauf nicht antworten 1 resp. 4.8%



Nein 1 resp. 4.8%



Es ist geplant 0 resp. 0%

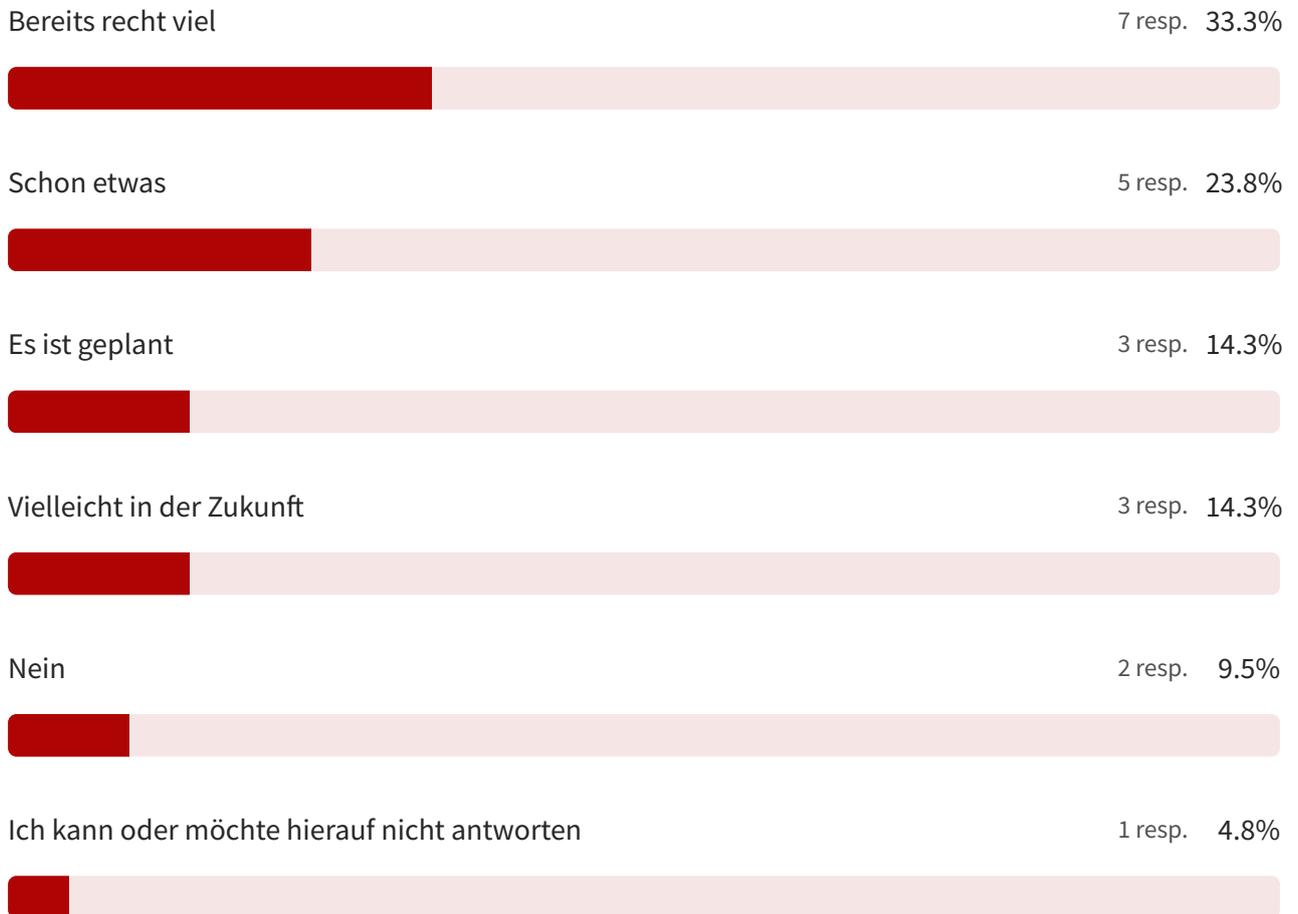


Vielleicht in der Zukunft 0 resp. 0%



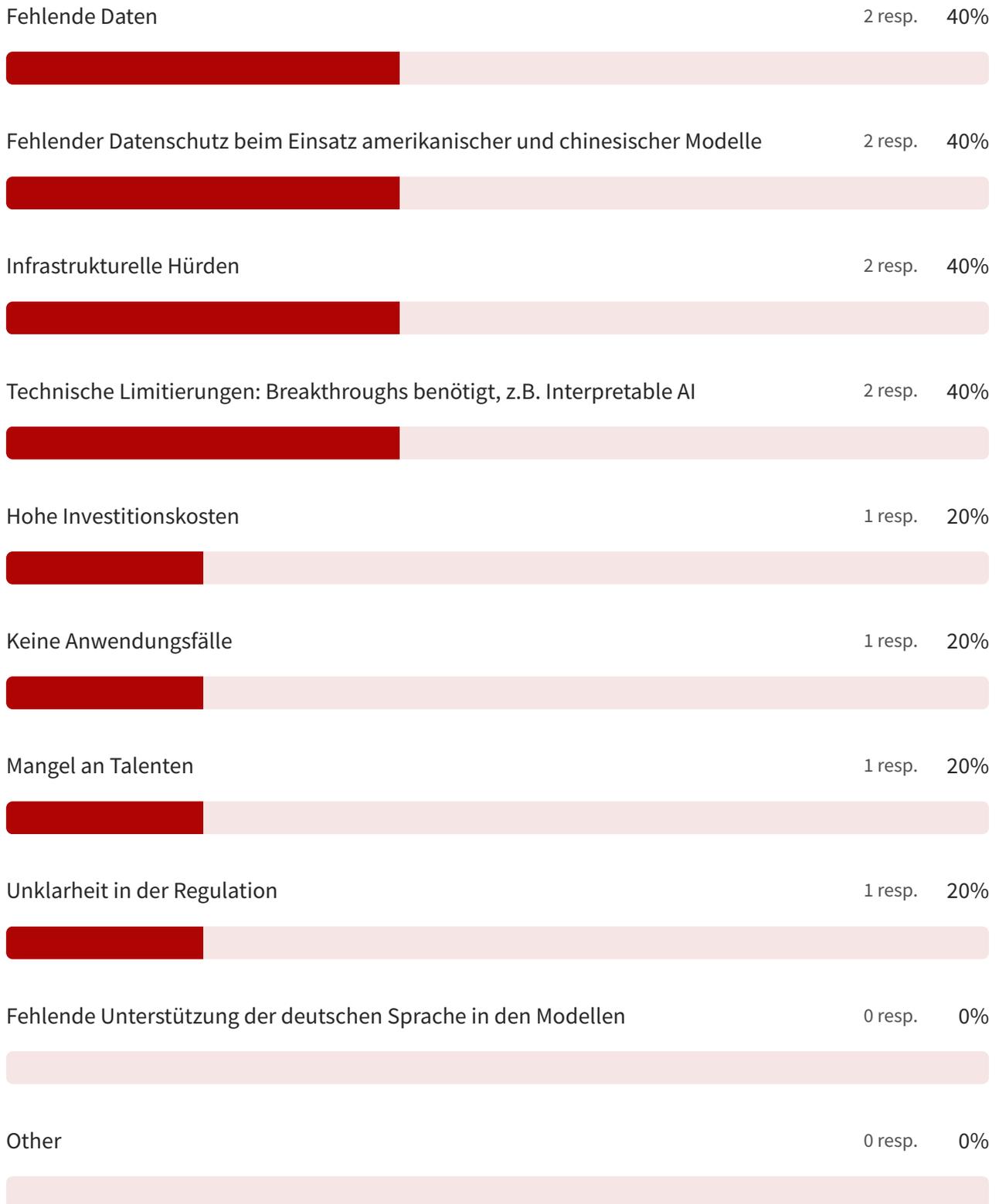
Arbeiten Sie bereits an oder wollen Sie mit Foundation-Modellen arbeiten?

21 out of 21 answered



Was hindert Sie daran, bereits heute Foundation-Modelle einzusetzen?

5 out of 21 answered



Wie hoch schätzen Sie die Wahrscheinlichkeit ein, Foundation-Modelle (intensiver) für Ihre Arbeit zu nutzen?

21 out of 21 answered

Hoch 9 resp. 42.9%



Eher hoch 5 resp. 23.8%



Weder hoch noch niedrig 5 resp. 23.8%



Ich kann oder möchte hierauf nicht antworten 1 resp. 4.8%



Niedrig 1 resp. 4.8%



Eher niedrig 0 resp. 0%



Beschäftigen Sie sich mit möglichen Bias, Diskriminierung und Misrepresentation in Daten?

21 out of 21 answered

Nein 12 resp. 57.1%



Ja 9 resp. 42.9%



Wissen Sie, wie Sie Bias, Diskriminierung und Misrepresentation hinreichend quantifizieren können, um diese zu adressieren?

7 out of 21 answered

Ja 4 resp. 57.1%

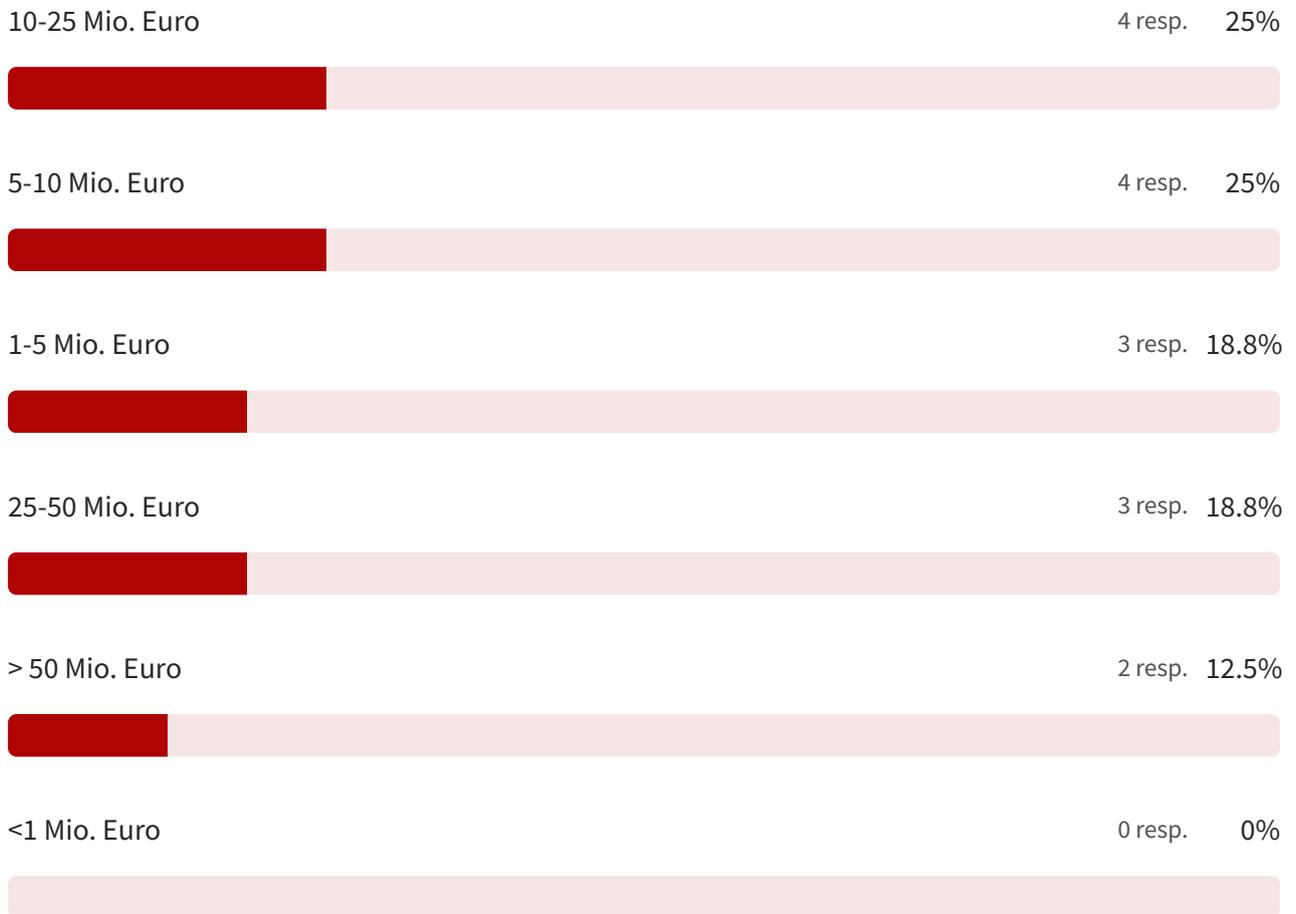


Nein 3 resp. 42.9%



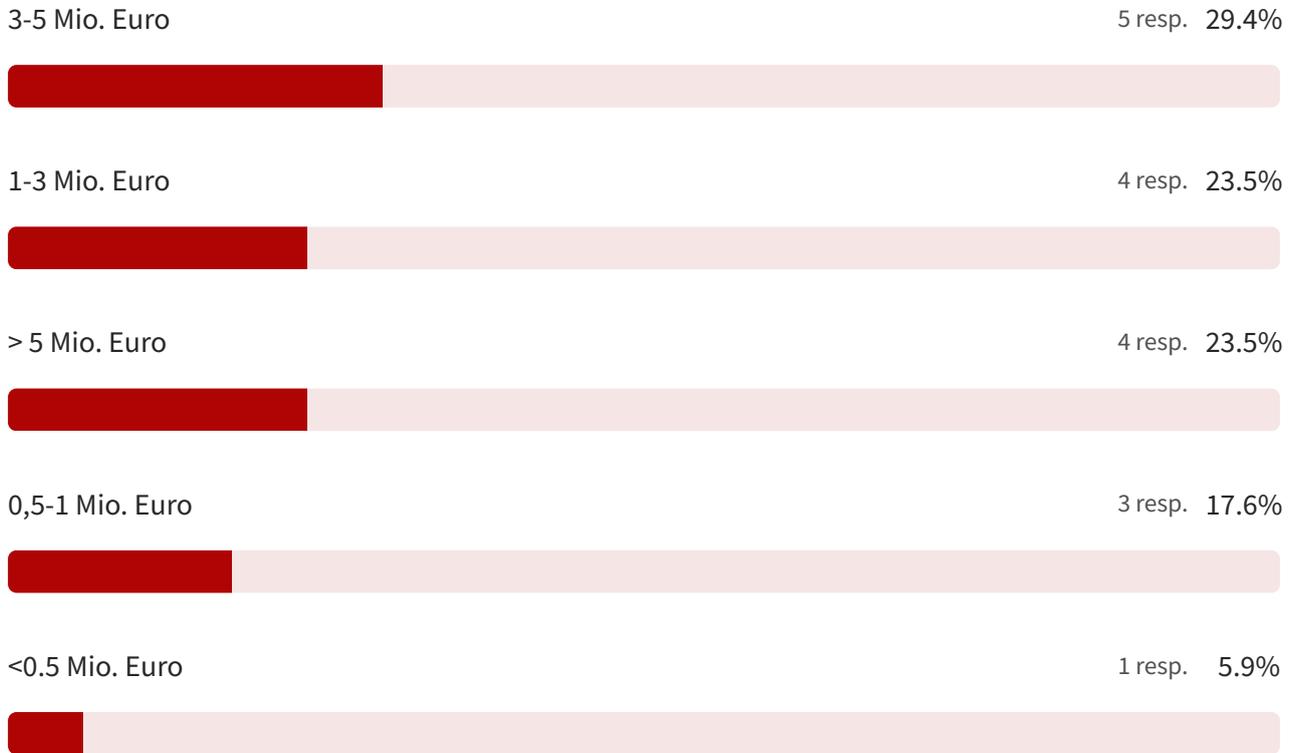
Welche Kosten halten Sie für die Entwicklung eines Foundation-Modells für realistisch?

16 out of 21 answered



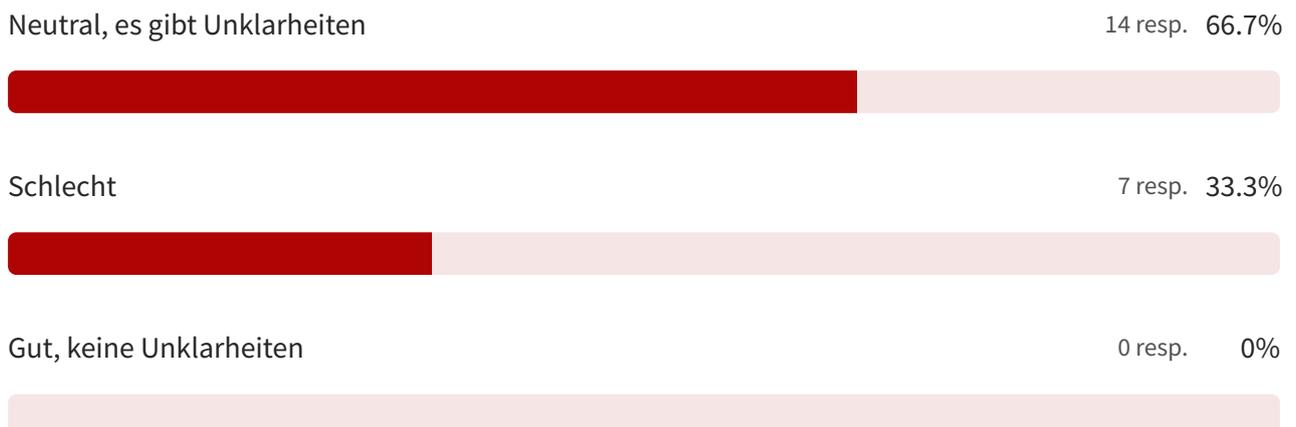
Welche Kosten halten Sie für das Training eines Foundation-Modells für realistisch?

17 out of 21 answered



Wie sind Sie über die Regulation der Entwicklung von Foundation-Modellen informiert?

21 out of 21 answered



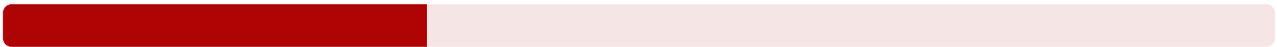
Empfinden Sie Ihr Wissen über die Regulation als Enabler oder Disabler für Ihren möglichen Einsatz von Foundation-Modellen?

12 out of 21 answered

Enabler 8 resp. 66.7%



Disabler 4 resp. 33.3%



Anhang B - Zusätzliche Information zu den Interviews

Anhang B.1 - Die Interviewpartner:innen

Wirtschaft

	Wolfgang	Hauner	Allianz SE
Dr.	Maik	Friedel	BASF SE
Dr.	Marion	Legler	Bayer AG
Dr.	Hans-Jörg	Vögel	BMW Group
Dr.	Michael	Fausten	Robert Bosch GmbH
	Jean-Paul	Schmetz	Burda Media
	Mario	Deng	BWI GmbH
Dr.	Corina	Apachiçe	Continental AG
Dr.	Matthias	Dorner	DATEV eG
	Stephan	Kaulbach	Deutsche Bahn AG
Dr.	Frank	Säuberlich	EnBW Energie Baden-Württemberg AG
Dr.	Sebastian	Kaiser	Ergo Group AG
	Thomas	Wolf	Hugging Face, Inc.
Dr.	Sabine	Donauer	Infineon Technologies AG
	Nico	Kelling	Infineon Technologies AG
	Rainer	Sträter	Ionos SE
	Christian	Spannbauer	Lufthansa Group
	Jochen	Kaiser	Mercedes-Benz Group AG
Dr.	Stephan	Meyer	Munich RE
Dr.	Michael	Müller-Wünsch	Otto GmbH & Co KG
Dr.	Armin	Kurrle	Porsche AG
Dr.	Lorenz	Determann	Rewe Group
Dr.	Feiyu	Xu	SAP SE
Dr.	Andreas	Wierse	sicos BW GmbH
Dr.	Michael	May	Siemens AG
Dr.	Dirk	Schlesinger	TÜV Süd AG
Dr.	Sebastian	Hallensleben	VDE Verband der Elektrotechnik Elektronik Informationstechnik e. V.
Dr.	Patrick	van der Smagt	Volkswagen AG
Dr.	Alexander	Borek	Zalando SE

Tabelle 23: Übersicht der Interviewpartner:innen im Bereich Wirtschaft

Rechenzentrums- und Hardwareexpert:innen

	Wolfgang	Christmann	christmann informationstechnik + medien GmbH & Co. KG
Dr.	Stefan	Rüping	Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS
Dr.	Stefan	Kesselheim	Forschungszentrum Jülich GmbH
Dr.	Helmut	Kreiser	GSI Helmholtzzentrum für Schwerionenforschung
	Jan	Seiler	Gigabyte Technology
	Andreas	Herden	Green Mountain Datacenter GmbH
Dr.	Bastian	Koller	Höchstleistungsrechenzentrum Stuttgart
	Dennis	Hoppe	Höchstleistungsrechenzentrum Stuttgart
	Oleksandr	Shcherbakov	Höchstleistungsrechenzentrum Stuttgart
	Volker	Ludwig	Interxion Deutschland GmbH
	Marco	Maslon	Northern Data AG
	Bedrettin	Altay	Noya Group Holding GmbH
	Volker	Meschonat	Nvidia Corporation
	Markus	Hacker	Nvidia Corporation
	Olaf	Dalmer	OneFiber Interconnect Germany GmbH
	Wolfgang	Dreyer	Oracle Corporation
Dr.	Thorsten	Henrich	PlusServer GmbH
	Max	Schulze	SDIA - Sustainable Digital Infrastructure Alliance e.V.
	Alexander	Hauser	TTSP HWP Planungsgesellschaft mbH

Tabelle 24: Übersicht der Interviewpartner:innen im Bereich Rechenzentrum und Hardware

Wissenschaft

Prof.	Alexander	Löser	Beuth Hochschule für Technik Berlin
Prof.	Sabine	Kirchmeier	European Federation of National Institutions for Language (EFNIL)
Dr.	Joachim	Köhler	Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS)
Prof.	Stefan	Wrobel	Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS)
Dr.	Narges	Ahmidi	Fraunhofer-Institut für Kognitive Systeme (IKS).
Dr.	Holger	Karl	Hasso-Plattner-Institut
Prof.	Alexander	Waibel	Karlsruher Institut für Technologie (KIT) und Carnegie Mellon University, Pittsburgh
Prof.	Jürgen	Schmidhuber	King Abdullah University of Science and Technology (KAUST)
Prof.	Hinrich	Schütze	Ludwig-Maximilians-Universität München
Prof.	Volker	Tresp	Ludwig-Maximilians-Universität München
	Dietmar	Harhoff	Max-Planck-Institut für Innovation und Wettbewerb
Dr.	Peter	Norvig	Stanford University, Google Inc., NASA
Prof.	Iryna	Gurevych	Technische Universität Darmstadt
Prof.	Kristian	Kersting	Technische Universität Darmstadt
Prof.	Andreas	Dengel	Technische Universität Kaiserslautern
Prof.	Daniel	Cremers	Technische Universität München
Prof.	Josef	van Genabith	Universität des Saarlandes
Prof.	Anette	Frank	Universität Heidelberg
Prof.	Sepp	Hochreiter	Universität Linz und Linz Institute of Technology (LIT)
Prof.	Leo	Wanner	Universitat Pompeu Fabra Barcelona (UPF)
Prof.	Ralf	Herbrich	Universität Potsdam

Tabelle 25: Übersicht der Interviewpartner:innen im Bereich Wissenschaft

Sonstige

	Philipp	Günther	Berlin Partner für Wirtschaft und Technologie GmbH
Dr.	Daniel	Gille	Agentur für Innovation in der Cybersicherheit GmbH
	Risto	Uuk	Future of Life Institute
Dr.	Tina	Klug	Hessische Staatskanzlei Ministerin für Digitale Strategie und Entwicklung
	Olly	Salzmann	KI Park e.V.
Dr.	Peter	Mendler	Ministerium für Wirtschaft, Arbeit und Tourismus Baden-Württemberg
	Christian	Dinnus	Ministerium für Wirtschaft, Industrie, Klimaschutz und Energie des Landes Nordrhein-Westfalen
	Sebastian	Ley	Ministerium für Wirtschaft, Industrie, Klimaschutz und Energie des Landes Nordrhein-Westfalen
	Philipp	Denker	PD - Berater der öffentlichen Hand GmbH
	Jörg	Schaub	Wirtschaftsinitiative FrankfurtRheinMain e.V.

Tabelle 26: Übersicht der sonstigen Interviewpartner:innen

Anhang B.2 - Die Leitfragen

Leitfragen für die Interviews mit Wirtschaftsvertreter:innen:

- 1) Inwiefern nutzen Sie bereits KI-Foundation-Modelle? Welche Modelle nutzen Sie und für welche Anwendungen?
- 2) Welche Bedeutung messen Sie Foundation-Modellen aktuell und in der Zukunft zu?
- 3) Wie wichtig ist aus Ihrer Sicht der Aufbau eines europäischen Ökosystems rund um Foundation-Modelle, inkl. eigener europäischer Modelle?
- 4) Würden Sie die Services der LEAM Initiative nutzen? In welchem Umfang?

Anhang B.3 - Ergebnisprotokolle

Interviewprotokoll Allianz

Interviewer:

Wolfgang Hauner, Head of Group Data Analytics, Allianz SE

Interviewer:

Jörg Bienert, Alexander Thamm GmbH

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

22. Dezember 2022

Zusammenfassung

1. KI-Foundation-Modelle

- Die Zukunft der KI liegt in Foundation-Modellen. Die Initiative LEAM ist daher sehr wichtig.
- Die Möglichkeiten der KI und der Foundation Modelle werden Geschäftsmodelle verändern, ähnlich der Situation beim Aufkommen des Internets und e-Commerce.
- Für interne Anwendung (bspw. in der Kommunikation und dem Kundenservice) können die bestehenden Modelle bereits gut eingesetzt werden.
- Die Nutzung von KI-Services über eine API (und der Transfer der Daten in die USA) ist aufgrund des Einsatzes sensibler Daten keine Option.
- Es gibt wichtige Anforderungen für versicherungsspezifische Modelle, die es aktuell aber noch nicht gibt, z.B. in den Bereichen:
 - Verbesserung des Kundenservices, indem Fragen zur Police automatisch ausgelesen werden oder
 - Generierung von individuellen Policen auf Basis der Foundation-Modellen (Hierfür müssen die Modelle aber verlässlicher und rechtlich abgesichert werden).
- Ohne die Verfügbarkeit dieser Modelle wird es mittelfristig zu Nachteilen für Versicherungsunternehmen im internationalen Wettbewerb kommen..
- V.a. bei der Verwendung sensibler Versicherungsdaten braucht es lokale Modelle, um den Datenschutz zu gewährleisten.
- Die Herausforderung liegt nicht im traditionellen Versicherungsmarkt, der stark reguliert ist. Stattdessen besteht die Gefahr, den Schritt bei neuen, digitalen Geschäftsmodellen zu verpassen, die keine nationalen Grenzen kennen. Ähnliche Entwicklungen gab es bspw. bei Amazon und Google.

2. Zusammenarbeit mit LEAM

- Einzelne Unternehmen in Deutschland sind nicht in der Lage die notwendigen Investitionen alleine zu tragen. Es braucht die Unterstützung der Politik.
- Der Möglichkeit einer PPP bzw. eines Joint Ventures steht die Allianz offen gegenüber. Eine Zusammenarbeit der DAX Unternehmen hierfür ist durchaus realistisch.

Interviewprotokoll BASF

Interviewter:

Maik Friedel, Principal Scientist for Artificial Intelligence Chemistry & Leiter der Initiative "Generative AI", BASF

Interviewer:

Jörg Bienert, Alexander Thamm GmbH

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

21. Dezember 2022

Zusammenfassung

1. KI-Foundation-Modelle

- BASF erforscht bereits intensiv den Einsatz von Foundation-Modellen und hat hierfür kürzlich die Initiative „Generative AI“ ins Leben gerufen
- Ziel ist, neben dem Model Tuning von bereitgestellten Modellen auch eigene Foundation Modelle zu tunen.
- Hierzu gehören neben großen Sprachmodellen auch domänenspezifische Modelle im Chemie Bereich
- BASF besitzt bereits eine eigene Supercomputing-Recheninfrastruktur (Quriosity) und plant diese weiter auszubauen. Der Vorteil einer eigenen Infrastruktur liegt in der Datensicherheit.
- Mitarbeiter nutzen bereits privat Dienste wie GPT-3 oder ChatGPT. Das ist kaum kontrollierbar und kann ein potentiell Daten-Sicherheitsrisiko darstellen.
- Das Trainieren bzw. Tunen von Modellen durch BASF auf Basis von eigenen Daten ist eine strategisch, wichtige Capability. Hierfür können außereuropäische Services (z.T. auch Cloud-Services generell) aus Datensicherheitsgründen nur bedingt, oder gar nicht in Anspruch genommen werden.
- Wenn bei BASF bzw. in Deutschland/Europa entsprechende Services und Foundation Modelle nicht bereitgestellt werden können, entwickelt sich hieraus ein großer Wettbewerbsnachteil.

2. Zusammenarbeit mit LEAM

- Aus den oben genannten Gründen wäre ein LEAM Service eine gute Unterstützung um eigene Foundation-Modelle zu erstellen und existierende Modelle zu tunen

Interviewprotokoll Bayer Pharma

Interviewter:

Marion Legler, Head of Decision Science & Advanced Analytics at Bayer Pharma

Interviewer:

Jörg Bienert, Alexander Thamm GmbH

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

20. Dezember 2022

Zusammenfassung

1. KI-Foundation-Modelle

- Bayer arbeitet intensiv mit Foundation-Modellen, v.a. im Bereich Sprache (Natural Language Processing) und der Bildverarbeitung (Computer Vision)
- Im Bereich Natural Language Processing greift Bayer auf den Medizinbereich spezialisierte Modelle wie SciBERT, BioBERT und PubMedBERT zurück.
- Aber: Die aktuellen Modelle erfüllen noch nicht alle Erwartungen, bspw. ist nicht die gesamte medizinische Terminologie eingebunden. Ebenso können die aktuellen Language Models die Nuancen der medizinischen Inhalte, die für Laien oft nicht erkennbar sind, noch nicht differenzieren. Größere Modelle im medizinischen Bereich könnten dort weiterhelfen.
- Ein *europäisches* Foundation-Modell für die Medizin wäre für Bayer sehr interessant. Dieses könnte Bayer dann speziell für seine Ansprüche, d.h. konkrete Indikationsgebiete von Interesse, anpassen.
- Aufgrund des Datenschutzes ist es für Bayer keine Option, Modelle von US-Providern auf amerikanischen Servern zu nutzen bzw. anzupassen, wenn dabei wertvolle (Patienten-)daten geteilt werden.
- Gleichzeitig besteht beim „finetuning“ die Gefahr, dass ein gewisser Bias, der aus dem Ursprungsmodell hervorgeht, bestehen bleibt. Insofern besteht hier ein großes Interesse an großen Sprachmodellen, die von der Basis aus auf europäischen Daten trainiert wurden und somit europäische Sprachgepflogenheiten aber auch unterschiedliche europäische Sprachen per se berücksichtigt.

- Neben Sprachmodellen spielen auch Foundation-Modelle im Bereich Computer Vision eine wichtige Rolle.
- Auch hier lassen sich Modelle, die bspw. auf amerikanischen Patientendaten trainiert wurden, nicht immer mit derselben Qualität auf europäische Patientendaten anwenden. Der Grund dafür liegt hier in den unterschiedlichen ethnischen Zugehörigkeiten der PatientInnen, die sich dann auch in minimalen Unterschieden bspw. der CT-Scans oder anderen Bilddaten niederschlagen, die jedoch für die Performance der Modelle ausschlaggebend sind. Modelle, die auf einer Population trainiert sind, sind somit meist nicht generalisierbar und auf andere PatientInnenpopulationen anwendbar. Aus diesem Grund besteht auch hier ein großes Interesse an europäischen Modellen.
- Bayer kann sich gut vorstellen, das Serviceangebot von LEAM im Bereich Model-Tuning und evtl. auch in der Erstellung von Foundation Modellen zu nutzen.

2. Zusammenarbeit mit LEAM

- Bayer könnte sich prinzipiell auch ein Engagement an LEAM im Rahmen eines Joint-Ventures / einer PPP vorstellen.
- Ein anzustrebendes Ziel wäre, gemeinsam mit anderen Akteuren aus dem Bereich Health ein Foundation-Modell speziell für Anwendungen in der Medizin zu entwickeln. Das Foundation Modell könnte dann als wertvolle Basis für firmen-spezifische Domän-Modelle dienen.

Interviewprotokoll Berlin Partner

Interviewter:

Philipp Günther, Berlin Partner

Interviewer:

Vanessa Cann, KI Bundesverband
Jörg Bienert, Alexander Thamm GmbH

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

30. November 2022

Zusammenfassung

1. Berlin Partner

- **Informationen zu Berlin Partner**
 - Die Berlin Partner für Wirtschaft und Technologie GmbH ist die Wirtschaftsförderung des Landes Berlin.

- Mit ca. 220 Mitarbeitenden unterstützt das Public-Private-Partnership Unternehmen, Wissenschaftseinrichtungen und NPOs am Standort.
- Im Innovationsfeld Künstliche Intelligenz unterstützt Berlin Partner Akteure in Förderfragen, vernetzt sie untereinander und in Anwendungsbranchen und sorgt für mehr Sichtbarkeit über Berlins Grenzen hinaus.
- **Entwicklung KI-Foundation Modelle in Europa**
 - Sie sehen die Entwicklung eigener Foundation-Modelle als entscheidende Grundlage für deutsche KI-Innovationen innerhalb der europäischen Wertegemeinschaft.
 - In ihrer Rolle als Wirtschaftsförderung unterstützen sie Akteure in der Ansiedlung, Identifikation von Fördermöglichkeiten, Stakeholder-Dialog, Kommunikation und Transfer.
 - Beispielprojekte sind:
 - ResKriVer unter Leitung des Fraunhofer Fokus
 - WHO Hub for Pandemic and Epidemic Intelligence
 - EU Testing and Experimentation Facilities (TEF) Health unter Leitung der Charité Berlin
 - www.ki-berlin.de: eine Plattform auf der News, Events und Erfolgsgeschichten aus dem Berliner KI-Ökosystem dargestellt werden.
 - Projekte zu Foundation-Modellen wurden bisher nicht enger betreut.
- **Governance**
 - Berlin Partner ist selbst eine PPP.
 - *aus dem Protokoll des vorherigen Gesprächs zum PPP-Modell:*
 - Berlin Partner bestreitet öffentliche Aufgaben, die vom Land vergeben werden. Sie erhalten eine institutionelle Zuwendung des Landes und keine Finanzierung über eine Projektstruktur.
 - Die Partner für Berlin Holding Gesellschaft für Hauptstadt-Marketing mbH ist zur Hälfte privat, zur Hälfte aus der Stadt finanziert. Sie hält 28% an Berlin Partner.
 - IBB (Investitionsbank Berlin) Unternehmensverwaltung AöR - 31,5%
 - Technologiestiftung Berlin - 30,0%
 - jeweils zu 3,5%: Handwerkskammer Berlin, IHK Berlin, Vereinigung der Unternehmensverbände Berlin und Brandenburg e.V.
 - Insgesamt hält der öffentliche Sektor weniger als 50% der Anteile.
 - Je mehr Partner im Projekt sind, desto mehr Kompetenzen gibt man auch ab.

- Eine Möglichkeit ist die Finanzierung über eine Betreiberkonzession, wie es bspw. bei Messengeländen der Fall ist.

2. Wie kann LEAM mit den Projekten zusammenarbeiten?

● Generelle Einschätzung zu LEAM

- LEAM kann eine Grundlage für KI-Innovationen in Berlin, Deutschland und Europa sein.
- Für Berlin Partner zeigt das Interesse des Ökosystems, dass LEAM unterstützt werden sollte. Dabei ist vor allem die europäische Perspektive interessant.
- Neben dem Interesse bestehender KI-Akteure in der Hauptstadtregion, gibt es auch Gespräche mit internationalen Akteuren, die nach Berlin kommen wollen, von LEAM profitieren könnten und sich vermutlich einbringen würden. LEAM würde auch die Attraktivität des Standorts weiter steigern.

● Der Standort Berlin

- Berlin bietet ein dynamisches und diverses KI-Ökosystem mit internationaler Strahlkraft sowie Akteuren, die auf Augenhöhe kooperieren und offen für neue Partner sind.
- Berlin ist Innovationsstandort. Auch Unternehmen, deren Hauptstandort woanders liegt, haben häufig Innovation Labs oder Entwicklungsteams in der Hauptstadt.
- In Berlin findet exzellente KI-Forschung statt – in den Grundlagen, sowie anwendungsnahe. Schwerpunkte sind u.a. NLP sowie erklärbare und vertrauenswürdige KI.
- Berlin ist KI-Startup-Hauptstadt.
- Berlin zieht als internationaler Hub viele internationale Talente an, die auch für LEAM benötigt würden.
- Mögliche Standorte und Rechenzentren:
 - Mit NTT und Penta Infra gibt es Akteure in Berlin, die ohne Zeitverzug ein LEAM Rechenzentrum integrieren und hosten könnten.
- Die 11 Zukunftsorte Berlins sowie Innovationsparks genießen Förderprivilegien, haben bereits relevante Infrastruktur und könnten ebenfalls interessante Partner und Standorte sein.
- Finanzierung:
 - Die Gesprächsbereitschaft der Berliner Landesregierung wird aus Sicht von Berlin Partner als hoch eingeschätzt. Ein enger Austausch mit der Senatsverwaltung für Wirtschaft, Energie und Betriebe zu diesem Thema besteht bereits.

- Generell gilt, dass eine Bundesförderung, an der sich ein Land beteiligen kann, die Möglichkeit einer schnelleren Förderung durch das Land deutlich begünstigt.
- Beispiel: AI Quality & Testing Hub / TEF Health
 - Berlin hat als erstes der involvierten Bundesländer Fördermittel für die bundesweite Initiative zugesagt.
 - Berlin Partner konnte außerdem EU-Förderung sowie Partner vermitteln, um ein EU-weites Konsortium zur Bewerbung für Testing and Experimentations Facilities (TEF) aufzustellen. Dabei stemmen im Erfolgsfall das Land Berlin, sowie die Charité Berlin die nationale Ko-Finanzierung.
- **Zusammenarbeit mit Berlin Partner**
 - Berlin Partner kann selbst keine finanzielle Unterstützung leisten.
 - Sie unterstützt aber bei:
 - der Suche nach geeigneten Standorten und entsprechenden Partnern (KI Park, Penta-Infra, Zukunftsorte uvm.).
 - der Initiierung und Begleitung der Kommunikation mit Ministerialverwaltungen auf Landesebene.
 - der Identifikation von weiteren Fördermöglichkeiten (Landes-, Bundes- und EU-Mittel).
 - der Vernetzung mit weiteren potentiellen Partnern und Kunden aus ihrem Netzwerk.
 - durch einen Erfahrungsaustausch zu einer möglichen Governancestruktur und dem Finanzierungsmodell eines PPP.
 - Stakeholder-Dialog auf regionaler Ebene.
 - Kommunikation & Transfer, über Area Managements und Auslandsbüros auch weltweit.

Interviewprotokoll BMW

Interviewter:

Hans-Jörg Vögel, Manager AI, Robotics, and Cognitive Systems, BMW Group

Interviewer:

Jörg Bienert, Alexander Thamm GmbH

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

22. Dezember 2022

Zusammenfassung

1. KI-Foundation-Modelle

- Die Zielsetzung von LEAM ist sinnvoll und sehr plausibel.
- BMW ist als Associate Partner in OpenGPT-X involviert.
- Es gibt sehr viele Anwendungsfälle für Foundation-Models z.B. im Bereich Benutzerhandbücher, Kunden- und Fahrerdialoge.
- Multimodale Modelle sind ebenfalls in Planung. Hier gibt es Forschung im Bereich Innenraumüberwachung und Sprachdialogsysteme. Die Fragestellung, inwiefern Foundation-Modelle / Transformer-Modelle zu Durchbrüchen auch beim Autonomen Fahren führen könnten, ist offen.
- Wichtig ist die Berücksichtigung von Datenschutz, Datensicherheit und Informationsschutz, GDPR-Compliance, Unterstützung von Internationalisierung (zweistellige Anzahl von Sprachen für Märkte weltweit) und die Bereitstellung einer mandantenfähigen Plattform. All dies kann durch US-amerikanische general-purpose Services (derzeit) nicht gewährleistet werden.
- Die Services von ChatGPT sind bereits jetzt schon kritisch, da Nutzung durch Mitarbeiter mit Unternehmensdaten schwer kontrollierbar.
- Für die digitale Souveränität in Deutschland ist es wichtig, die Modelle zu verstehen und über die Bereitstellung von entsprechender Infrastruktur und Möglichkeiten Experten und Talente im Land zu halten. Forschung und Transfer der Forschungsergebnisse in die Wirtschaft ist dabei essenziell.
- Die intensive Nutzung von KI und die Möglichkeit, Foundation Modelle zu nutzen, ist sehr wettbewerbsrelevant. Im Moment gibt es hierfür in Deutschland nicht die erforderliche Infrastruktur.

2. Zusammenarbeit mit LEAM

- BMW ist auf jeden Fall ein potenzieller Nutzer von LEAM. Betriebssicherheit, Qualität und wirtschaftliche Wettbewerbsfähigkeit der Services sind dafür eine Voraussetzung.
- Die Beteiligung von BMW an einen Joint Venture / einer PPP kann diskutiert werden, passende Rahmenbedingungen vorausgesetzt.

Interviewprotokoll Bosch Center for AI

Interviewter:

Michael Fausten, SVP AI and Systems at Robert Bosch GmbH, Bosch

Interviewer:

Jörg Bienert, Alexander Thamm GmbH

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

15. Dezember 2022

Zusammenfassung

1. KI-Foundation-Modelle

- KI-Foundation-Modelle sind ein relevantes Thema bei Bosch. Aktuell klärt Bosch die Potenziale der Technologie.
- Neben Sprach- und Bilddaten sind auch weitere Industriedaten spannend.
- Bei der Nutzung amerikanischer Modelle kann es zu Abhängigkeiten von den USA kommen. Darüber hinaus gibt es das Problem, dass aufgrund geopolitischer Entwicklungen die amerikanischen Modellen eventuell in anderen Märkten - bspw. China - nicht mehr nutzbar sind.
- Europäische Modelle müssen genauso leistungsstark sein wie die amerikanischen.

2. Zusammenarbeit mit LEAM

- Bosch betreibt selbst ein Rechenzentrum. Für die Entwicklung von Foundation Modellen ist ein LEAM Service durchaus attraktiv. Daneben besteht vor allem Interesse an der Möglichkeit des Modell Tunings.
- Bosch steht einer finanziellen Beteiligung bei LEAM grundsätzlich offen gegenüber. Dafür müssten jedoch noch einige offene, insbes. kommerzielle Fragen geklärt werden. Ein Konsortium aus mehreren Unternehmen klingt nach einem denkbaren Weg.
- Eine zeitnahe Realisierung ist entscheidend. In fünf Jahren sind wir bereits zu spät.

Interviewprotokoll BWI GmbH

Interviewer:

Mario Deng, Lead Service Manager Data Analytics, BWI GmbH

Interviewer:

Jörg Bienert, Alexander Thamm GmbH

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

10. Januar 2022

Zusammenfassung

1. KI-Foundation-Modelle

- Die Services, die auf Basis von KI-Foundation-Modellen entwickelt werden, haben einen Mehrwert. Die BWI möchte diese Anwendungen in ihr Portfolio übernehmen.
- Die Nutzung von US-amerikanischen AI-Services über API ist für die BWI nicht realistisch, da bei sicherheitskritischen Daten besondere Schutzvorkehrungen gelten. Die Technologie dahinter ist aber nutzungswert.
- Es müssen Use Cases entwickelt werden, die über das Thema Chatbot hinausgehen.
- Es ist ineffizient, wenn jede Organisation ihr eigenes Rechenzentrum baut und KI-Foundation-Modelle berechnet. Stattdessen sollten gemeinsam wenige Foundation-Modelle berechnet werden, die dann individuell angepasst werden können.

2. Zusammenarbeit mit LEAM

- Aufgrund hoher Sicherheitsanforderungen darf die BWI viele Anbieter nicht nutzen.
- Die BWI hat Bedarf an den Modellen und den von LEAM geplanten Services.
- Entscheidend sind die Sicherheitsvorkehrungen der Infrastruktur sowie die Sicherheit beim Transfer der Daten.
 - Es muss ausgearbeitet werden, wie diese Sicherheitsvorkehrungen auszusehen haben.
- Daneben benötigt das BWI für eine Nutzung der Infrastruktur folgende Voraussetzung:
 - Eine saubere Dokumentation bei der Entwicklung der Modelle
 - Transparenz bei den für Modelle genutzten Daten und den beteiligten Personen
 - Ein Verrechnungsmodell, das eine einfache Kostenplanung ermöglicht.

Interviewprotokoll Continental

Interviewter:

Dr. Corina Apachițe, Programm-Leiterin für Künstliche Intelligenz und Daten

Interviewer:

Jörg Bienert, Alexander Thamm GmbH

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

14. Dezember 2022

Zusammenfassung

1. Foundation-Modelle

- Die Fähigkeiten von großen Sprachmodellen wie GPT-3 sind faszinierend.
- Diese Fähigkeiten sind auch für Continental wichtig, allerdings müsste hierzu ein domainspezifisches Tuning auf Basis von „Ingenieursprache und spezifischen Inhalten“ erfolgen. Dies kann derzeit nicht über GPT-3 oder ähnliche Modelle abgewickelt werden.
- Die Befähigung ein eigenes große Sprachmodell zu erstellen bzw. zu tunen ist sehr wichtig für die Zukunftsfähigkeit der deutschen und europäischen Digitalindustrie.

2. Multimodale Modelle

- Es besteht ein großer Bedarf auch an neuen multimodalen Modellen u.a. für das autonome Fahren.
- Auf Basis bestehender Modelle sind derzeit eher inkrementelle Verbesserungen möglich.
- Wenn ein Anbieter in der Lage ist, auf Basis von großen Foundation Modellen ein sehr viel leistungsfähigeres autonomes Fahren anzubieten, würde dies den Markt stark verändern. Gerade in Deutschland müssen wir uns für diese Disruption stärken und die nutzbare Datenbasis vergrößern.

3. Gemeinsame Aktivitäten

- Die Industrie muss zusammenarbeiten und befähigt werden, unter Einhaltung von Regulierung und Gesetzen Foundation Modelle zu entwickeln bzw. bereitgestellte Foundation Modelle zu tunen.
- Ein einzelnes Unternehmen kann diese Herausforderungen aufgrund der unzureichenden Datenmenge, Mangel an Personal und Erfahrung nicht stemmen.
- Die Erfahrungen aus bestehenden Projekten (bspw. Catena-X) muss genutzt werden.
- Wir brauchen in Deutschland eine neu aufgebaute „Supply-Chain“ für Modelle und Daten.
- Die Möglichkeit, große Modelle zu erstellen, wird auch ein Katalysator für den Aufbau von Datenpools darstellen.
- Ebenfalls kann durch die Befähigung, große Modelle zu erstellen und durch führende Forschungsaktivitäten in diesem Bereich ein attraktives Betätigungsfeld für Top-Talente gestaltet werden und so einem „Brain-Drain“ entgegen gewirkt.

- All diese Punkte sind enorm wichtig für das geostrategische Setup und die Wettbewerbsfähigkeit der deutschen Wirtschaft und Gesellschaft.

Interviewprotokoll Cyberagentur

Interviewer:

Daniel Gille, Leiter Sicherheit durch KI und Sicherheit für KI, Cyberagentur
Michael Lindner, Pressesprecher, Cyberagentur

Interviewer:

Jörg Bienert, Alexander Thamm GmbH
Jakob Tesch, Ubermetrics

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

09.12.2022

Zusammenfassung

1. Cyberagentur

- Die Cyberagentur ist eine GmbH des Bundes. Alleinige Gesellschafterin ist die Bundesrepublik Deutschland, vertreten durch das Bundesministerium des Innern und für Heimat sowie das Bundesministerium der Verteidigung.
- Sie forscht nicht selbst, sondern beauftragt Grundlagenforschung im Bereich disruptiver Technologieansätze mit Bezug zu Innerer und/oder Äußerer Cybersicherheit.
- Künstliche Intelligenz ist aus Sicht der Cyberagentur eine Schlüsseltechnologie und wird entsprechend in der Forschungsstrategie als eines der relevanten Themenfelder behandelt.
- Die SPRIND, die im Auftrag des Bundesministeriums für Wirtschaft und Klimaschutz sowie des Bundesministeriums für Bildung und Forschung arbeitet, ist die "Schwesteragentur" der Cyberagentur.

2. Nutzung von LEAM

- Es lässt sich aktuell noch nicht sagen, ob die Cyberagentur selbst die Kapazitäten von LEAM bräuchte, insbesondere da sie selbst nicht im KI-Bereich forscht und entwickelt. Für Forschungs- und Entwicklungsaufgaben mit Sicherheitsbezug müssten in jedem Fall hohe Security-Anforderungen erfüllt werden. Insbesondere müssten LEAM-Infrastrukturen relevante VS-Kriterien erfüllen.
- Eine zukünftige Nutzung hochperformanter KI-Infrastrukturen durch forschende Auftragnehmer:innen im Rahmen der Umsetzung ihres Forschungsvorhabens ist

durchaus vorstellbar. Allerdings kann die Cyberagentur ihren Auftragnehmer:innen hinsichtlich zu nutzender externer Ressourcen keine konkreten Empfehlungen aussprechen oder Vorgaben machen.

- Das Thema KI-Foundation-Modelle ist in seiner Relevanz erkannt und auf der Forschungsroadmap entsprechend hoch priorisiert. Seine Bedeutung wird in der Zukunft noch steigen.
- Für die Cyberagentur ist bei der Betrachtung von Security-Fragestellungen jeder Schritt im ML-Lebenszyklus relevant.
- Die Nutzung nicht-europäischer Modelle kann unter Umständen ein Sicherheitsrisiko für die Bundesrepublik Deutschland darstellen. Aktuell stehen wir noch am Anfang der Entwicklung, aber bereits in wenigen Jahren sind möglicherweise viele Anwendungsbereiche, Wertschöpfungsprozesse und Geschäftsmodelle zumindest in Teilen auf Foundation Models angewiesen, über deren Funktionsweise, Entwicklung, Trainingsdaten etc. nur unzureichende Transparenz besteht. Als mahndes Beispiel sei auf die Hintertüren-Diskussion um 5G und Huawei verwiesen.

3. Infrastruktur

- Der Bedarf nach einer übergreifenden KI-Infrastruktur ist nach persönlicher Einschätzung DG gegeben, um die technologische Souveränität bei der Entwicklung und Anwendung großer KI-Modelle sicherzustellen.

Interviewprotokoll Deutsche Bahn

Interviewer:

Stephan Kaulbach, Head of Data Intelligence Center, Deutsche Bahn AG

Interviewer:

Jörg Bienert, Alexander Thamm GmbH

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

21. Dezember 2022

Zusammenfassung

1. KI-Foundation-Modelle

- Die Bedeutung des Themas Foundation-Modelle wird in der Zukunft wachsen.
- Aktuell nutzt DB KI-Modelle v.a. für Optimierungen. Es gibt aber erste Bestrebungen, Foundation-Modelle bspw. bei der Angleichung von Handbüchern

für Lokführer und die Instandhaltung. Im Bereich Customer Support (bspw. Chatbots) wird das Thema ebenfalls wichtiger werden.

- Aktuell kann niemand abschätzen, was mit den Daten passiert, die für die Nutzung der amerikanischen Modelle über API in deren Cloud-Services übertragen werden. geschieht. Das stellt die Deutsche Bahn vor Herausforderungen, da sie nicht verhindern können, dass die Mitarbeiter bereits jetzt Anwendungen wie ChatGPT nutzen.
- Die DB arbeitet bereits mit dem deutschen Startup Deepl zusammen. Sie haben in ihren Sprachendienst das Bahnlexikon integriert. Bei dieser Kooperation wurde vor allem Wert auf die Einhaltung von Datenschutzstandards gelegt..
- Generell sind europäische Modelle aufgrund des Datenschutzes einfacher zu integrieren als amerikanische Modelle.

2. Zusammenarbeit mit LEAM

- Die Initiative LEAM ist unterstützenswert. Die DB wäre sehr an der Nutzung von LEAM-Modellen interessiert.
- Die Beteiligung an einem Joint Venture bzw. einer PPP ist grundsätzlich sinnvoll. Die Umsetzung müsste, vor allem hinsichtlich der Struktur der DB, geklärt werden.
- Entscheidend ist, dass beim Thema KI-Foundation-Modelle zeitnah etwas passiert.

Interviewprotokoll EnBW

Interviewer:

Dr. Frank Säuberlich, Chief Data Officer, EnBW Energie Baden-Württemberg AG

Interviewer:

Jörg Bienert, Alexander Thamm GmbH

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

20. Dezember 2022

Zusammenfassung

1. KI-Foundation-Modelle

- Aktuell werden KI-Modelle bereits intensiv in klassischen Vorhersagen wie Predictive Maintenance etc. eingesetzt.
- Große Sprachmodelle könnten zukünftig ein wichtiges Thema u.a. zur Analyse/Verarbeitung von unstrukturierten Daten werden.

- Es gibt auch erste Ansätze, wie mehr Rechenkapazität bzw. Quantencomputing die aktuell genutzten Modelle verbessern könnten, hier denken wir vor allem in Kontext von Optimierungsmodellen, bspw. für Layout von Offshore Windparks.
- Hierfür sind leistungsfähige Rechner erforderlich. Als Anbieter kritischer Infrastruktur ist die Nutzung von amerikanischen Cloud-Service Providern hier mit entsprechenden Risiken / Unwägbarkeiten verbunden. Das Thema Datensicherheit ist wichtig. Die EnBW schaut sehr genau auf diese Themen, wenn Modelle außerhalb Europas gehostet werden.
- Eine europäische Alternativen würde hierbei sehr helfen, daher wird die LEAM Initiative begrüßt.
- Neben Rechenkapazität spielen auch hochwertige Daten und ausreichend Personal eine Rolle.

2. Zusammenarbeit mit LEAM

- Einer Beteiligung am Projekt LEAM steht die EnBW offen gegenüber. Für eine Zusage müssten allerdings weitere Personen involviert und die praktische Umsetzung konkretisiert werden.

Interviewprotokoll Ergo

Interviewer:

Sebastian Kaiser, Head of Machine Learning, ERGO Group AG

Interviewer:

Jörg Bienert, Alexander Thamm GmbH

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

16. Dezember 2022

Zusammenfassung

1. KI-Foundation-Modelle

- Das Thema KI-Foundation-Modelle ist von strategischer Bedeutung. Die hierdurch realisierbaren Produktivitätssteigerungen und Service-Verbesserungen können zu Wettbewerbsvorteilen führen.
- Die größte Herausforderung bei der Nutzung amerikanischer Modelle ist der Datenschutz. Der Zugang zu GPT-3 über API auf einen Rechner in den USA erlaubt keine Verarbeitung von sensiblen Versicherungsdaten .

- Ein großer Nachteil der aktuellen Foundation-Modelle ist, dass sie nicht auf versicherungsspezifischen Daten trainiert sind. Andere Sprachmodelle (z.B. Wu Dao) enthalten mehr versicherungsspezifische Dokumente und sind daher besser auf die Domäne anwendbar.

2. Zusammenarbeit mit LEAM

- Das Serviceangebot von LEAM trifft genau die Bedürfnisse der Ergo hinsichtlich der Nutzung von großen KI-Modellen.
- Wichtig ist aber auch der Vergleich der Kosten ggü. Anbietern (auch aus den USA).
- Ein kompetitives europäisches Foundation-Modell würde die Diskussionen rund um den Datenschutz enorm vereinfachen.
- Eine mögliche finanzielle Beteiligung der ERGO im Rahmen eines Joint Ventures / Public Private Partnership ist prinzipiell denkbar
- Die Muttergesellschaft der Ergo, Munich Re, ist der Zusammenarbeit mit anderen Organisationen in diesem Bereich grundsätzlich aufgeschlossen. Allerdings liegt die Entscheidung hier bei der Munich Re.

Interviewprotokoll Forschungszentrum Jülich

Interviewter:

Stefan Kesselheim, Head of AI Consultants Team, FZ Jülich

Interviewer:

Jörg Bienert, Alexander Thamm GmbH

Jakob Tesch, Ubermetrics

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

14.11.2022

Zusammenfassung

1. Governance des FZ Jülich

- **Finanzierung:**
 - Die Finanzierung erfolgte über zwei Stränge:
 - zu ca. 50% über die EU über die Organisation PRACE - Partnership for Advanced Computing in Europe
 - zu ca. 50% über das BMBF und das Land NRW über den Verein Gauss Centre for Supercomputing (GCS)
 - PRACE wird von EuroHPC abgelöst

- weitere Communities (z.B. Helmholtz-Gemeinschaft) sind z.T. an den Rechnern beteiligt und können Ressourcen frei nutzen
- **Organisationsform:**
 - Das FZ Jülich ist eine GmbH des öffentlichen Rechts.
 - Sie wirtschaftet daher nicht wie eine standardmäßige GmbH.
- **Die Recheninfrastruktur ist Eigentum des FZ Jülich**
 - Aber das FZ Jülich muss den größten Teil der Rechenzeit für externe Projekte zur Verfügung stellen.
 - Und nur ein kleiner Teil steht dem FZ Jülich zur freien Verfügung.
- **Zugang zu Rechenpower:**
 - Halbjährlich werden Projektaufträge auf Rechenzeit veröffentlicht.
 - Projekte werden anhand eines objektiven, peer-review Verfahrens bewertet.
 - Bei kleineren Tier 2 Zentren (bspw. NHR-Verbund) gibt es rollierende Aufträge.
 - In der Regel laufen die Jobs max 24 Stunden
- **Personal:**
 - Eine Aussage zur Anzahl des Personals ist schwierig, da es sich in viele verschiedene Bereiche zergliedert.
 - Insgesamt arbeiten am JSC ca. 300 Personen. Ein kleiner Teil ist direkt mit dem Rechnerbetrieb befasst. Ein Großteil betreibt angewandte Forschung z.B. zu großskaligen Simulation und künstlicher Intelligenz. Diese Expertise ist ein Schlüsselfaktor für den Erfolg.
 - Alle Mitarbeiter sind über den öffentlichen Dienst eingestellt. Das bringt einige Herausforderungen mit sich:
 - Es gibt wenig Flexibilität beim Gehalt,
 - Kündigungen sind so gut wie ausgeschlossen und
 - es ist schwierig, Dauerstellen zu schaffen.
 - Wichtig ist, dass auch das Betreibermodell KI-Expertise besitzen muss.
- **Kunden:**
 - Im FZ Jülich werden v.a. Simulation für wissenschaftliche Domänen wie Klimaforschung und Quantenphysik berechnet.
 - Industriepartner spielen nur eine kleine Rolle. Sie zahlen für die Nutzung.

2. Technik

- **Nutzung**
 - In der Regel liegt die Auslastung über 90%. Die Verfügbarkeit ist typischerweise über 80%. Dies ist auf geplante Wartungen sowie gelegentlich auftretende Schwierigkeiten mit der Hardware zurückzuführen.

- Zeitweise liegt die Auslastung bei weniger als 90%. Dies ist vor allem darauf zurückzuführen, dass Projekte gleichzeitig starten, aber zum Start oft noch keine Rechenzeit benötigen.
- **Limitierung:**
 - Es gibt eine maximale Berechnungszeit von 24 Stunden. Dies ist für eine gleichmäßig hohe Auslastung sehr entscheidend. In Ausnahmefällen und größeren Projekten sind Jobketten und Reservierung möglich.
- **Managementsoftware**
 - Die Jobvergabe erfolgt über das Queuing-System SLURM.
 - Die verwendete Managementsoftware ist eine Eigenentwicklung des FZ Jülich.

Interviewprotokoll Future of Life Institute

Interviewer:

Risto Uuk, Policy Researcher, Future of Life Institute

Interviewer:

Jörg Bienert, Alexander Thamm GmbH

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

05.12.2022

Zusammenfassung

Hinweis: Das Interview wurde auf Englisch geführt.

1. Thema KI-Foundation-Modelle

- Das Future of Life Institute hat einen Artikel veröffentlicht zum Thema: [Emerging Non-European Monopolies in the Global AI Market](#)
- Aktuell entwickelt lediglich AlephAlpha Foundation-Modelle in Europa. Sie können aber nicht mit den amerikanischen Modellen konkurrieren.
- Europa braucht eigene Modelle. Die Hoffnung ist, dass diese vertrauenswürdiger sind, da sie auf GDPR-konformen Daten und weiteren vertrauenswürdigen EU Gesetzen und Guidelines beruhen.
- Das Rennen um die besten Modelle hat negative Auswirkungen auf die Sicherheit. Es geht den Unternehmen darum, schnellstmöglich neue Modelle zu produzieren und nicht das sicherste Modell.
- Die Sicherheitsvorkehrungen bei GPT-3 sind bspw. einfach zu umgehen.

- Im AI Act gibt es aktuell keine Gründe für US-Unternehmen, ethische Modelle zu entwickeln.
- Auch Open Source Modelle haben Probleme. Sie sollten nicht ungetestet auf die Welt losgelassen werden. Es bringt auch niemand Autos ohne Bremsen auf die Straße und sagt, probiert mal aus. Der Provider muss sicherstellen, dass jedes Modell so sicher wie möglich ist.
- Lediglich große (amerikanische) Unternehmen haben die Ressourcen, um Foundation-Modelle zu entwickeln. Selbst bei einer Zusammenarbeit der Open-Source Community, dem Mittelstand und anderen Unternehmen, wird es schwierig, diese Vormachtstellung zu brechen.

Interviewprotokoll Höchstleistungsrechenzentrum Stuttgart

Interviewer:

Bastian Koller, Geschäftsführer, HLRS

Dennis Hoppe, Leiter Künstliche Intelligenz und Quantum Computing, HLRS

Interviewer:

Jörg Bienert, Alexander Thamm GmbH

Jakob Tesch, Ubermetrics

Patrick Bunk, Ubermetrics

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

22. November 2022

Zusammenfassung

1. Governance

- **Organisationsform**

- Das HLRS wurde 1996 als Bundeshöchstleistungsrechenzentrum gegründet und ist seit mehr als 25 Jahren eine zentrale Einrichtung der Universität Stuttgart. Das HLRS bietet ein umfassendes Paket an Ressourcen und Services für Hochleistungsrechnen (HPC), Datenanalyse, Künstliche Intelligenz, Visualisierung und verwandte Technologien. Das HLRS unterstützt Wissenschaftler, Ingenieure und Nutzer:innen aus vielen Forschungs- und Anwendungsbereichen durch die Bereitstellung von HPC-Tools und Fachwissen, die für Forschung, Entwicklung besserer Produkte,

Lösung komplexer und sehr rechenintensiver Probleme oder die Umsetzung neuer Ideen erforderlich sind.

- Das Gauss Centre for Supercomputing e.V. (GCS) dient dem Zusammenschluss der drei Bundeshöchstleistungsrechenzentren in Deutschland: HLRS (Höchstleistungsrechenzentrum der Universität Stuttgart), FZJ (Forschungszentrum Jülich) und LRZ (Leibniz Rechenzentrum, Garching bei München). Die drei Zentren stimmen sich im Rahmen des GCS ab, agieren aber im operativen Bereich autonom.

- **Kunden**

- Das HLRS unterstützt sowohl akademische Nutzer:innen als auch Kund:innen aus der Industrie.
- Kunden aus der Industrie sind in Anzahl und Nutzung der Systeme der Bundeshöchstleistungsrechenzentren ein Alleinstellungsmerkmal des HLRS. Hier können, im Schnitt, bis zu 10% der Rechenkapazität durch Industriekunden, auf Basis kostendeckender Preise, genutzt werden.
- Um die Nutzung der Rechenressourcen für Produktionsläufe durch die Industrie zu ermöglichen wurde 1995 eine Public Private Partnership mit den Industriepartnern Daimler und Porsche gegründet. Diese PPP, die HWW GmbH, ist auch heute noch aktiv; aktuelle Gesellschafter sind Porsche, T-Systems, das Land Baden-Württemberg, die Universität Stuttgart sowie das Karlsruher Institut für Technologie. Die HWW agiert primär als Vermittler von Rechenzeit.
- Aktuell rechnen zahlreiche KMUs und Großunternehmen regelmäßig am HLRS.
- Akquise erfolgt entweder direkt oder bspw. über die SICOS BW GmbH, die potenzielle Kund:innen anspricht. Oft kommt es am Anfang zu Testläufen, Proof-of-Concepts und dann im Idealfall zur Gewinnung neuer Kunden.
- Das HLRS bietet seine Rechen- und Speicherressourcen seinen Kunden zu kostendeckenden Preisen an.
- Um Industriekunden zu gewinnen, bietet das HLRS umfangreichen Support, Sicherheit (über ISO-Zertifizierung) und einen Fokus auf Nachhaltigkeit (bspw. Zertifizierung über den Blauen Engel). Dies sind Aspekte, die ein gängiger Cloud-Anbieter nicht bietet und Vertrauen schaffen.

- **Personal**

- Das HLRS hat feste Stellen im Budget der Universität Stuttgart, der Rest der Mitarbeitenden sind auf Drittmittel eingestellt.

- Im Rahmen des deutschen SiVeGCS Projektes (das Projekt, das u.a. die Rechner der drei Gauss Zentren finanziert), werden u.a. zwei weitere Stellen zur Interaktion mit der Industrie auf Landesebene gefördert. Die Förderung von SiVeGCS erfolgt durch den Bund (BMBF), sowie das Land Baden-Württemberg.
- Bei der Beschaffung neuer Flaggschiff-Systeme liegt oft ein Augenmerk auf eine weitergehende Kooperation mit dem Anbieter, der den Zuschlag bekommt. Innerhalb solcher Kooperationen ist auch Personal des Herstellers vor Ort anzustreben, sowie Mitarbeitende, die in der Kooperation mit dem HLRS zusammenarbeiten, um die HPC-Services zu stärken. Diese Personen unterstützen dann das HLRS und den Produktionsbetrieb.
- **Zugang zu Rechenressourcen**
 - Die Wissenschaft bewirbt sich bei Großprojekten zentral über das GCS und deren durch einen Lenkungsausschuss begutachtete peer-reviewte Calls.
 - Beim HLRS sind jederzeit Anfragen für Projekte auf den Systemen möglich, die von einem wissenschaftlichen Lenkungsausschuss begutachtet werden. Ein Test-Zugang kann individuell nach Absprache zugänglich gewährt werden.
 - Für Entwicklungsprojekte der Wirtschaft gibt es einen klaren Zugangsprozess und gegebenenfalls die Möglichkeit der Priorisierung von Rechenjobs.

2. Ausgewählte Projekte mit Industriebezug

- **Fortissimo - Factories of the Future Resources, Technology, Infrastructure and Services for Simulation and Modelling**
 - Das aktuelle Projekt (FF4EuroHPC) wird über die EuroHPC Joint Undertaking gefördert.
 - Es ermöglicht KMUs den Zugang zu HPC-Systemen in Europa, wie dem HLRS, um sogenannte Businessexperimente durchzuführen.
 - Für die Nutzung müssen Projektanträge geschrieben werden, die von Experten begutachtet werden.
 - Ergebnisse sind Case Studies, die veröffentlicht werden.
- **CATALYST**
 - Das Projekt läuft nach 5 Jahren Ende des Jahres 2022 aus; es wurde durch das MWK Baden-Württemberg gefördert.
 - Das Projekt stellt Kontingente (Personal und Rechenbudget) sowohl für Wissenschaft und Industrie zur Verfügung, um Technologien und Potentiale zu validieren.

- Eine geringe Einstiegshürde (bspw. gibt es keine lange Antragsphase) ermöglicht eine hohe Akzeptanz des Förderprojektes.
- Projekte haben in der Regel eine Laufzeit zwischen drei und zwölf Monaten. Ergebnis ist eine Success Story oder mindestens eine gemeinsame wissenschaftliche Publikation, die veröffentlicht wird.
- Das HLRS hat darüber auch neue Kunden gewonnen.
- **Solution Center**
 - Die Solution Center sind eigenständige Vereine, die den Mitgliedern als Wissensplattform dienen.
 - Anschubfinanzierung erhalten sie vom Land Baden-Württemberg. Mittelfristig müssen sie sich aber über Mitgliedsbeiträge und Förderprojekte finanzieren.
 - Für das HLRS ermöglichen sie den Kontakt in verschiedene Branchen und sind eine wichtige Säule im Wissens- und Technologietransfer.
 - Satellitenprojekte der Solution Center werden unter anderem auf den HLRS-Systemen gerechnet.

3. Technik

- Verlässlichkeit ist für das HLRS wichtig. Daher:
 - beschafft das HLRS seine neuen Systeme meist basierend auf ausgereiften Konzepten und Technologien, um Stabilität und Sicherheit zu garantieren; dies schließt jedoch den Einsatz neuester Technologien nicht aus, die als Teil des Gesamtsystems eingebunden werden können
 - bietet das HLRS einen vollumfänglichen Service für Kunden an.
 - unterzieht sich das HLRS regelmäßig relevanten Zertifizierungen wie EMAS, Blauer Engel oder Sicherheitsstandards wie TISAX oder im ersten Halbjahr 2023 der ISO 27001 Zertifizierung.

4. LEAM

- Bei LEAM müssen Wissenschaft und Industrie klar getrennt werden.
- Die 10% Nutzung für die Industrie ist eine Sonderregelung, die andere HPC-Zentren in Deutschland nicht so einfach nutzen können.
- Eine individuelle Kooperation zwischen LEAM und dem HLRS ist zu begrüßen.

Interviewprotokoll

Höchstleistungsrechenzentrum Stuttgart

Interviewer:

Oleksandr Shcherbakov, Wissenschaftlicher Mitarbeiter, HLRS

Interviewer:

Jörg Bienert, Alexander Thamm

Jakob Tesch, Ubermetrics

Hauke Timmermann, eco Verband

Protokollant:

Alex Dickmann, KI Bundesverband

Kim Lambers, eco Verband

Datum:

02. Dezember 2022

Zusammenfassung

1. Hardware

- In Stuttgart sind insgesamt 24 KI-Knoten mit jeweils 8 NVIDIA A100 GPUs vorhanden (= 192 GPUs). Die KI-Knoten sind in das HPC-System integriert, um hybride HPC/KI Workflows zu ermöglichen. Neben diesen KI-Knoten existiert noch ein weiteres KI-System (CS-Storm) mit insgesamt 64 NVIDIA V100 GPUs und lokalen SSDs. Das reicht für LEAM eher nicht aus.
- Im Jahr 2024 soll ein neues HPC-System installiert werden, welches aller Voraussicht nach ebenfalls über Beschleunigertechnologie verfügen wird. Weitere Informationen können erst Ende 2023/Anfang 2024 bekannt gegeben werden. Das System wird öffentlich ausgeschrieben.
- Der HAWK-Cluster wird wassergekühlt; allgemein geht der Trend aufgrund von Energieeffizienz und Abwärmenutzung in Richtung Wasserkühlung.
- Gekühlt wird direkt auf der CPU und dem Arbeitsspeicherriegel.
- Co-Location ist zunächst am HLRS nicht vorgesehen; über kleinere Systeme kann individuell diskutiert werden.
- Linkliste:
 - [Entgeltordnung](#)
 - <https://www.hlrs.de/solutions/systems/hpe-apollo-hawk>
 - <https://www.hlrs.de/solutions/systems/cray-cs-storm>
 - [https://kb.hlrs.de/platforms/index.php/Batch_System_PBSPro_\(vulcan\)#Node_types](https://kb.hlrs.de/platforms/index.php/Batch_System_PBSPro_(vulcan)#Node_types)

Interviewprotokoll Hubert Burda Media

Interviewter:

Jean-Paul Schmetz, Chief Scientist, Hubert Burda Media

Interviewer:

Jörg Bienert, Alexander Thamm GmbH

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

14. Dezember 2022

Zusammenfassung

1. KI-Foundation-Modelle

- Burda sieht eine hohe Bedeutung von Sprachmodellen.
- Derzeit wird u.a. in Vietnam ein eigenes Sprachmodell entwickelt.
- Ein deutsches/europäisches Foundation-Model würde von Burda zum Tuning und Erstellung unterschiedlicher Anwendungen genutzt werden.

2. Daten-Problematik

- Problematisch ist vor allem die Verfügbarkeit von qualitativ hochwertigen Daten. Es reicht nicht aus, einen "Sumpf an Daten" zu haben, diese müssen auch gerankt, kommentiert und überprüft werden. Das kostet viel Arbeitszeit.
- Investitionen sind neben HW vor allem in die Aufarbeitung der Daten erforderlich
- Burda besitzt qualitativ hochwertige Datensätze, die sie Kunden anbieten.

3. Zusammenarbeit mit LEAM

- Burda könnte LEAM prinzipiell einen gut kuratierten Datensatz gegen Lizenzgebühren zur Verfügung stellen..
- Es ist unbestritten, dass jeder Anwender ein europäisches LEAM Modell benutzen würde, wenn es kompetitiv zu den amerikanischen Modellen ist.

Interviewprotokoll Hugging Face

Interviewter:

Thomas Wolf, Co-Founder Hugging Face

Carlos Munoz Ferrandis, Hugging Face

Interviewer:

Jörg Bienert, Alexander Thamm GmbH
Alexander Löser, Berliner Hochschule für Technik (BHT)

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

20. Dezember 2022

Zusammenfassung

Hinweis: Das Interview wurde auf Englisch geführt.

AI Foundation Models

- The Project LEAM is a good initiative. There is a need for a data center that specializes in AI.
- The danger is that in the future all foundation models will come from American private companies.
- Hugging Face has trained their model on the French computer Jean Zay. Their experience shows that it is optimized for other workflows and might be challenging to train an AI foundation model with the standardized workflows of the Jean Zay.
- At Hugging Face, there are already initial considerations to build their own computing cluster. However, the goal is not economic profit, but to support the mission of open models.
- In addition, high quality data is critical. There needs to be public support for labeling data.
- Hugging Face is currently investigating what makes a good dataset. The second step is to produce these datasets on a large scale.

Collaboration with LEAM

- Hugging Face would be interested in using the LEAM computing center.

Interviewprotokoll Ionos

Interviewer:

Rainer Sträter, SVP Cloud Services and Global Platform Hosting, Ionos

Interviewer:

Jörg Bienert, Alexander Thamm GmbH

Hauke Timmermann, eco Verband

Jakob Tesch, Ubermetrics

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

14. Dezember 2022

Zusammenfassung

1. Verfügbarkeiten Ionos

- Ionos betreibt aktuell drei Rechenzentren in Deutschland.
- In Frankfurt wird aktuell ein neues Rechenzentrum errichtet. Der Bau beginnt 2023. Das Rechenzentrum wird 2024/2025 betriebsbereit sein. Das Ziel ist es, carbon negative zu sein.
- Energiekosten machen aktuell bereits ein Drittel aller Kosten aus. Die Kosten werden in Zukunft sicherlich weiter steigen. Energieeinsparung ist daher höchst relevant.

2. Zusammenarbeit mit LEAM

- Die Kalkulation für den Aufbau und Betrieb eines Supercomputing RZ für LEAM sieht realistisch aus. Unter den angegebenen Parametern lässt sich das Projekt realisieren.
- Beschaffung und Aufbau der Infrastruktur stellt für Ionos kein Problem dar. Die Lieferung und der Aufbau der Boxen alleine dauert aber sicherlich ein halbes Jahr.
- Entscheidend ist für Ionos eine Zusage zur Abnahme von Rechenzeit über 24 Monate. Die genaue Höhe muss im operativen Betrieb geklärt werden.
- V.a. das neue Rechenzentrum in Frankfurt bietet sich für LEAM an. Die anderen Rechenzentren in Berlin und Karlsruhe eher weniger.
- Der Preis richtet sich vor allem nach den Kosten für den Aufbau und Betrieb.
- Eine genauere Ausgestaltung kann im zweiten Schritt diskutiert werden.

3. Finanzierung

- Für Ionos ist die Unterstützung von LEAM eine ernsthafte Option.
- Die Finanzierung könnte u.a. erfolgen durch
 - Aufbau und Betrieb eines eigenen RZ und Vereinbarung einer garantierten Abnahmemenge (z.B. 60%-70%) der Kapazitäten durch den Bund
 - Aufbau und Betrieb eines eigenen RZ mit Anschubfinanzierung durch die öffentliche Hand und Bereitstellung von Rechenkapazität für Wissenschaft/Startups/Public.
- Ionos hat umfangreiche Erfahrungen mit öffentlichen Ausschreibungen.
- Bei beiden Vorgehen gilt, dass das Konsortium so klein wie möglich gehalten werden sollte. Ansonsten läuft das Projekt Gefahr, in Abstimmungsschleifen zu verlaufen.

4. Kühlung

- Die A100 besitzen serienmäßig keine Liquidkühlung. Es gibt aber Anbieter, die diese anpassen. Die neue H100 Generation hat eine Standardmäßige Liquidkühlung. Deren Nutzung wird daher empfohlen.
- Die Power Usage Effectiveness (PUE) liegt in neuen Rechenzentren bei unter 1,1. Es wird also wenig Energie für andere Zwecke als den Betrieb der Server benötigt.
- Wirkliche Effizienz entsteht aber erst, wenn die Abwärme auch großflächig genutzt und nicht "in die Umwelt gepustet" wird. Hier lassen sich auch neue Geschäftsfelder erschließen.

Interviewprotokoll KI Park

Interviewer:

Olly Salzmann, Stellvertretender Vorstandsvorsitzender

Interviewer & Protokollant:

Vanessa Cann, KI Bundesverband

Datum:

21. Dezember 2022

Zusammenfassung

1. Warum unterstützt der KI Park eine Initiative wie LEAM?

- KI Park Mission: Wir wollen einen Beitrag leisten, um KI-Innovationen "made in Germany and Europe" zu beschleunigen und damit Deutschland und die EU bis 2030 zu einem global führenden Innovationsstandort für KI zu machen
- Ziel der Aktivitäten des KI Parks ist die Schaffung von Voraussetzungen für die erfolgreiche Erforschung und Entwicklung zukunftsweisender KI-Technologien. Der Schwerpunkt des KI Parks liegt auf Deutschland bzw. Europa und in klarer Abgrenzung zu den USA und China. Daraus abgeleitet die Notwendigkeit einer deutschen bzw. europäischen KI-Souveränität anerkennt und ermöglichen möchte.
- LEAM bietet hier eine gute Möglichkeit den Technologiestandort Deutschland bzw. Europa wieder auf die Weltkarte zu bringen. Diese Ambition wird durch das Ökosystem des KI Park in der Hauptstadt Berlin unterstützt.
- Der KI Park existiert und definiert sich über seine Mitglieder und repräsentiert eine ausgewählte Gruppe an global führenden Unternehmen wie z.B. Deloitte, VW, Schaeffler oder Celonis und Forschungseinrichtungen wie das ZUSE Institute, die Friedrich-Alexander Universität, Humboldt Innovation oder ISST Fraunhofer. Die Mitglieder des KI Parks können sowohl von den neuesten Technologien und Lösungen profitieren, als auch vom kuratierten Zugang zu Wissen, Fähigkeiten und Erfahrungen. Das liefert auch einen Beitrag zum Gelingen von Initiativen wie

zB LEAM.

2. Warum eignet sich der KI Park als Standort für LEAM?

- Im September 2022 wurde das neuste und modernste Data Center in Deutschland direkt neben der KI Park-Geschäftsstelle im Marienpark eröffnet und wird in den nächsten Monaten noch kontinuierlich erweitert. Es besteht daher die Möglichkeit für LEAM auf modernste aber bereits bestehende Infrastruktur zurückzugreifen zu können. Das bestehende Data Center zeichnet sich insbesondere auch durch seine Nachhaltigkeit aus, so wird die Abwärme des Data Center zum Heizen des Marienparkcampus verwendet und der Campus hat seinen eigenen Solarpark zur Stromproduktion. Neben der bestehenden bzw. im Bau befindlichen (Server) Infrastruktur bietet die Örtlichkeit des Marienpark Campus in Berlin auch die Möglichkeit, weitere Firmen und andere Organisationen vor Ort anzusiedeln und somit die Infrastruktur und Ergebnisse von LEAM zu konsumieren bzw. darauf aufzubauen weitere Produkte entwickeln zu können.
- Der Community Gedanke wird allein durch physische Nähe des KI Park zu den Schaltzentralen des deutschen Staates (u.a. Regierung, Parlament), Kompetenzzentren der Industrie (z.B. Siemens City, AWS Research, Innovationshubs, etc.) und Forschungseinrichtungen (drei Universitäten, DFKI, Fraunhofer, etc) weiter unterstützt und ausgebaut. Es kommt so also zu einer Verknüpfung von Infrastruktur und Innovations-Community in nächster Umgebung des deutschen Startups und Innovationszentrums Berlin im Rahmen des Marienpark Technologie- und Innovationscampus.
- Ganz unabhängig vom KI Park in Berlin, bietet sich die Umsetzung eines Data Centers für Projekte wie LEAM allein aus Gesichtspunkten des Energiebedarfs und den damit verbundenen Nachhaltigkeitsaspekten im Norden von Deutschland und insbesondere in Küstennähe an. Da hier zum einen ausreichend Strom erzeugt wird und zumindest für den Norden bereits ein ausreichend großes Verteilernetz existiert, wohingegen die Leitungen in Richtung Süden noch fehlen.

3. Welche Anknüpfungspunkte gibt es zu bestehenden Aktivitäten des KI Park?

- Es gibt bereits verschiedene Initiativen im Rahmen des KI Park, die sowohl als Grundlage für LEAM dienen können, bzw. sich gut mit LEAM ergänzen. Zum einen ist der KI Park physisch Teil des im Aufbau befindlichem Marienparks, ein in Berlin Tempelhof-Schöneberg gelegener, rund 360.000 m² großer Gewerbepark, in dem ein innovatives Ökosystem mit Unternehmen und Start-ups aus zukunftsweisenden Bereichen wie KI, Critical Infrastructure, Additiver Fertigung

oder auch Laboren für die physische Testung neuer Technologie z.B. im Bauhaus der Erde. Es besteht also noch ausreichend Platz und Infrastruktur für die Ansiedlung weiterer Unternehmen, Initiativen für den Aufbau eines Innovations-Campus und einer (physischen) KI-Community.

- Die Notwendigkeit für viele kleine wie großen Unternehmen einen rechtlichen Rahmen für schnelles Experimentieren und Forschung auch mit externen Partnern zur Verfügung zu haben ist ein Kern des KI Park Ökosystems. Kleine Unternehmen scheuen oft die notwendigen jedoch meist riskanten Investitionen in Technologie und Infrastruktur; bei großen Unternehmen steht meist organisatorische Komplexität und generelle Widerstände, die mit zunehmender Unternehmensgröße entstehen, im Wege neuer Innovationen durch schnelles Experimentieren und Entwicklung Raum zu geben. In Bezug auf LEAM bietet der KI Park eine rechtliche und organisatorische Umgebung für die erfolgreiche Realisierung des LEAM Projektes, sowie eine möglichst geringe Barriere bzw. eine Vielzahl von Anknüpfungspunkten für die anschließende (kommerzielle) Nutzung von riesigen Sprachmodellen.
- Einer der zentralen Säulen des KI Park ist KI-Ethik oder verantwortungsvolle KI bzw. Wert getriebener Einsatz von Technologie. Da das LEAM Projekt in der Projektbeschreibung auch in Abgrenzung an andere, vergleichbare Initiativen den Datenschutz, europäische Werte und Open-Source in den Vordergrund stellt, bietet sich hier für den KI Park eine Reihe von Anknüpfungspunkten. Ergänzend ist hier auch die Unterstützung eines 60 Millionen Euros schweren, EU- weiten Konsortium zur Erforschung von bildgebenden Verfahren und vertrauensspendender KI im Gesundheitswesen durch den KI Park zu erwähnen (TEF Health via FAU).
- Das LEAM Projekt mit der dazugehörigen Infrastruktur ergänzt sich perfekt mit den bereits angelaufenen Initiativen des KI Parks im Marienpark seinen Mitgliedern Zugang zu 5G bzw. 6G Mobilfunknetzwerke und Quanten Computer zu Testzwecken bzw. in Zukunft bei entsprechendem Erfolg auch im Regelbetrieb zur Verfügung zu stellen.

4. Wie ließe sich LEAM im Rahmen des KI Park organisatorisch umsetzen und finanzieren?

- Die Betreibergesellschaft für LEAM und dessen Kommerzialisierung könnte als Teil des wirtschaftlichen Geschäftsbetriebs den KI Parks realisiert werden. Idee: ohne Profitmaximierung aber kostendeckend bzw. zur Finanzierung weiter Forschung bzw. Weiterentwicklung
- Mitglieder des KI Parks können die Umsetzung des LEAM Projektes sowohl finanziell wie auch organisatorisch unterstützen. Der KI Park kann hier als Koordinierungsstelle und erster Ansprechpartner zur Verfügung stehen.

- Unter den an LEAM interessierten Mitgliedern und Freunden des KI Park sind neben DAX 30 Unternehmen, deutschen Unicorns auch die weltweit größte Professional Service Firm (Deloitte), der wichtigste KI-Infrastruktur Anbieter für riesige KI-Modelle schlechthin (NVIDIA), ein hochinnovativer Projektentwickler mit Schwerpunkt Technologie und Innovation (Investa), sowie einer der global führenden Data Center Entwickler und Betreiber (NTT). So besteht hier sowohl die Bereitschaft mit eigenen Kräften, Technologie und Wissen zur Verfügung zu stellen, als auch ein Interesse sich für die deutsche bzw. europäische KI-Souveränität zu engagieren.

Interviewprotokoll Lufthansa

Interviewer:

Christian Spannbauer, CTO, Lufthansa Group Digital Hangar

Interviewer:

Jörg Bienert, Alexander Thamm GmbH

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

22. Dezember 2022

Zusammenfassung

1. KI-Foundation-Modelle

- Aktuell arbeitet Lufthansa im Servicebereich mit Foundation-Modellen.
- Generell sind die Bereiche Optimierung, Sustainability, Efficiency u.ä. interessanter. Dort sollen spezifische Anwendungsfälle identifiziert werden
- Bei sensiblen Daten ist die Nutzung amerikanischer Modelle schwierig. Bei operativen Daten (bspw. Flugdaten, Wetterdaten, u.ä.) ist die Problematik nicht so groß.
- Die Herausforderung ist aktuell nicht die Technologie, sondern die internen Möglichkeiten. Ohne externe Unterstützung ist eine Implementierung nicht möglich.

2. Zusammenarbeit mit LEAM

- Interesse an der Zusammenarbeit ist da und Lufthansa ist bereit, sich in dem Bereich zu engagieren.
- Eine Beteiligung wäre im Prinzip möglich, ist aber unter den aktuellen gesamtwirtschaftlichen Rahmenbedingungen zu prüfen und müsste sich für die Lufthansa aber betriebswirtschaftlich rechnen.

Interviewprotokoll Mercedes-Benz

Interviewer:

Jochen Kaiser, Chief Data Officer, Mercedes-Benz Group AG

Interviewer:

Jörg Bienert, Alexander Thamm GmbH

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

15. Dezember 2022

Zusammenfassung

1. Allgemeines

- Die Entwicklung und Nutzung von KI ist vor allem dann wichtig, wenn sich hieraus ein Wettbewerb differenzierender Faktor ergibt.
- China ist derzeit dominierend (56% der Patente) gefolgt von USA (~20%), Europa liegt mit ca. 8% schon sehr weit hinten.
- Einsatzgebiet von KI sind neben dem autonomen Fahren vor allem Fahrzeugsteuerung und Fahrzeugintegration.
- Derzeit läuft ein Programm, um mehrere hundert Mitarbeiter im Bereich Data&AI auszubilden.

2. KI-Foundation-Modelle

- Die Services von OpenAI werden aktuell genutzt. Prinzipiell ist es sinnvoller, große Modelle selbst zu entwickeln oder auf Basis großer Modelle eigene Modelle zu tunen (auch unter Gesichtspunkten von Bias, Ethik und GDPR)
- Foundation-Modelle sollten idealerweise Open Source und über die gesamte Industrie angeboten werden.
- Mercedes-Benz betreibt bereits ein eigenes Rechenzentrum in Norwegen. Das Thema Nachhaltigkeit ist dabei ein treibender Faktor.
- Neben der Rechenzeit sind vor allem die Themen Personal und Daten eine große Herausforderung für die Entwicklung von Foundation Modellen. Der Data Act der EU kann dabei helfen, diese frei zugänglich zu machen.

3. Zusammenarbeit mit LEAM

- Mercedes-Benz kann die geplanten LEAM Services nutzen. Bei amerikanischen Services behindern juristische Vorgaben oft die Anpassung und Nutzung bzw. verlangsamen den Start von Projekten durch längliche Klärung von juristischen Klauseln.

- Synergien mit anderen Unternehmen u.a. auch im Rahmen von Catena-X sind sinnvoll.
- Opendata und der European Data Act können sich mit der Entwicklung von Foundation Modellen gegenseitig positiv beeinflussen.
- Die Beteiligung in einem Konsortium zur Finanzierung von LEAM ist gut vorstellbar und sinnvoll. Dabei sollte auch auf das Thema Daten geachtet werden. Die Herausforderung ist die Geschwindigkeit, mit der das Projekt realisiert wird.
- Die Initiative ist v.a. auch für kleinere Akteure wie Zulieferer entscheidend. Die werden das Thema Nutzung von Foundation-Modelle nicht alleine umsetzen können.

Interviewprotokoll Ministerium für Wirtschaft, Arbeit und Tourismus Baden-Württemberg

Interviewer:

Dr. Peter Mendler, Leitung des Referats „Industrie- und Technologiepolitik, Digitalisierung“, Stv. Abteilungsleiter im Ministerium für Wirtschaft, Arbeit und Tourismus Baden-Württemberg

Interviewer:

Jörg Bienert, Alexander Thamm
Vanessa Cann, KI Bundesverband

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

30. November 2022

Zusammenfassung

1. Aktivitäten in Baden-Württemberg

- Aleph Alpha als KI Champion 2021 war der erste öffentlich sichtbare Kontaktpunkt mit dem Thema KI-Foundation Modelle.
- IPAI & KI-Exzellenzzentren sind weitere Anknüpfungspunkte mit dem Thema.
- Es gibt verschiedene Förderprojekte und –programme zum Thema KI, aber Foundation Modelle laufen unter anderen Dimensionen.
- Ziel des Landes Baden-Württemberg ist es, Ökosysteme zu schaffen, die auch physisch angesetzt sind.

2. Warum ist LEAM für Baden-Württemberg interessant?

- Es passt in die KI-Strategie der Landesregierung. Generell ist das Thema KI hoch auf der Agenda des Landes angesetzt.
- Baden-Württemberg hat seine Aktivitäten bei KI in den letzten Jahren mit hohem Tempo entwickelt. Es gibt Synergien mit:
 - dem IPAI und den KI-Exzellenzzentren,
 - den Universitäten und Forschungseinrichtungen im Land, insbesondere auch mit dem CyberValley,
 - den zahlreichen Unternehmen mit KI-Aktivitäten einschließlich der zahlreichen KI-Startups
 - dem bestehenden KI-Ökosystem und
 - verschiedenen Wirtschaftsinitiativen wie bspw. die europäische Testing and Experimentation Facility (TEF Manufacturing) oder die Vanguard Initiative.
- Aleph Alpha gilt als Leuchtturm für das Thema.
- LEAM würde dem rasch wachsenden KI-Ökosystem in Baden-Württemberg zusätzliche Dynamik geben und Baden-Württemberg könnte ggf. mit hoher Wahrscheinlichkeit rasch Flächen für LEAM mobilisieren. Dadurch könnten sich zusätzliche nationale und europäische Kooperationen entwickeln.

3. LEAM

- Die Initiative wird ein Erfolg, wenn es eine Zusammenarbeit zwischen privaten und öffentlichen Akteuren ist. Dafür müssen beihilferechtliche Fragestellungen geklärt werden.
- Für die Finanzierung über ein IPCEI-Projekt müssten zuerst die beihilferechtlichen Voraussetzungen geschaffen werden.
- Es braucht ein tragfähiges Finanzierungskonzept und ein nachhaltiges Geschäftsmodell, damit das Projekt wirtschaftlich erfolgreich sein wird.

4. Finanzierung:

- Die Landesregierung Baden-Württemberg möchte bei KI im internationalen Innovationswettbewerb vorne dabei sein.
- Das Land würde sich dafür einsetzen, dass LEAM nach Baden-Württemberg kommt.
- Private Finanziere müssen in das Projekt eingebunden werden. Hier stellt das Land bei Bedarf gerne Kontakte her.

Interviewprotokoll Ministerium für Wirtschaft, Industrie, Klimaschutz und Energie des Landes Nordrhein-Westfalen

Interviewter:

Christian Dinnus, Referatsleiter Digitale Wirtschaft, Digitale Geschäftsmodelle (MWIKE NRW)

Sebastian Ley, Digitale Wirtschaft, Digitale Geschäftsmodelle (MWIKE NRW)

Dr. Dirk Hecker, Managing Director Fraunhofer Allianz Big Data, Fraunhofer IAIS

Interviewer:

Vanessa Cann, KI Bundesverband

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

06.12.2022

Zusammenfassung

Wichtiger Hinweis: Die Interviewten äußern ihre fachliche Einschätzung, die im Wesentlichen auf öffentlich verfügbaren Informationen beruhen. Zusagen oder ähnliches sind mit den Äußerungen nicht verbunden.

1. Thema KI-Foundation-Modelle

- Die Größe der Modelle und Recheneinheiten spielt eine Rolle. Je größer diese sind, desto mehr Wirkung lässt sich erzielen. Um Basismodelle entstehen Ökosysteme, die darauf Zugriff haben möchten. Diese Ökosysteme sind das, was wir in Deutschland und Europa erreichen sollten. Wichtig ist entsprechend eine Anwendbarkeit der Basismodelle (Anpassen für konkrete Anwendungsfälle).

2. Standort NRW

- Das Rheinland liegt ideal zwischen den Welt-Internetknoten in Frankfurt und Amsterdam, bei hoher Stromversorgungssicherheit und eingebettet zwischen den Städten Köln, Düsseldorf, Aachen, Bonn, mit hohem Potenzial an Fachkräften (vgl. Machbarkeitsstudie Dateninfrastrukturen im Rheinischen Revier unter www.dateninfrastruktur.nrw).
- Die Region ist auch ein starker Forschungsstandort mit relevanten Forschungsfeldern, bspw. RWTH Aachen, Universität Bonn, Universität Köln, Fraunhofer IAIS, FZ Jülich (Quanten- und Super-Computing)
- Nordrhein-Westfalen verfügt über ein stark aufstrebendes Start-up Ökosystem, dass gerade im Rheinland (Aachen, Bonn, Köln, Düsseldorf) besonders stark ausgeprägt ist (vgl. www.wirtschaft.nrw/sites/default/files/documents/nrw_start-

[up-report_2020.pdf](#)). Im speziellen gibt es im Umfeld von KI und Sprachmodellen besonders hervorzuhebende Start-ups, wie bspw. DeepL, das erste NRW-Unicorn mit Sitz in Köln oder Cognigy aus Düsseldorf.

- Nordrhein-Westfalen hat die eigene Kompetenzplattform KI.NRW als zentrale Anlaufstelle für Künstliche Intelligenz in Nordrhein-Westfalen. Neben konkreten Projekten zum Transfer aus der Forschung in die Wirtschaft ist das Thema „KI-Zertifizierung“ ein wichtiges Thema in NRW.
- Im Rheinischen Revier (Hürth) entsteht das Projekt AI Village.
- Das Gaia-X-Projekt Open GPTX für die Entwicklung großer KI-Sprachmodelle hat mehrere starke Teilnehmer und Partner aus NRW, so Westdeutscher Rundfunk (WDR), Forschungszentrum Jülich, Fraunhofer IAIS (und mehr).
- Auf Basis der genannten Voraussetzungen werden Maßnahmen umgesetzt, die das Rheinland zur Digitalregion weiterentwickeln, insbesondere sollen Digitalparks entstehen.
- Ein Digitalpark ist eine für Unternehmen der Digitalwirtschaft optimierte Gewerbefläche, in räumlicher Nähe zu großen Rechenzentren und Internetknoten (Grund: sehr schnelle Reaktions- bzw. Latenzzeiten). Die Finanzierung erfolgt über private Investoren.
- In einem Digitalpark steht immer auch ein Rechenzentrum. Hier werden Synergien und große Chancen zu dem Vorhaben „KI-Rechenzentrum“ gesehen.

3. Governance

- Die Umsetzung im Rahmen eines PPP-Modells ist eine Möglichkeit, die im Detail zu prüfen wäre.
- Die Höhe des Finanzierungsbedarfs lässt den Bund als natürlichen Partner dieses Projekts erscheinen. Fachlich werden starke Anknüpfungspunkte zu nordrhein-westfälischen Initiativen gesehen (siehe „Standort NRW“, sowie „Housing“).

4. Housing

- Standorte für Digitalparks gibt es in der Region, und damit auch für ein Rechenzentrum in der für das KI-Rechenzentrum angedachten Größenordnung.
- Die Anforderungen (Infrastruktur, Größe, Lage) an einen Standort und die Flächensuche sind sehr wichtig und sollten bei der Machbarkeitsstudie mit bedacht werden. Kompetenzen bei der Flächensuche für Rechenzentren liegen in NRW vor.
- Eine gute Stromversorgung auf der einen Seite und eine gute Nutzung entstehender Abwärme auf der anderen Seite sind für Rechenzentren wesentlich. Insgesamt sollte das Thema „Nachhaltigkeit“ in all seinen Facetten berücksichtigt werden (Energieeffizienz, erneuerbare Energien, Wassereffizienz, Recycling).

Interviewprotokoll MPI

Interviewer:

Prof. Dietmar Harhoff, Geschäftsführender Direktor, Max-Planck Institut für Innovation und Wettbewerb

Interviewer:

Jörg Bienert, Alexander Thamm GmbH

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

21. Dezember 2022

Zusammenfassung

1. KI-Foundation-Modelle

- Herr Prof. Harhoff stimmt der Analyse der LEAM-Initiative definitiv zu. Europa darf bei diesem Thema nicht in die Abhängigkeit der Amerikaner und Chinesen geraten.
- Es ist vor allem eine Frage der Ressourcen. Die Entwicklung von GPT-3 zu ChatGPT war bspw. eine Frage des Aufwands und kein technologischer Durchbruch.
- Die Frage nach verfügbaren Daten wird entscheidend sein. Zu viele europäische Organisationen halten ihre Daten noch verdeckt.

2. Next-Steps

- Er empfiehlt der LEAM- Initiative weitere Gespräche.

Interviewprotokoll Otto

Interviewer:

Dr. Michael Müller-Wünsch, CIO, Otto GmbH & Co KG

Interviewer & Protokollant:

Jörg Bienert, Alexander Thamm GmbH

Datum:

22. Dezember 2022

Zusammenfassung

1. KI-Foundation-Modelle

- Bis 2030 könnten 50% der IT-Anwendungen durch KI-Modelle geprägt sein.
- Die Gefahr ist, dass alle Daten, die in US-Cloud Services gespeichert und verarbeitet werden, sich nicht mehr im EU-Raum befinden. Letztendlich machen wir Amerika damit stark.
- Otto evaluiert derzeit KI-Modelle, u.a. neben Sprachmodellen auch im Bereich Bild- und Videobearbeitung. Große KI Modelle werden in diesem Umfeld zunehmend an Bedeutung gewinnen.
- Otto muss prüfen, inwiefern sie US-Services aus Datenschutz und Datensicherheitsaspekten nutzen können.

2. Zusammenarbeit mit LEAM

- Ein unabhängiger deutscher Service zur Erstellung und Nutzung von Foundation Modellen ist hilfreich und zu begrüßen. Eine Initiative wie LEAM ist politisch sinnvoll.
- Das Angebot muss allerdings wettbewerbsfähig im Hinblick auf Servicequalität und Kosten sein.
- Eine Zusammenarbeit mit LEAM und eine Nutzung der Services wird gewünscht.
- Eine mögliche Investition in ein Joint-Venture wäre sinnvoll, ist aber kurz- und mittelfristig wegen der angespannten wirtschaftlichen Situation nicht realistisch.

Interviewprotokoll REWE

Interviewer:

Lorenz Determann, Bereichsleiter Analytics, REWE Group

Interviewer & Protokollant:

Jörg Bienert, Alexander Thamm GmbH

Datum

15.12.2022

Allgemeine Einschätzung

- Allgemein wird das LEAM Vorhaben begrüßt, vor allem unter dem Aspekt, alternative Angebote für General Purpose Modelle aus Deutschland zu bekommen
- Eine Nutzung der LEAM Services bzw. Foundation Modelle wird stark abhängig sein, von dem Mehrwert und wirtschaftlichen Nutzen – vor allem hinsichtlich

- Den Kosten der bereitgestellten Services/Modelnutzung im Wettbewerbsvergleich
- Der Qualität der Modelle

KI bei Rewe

- Derzeit werden in unterschiedlichen Bereichen eher kleine / mittelgroße KI-Modelle selbst entwickelt (u.a. auch auf Basis von Bert)
- Bei zentralen Modellen ist eine Eigenentwicklung wichtig, um mit eigenen Daten und Berechnungen Wettbewerbsvorteile erzielen zu können
- In allgemeineren, unkritischen Bereichen ist auch eine Nutzung von allgemeinen Modellen denkbar

Cloud Infrastruktur

- Rewe greift u.a. auch im Data&Analytics / AI Bereich auf die Cloud Services von Google zurück
- Eine mögliche Abhängigkeit von amerikanischen Providern im Bereich der KI Applicationen / KI Foundation Modellen wird in diesem Kontext nicht als besonders kritisch gesehen.

Interviewprotokoll SAP

Interviewter:

Dr. Feiyu Xu, Senior Vice President, Global Head of Artificial Intelligence, SAP

Interviewer:

Dr. Sven Schmeier, DFKI

Dr. Gerhard Paass, Fraunhofer IAIS

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

13. November 2022

Zusammenfassung

1. KI-Foundation-Modelle

- SAP nutzt Foundation-Modelle bereits, v.a. in der Dokumentenverarbeitung. Dabei spielen auch multimodale Modelle eine Rolle, wenn bspw. Rechnungen per Foto geschickt werden und weiterverarbeitet werden müssen. Im Bereich Prozessplanung machen wir auch Experimente mit Foundation-Modellen.

- Entscheidend für einen erfolgreichen Einsatz von Foundation-Modellen sind folgende Aspekte:
 - (1) Konzepte für den sicheren Zugang zu den Daten;
 - (2) Eine Infrastruktur zur Datenhaltung und Rechenkapazität, die auf grüner Energie basiert;
 - (3) Realisierung von „Data to Value“: Wie können wir domänenspezifische Modelle trainieren und damit wirtschaftlichen Mehrwert schaffen, z.B. für Geschäftsprozesse, Prozesse in Bereichen wie Manufacturing oder Supply Chain etc. bzw. auch für industriespezifische Anpassungen.
- Wichtig in diesem Kontext ist, dass die deutsche Industrie die KI-Foundation-Modelle nutzen kann. Viele deutsche Firmen haben keinen Zugriff auf die notwendigen menschlichen Ressourcen, wie Data Scientists, und auf die notwendige KI-Infrastruktur. Deshalb muss der Zugang zu KI-Technologien, insbesondere zu Foundation-Modellen, auch möglichst einfach gestaltet werden. In der Planung von LEAM muss man unterschiedliche Personas und Stakeholders für die Entwicklung und Anwendung der Foundation-Modelle identifizieren und ihre Rollen spezifizieren, als Beitrag zur Entwicklung einer holistischen Strategie für die Deutsche Forschung, Wirtschaft und Industrie.

2. Zusammenarbeit mit LEAM

- Ein Ansatz wie LEAM hilft, um ähnliche Modelle parallel an verschiedenen Orten und in verschiedenen Anwendungskontexten berechnen und nutzen zu können.
- Das Prä-Investment via LEAM ist notwendig, damit KI-Foundation-Modelle auch in der Praxis und mit wirtschaftlichem Mehrwert einsetzbar sind.
- Man braucht dazu ein Kollaborationsmodell zwischen der Wirtschaft und der Wissenschaft.
- LEAM sollte auf jeden Fall auch die Möglichkeiten für Inferenzen über KI-Foundation-Modelle zur Verfügung stellen. Dadurch bekommen kleinere Akteure die Ressourcen, um mit Foundation-Modellen gewinnbringend zu arbeiten.

Interviewprotokoll Siemens

Interviewter:

Michael May, Head Company Core Technology Data Analytics & Artificial Intelligence, Siemens AG

Interviewer:

Jörg Bienert, Alexander Thamm GmbH

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

22. Dezember 2022

Zusammenfassung

1. KI-Foundation-Modelle

- Foundation-Modelle sind ein wichtiges Thema. Mit GPT-3 wurde bereits experimentiert, bspw. für den Servicebereich.
- Für die Siemens-Geschäftsbereiche ist v.a. das Thema Industrial AI wichtig. Hier sind uns im Bereich Foundation-Modelle noch keine echten Anwendungen bekannt. Ein wichtiger Punkt ist, dass die Datensätze tendenziell in diesem Bereich kleiner bzw. fragmentiert sind.
- Siemens arbeitet gemeinsam mit der LMU und TUM an Grundlagenforschungsthemen, z.B. im Rahmen von Doktorarbeiten, und evaluiert parallel konkrete Anwendungsfälle für Foundation Models.
- Siemens hat derzeit keine eigene hausinterne Infrastruktur zum Trainieren von Foundation Models.

2. Zusammenarbeit mit LEAM

- Ein deutsches bzw. europäisches Angebot ist wünschenswert und hätte viele Vorteile.
- Es besteht ein hohes Interesse, die LEAM-Initiative zu unterstützen
- Projekte zur Erstellung von Foundation-Modellen können dabei unterstützen, das Thema Datenteilung (innerhalb und zwischen Unternehmen) neu zu beleben und ihm eine neue Relevanz zu verleihen.
- Das wird v.a. kleineren Unternehmen helfen, die aufgrund fehlender finanzieller und personeller Ressourcen weniger Möglichkeiten haben als Großunternehmen.

Interviewprotokoll TÜV Süd

Interviewer:

Dirk Schlesinger, Chief Digital Officer, TÜV Süd

Interviewer:

Jörg Bienert, Alexander Thamm

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

06.12.2022

Zusammenfassung

1. Thema LEAM

- Wir brauchen eine gemeinsame Aktion!

- Das mögliche de-facto Monopol der Marktbegleiter aus den USA verträgt sich nicht mit unserem europäischen Wertekanon.
- Wir müssen Basisfähigkeiten aufbauen, um nicht abhängig und abgehängt zu werden.
- Ein Gegengewicht aus geostrategischer Sicht ist notwendig – vgl. supply chain für Gas, seltene Erden, usw.

2. Rechenzentrum

- Ja, wir brauchen ein gemeinsames Rechenzentrum.
- Mit den Gauss Centre for Supercomputing haben wir in Deutschland die Erfahrung, wie das Thema Rechenzentrum geht – wir wissen, wie Großforschung geht (DESY, Jülich...).
- Capex ist besser als Opex, weil das Budget nicht jedes Jahr neu verhandelt werden muss.

3. Datenmanagement

- Es gibt offene Fragen zu Themen wie GDPR, homomorpher Verschlüsselung, Federated Learning und weiteren. Die Gefahr ist, dass viele Unternehmen ohne Garantien für ihre IP nicht bereit sind, ihre Daten zu teilen.
- GDPR-Konformität kann dabei ein Wettbewerbsvorteil ggü. USA und China sein.
- Federated Learning as a Service ist eine großartige Idee. Vor allem, wenn Partnern Rechtsunsicherheiten genommen werden können.

4. Was macht der TÜV?

- Der TÜV SÜD nutzt KI selbst in verschiedenen, Projekten (bspw. Visual Analytics).
- Wir (TÜV SÜD) nutzen bspw. NLP für Konsistenzprüfungen in Dokumenten und bauen darauf aktuell ein Document Service Layer auf.
- Der TÜV verwendet große Modelle bisher nicht! Es ist aktuell auch nicht geplant, diese zu nutzen. Die größte Herausforderung ist dabei sicherlich die Einführung der Modelle innerhalb der Organisation.
- TÜV nutzt Explainability Verfahren.
- Kuratierte, große Datensätze werden aber ein wichtiges Thema, v.a. in Hinblick auf Marktzulassung von Produkten (benchmarks).
 - Case: Simulation von einem autonom fahrenden Auto braucht Daten über die Straßenbeschaffenheit, das Wetter, die Lichtverhältnisse, Fahrphysik, etc.
- Sobald der AI Act der Europäischen Union umgesetzt ist, wird sicherlich auch der TÜV auf dem Gebiet der Qualitätszertifizierung von KI werden.

5. Governance

- AI Lab GmbH:

- TÜV Gesellschaften arbeiten z.Zt, als Arbeitsgruppe des Verbandes vorwettbewerblich und gemeinschaftlich zusammen, um das Thema KI – Qualitätskontrolle voranzutreiben, v.a. auf technischer Ebene.
- Die Gründung einer AI Lab GmbH ist angestrebt, vorbehaltlich einer Genehmigung durch das Bundeskartellamt. Pilotprojekte werden aber bereits abgehandelt.
- Ein Vertrieb der entwickelten Anwendungen ist hier aber nicht ohne weiteres möglich, bzw. nicht angestrebt. Rolle der AI Lab GmbH ist die eines internen R&D-Dienstleisters
- Die TÜVe haben über viele Projekte dazugelernt. Einen allgemeinverbindlichen Blueprint gibt es nicht.
- AI Quality und Testing Hub:
 - Eine Landesgesellschaft des Landes Hessen und des VDE als erste Shareholder.
 - Herausforderung: Zusammenarbeit mit dem Land, das anders plant und arbeitet als eine Firma, der VDE oder der TÜV.
 - In NRW ist das AI Quality und Testing Hub ein klassisches Förderprojekt.
 - AIQs sind komplementär zu den AI Labs. Die AI Labs sind interne R&D Stellen, die Quality und Testing Hub bieten Kunden Services an.
 - Aktuell gibt es die Hubs nur ‚auf dem Papier‘ sie sind noch nicht final gegründet, obwohl dies die nächsten Wochen / Monate geschehen dürfte.
- Testing & Experimentation Facility Healthcare:
 - Charite als Konsortialführer. Läuft sehr gut.
 - Europäisches Projekt – 30 plus Partner
 - Ziel: Entwicklung und agile Zertifizierung von ‚echten‘ Medizinprodukten mit KI
- FSD GmbH als Analogie für AI Lab GmbH:
 - entwickelt Prüfmittel für Hauptuntersuchung der Autos
 - Bund beleihet FSD und ist daher auch beteiligt. Nicht immer sind die Interessen der öffentlichen Hand dieselben wie von Wirtschaftsunternehmen – Ausgleich und Absprache vorab wichtig.

Interviewprotokoll VDE

Interviewer:

Sebastian Hallensleben, Head of Digitalisation and AI, VDE Verband der Elektrotechnik Elektronik Informationstechnik e. V.

Interviewer:

Jörg Bienert, Alexander Thamm GmbH

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

22. Dezember 2022

Zusammenfassung

1. KI-Foundation-Modelle

- KI-Foundation-Modelle sind eine Schlüsseltechnologie. Die europäische Wirtschaft und Gesellschaft braucht eigene Modelle.
- Die Auswirkungen auf Geschäftsmodelle und Jobs wird sehr groß, z.B. Redenschreiber, Analysten, Standard-Journalismus, Kundenservice, Briefings, etc.
- Das Training eigener Foundation-Modelle ist eine Frage der europäischen digitalen Souveränität. Europa braucht die Infrastruktur und die Kompetenz, um beim Thema nicht abgehängt zu werden.
- Der Bedarf an den Modellen steigt.
- Das Thema der Qualitätskontrolle dieser Modelle wird wichtiger. Eine Qualitätskontrolle ist aber nur möglich, wenn wir die Technologie selbst beherrschen.
- Derjenige, der die Technologie beherrscht, wird auch die Standards setzen und die Regulierung steuern. Wenn wir nicht in der Lage sind, hier mitzuwirken, werden wir die digitale Souveränität auch in diesem Bereich verlieren.
- Die Verfügbarkeit von Rechenkapazitäten sollte Teil der staatlichen Infrastruktur / Daseinsvorsorge sein.

2. AI Quality und Testing Hub

- Gemeinsam mit dem Land Hessen hat die VDE ein AI Quality und Testing Hub gegründet.
- Inhaltlich beschäftigt es sich mit Qualitätsmanagement und KI. Dafür sollen Prüfwerkzeuge zu einem Toolkit zusammengeführt und die Trainingsdaten-Qualität überprüft werden.
- Der Hub soll in den nächsten Jahren organisch wachsen.
- Für eine Zusammenarbeit mit dem Land Hessen mussten einige Punkte im Beihilferecht beachtet werden. Bspw. tritt Hessen als kommerzieller Investor auf, es gab Vorgaben für den Aufsichtsrat und das Finanzministerium muss den Business Plan absegnen.

Interviewprotokoll Volkswagen

Interviewter:

Patrick van der Smagt, Director of AI Research, Volkswagen Group Machine Learning Research Lab

Interviewer:

Jörg Bienert, Alexander Thamm GmbH

Protokollant:

Alex Dickmann, KI Bundesverband e.V.

Datum:

21. Dezember 2022

Zusammenfassung

1. KI-Foundation-Modelle

- Die Anwendung von KI-Foundation-Modellen bei VW befindet sich aktuell in der Evaluation. Dabei sind viele Anwendungen, auch im Bereich End-User (Kommunikation im Auto, Chatbots in der Kundenkommunikation, Digitalisierung von Handbüchern, etc.) denkbar.
- Dabei denkt VW sowohl über die Nutzung und Tuning bestehender Modelle als auch das Training eigener Foundation Modelle nach.
- VW besitzt kein eigenes Rechenzentrum, um Foundation-Modelle zu trainieren. Sie müssten hier auf externe Anbieter zurückgreifen.
- Für die Entwicklung und Nutzung von Foundation Modellen sind für VW sind auch die Themen Explainability, Trustworthy AI und die juristischen Rahmenbedingungen wichtig.
- Die Umsetzung und Nutzung von Modellen auf Basis einer durch einen Cloud-Service bereitgestellten API ist aus Gründen der Datensicherheit problematisch. VW arbeitet nur mit Modellen, die für sie kontrollierbar sind und idealerweise auch vom Unternehmen gehostet werden.
- Daneben ist das Thema Datenverfügbarkeit wichtig. VW besitzt eigene, spezielle Datensätze auf denen Modelle trainiert werden müssten.

2. Zusammenarbeit mit LEAM

- VW begrüßt die Initiative LEAM und möchte gerne weiter unterstützen.
- LEAM Services würden im Bereich Modell Tuning, aber auch potentiell in der Erstellung von eigenen Foundation Modelle genutzt.
- Die Beteiligung an einem Joint Venture bzw. einer PPP ist grundsätzlich sinnvoll. VW steht der Idee offen gegenüber.

Interviewprotokoll Zalando

Interviewer:

Alexander Borek, Director of Data Analytics, Zalando SE

Interviewer:

Jörg Bienert, Alexander Thamm GmbH

Protokollant:

Alex Dickmann, KI Bundesverband

Datum:

15. Dezember 2022

Zusammenfassung

1. KI-Foundation-Modelle

- Nachteil der amerikanischen Modelle ist, dass sie nicht öffentlich zugänglich sind und individuell angepasst werden können.
- Darüber hinaus spielt der Datenschutz eine wichtige Rolle. Vor allem Kundendaten sind sehr sensibel und für Zalando auch wettbewerbskritisch.
- Wenn bestimmte Foundation-Modelle aufgrund von Datenschutzbedenken nicht genutzt werden können und es keine Alternativen gibt, ist das ein großer Wettbewerbsnachteil für Zalando.
- Für Zalando sind vor allem auch die europäischen Sprachen relevant.
- Entscheidend ist, dass die Anwendungen auf Basis der Modelle beim Kunden funktionieren.

2. Zusammenarbeit mit LEAM

- Die geplanten LEAM Services sind interessant für Zalando. Eine Nutzung ist wahrscheinlich.
- Es ist besser, Einfluss auf die Technologie zu haben bzw. die Technologie selber betreiben zu können, als über API auf Fremd-Services aus Übersee zugreifen zu müssen.
- Mögliche Anwendungsfälle sind vor allem Chatbots, Verbesserungen in der Suche, aber (in geringerem Maße) auch die Image-Verarbeitung

Anhang C - Übersicht Cloud-GPU Anbieter

Die aufgeführten Kosten haben keinen Anspruch, alle Anbieter abzudecken. Sie dienen lediglich als Orientierung für marktübliche Preise.

Tabelle 27 gibt einen groben Eindruck über die Preise anderer Cloud Computing Anbieter. Die Preise pro GPU Stunde liegen ungefähr zwischen 1,94 EUR und 3,73 EUR. Somit sind die geplanten 2,21 EUR pro GPU Stunde für das LEAM-Hochleistungsrechenzentrum ein wettbewerbsfähiger Preis. In der Tabelle wurden die Preise für eine Stunde Rechenleistung der NVIDIA A100 Tensor Core GPU 80 GB verglichen, eines der stärksten, marktrelevanten GPUs. Die Preise variieren stark, je nach Anzahl der GPUs, weitere Hardware-Kapazitäten (Anzahl CPUs, RAM etc.) und Dauer der Buchung. Zum Beispiel bieten Amazon Web Services acht A100 GPUs für 1,38EUR pro GPU Stunde an, wenn diese für drei Jahre reserviert werden. Da das Hochleistungsrechenzentrum hauptsächlich für das Training von einzelnen Modellen gebucht wird, sind Buchungszeiträume von mehreren Monaten am ehesten vergleichbar und werden somit als Referenzwert genutzt, falls vorhanden (entsprechende Anbieter sind markiert). Allerdings muss bedacht werden, dass sich die angegebenen Preise auf Kosten für einen GPU belaufen. Es ist möglich, dass manche Anbieter größeren Projekten Rabatte anbieten können. Außerdem bieten manche Anbieter Spot-Preise. Dies sind stark reduzierte Preise für Rechenleistungen, welche zu jeder Zeit abgebrochen werden können, wenn die Nachfrage zu stark steigt. Der Spot-Markt ist ungeeignet für das hochkomplexe Training von Foundation-Modellen und somit sind die Preise nicht vergleichbar.

Anbieter	Preis
Google Cloud	3,73 EUR
Microsoft Azure	3,63 EUR
Paperspace	2,93 EUR
Amazon AWS EC2	2,90 EUR
Northern Data	2,59 EUR
Vultr	2,31 EUR*
Coreweave	2,10 EUR
Datacrunch	2,09 EUR
RunPod	1,98 EUR
FluidStack	1,94 EUR*

Tabelle 27: Kosten für eine GPU-Stunde auf einer NVIDIA A100 Tensor Core GPU 80 GB nach Anbieter. Preise in US-Dollar wurden in Euro umgerechnet zu einem Kurs von \$1 = 0,948768EUR (Dollarkurs am 06.12.2022)

* Monatsraten ansonsten On-Demand-Preise

MITWIRKENDE



MACHBARKEITSSTUDIE

zum Aufbau und Betrieb eines dedizierten
KI-Hochleistungsrechenzentrums für das
Trainieren großer KI-Modelle in Deutschland



2023