Large European AI Models

LEAM:AI

Große KI-Modelle als Basis für Erfolg in Forschung und Wirtschaft.
Open Source. Wertebasiert. Universell.

Konzeptpapier:

Large European AI Models (LEAM) als Leuchtturmprojekt für Europa

In den letzten Jahren haben Unternehmen und Forschungseinrichtungen in den USA und China bahnbrechende Ergebnisse mit sogenannten großen KI-Modellen erzielt. Trainiert mit riesigen Datenmengen zeigt diese neue Technologie-Generation zum ersten Mal ein tiefes Sprachvermögen gepaart mit einer Fähigkeit zu kognitiven Leistungen, die in relevanten Szenarien von menschlicher Intelligenz nicht unterscheidbar sind. Die großen KI-Modelle revolutionieren den Markt für Künstliche Intelligenz und werden sich disruptiv auf die gesamte Wirtschaft auswirken.

Erfreulicherweise gibt es in Europa bereits erste Projekte, die einen Fokus auf große Sprachmodelle legen. Aber die Entwicklungen in Europa sind viel zu zaghaft und nicht annähernd angemessen, um der gesellschaftlichen und wirtschaftlichen Bedeutung der Technologie gerecht zu werden. In anderen Teilen der Welt wird die Entwicklung mit höchsten Investitionen und in rasanter Geschwindigkeit vorangetrieben, während es den europäischen KI-EntwicklerInnen an verfügbarer und speziell für die KI-Entwicklung abgestimmter Infrastruktur fehlt.

Damit nicht erneut amerikanische Unternehmen Monopole aufbauen, europäische Unternehmen von diesen abhängig werden und damit Europa die Chancen der zukunftsprägenden Schlüsseltechnologie Künstliche Intelligenz für Wissenschaft, Wirtschaft und Gesellschaft nutzen kann, sind massive Investitionen in die Entwicklung und Bereitstellung großer KI-Modelle sowie KI-Infrastrukturen erforderlich. Es wäre gefährlich und fahrlässig, die technologische Entwicklung den großen Tech-Konzernen in den USA und China zu überlassen. Wir dürfen jetzt nicht zögern und die Chance vergeben, die Zukunft mitzugestalten!

Wir - einige der führenden deutschen Unternehmen, der renommiertesten ForscherInnen und Institutionen in der KI-Forschung - haben daher LEAM ins Leben gerufen. Wir planen eine verantwortungsvolle Entwicklung und den nachhaltigen Betrieb von großen KI-Modellen in und aus Europa möglich zu machen. Um dieses Ziel zu erreichen, werden Investitionen im mittleren dreistelligen Millionenbereich nötig werden.

Dieses Konzeptpapier ist das Ergebnis intensiver Beratungen der LEAM UnterstützerInnen. Es gibt einen ersten Einblick in ein mögliches KI-Hochleistungszentrum, eine Organisation, die dieses betreibt, sowie erste Modelle, die unter LEAM entwickelt werden können. Dabei sollte dieses Papier als ein erster Aufschlag verstanden werden. An vielen Stellen sind noch Fragen offen und die LEAM UnterstützerInnen arbeiten weiterhin intensiv an der Realisierung des Leuchtturmprojekts.

Inhaltsverzeichnis

Große KI-Modelle revolutionieren den Markt für Künstliche Intelligenz

Eine wesentliche Schlüsseltechnologie der Künstlichen Intelligenz sind große KI-Modelle, bei deren Entwicklung und Anwendung China und die USA in den letzten Jahren bahnbrechende Fortschritte erzielt haben. Die Entwicklung einer Künstlichen Intelligenz beruhte in der Vergangenheit vor allem auf dem Trainieren individueller und für eine einzige Aufgabe spezialisierter KI-Modelle. Vor einigen Jahren gelang dann ForscherInnen amerikanischer Industrielabore der entscheidende Durchbruch. Neue, generelle KI-Modelle, die mit riesigen Text- und Datenmengen trainiert wurden, entwickelten ein grundlegendes Verständnis von Sprache und Welt. Ohne extra darauf trainiert zu werden, lösten diese Modelle auf erstaunliche und effiziente Weise diverse Probleme der KI-Forschung.

Heute ermöglichen große KI-Sprachmodelle nicht nur das automatisierte Schreiben von Texten in nahezu menschenähnlicher Schreibqualität, sondern können auch selbstständig programmieren. KI-Sprachmodelle sind also nicht nur wichtige Grundlage für die Mensch-Maschine-Kommunikation, sondern in Zukunft auch für die Automatisierung weitreichender Wirtschaftsprozesse sowie als praktisch unendlich skalierbare Software-Entwicklungsressourcen einsetzbar.

Multimodale Modelle, die Text- und Bilddaten miteinander kombinieren, erstellen bereits Fotos aus Texten oder erkennen den Inhalt eines Videos. Fähigkeiten, die traditionelle KI-Modelle nur schwer entwickeln konnten.

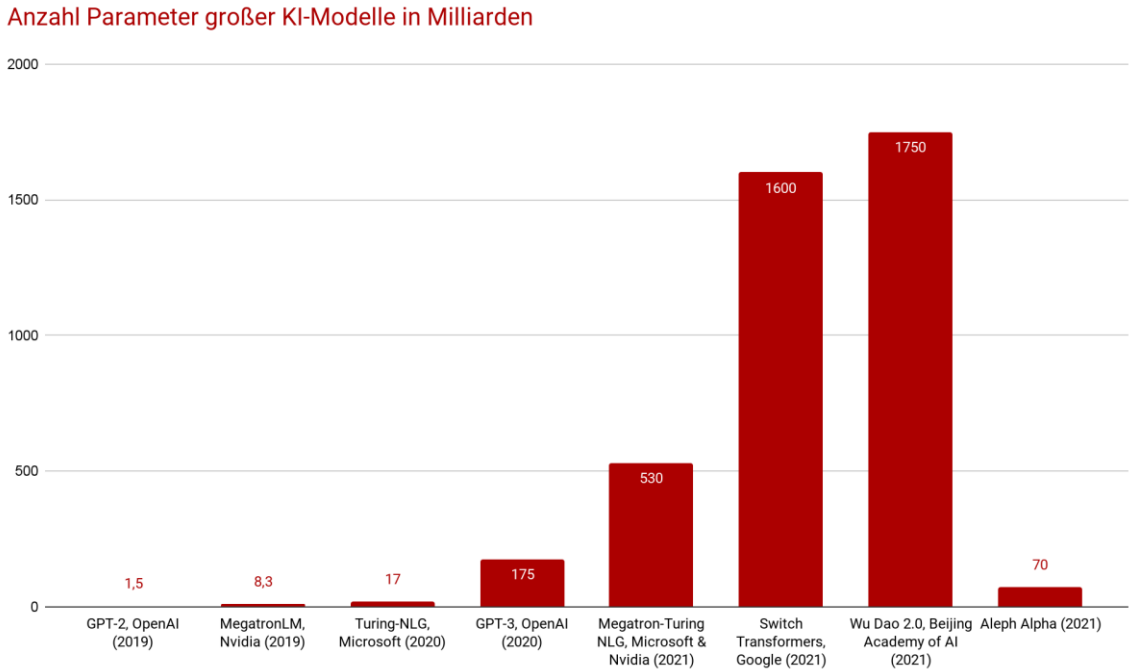
Der Vorteil der großen KI-Modelle ist, dass sie nicht hochspezialisiert sind und damit für jede Anwendung neu entwickelt werden müssen, sondern mit entsprechenden Anpassungen vielseitig einsetzbar sind.

In Europa fehlt es an der passenden Infrastruktur

Aktuell ist Europa bei der Entwicklung großer KI-Modelle ins Hintertreffen geraten, während China und Nordamerika mit massiven staatlichen und privatwirtschaftlichen Investitionen ihre vorherrschende Stellung im Bereich der Künstlichen Intelligenz weiter ausbauen. Alle paar Monate präsentieren KI-Labore der IT-Konzerne noch größere und mächtigere KI-Modelle. Die Folge ist ein stetig steigender Einfluss auf die sich in der Transformation befindenden, klassischen Wirtschaftszweige sowie die durch die

Digitalisierung neu entstehenden Wirtschaftszweige der Zukunft. Mit jeder dieser Neuentwicklungen wächst unser Rückstand und hat nun bereits gefährliche Ausmaße angenommen.

Bemühungen der öffentlichen Hand in dem Bereich, wie das vom BMWK geförderte Projekt OpenGPT-X oder die Förderbekanntmachung zu KI-Rechenzentren, zielen aktiv darauf ab, diesen Rückstand zu schließen. Es fehlt ihnen aktuell aber an ausreichender und für die Entwicklung großer KI-Modelle dedizierter Infrastruktur, um eine erfolgreiche Aufholjagd zu starten.



Seit der Veröffentlichung von GPT-2 im Februar 2019 hat sich die Anzahl der Parameter mit denen große KI-Modelle trainiert werden mehr als vertausendfacht. Das größte bis heute bekannte europäische Modell mit ca. 70 Milliarden Parametern ist klein im Vergleich mit den neuesten Modellen amerikanischer und chinesischer EntwicklerInnen mit mehr als 1,6 Billionen Parametern - Tendenz steigend.

Wir befürchten daher eine Abhängigkeit von nicht-europäischen KI-Lösungen. Mittelfristig werden europäische Unternehmen vor die Wahl gestellt, nicht-europäische KI-Anwendungen zu nutzen und damit wertvolle Unternehmensdaten preiszugeben oder sich dagegen zu entscheiden und nicht in vollem Umfang an den Vorteilen der Künstlichen Intelligenz zu partizipieren. Schließlich werden KI-Monopole in den Händen großer US-amerikanischer und chinesischer Technologiekonzerne entstehen.

Sicherheitspolitisch muss Europa im Wettlauf um neueste KI-Technologien mithalten, damit Europa nicht in wenigen Jahren schutzlos einer unvorstellbaren Flut von Desinformation durch intelligent kommunizierende, simulierte Akteure gegenübersteht. Europa muss daher ein Gegengewicht zu den KI-Modellen aus den USA und China stellen, um seine digitale Souveränität zu sichern und um eine Spitzenposition in dieser zentralen Schlüsseltechnologie einzunehmen.

LEAM - Die europäische Initiative zur Entwicklung großer KI-Modelle

LEAM (Large European AI Models) ist eine Initiative aus Wirtschaft und Forschung, die Deutschland und Europa befähigen will, den Anschluss an die aktuellen bahnbrechenden Innovationen in Forschung und Anwendung von KI-Methoden wiederzuerlangen und entsprechend europäischen Werten und Bedarfen weiterzuentwickeln. Daher strebt die LEAM-Initiative mit allerhöchster Dringlichkeit an, eine Infrastruktur aufzubauen, auf der große KI-Modelle entwickelt werden können sowie eine Organisation zu schaffen, die diese Entwicklung durchführt und die Ergebnisse bereitstellt.

Unser Ziel ist der Erhalt der europäischen Souveränität beim Thema Künstliche Intelligenz. Es ist zwingend notwendig, dass in Europa KI-Entwicklung auf höchstem Niveau stattfindet. Nur so ist sichergestellt, dass die europäische Wirtschaft an der neuen Welle von Innovationen partizipiert und dass die Bedürfnisse der europäischen Gesellschaften berücksichtigt werden.

LEAM fußt dabei auf der Überzeugung, dass KI-Entwicklung auf Basis europäischer Normen und Werte sowie aller europäischer Sprachen stattfinden muss. Die Entwicklung großer europäischer KI-Modelle soll deshalb den einen klaren Fokus haben:

Um die Ziele zu erreichen, sind der Aufbau einer kompetitiven Infrastruktur, die Zusammenstellung großer Korpora von Trainingsdaten nach europäischen Bedürfnissen

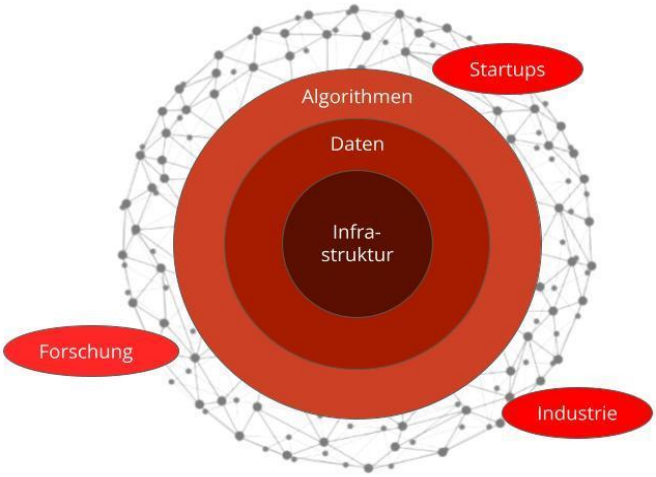
und Werten sowie das Training und die Bereitstellung großer KI-Modelle geplant. Basierend auf der Expertise und den Erfahrungen der UnterstützerInnen müssen für eine erfolgreiche Umsetzung dieses Vorhabens rund 400 Mio. Euro investiert werden. Die Initiative verspricht sich davon breite positive volkswirtschaftliche Effekte, u.a. die Sicherung und Stärkung der europäischen KI-Industrie, die Aufrechterhaltung der Wettbewerbsfähigkeit der europäischen Großunternehmen und eine Welle neuer Unternehmensgründungen im KI-Bereich.

Die Initiative gibt den unterstützenden Firmen unkomplizierten Zugang zu wissenschaftlichen und technischen Fachkenntnissen und fördert den Know-How-Transfer. Immer mit dem Ziel, Start-ups, KMUs und etablierten Firmen die Entwicklung neuer und innovativer KI-Produkte zu ermöglichen. Hierzu unterstützt LEAM die Wirtschaft aktiv bei der Prototypenentwicklung. Das gegenseitige aufeinander Einwirken und der Austausch über erfolgversprechende Anwendungsfälle zwischen LEAM und Wirtschaft sind zentral für LEAMs langfristigen Erfolg.

Durch öffentliche Veranstaltungen und Workshops wird das in LEAM vorhandene und erarbeitete Wissen über KI aktiv an die verschiedenen digitalen Communities weitergegeben. Neben technischen, sollen auch gesellschaftliche, ethische und rechtliche Aspekte vermittelt werden. LEAM kann so erheblich zur Vernetzung und Stärkung der europäischen Digitalwirtschaft beitragen.

Fünf Meilensteine, die erreicht werden müssen

Um die Ziele von LEAM zu erreichen, sind eine Reihe von Maßnahmen erforderlich.



1.) Etablierung eines KI-Hochleistungs-Rechenzentrums

Die LEAM-Initiative empfiehlt den Aufbau und Betrieb eines dedizierten KI-Rechenzentrums, um Wirtschaft und Wissenschaft Konkurrenzfähigkeit im Bereich der KI-Entwicklung zu ermöglichen und um Europa selbst zum Motor für Innovationen in diesem Bereich zu machen. Um das Ziel dieser beträchtlichen Investition nicht zu gefährden, muss die technische Infrastruktur von Anfang an in Hardwareauswahl und -konfiguration sowie in Rechen- und Speicherkapazität konkurrenzfähig mit den Laboren sein, in denen die Durchbrüche der letzten Jahre erzielt wurden.

2.) Aufbau von Kompetenzen und Personalkapazitäten

Parallel muss personelle Kompetenz in hinreichender Kapazität aufgebaut werden, um die ersten großen europäischen KI-Modelle zu schaffen. Dabei müssen auch Forschung und Technologieentwicklung konzentriert so gefördert werden, dass eine faire Chance besteht, im internationalen Wettbewerb zu bestehen und in wichtigen Anwendungsgebieten die Führung zu übernehmen. Aufgaben des Teams sind die Anpassung und Entwicklung von Algorithmen, die Kuratierung und Aufbereitung von Daten sowie das komplexe Training der Modelle einschließlich Evaluation und Optimierung.

3.) Sammlung von Daten und Entwicklung von Algorithmen

Bei der Auswahl der Daten und Algorithmen müssen die besonderen Bedürfnisse der europäischen Gesellschaft berücksichtigt werden, dazu gehören neben der Abdeckung von priorisierten Themenbereichen der europäischen Wirtschaft auch europäische Mehrsprachigkeit, Einhaltung ethischer Normen und Verlässlichkeit in kritischen Wissensbereichen.

4.) Etablierung einer eigenständigen Organisationseinheit

Als eine zentrale Maßnahme sieht die LEAM-Initiative auch die Schaffung einer Organisation, welche die entstandenen Modelle für die europäische Industrie und Forschung zur Verfügung stellt. Dazu gehört die Bereitstellung für Nutzer aus Wirtschaft, Wissenschaft und Verwaltung auf hinreichend mächtigen Cloudstrukturen, die das Training der Modelle für spezielle Anwendungen (Fine-Tuning), den Einsatz der Modelle in fertigen Anwendungen (Inferencing) und das Experimentieren für die Verbesserung der Modelle und die Erforschung neuer Anwendungen ermöglichen, hohe

Leistungsanforderungen ergeben sich für diese Funktionen aus der Größe der Modelle, der Zahl der Nutzer, den Effizienzanforderungen der kommerziellen Anwender sowie der Sicherheit und Vertraulichkeit der Daten. Die Organisation muss rechtlich und betriebswirtschaftlich in der Lage sein, die Kosten für die Verfügbarmachung und die damit verbundenen Dienstleistungen auf die Nutzer umzulegen.

5.) Berücksichtigung von europäischen Werten und Nachhaltigkeit

Vor allem im Bereich der großen Sprachmodelle gilt es, eine möglichst breite Unterstützung für verschiedene Sprachen aufzubauen und europäische Werte bei der Qualität der KI-Modelle (z.B. durch Optimierung hinsichtlich evtl. auftretenden Bias) zu berücksichtigen

Eine zentrale Anforderung an alle Maßnahmen ist die Maxime der Nachhaltigkeit. Die Planung und der Einsatz der Infrastruktur, die Auswahl der Algorithmen, die Gestaltung der Trainingsprozesse und auch die Organisation und technische Ausführung der Verfügbarmachung müssen so angelegt sein, dass eine maximal mögliche Energieeffizienz erreicht wird. Hierbei werden offene Infrastruktur-Architekturen z.B. Gaia-X ebenfalls eine zentrale Rolle spielen.

Nutzen von LEAM für die europäische Digitalwirtschaft

LEAM soll ein Zugpferd für KI-Innovationen in Europa werden: Ein Fokussierungspunkt, um den herum sich ein leistungsfähiges, vielfältiges Ökosystem aus Wirtschaft, Forschung, neuen Geschäftsmodellen und Startups bildet.

Als zentrales Leuchtturmprojekt bildet LEAM in enger Abstimmung und Zusammenarbeit mit bestehenden Forschungseinrichtungen, Unternehmen, Verbänden und Initiativen den Kristallisationspunkt für die zielgerichtete Sammlung von Daten (OpenData), breit gefächerte Forschungsvorhaben, die Entwicklung von wertschöpfenden Anwendungen für alle Branchen und wirtschaftliches Wachstum.

Der konkrete Nutzen für die Forschungslandschaft ist vielfältig. Die Initiative arbeitet an aktuellen wissenschaftlichen Fragestellungen wie z.B. der Minimierung von BIAS, der Effizienzsteigerung von KI-Modellen, Urheberrechts-Fragestellungen, der Nutzung von neuen Rechner-Architekturen oder auch GDPR-Fragestellungen. Gleichzeitig werden Arbeiten rund um den Betrieb großer KI-Infrastrukturen eine Reihe neuer Fragestellungen und Optimierungspotentialen aufwerfen.

Privatwirtschaftliche Unternehmen werden in der Lage sein, auf der Basis von LEAM große Standard-KI Modelle zu nutzen, ohne dass seine Daten das Unternehmen oder den EU-Bereich verlassen müssen. Gleichzeitig können individuell angepasste große KI-Modelle erstellt werden auf deren Basis neue Datenprodukte und sogar Geschäftsmodelle entwickelt werden.

Wertschöpfende Anwendungen und wissenschaftliche Fragestellungen

Die Liste der wertschöpfenden Anwendungen von großen KI-Modelle ist lang und wächst kontinuierlich. Da es für jede dieser Anwendungsklassen eine Vielzahl von Einsatzfeldern in Wirtschaft und Gesellschaft gibt, ist das wirtschaftliche Potential einer Basistechnologie mit hinreichender Kompetenz unermesslich.

In Deutschland gibt es bereits eine ständig wachsende Zahl von KI Unternehmen, die Anwendungen aus diesen Klassen in ihrem Produktportfolio haben. Durch die Verfügbarkeit der geplanten Modelle wird diese Zahl sicher noch signifikant zunehmen. Hier seien nur einige Beispiele aufgezeigt:

Anwendungs-Beispiele großer KI-Modellen

Document-Processing	Übersetzung	Chatbots	Textgenerierung	Programmcode-Generierung
Sprach Ein-/ Ausgabesysteme	Erkennung von Desinformationen in Videos	Proteinfaltungen	Drug-Design	Generierung von Bildern

Gleichzeitig gibt es eine Vielzahl von Fragestellung im Bereich der Forschung, die von einer Reihe von Universitäten und Forschungseinrichtung mit Unterstützung der LEAM Infrastruktur angegangen werden können. Hierzu gehören bspw.:

Beispiele wissenschaftlicher Fragestellungen

Minimierung von Bias	Urheberrechts-Fragestellungen	Einsatz großer Modelle bei kleinen Sprach-communities	Effizienzsteigerung von Modellen
Nutzung neuer Rechner-architekturen	GDPR-Fragestellungen	Nachvollziehbarkeit von KI Anwendungen	Senkung des Energieverbrauchs

Liste aller Unterstützer

LEAM wird bereits von folgenden Institutionen, Verbänden und Unternehmen unterstützt:

Organisation	Ansprechpartner
Alexander Thamm GmbH	Alexander Thamm
Bayer AG	Dr. Marion Legler
Beuth Hochschule für Technik Berlin	Prof. Dr. Alexander Löser
Bosch Center for Artificial Intelligence	Dr. Michael Fausten
Bundesdruckerei GmbH	Dr. Manfred Paeschke
Cloud&Heat Technologies GmbH	Dr. Ronny Reinhardt
Continental AG	Dr. Corina Apachiţe
DFKI - Deutsches Forschungszentrum für Künstliche Intelligenz GmbH	Dr. Antonio Krüger
E.ON Energie Deutschland GmbH	Dr. Christian Essling
eco - Verband der Internetwirtschaft e.V.	Hauke Timmermann
Fraunhofer IAIS	Dr. Joachim Köhler
FZI Forschungszentrum Informatik	Jan Wiesenberger
GI - Gesellschaft für Informatik e.V.	Daniel Krupka
GIANCE Technologies GmbH	Prof. Dr. Hans Uszkoreit
Humboldt-Universität zu Berlin	Prof. Dr. Alan Akbik
inovex GmbH	Hans-Peter Zorn

Organisation	Ansprechpartner
IW - Institut der deutschen Wirtschaft Köln e.V.	Dr. Hans-Peter Klös
KI Bundesverband e.V.	Jörg Bienert
KI.I.E.Z. - Künstliche Intelligenz Entrepreneurship Zentrum	Dr. Tina Klüwer
Labs Networks Industrie 4.0 e.V.	Thomas Hahn
Lufthansa Industry Solutions	Susan Wegner
Mediengruppe RTL Deutschland GmbH	Valeria Klassen
Merantix AG	Dr. Johannes Otterbach
Merck KGaA	Dr. Helmut Linde
msg systems ag	Werner Achtert
OmniBot GmbH	Jascha Stein
REWE Digital GmbH	Kai-Uwe Reimers
SAP	Dr. Felix Sasaki
T-Systems GmbH	Roman Kwasny
Technische Universität Darmstadt	Prof. Dr. Kristian Kersting
TUI	André Exner
Vattenfall Eurofiber GmbH	Dr. Birger Ober
VDE - Verband der Elektrotechnik Elektronik Informationstechnik e. V.	Emmanuel Kahembwe
WDR - Westdeutscher Rundfunk	Dr. Dirk Maroni
Wilhelm Büchner Hochschule	Prof. Dr. Helge Nuhn
2txt GmbH	Johannes Bubenzer

KI-Hochleistungszentrum

Die erste Maßnahme von LEAM ist der Aufbau eines Hochleistungsrechenzentrums speziell für die Entwicklung von Künstlicher Intelligenz. Nur mit einer eigenen Infrastruktur können die technologische Souveränität und rechtlichen Rahmenbedingungen in Europa zweifelsfrei eingehalten werden.

Für die Beschaffung und den Aufbau der IT-Infrastruktur verfolgt LEAM einen technologie-agnostischen Standpunkt. Wichtig ist, dass die Infrastruktur zum Zeitpunkt der Beschaffung mindestens dem Stand der Technik entspricht oder diesem optimalerweise sogar voraus ist. Durch die Weiterentwicklung und die damit einhergehende Steigerung der Leistungsfähigkeit von Hardware für KI-Anwendungsfälle am Markt ist eine Erneuerung der Infrastruktur nach fünf Jahren geplant.

Weitere entscheidende Faktoren für den Aufbau eines KI-Hochleistungszentrums sind in diesem Kapitel aufgeführt.

Warum wir ein dediziertes KI-Hochleistungszentrum brauchen

Deutschland hat mit dem JUWELS-Hochleistungsrechner am Forschungszentrum Jülich bereits einen der potentesten Supercomputer der Welt. In Italien und Frankreich stehen weitere Rechenzentren, die zu den stärksten der Welt zählen. Dennoch schlägt LEAM den Neubau eines KI-Hochleistungszentrum speziell für die Entwicklung von Künstlicher Intelligenz vor. Dies hat mehrere Gründe.

Zum einen bedeutet ein aktuell gutes Ranking zwischen den stärksten Computern der Welt nicht, dass dieses Ranking auch weiterhin bestehen bleibt. In der Tat sehen wir weltweit den Aufbau neuer Hochleistungsrechenzentren, die immer mehr Petaflops besitzen. Zuletzt kündigt das amerikanische Unternehmen Meta an, ein neues Rechenzentrum zu bauen, das 50-mal schneller als JUWELS in Jülich rechnen soll. Europa muss weiter investieren, um den Anschluss nicht zu verlieren.

Zum anderen sind die europäischen Hochleistungszentren nicht speziell für die KI-Entwicklung ausgelegt. KI-EntwicklerInnen müssen sich daher gemeinsam mit Physikern, Klimaforschern oder auch Ärzten auf Rechenzeit bewerben. Das bremst nicht nur die Entwicklung neuer KI-Modelle, sondern verhindert die Entwicklung im Zweifelsfall komplett. Ein dediziertes KI-Hochleistungszentrum löst dieses Problem.

Ein dediziertes KI-Hochleistungszentrum kann außerdem speziell an die Herausforderungen der KI-Entwicklung angepasst werden. So reicht in der KI-Entwicklung bspw. eine Genauigkeit von 16 Bit, im Gegensatz zu den 64 Bit, die andere Supercomputer benötigen. Das hat nicht nur den Vorteil, dass KI-Modelle günstiger und schneller berechnet werden können als auf allgemeinen Supercomputern, sondern die Entwicklung von KI-Modellen auch nachhaltiger ist.

Wir sind uns bewusst, dass die existierenden Hochleistungsrechner eine hohe Marktmacht haben und es einer gewaltigen Anstrengung bedarf, diese zu brechen. Fest steht aber auch, dass die europäische Wirtschaft und Forschung in Zukunft auf diese Infrastrukturen angewiesen sein wird. Die europäischen Staaten müssen sich daher die Frage stellen, ob man diese Infrastruktur selbst betreibt oder nicht-europäische Infrastruktur nutzt. Für LEAM ist die Antwort auf diese Frage klar: Nur eine eigenständige Infrastruktur ermöglicht KI-Entwicklung unabhängig von den USA und China sowie die volle Einhaltung europäischer Werte und Gesetze.

Ökologische Nachhaltigkeit

Um den Klimazielen nachzukommen und den Leuchtturmcharakter des Vorhabens herauszustellen, ist eine Berücksichtigung ökologischer Nachhaltigkeit ein unabdingbares Kriterium für LEAM. Die Infrastruktur muss daher hohe Anforderungen an Energieeffizienz und ökologische Nachhaltigkeit erfüllen.

LEAM plant daher eine Versorgung des Rechenzentrums mit 100% Strom aus erneuerbaren Energiequellen, der optimalerweise direkt bezogen werden kann. Es sollen außerdem energieeffiziente Kühlungslösungen genutzt und entstehende Abwärme nachgenutzt werden. Darüber hinaus sollen entstehende Treibhausgas-Emissionen gemessen und transparent kommuniziert werden. Bei der Planung der Infrastruktur ist die Reduktion von Emissionen entscheidender Faktor.

Gebäudeinfrastruktur

Einer der wichtigsten Erfolgsfaktoren für die Initiative ist eine zeitnahe Umsetzung von LEAM. Deshalb soll die Gebäudeinfrastruktur für die IT-Hardware so gewählt werden, dass diese einer zeitnahen Umsetzung nicht im Weg steht. Die Partner des LEAM-Initiative halten deshalb einen kompletten Neubau eines Rechenzentrums nicht für erstrebenswert.

Mögliche Lösungen können stattdessen die Integration in ein Bestandsgebäude oder in

ein bereits geplantes Neubauvorhaben, die Erweiterung bestehender Groß-Rechenzentren oder die mobile und flexible Nutzung von Container-basierten Lösungen sein.

Weitere Aspekte, die bei der Planung der Gebäudeinfrastruktur berücksichtigt werden sollten, sind:

- Gewährleistung der Homogenität der IT-Infrastruktur für das Modell-Training, das heißt die Bündelung der Ressourcen für das Training des Modells an einem Standort
- Präferenz eines Rechenzentrums-Standortes in Deutschland
- Trennung der Infrastruktur in vorwettbewerblichen und wettbewerblichen Teil in Abstimmung mit den zu definierenden Anwendungsfällen und Business Cases
- Teile der IT-Infrastruktur für LEAM, wie die Ressourcen für das Data-Pre-Processing, können zur Beschleunigung des Vorhabens auch von Cloud-Anbietern als Cloud-Ressource beschafft werden, sofern die Anforderungen erfüllt werden können.

Software-Stack

Zu diesem Zeitpunkt hat sich LEAM noch nicht für einen spezifischen Software-Stack entschieden. Es ist aber allen Beteiligten klar, dass bei der Stack-Entwicklung auf Open Source Software gesetzt werden soll. Die Nutzung gewerblicher Software und Lizenzen soll soweit wie möglich vermieden werden. Diese Entscheidung ist nicht nur im Einklang mit LEAMs eigenem Ziel, Open Source Modelle zu entwickeln. Vielmehr sind Open Source Stacks aktuell die besten Lösungen am Markt.

Im nächsten Schritt sollen Best Practice-Beispiele analysiert und deren Konzepte zum Software-Stack auf Adaptierbarkeit für LEAM geprüft werden.

Datenschutz und Datensicherheit

LEAM hat sich höchsten Datenschutzstandards verschrieben. Das Gesamtkonzept (inkl. Betriebskonzept) muss daher höchstmöglichen Datenschutz- und Datensicherheitsanforderungen erfüllen.

Die Erfüllung von Datenschutz- und Datensicherheitsanforderungen muss bei der Planung der IT-Infrastruktur mit einbezogen werden. Dafür ist klar herauszustellen, an welcher Stelle personenbezogene Daten verarbeitet werden. Eine Möglichkeit, den

Anforderungen nachzukommen, ist die Unterscheidung verschiedener Sicherheitsstufen, auch bei der Planung der IT-Infrastruktur.

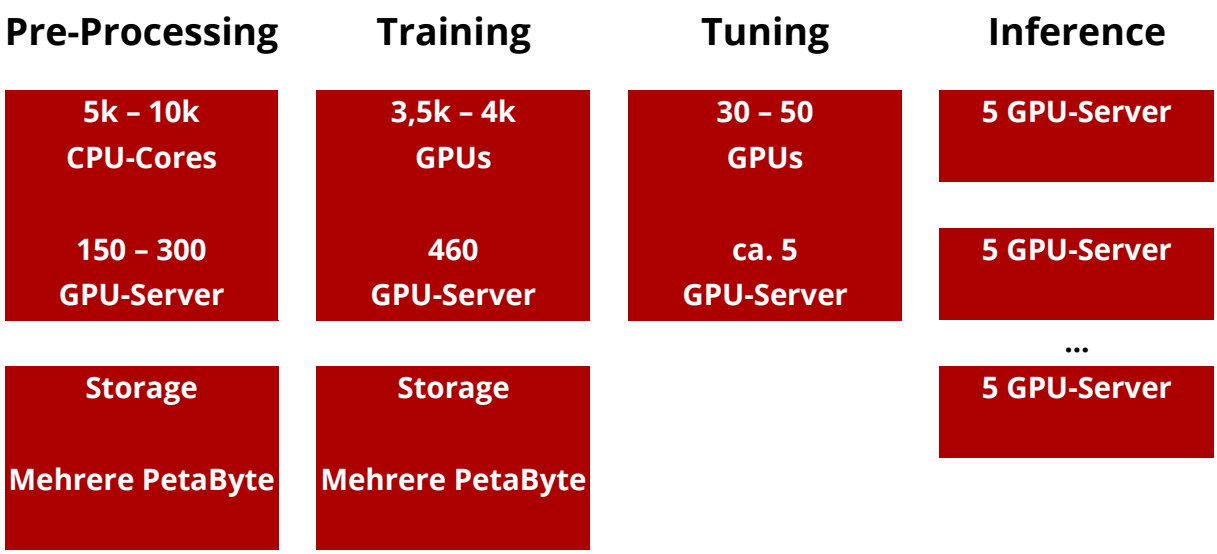
KI-Supercomputer-Hardware

Die Abschätzung der benötigten Ressourcen basiert auf Erfahrungswerten für die Entwicklung bestehender großer KI-Modelle wie GPT-3.

Für das Data-Pre-Processing eines multilingualen Modells (z. B. Data Cleaning wie HTML-Splitting) werden ca. 5.000 bis 10.000 CPU-Cores bzw. 150-300 CPU-Server benötigt. Das Training von GPT-3 nahm 355 GPU-Jahre in Anspruch. Um den Stand der Technik nicht nur aufzuholen, sondern auch darüber hinaus Fortschritte zu erreichen und Freiräume zum Experimentieren für echte Innovationen im KI-Bereich zu gewährleisten, wird im LEAM-Vorhaben mit einer Größenordnung von ca. 460 spezialisierten KI-Servern kalkuliert. Dabei sind auch Storage-Ressourcen im Umfang von ca. 10 TB pro KI-Modell einzukalkulieren.10 TB

Beim GPT-3-Modell wurden 20-50 GPUs pro Modell für die Inferenz benötigt.

Für die Gestaltung des Tunings und der Inferenz des Modells gibt es verschiedene Optionen, die in der weiteren Planung des Vorhabens noch detailliert zu definieren sind. Es kann eine Architektur gewählt werden, mit welcher den Nutzern des LEAM-Modells eine Inferenz auf ca. 30 bis 50 GPUs als Service angeboten wird. Es können aber auch Optionen angebunden werden, bei denen die Inferenz oder das Tuning auf der eigenen IT-Infrastruktur der Nutzer (on premise) erfolgt. Für ein eigenes Finetuning bei Nutzern wären mehrere 100 GPUs erforderlich. Bei Use Cases mit einem kleinen, robusten Modell, welches das Tuning bereits durchlaufen hat, könnte auch ein kleinerer Umfang an GPUs ausreichen. Das Schaubild 1 verdeutlicht eine mögliche Verteilung der Ressourcen des KI-Supercomputers für die abzudeckenden Bereiche von LEAM.



Mögliche Ressourcenverteilung innerhalb des KI-Supercomputers

Betriebsaufwand

Anforderungen an den Betrieb und die damit verbundenen Kosten sind genauso relevant wie die initialen Kosten für die Infrastruktur und müssen deshalb ebenfalls im Detail betrachtet werden.

Die Gesamtkosten werden sich aus der initialen Investition in IT-Hardware und Infrastruktur-Komponenten sowie aus laufenden Kosten für den Betrieb zusammensetzen. Wichtig ist dabei, dass die Betriebskosten von Anfang an mit eingeplant werden. Die genaue Aufschlüsselung ist im Anhang zu finden.

Kostenschätzung

In der Anlage C: Kosten Infrastruktur werden die getroffenen Annahmen für die Inputgrößen und Dimensionierung transparent gemacht und eine erste Kostenschätzung vorgenommen.

Unter einer GPU-Server-Einheit wird dabei eine Gesamtlösung inklusive aller erforderlichen Bestandteile (GPUs, Netzwerk und Storage) verstanden. Für die Abschätzung der Kosten für das Pre-Processing wird angenommen, dass die CPU-Sever als Cloud-Lösung von einem Anbieter bezogen werden.

In der weiteren Planung für das LEAM-Vorhaben ist der Ausbaupfad noch zu definieren, sodass auch die Kosten für Ausbau und die Reinvestition explizit aufgenommen werden können.

Eine erste Abschätzung der Kosten (Details s. Anlage C: Kosten Infrastruktur) ergibt folgendes Szenario:

Hardware Investitionskosten: ca. **300.000.000 €**
Hardware Betriebskosten pro Jahr: **15.000.000 € bis 20.000.000 €**

Daraus ergeben sich über einen Zeitraum von 5 Jahren Gesamtkosten von **min. 375 Mio. Euro** und Betriebskosten pro Stunde für das Gesamt-Cluster von **ca. 8.000 bis 9.000 Euro**. Vor allem in Anbetracht steigender Preise für Hardware sowie der Stromnutzung, sollen diese Zahlen nur als Orientierungswerte verstanden werden.

Die ersten LEAM-Modelle

LEAM hat das Ziel diverse Arten großer KI-Modelle zu realisieren. Dennoch soll der Fokus zunächst auf großen multilingualen europäischen Sprachmodellen liegen, die mit explizitem Wissen aus großen Wissensbasen angereichert sind. Bereits während der Realisierung dieser Sprachmodelle sollen andere Arten großer europäischer KI-Modelle geplant und vorbereitet werden. Diese können z.B. aus den Bereichen Biomedizin, Unternehmensprozesse oder Logistik kommen.

Priorisierung auf Sprachmodelle

Die anfängliche Konzentration auf Sprachmodelle hat mehrere Gründe:

- **Stand der Technologie**: die Forschung der letzten Jahre hat Lernmethoden und Architekturen hervorgebracht, deren Reifegrad einerseits eine hohe Erfolgswahrscheinlichkeit garantiert und die andererseits aber auch noch ein hohes Potential für weitere Sprünge in der Technologieevolution bieten.

- **Unmittelbarer Bedarf**: große neuronale Sprachmodelle haben ein großes Anwendungspotential, weil sie sich für die Verbesserung von vielen wichtigen Anwendungen eignen, z.B. Suchtechnologie, Textgenerierung, maschinelle Übersetzung, Wissensextraktion, Informationssysteme, persönliche Assistenten, Content Qualitätsverbesserung, IPR Monitoring.

- **Verfügbarkeit von Daten**: Neben den offen verfügbaren großen Sprachdatensammlungen gibt es zusätzliche geeignete große Sprachdatenbestände für europäische Sprachen, die in die amerikanischen und chinesischen Sprachmodellen noch nicht eingegangen sind.

- **Europäische Interessen**: Die US-amerikanischen und chinesischen Sprachmodelle sind für die europäischen Industrien und Gesellschaften nur sehr schwierig zugreifbar und kaum kommerziell nutzbar. Desweiteren erfüllen sie in der Abdeckung der Sprachen und Anwendungsbereiche sowie in der Einhaltung von ethischen Standards nicht wirklich die europäischen Anforderungen.

Eigenschaften der Sprachmodelle

Im Folgenden werden die ersten drei Generationen von LEAM Sprachmodellen vorgeschlagen, die wir hier als LEAM-1, LEAM-2 und LEAM-3 bezeichnen. Sie werden sich

von den größten derzeit bestehenden Sprachmodellen durch mehrere Eigenschaften unterscheiden:

- Eine bessere und gezielte Abdeckung europäischer Sprachen, später auch außereuropäischer Sprachen, durch eine Einbeziehung von zusätzlichen existierenden europäischen mono- und multilingualen Sprachdatenbeständen.
- Eine gezielte Ergänzung der Web-Crawls und Buchkollektionen um bestehende Datenbestände aus Textsorten und Genres, die wichtig für Wirtschaft und Gesellschaft sind.
- Eine systematische Anreicherung der Sprachdaten um Wissensbestände aus großen offenen Wissensgraphen wie Wikidata und DBpedia.
- Eine bestmögliche Adaption des Modells auf die Werte und Anforderungen der europäischen Gesellschaft durch geeignete Datengewichtungen, Vermeidung von ungewollten Doubletten in den Trainingsdaten und eine dedizierte überwachte zweite Prätrainingsphase.

Darüber hinaus sollen die Trainingsdaten für LEAM-2 auch hinreichend große Anteile an Sprachdaten für wichtige außereuropäische Sprachen beinhalten, um diese Sprachen in Anwendungen aufzunehmen.

LEAM-3 soll dann auch zusätzlich auf Multimediadaten trainiert werden, insbesondere auf kombinierten Sprach- und Bilddaten, sowie auf annotierten Bilddaten. Wenn es die Ressourcen an Daten und Rechenkapazität erlauben, können unter Umständen auch Videodaten einbezogen werden.

Da als Ausgangspunkt der Datenkuratierung die gigantischen offenen Datenbestände verwendet werden, die in GPT 3, Megatron Turing NLG 530B und Google MT5 eingeflossen sind, sollten bereits die LEAM-1 Modelle durch die geplante Hinzunahme der zusätzlichen Daten größer und durch die Art und Qualität dieser Daten auch erwartbar leistungsstärker sein. Dabei wird aktuell noch nach Lösung gesucht, die Open-Crawl Daten der amerikanischen Modelle datenschutzkonform zu verwenden.

Wegen der Reife der existierenden Architekturen und Lernverfahren für große Sprachmodelle und auch um in der ersten Phase der LEAM Entwicklung eine gewisse internationale Vergleichbarkeit zu erreichen, soll zumindest in dieser Anfangsphase eine generative Transformer-Architektur eingesetzt werden. Ein Ansatz der sich in den GPT Modellen von OpenAI und den T5 und Gopher Modellen von Google bewährt hat.

Diese Auswahl an Modellart, Datenauswahl, Architektur und Lernstrategie erfüllt die folgenden Anforderungen, die für den Erfolg von LEAM ausschlaggebend sein werden:

- **Pragmatismus und Geschwindigkeit**: Der bestehende Vorsprung erfordert eine erste Technologieentwicklung auf der Basis bestehender bewährter Methoden. Parallel zu dieser Entwicklung und dann besonders unter Nutzung der ersten LEAM Modellgeneration muss aber natürlich deutsche und europäische Forschung der Verbesserung dieser Methoden und der Schaffung neuer Methoden gewidmet sein.

- **Originalität und Innovation**: Obwohl auf bestehenden Paradigmen des neuronalen Lernens aufgesetzt wird, werden die ersten LEAM Modelle durch die Schwerpunkte Multilingualität und Wissen verbesserte und auch neue Anwendungspotenziale aufweisen. Sie werden durch diese Schwerpunkte auch neuen Erkenntnisgewinn ermöglichen und somit die Forschung bereichern. Auch die vorgesehene massive Anwendung und Weiterentwicklung neuester Methoden zur Reduktion von Bias und Toxicity wird zu neuen originären Forschungsresultaten führen.

- **Bedarf und Anwendungspotenziale**: Bei der Evaluation von GPT 3, Megatron Turing NLG 530B und Google MT5 zeigte sich, dass die vortrainierten Sprachmodelle eine große Breite von Anwendungen unterstützen und in fast jeder dieser Anwendungen die Leistungsfähigkeit früherer Technologien deutlich hinter sich lassen. Zu diesen Anwendungsklassen gehören: Textklassifikation, Informationsextraktion, semantische und multilinguale Suche, Textgenerierung, Textzusammenfassung, Chatbots/Digitale Assistenten und automatische Übersetzung. Einige dieser Technologien werden bereits praktisch eingesetzt, erfüllen aber noch nicht die hohen Anforderungen an Korrektheit und Verlässlichkeit des produktiven Einsatzes in Wirtschaft und Verwaltung. Durch die geplante systematische Erweiterung der Modelle um hochwertige Daten aus diesen Bereichen sollen neue Anwendungsfelder eröffnet werden. Auch durch die Erweiterung auf zusätzliche Sprachen erhöht sich das Anwendungspotenzial. Im Gegensatz zu den bestehenden Sprachmodellen in den USA und in China, werden die LEAM Modelle der europäischen Wirtschaft und Forschung zu besten Bedingungen zur Verfügung stehen. Selbst für mittelständische Unternehmen und für Startups wird die Nutzung unproblematisch und erschwinglich sein.

- **Europäischer Fokus**: Zur Mission von LEAM gehört neben der Nutzung der Resultate durch die europäische Wirtschaft, Forschung und andere Bereiche der Gesellschaft auch die Fokussierung auf die Stärken und besonderen Prioritäten

der deutschen und der europäischen Gesellschaft. Die angestrebte Mehrsprachigkeit fördert die Inklusion, die Bewahrung der kulturellen Vielfalt und die Unterstützung von Mitgliedsländern, die eine solche technologische Entwicklung allein nicht leisten können. Durch die zusätzliche Konzentration auf wirtschaftliche und technische Daten und Anwendungen und mit der Hinzunahme von Wissensdaten stärken wir auch die Teilbereiche der heimischen KI-Industrie, in denen Deutschland und Europa noch sehr große Chancen haben: Enterprise AI, Smart Manufacturing, Health, Education und KI für öffentliche Verwaltungen. Der Schwerpunkt auf strukturierten Wissensbeständen entspricht auch einem europäischen Schwerpunkt, denn die fortgeschrittensten Wissenstechnologien sind meist europäischen Ursprungs. Letztendlich ergibt sich ein weiterer europäischer Fokus durch den Schwerpunkt auf der Verwendung von KI-Methoden zur Reduktion von Bias und Toxizität in den auf Massendaten trainierten Modellen.

Weitere Informationen zu den ersten LEAM Sprachmodellen finden sich in Anlage A: Große KI-Sprachmodelle.

Governance und Finanzierung

Um in Deutschland große KI-Modellen zu entwickeln und zu implementieren ist eine eigene Organisationseinheit erforderlich, die unter anderem ein KI-Supercomputing Rechenzentrum aufbaut und betreibt.

Im Folgenden wird diese Organisationseinheit zur besseren Lesbarkeit als LEAM Betreibergesellschaft oder kurz LBG bezeichnet.

LBG Bereiche – Übersicht

Die Tätigkeiten, Services und Organisationseinheiten der LBG lassen sich in folgende Bereiche aufteilen:

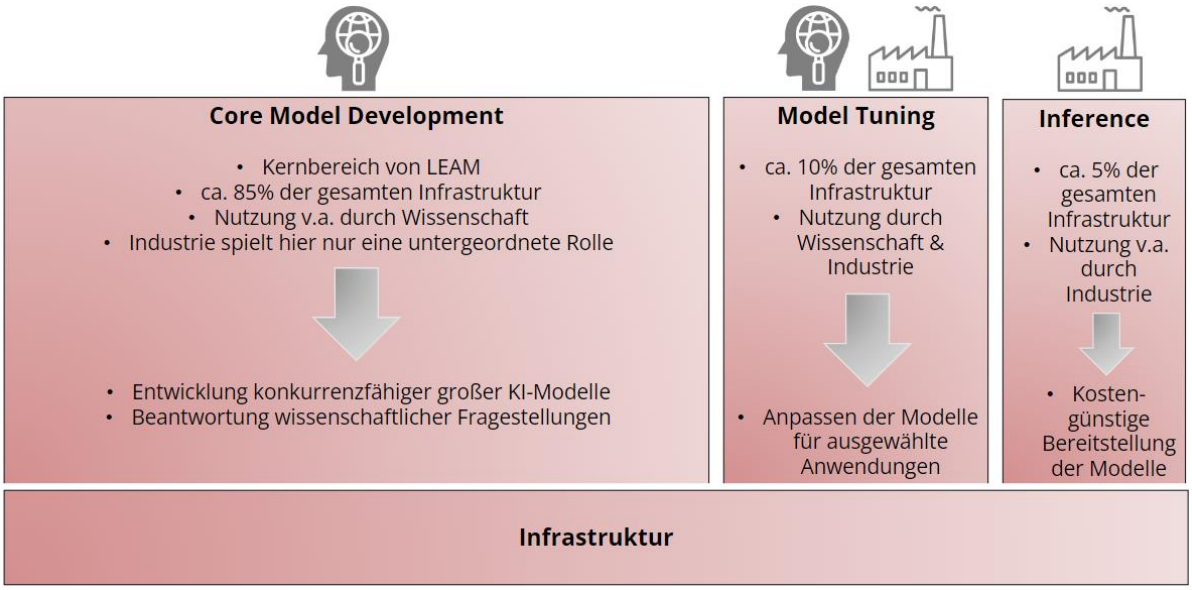
1. Bereitstellung einer Basis Infrastruktur
2. Erstellung großer KI-Modelle / Core Model Development
3. Tuning und Anpassungen großer KI-Modelle / Model Tuning
4. Bereitstellung der Services auf Basis großer KI-Modelle / Inference

Das Schaubild verdeutlicht diese Aufteilung. Die Organisationseinheit Infrastruktur ist dabei das Fundament der gesamten LBG. Auf dem von dieser Organisationseinheit bereitgestellten KI-Hochleistungszentrum werden die Modelle der anderen Organisationsbereiche entwickelt und bereitgestellt werden.

Der mit Abstand größte Teil der Infrastruktur - rund 85% - sollen dabei für das Erstellen großer KI-Modelle genutzt werden. Aktuell gehen die LEAM Partner davon aus, dass bis zu 90% der in diesem Cluster zur Verfügung stehenden Rechenkapazität für öffentliche Projekte wie beispielsweise Forschungsarbeiten genutzt werden. Nur rund 10% würden dann von kommerziellen Kunden genutzt. Aufgrund dieser Aufteilung kann eine Gemeinnützigkeit der LBG gewährleistet werden.

In den anderen beiden Bereiche, Model Tuning und Inference, spielen kommerzielle Kunden wiederum eine größere Rolle. Vor allem im Bereich Inference wird erwartet, dass kommerzielle Abnehmer mit rund 90% die Hauptabnehmer sind. Dies hat den Grund, da die von der LBG entwickelten KI-Modelle vor allem der Wirtschaft zugutekommen sollen. Es bleibt aber zu beachten, dass diese beiden Organisationseinheiten nur rund 10% bzw. 5% der Gesamtinfrastruktur belasten sollen. Der Fokus liegt damit eindeutig auf der forschungsbasierten Entwicklung und temporären Bereitstellung großer KI-Modelle.

Es sollte außerdem erwähnt werden, dass das Tuning- und das Inference-Cluster-Team lediglich eine Supportfunktion für die Abnehmer übernimmt und kein vollständiges Projektgeschäft. Stattdessen wird die LBG versuchen, die Kunden an KI-Unternehmen oder Forschungsinstitutionen zu vermitteln.



Aufteilung der Organisationseinheiten der LEAM Betreibergesellschaft (LBG)

Im Folgenden werden die vier Organisationseinheiten näher beleuchtet. Bei allen vier Organisationseinheiten werden die Aufgaben, Kosten und Einnahmemöglichkeiten der LBG aufgezeigt. Beispielszenarios zu den Kosten und Einnahmemöglichkeiten sind in den Anlage D: Bilanz Core Model Development bis G zu finden.

LEAM Bereich - Infrastruktur

Ein zentrales Ziel der LEAM-Initiative ist es, eine international konkurrenzfähige Rechen-Infrastruktur aufzubauen, die als Basis dient, skalierbare KI-Modelle trainieren zu können.

Die LBG wird diese Infrastruktur warten und einer kontinuierlichen Erneuerung unterziehen, so dass sie stets auf dem neuesten Stand bleibt. Die verfügbare Rechenzeit wird Forschungsinstitutionen und zu einem kleinen Teil privatwirtschaftlichen Unternehmen über ein noch zu definierendes Verfahren (abhängig vom Betriebs- und Finanzierungsmodell) zur Verfügung gestellt.

Aufgaben

Im Bereich der Infrastruktur kommen der LBG drei Aufgabenbereiche zu.

Zunächst gilt es, das KI-Rechenzentrum anhand der in Kapitel 2 aufgestellten Kriterien aufzubauen. Der LBG kommen dabei folgende Aufgaben zu:

1. Detaillierte Konzeption der Infrastruktur
2. Auswahl von Standorten und Housing
3. Ausschreibung der Infrastruktur und Verhandlungen mit Vendoren
4. Kauf, Installation und Inbetriebnahmen des Rechenzentrums
5. Aufbau einer Betriebsmannschaft
6. Aufbau eines LCM (Life-Cycle Management)
7. Auswahl eines Cloud-Ökosystems (unter Berücksichtigung von Gaia-X)

Nach dem Aufbau des Rechenzentrums, muss der Betrieb sichergestellt werden. Die LBG muss dafür u.a. folgende Aufgaben wahrnehmen:

1. Fortlaufende Wartung und Pflege der Infrastruktur
2. Implementierung und Aktualisierung des kompletten Software-Stacks (Infrastruktur, Cloud etc.)
3. Betrieb des Rechenzentrums
4. Entwicklung eines Servicekatalogs (Web-basierters Portal für die Nutzung der Infrastruktur)
5. Bereitstellung von flexiblen Infrastructure as a Service - Diensten
6. Entwicklung und Betrieb von Abrechnungsmodellen

Gleichzeitig gilt es, die Datensammlung aufzubauen und zu betreiben:

1. Konzeption der Infrastruktur für die Datensammlung
2. Etablierung von Prozessen für die Abbildung der der Datengewinnung in der Infrastruktur
3. Verwaltung von Storage-Diensten für die Speicherung und Veredelung von Daten
4. Bereitstellung der Daten incl. eventueller Abrechnungsmodelle

Kosten

Die hier aufgeführten Kosten für den Aufbau und den Betrieb einer LEAM Infrastruktur stammen aus den Erfahrungswerten der LEAM Partner. Eine genaue Übersicht findet sich in Anlage C: Kosten Infrastruktur.

Die initialen Ausgaben zum Aufbau der Infrastruktur bestehen aus den Kosten für die eigentliche Hardware sowie aus Kosten für Gebäude und Infrastruktur:

Investitionskosten	
Gebäude und Infrastruktur	50.000.000 € bis 80.000.000 €
Anfängliche Hardware	230.000.000 €
CAPEX gesamt	**280.000.000 € bis 310.000.000**

Die Betriebsausgaben setzen sich aus Stromkosten, Kosten für die Instandhaltung sowie Personal und sonstigen Posten zusammen:

Betriebskosten	
Instandhaltung	6.000.000 €
Stromkosten	6.000.000 € bis 8.000.000 €
Personal (10 MA) + Sonstiges	3.000.000 € bis 5.500.000 €
OPEX jährlich	**15.000.000 € bis 19.500.000 €**

Einnahmemöglichkeiten

Der Bereich Infrastruktur stellt die IT-Infrastruktur als Grundvoraussetzung für die anderen Bereiche bereit und generiert damit keine direkten Einnahmen. Ob eine interne Verrechnung mit den anderen Organisationseinheiten wünschenswert oder sogar benötigt wird, soll zu einem späteren Zeitpunkt diskutiert werden.

Ein Potenzial für zusätzliche Einnahmen bzw. Kostenminderung der Infrastruktur besteht in der Nutzung und Weitervermarktung der entstehenden Abwärme.

Ausblick nach fünf Jahren

Die LEAM Partner gehen davon aus, dass es innerhalb eines Zeitraumes von fünf Jahren notwendig werden wird, die Hardware auf den dann aktuellen Stand der Technik zu verbessern und defekte Hardware zu ersetzen. Die erforderlichen Reinvestitionen können verringert werden, in dem ein Teil veralteter Bestandshardware weiter für den Bereich Inference genutzt wird.

LEAM Bereich – Core Model Development

Neben dem Aufbau und Betreiben der Infrastruktur ist die Entwicklung von und die

Forschung an großen KI-Modellen die Kernaufgabe der LBG. Die LBG wird die Erstellung der KI-Modelle koordinieren und organisieren, aber auch selbstständig umsetzen. Externe Partner werden Beratungsservices in Anspruch nehmen und Modellentwicklung in Auftrag geben können.

Aufgaben

Die Aufgaben der LBG im Bereich Core Model Development lassen sich in drei Bereiche unterteilen: Entwicklung und Training großer KI-Modelle, Ressourcenplanung und Koordination der Entwicklung durch externe Partner.

Im Arbeitsbereich Entwicklung großer KI-Modelle gilt es für die LBG zunächst, einen globalen Trainings-Daten-Pool aufzubauen und diesen fortwährend entsprechend engster Datenschutzvorgaben zu pflegen. Darauf aufbauend können eigene Modell-Trainings entwickelt und durchgeführt werden.

Die LBG wird Wert darauflegen, dass die Recheninfrastruktur zu jeder Zeit optimal ausgelastet ist. Die Planung und Ausnutzung der Ressourcen ist daher eine weitere Aufgabe des Bereiches Core Model Development. Der Arbeitsbereich muss bspw. dafür sorgen, dass alle Teilbereiche der Modellskalierung beachtet und vorangetrieben werden.

Darüber hinaus wird die LBG ihre Services externen Forschungsinstituten und zu einem kleinen Teil auch privatwirtschaftlichen Unternehmen zur Verfügung stellen. Hierbei kommt der LBG die Aufgabe der Koordination und Organisation der Modell-Erstellung zu. Externe Partner werden bei der Erstellung von KI-Modellen unterstützt.

Kosten

Bei der Entwicklung der Core-Models kommt ein Trichtermodell zum Einsatz. Zunächst berechnet die LBG viele kleinere Modelle, ehe die Ergebnisse daraus benutzt werden, um wenige größere Modelle zu berechnen. Dieser Vorgang wird wiederholt, bis ein voll ausgereiftes großes KI-Modell bereitsteht.

Die Erfahrungswerte der LEAM-Partner zeigen, dass die Vorbereitung und Umsetzung eines solchen Trainingslaufs mindestens drei Monate benötigt. Demnach plant LEAM aktuell, drei oder vier große KI-Modelle pro Jahr zu erstellen.

Der Fokus des Bereiches Core Model Developments soll außerdem auf der Forschung liegen. Nur etwa 10% des Service werden durch private Unternehmen in Anspruch

genommen. Kosten für den Bereich Core Model Development entstehen durch Personalkosten (incl. Personalnebenkosten).

Einnahmemöglichkeiten

Der Bereich Core Model Development stellt die entwickelten Modelle möglichen Anwendern open source und damit kostenlos zur Verfügung. Es sind damit keine direkten Einnahmen vorgesehen. Jedoch schließt die LBG eine privatwirtschaftliche Verwertungsmöglichkeit nicht vollkommen aus. Rund 10% der Infrastruktur sollen an privatwirtschaftliche Unternehmen vermietet werden.

LEAM Bereich – Model Tuning

In der Regel benötigen große KI-Modelle für spezielle Anwendungen ein Finetuning. Die LBG wird diesen Bereich ebenfalls anbieten.

Aufgaben

Die LBG wird das Tuning und die Customization großer KI-Modelle anbieten und bei Auftragserteilung - ggf. in Kooperation mit dem Kunden - selbstständig vornehmen. Darüber hinaus wird sie externe Kunden beim Tuning beraten und unterstützen.

Kosten und Einnahmemöglichkeiten

Kosten für den Bereich Model Tuning entstehen durch Personalkosten.

Erlöse erzielt der Bereich Model Tuning durch die Berechnung von Beratungs- und Unterstützungsservices an externe Partner sowie Lizenzeinnahmen im Rahmen dedizierter Modell-Erstellung.

LEAM Bereich – Inference

Ein wesentlicher Beitrag der LEAM-Initiative zur Erhaltung der digitalen Souveränität Europas sowie zur Standortförderung ist es, international konkurrenzfähige, große KI-Modelle zu trainieren und diese sehr günstig als Service per API öffentlich verfügbar zu machen.

Aufgaben

Die LBG wird zunächst ein API-Abrechnungsmodell unter Berücksichtigung der jeweiligen Komplexität für die SAAS-Nutzung der allgemeinen prätrainierten Modelle entwickeln und dieses umsetzen. Sie wird außerdem das Hosting und den Betrieb individueller KI-Modelle für dedizierte Kunden übernehmen.

Kosten und Einnahmemöglichkeiten

Die Kosten für den Betrieb der KI-Modelle bestehen im Wesentlichen aus Personalkosten für Forschung und Implementierung. Darüber entstehen zusätzliche Hardwarekosten für das Hosten.

Einnahmen erzielt der Bereich Inference durch die Bereitstellung der Services über API. Hierbei soll ein Preis pro genutztes Token abgerechnet werden. Der Preis pro Token soll die entstandenen Kosten decken, aber so günstig wie möglich angeboten werden.

Zusammenfassung

Die Initiative LEAM schlägt eine LEAM Betreibergesellschaft vor, die sich in vier Organisationseinheiten aufteilt. Den Bereichen Infrastruktur und Core Model Development kommen dabei die größte Bedeutung zu.

In der aktuellen Kalkulation gehen die LEAM Partner von einer Investition in Höhe von rund **300.000.000 €** für die Infrastruktur aus. Hinzu kommen jährliche Betriebskosten für das Betreiben der Infrastruktur in Höhe von **15.000.000 € bis 20.000.000 €** sowie Kosten für ein kontinuierliches Update der Infrastruktur.

Zum aktuellen Zeitpunkt ist es für die LEAM Partner nicht ohne weiteres möglich, eine verlässliche Bilanz für die drei Organisationseinheiten Core Model Development, Modeltuning und Inference aufzustellen. Verschiedene Szenarien sind in den Anlage D: Bilanz Core Model Development aufgelistet. Diese Zahlen stellen keine endgültige Aussage über Kosten und Einnahmen dar, sondern dienen lediglich als eine erste Orientierung.

Die Kosten bestehen vor allem aus Mitarbeiterkosten für den Betrieb der LBG und der Entwicklung der großen KI-Modelle. Es ist anzumerken, dass einige Kostenpunkte (bspw. Anzahl der Mitarbeiter) lediglich auf Erfahrungswerten und Annahmen der LEAM Partner beruhen.

Bei den Einnahmen ist vor allem die Marktgröße für die Bereiche Model Tuning und Inference unbekannt. GPT3 wurde in den ersten neun Monaten bereits von mehr als 300

Applikationen genutzt. Ob eine ähnliche Adaptionsrate in Deutschland bzw. Europa zu erwarten ist, lässt sich nicht vorhersagen. Erfahrungen des KI Bundesverbandes zeigen aber, dass ein Großteil der KI Unternehmen in Deutschland von einem solchen Angebot Gebrauch machen würden.

Organisationsformen

Eine der wesentlichen Fragen zu diesem Zeitpunkt ist die Organisationsform bzw. Rechtsform der zukünftigen LEAM Betreibergesellschaft. Als mögliche Rechtsformen kommen potenziell folgende Modelle in Frage:

- Eingetragener Verein
- GmbH
- Gemeinnützige GmbH
- AG
- Stiftung

Auch Mischformen dieser Organisationen sind möglich. Zum jetzigen Zeitpunkt legen sich die LEAM Unterstützer nicht auf eine Organisationsform fest. Stattdessen sollen alle Optionen genau geprüft und die bestmögliche gewählt werden.

Finanzierungsmodelle

Es gibt eine Reihe von potenziellen Finanzierungsmodellen, die unterschiedlich ausgestaltet und kombiniert werden können.

Die Konzeption der Organisations- bzw. Rechtsform inkl. zugehöriger Finanzierungsmodelle ist eine der wesentlichen Aufgaben im Rahmen der nächsten Schritte zur Detaillierung der LEAM Konzeption.

Als Diskussionsgrundlage werden hier grob drei prinzipielle Szenarien vorgestellt.

Szenario 1: Öffentliche Finanzierung

Das Projekt LEAM dient der Forschung an und der Entwicklung von großen KI-Modellen mit dem Ziel dem gesamten Wirtschaftsstandort Deutschland und der EU zu helfen. Langfristig soll die digitale Souveränität der EU im Bereich im KI gesichert werden. Das Projekt dient damit eindeutig Staatszielen und eine Finanzierung durch öffentliche Mittel ist möglich.

Die Organisationsform kann dabei eine private sein, solange die Gemeinnützigkeit des Projektes gesichert ist. Die zukünftige LBG müsste sich dann aktiv auf Förderprogramme bewerben.

Der größte Vorteil dieser Art der Finanzierung ist ein sicherer und stabiler Geldfluss, solange die LBG Fördergelder bezieht. Darüber hinaus gibt es keine Gewinnverpflichtung. Dies ist insbesondere hilfreich, falls LEAM längerfristig nicht profitabel betrieben werden kann.

Gleichzeitig stellt eine hundertprozentige öffentliche Förderung die LBG auch vor Herausforderungen. So ist unter diesem Szenario eine finanzielle Verwertung erschwert und es gibt weitere enge Richtlinien. Darüber hinaus würde sich die Initiative in die Abhängigkeit von politischen Entscheidungsträgern geben. Schließlich müsste geklärt werden, welche Förderprogramme für LEAM in Frage kommen und was mit der LBG nach der öffentlichen Förderung geschieht.

Für den Erfolg von LEAM wird es außerdem entscheidend sein, dass die Wirtschaft die entwickelten KI-Modelle übernimmt. Sollte es kein finanzielles Engagement der Wirtschaft geben, ist eine geringe Adaption der Modelle wahrscheinlicher.

Szenario 2: Private Finanzierung

Ein zweites Szenario ist die Finanzierung komplett über privates Kapital. Die LBG könnte bspw. als Joint Venture verschiedener Konzerne aufgesetzt oder über Venture Capital finanziert werden.

Eine solche Finanzierung hat den Vorteil, unabhängig politischer Einflüsse zu sein und sich vollkommen auf den wirtschaftlichen Nutzen fokussieren zu können. Gleichzeitig wären Akteure aus der Wirtschaft von Anfang an in der Initiative involviert.

Dennoch scheidet eine private Finanzierung nach Ansicht der LEAM-Partner zu diesem Zeitpunkt aus. Die hohen Investitionskosten gepaart mit einer unklaren Situation hinsichtlich der Profitabilität machen die Initiative für Investoren uninteressant.

Darüber hinaus besteht bei einer rein privaten Finanzierung die Sorge, dass der Bereich der Forschung nicht ausreichend berücksichtigt würde.

Szenario 3: Public Private Partnership

Bei einer Public-Private-Partnership (PPP) wird ein öffentliches Vorhaben teilweise oder ganz durch private Unternehmen finanziert. Hierbei werden öffentliche Leistungen in ursprünglich staatlicher Verantwortung nun in Kooperation mit der privaten Wirtschaft realisiert.

Für die Umsetzung einer solchen Partnerschaft gibt es verschiedene Modelle. Die LEAM-Partner halten dabei die Gründung einer gemeinschaftlichen Gesellschaft zwischen öffentlichen und privaten Akteuren für die beste Lösung. Je nach Ausgestaltung der Gesellschaft können so auch Fördermittel beantragt werden.

Für LEAM wäre bspw. eine gGmbH möglich, die primär als Forschungseinrichtung agiert und so 100% Förderfähigkeit erreichen kann. Für gewinnbringende Tätigkeiten könnte eine GmbH gegründet werden.

Eine solche Public Private Partnership hat den Vorteil der gegenseitigen Unterstützung von öffentlichen und privaten Akteuren. So kann die Finanzierung gesichert und gleichzeitig ein Engagement der privaten Wirtschaft erwartet werden. Darüber hinaus kann LEAM über eine PPP eine größere Wahrnehmung in der Öffentlichkeit erzeugen.

Bewertung und Empfehlung

Eine Finanzierung durch rein privates Vorgehen scheint ungeeignet, da die Investitionssumme des Projektes hierfür zu hoch ist und eine langfristige Profitabilität nicht gewährleistet werden kann. Eine rein öffentlich gestaltete Finanzierung erweist sich ebenfalls als schwierig u.a. wegen der offenen rechtlichen Fragen vor allem in Bezug auf Beihilfe und Kartellrecht.

Am wahrscheinlichsten ist nach derzeitigem Kenntnisstand eine Public Private Partnership. Hier kann ein öffentliches Vorhaben teils privat und teils öffentlich finanziert werden. Die genaue Ausgestaltung einer solchen PPP soll in Gesprächen mit der Wirtschaft und öffentlichen Akteuren genauer erörtert werden. Es scheint aber realistisch, dass der größte Teil des Projektes von der öffentlichen Hand über Forschungsgelder finanziert werden wird, während private Akteure sich in bestimmten Teilbereichen beteiligen können.

Eine alternative Idee der LEAM UnterstützerInnen ist es, zunächst eine Finanzierung über öffentliche Gelder zu erhalten, um die LBG zu etablieren. Nach einigen Jahren des Betriebs soll dann ein Umstieg auf private Kapitalgeber erfolgen.

Die Initiative LEAM steht zu dieser Fragestellung aktuell im Austausch mit relevanten Akteuren.

Zusammenarbeit mit weiteren Initiativen

In Europa und weltweit gibt es eine Reihe von Initiativen zur Demokratisierung von großen KI-Modellen. LEAM wird sich intensiv mit diesen auseinandersetzen und Kooperationen anstreben.

OpenGPT-X

Im Januar 2022 startete das vom Bundesministerium für Wirtschaft und Klimaschutz (BMWK) geförderte Projekt OpenGPT-X. Die zehn Partner aus Wirtschaft, Forschung und Medien haben das Ziel in drei Jahren große KI-Sprachmodelle zu entwickeln. Diese Modelle sollen Anwendern im Gaia-X Ökosystem zur Verfügung gestellt werden. Anders als kommerzielle KI-Sprachmodelle, wie das amerikanische GPT-3 oder das chinesische Wu Dao 2.0 verspricht OpenGPT-X die Orientierung an europäischen Wertestandards wie beispielsweise den Datenschutz und die Implementierung mehrerer europäischer Sprachen.

Inhaltlich bewegt sich das Projekt OpenGPT-X damit nah an LEAM. LEAMs Ziel geht allerdings weit über die Entwicklung und Bereitstellung von Sprachapplikationen hinaus. Bei LEAM geht es um den Aufbau und die Bereitstellung einer neuen Infrastruktur sowie der Entwicklung vielfältiger großer KI-Modelle. OpenGPT-X hat nicht die Mittel, diese Ziele zu erreichen. Aber OpenGPT-X wird als ein wichtiger Schritt gesehen, Sprachmodell-Technologien für neue AnwenderInnen nutzbar zu machen.

Einige der Partner im Projekt Open GPT-X sind bereits Unterstützer der Initiative LEAM. Dazu gehören bspw. der KI Bundesverband sowie einige seiner Mitglieder, das Fraunhofer Institut IAIS und das DFKI. Der Austausch mit dem Projekt Open GPT-X ist daher bereits weit fortgeschritten. OpenGPT-X kann somit als erster Schritt für große europäische KI-Modelle gesehen werden und bietet vor allem einen großen Erfahrungsschatz, aus dem die Initiative LEAM lernen kann.

KI-Servicezentren

Anfang 2022 fördert das Bundesministerium für Bildung und Forschung (BMBF) die Etablierung von KI-Servicezentren in Deutschland. Diese KI-Servicezentren werden mit Hardware, Software und Personal ausgestattet, um Forschung und einen Transfer in die Wirtschaft zu ermöglichen.

LEAM plant mit diesen Servicezentren zusammenzuarbeiten. Dabei soll diese Zusammenarbeit nicht nur auf informationeller Ebene stattfinden, stattdessen können die KI-Servicezentren gemeinsam mit LEAM auch kleine oder mittelgroße KI-Modelle berechnen und bereitstellen.

Hugging Face

Hugging Face ist ein open-source und Plattform Provider im Bereich Machine Learning. Mit einem Community-Ansatz hat sich das Unternehmen seit der Gründung 2016 in New York zu einem der weltweit führenden NLP Startups entwickelt. Ihr Ziel: NLP voranzubringen und zu demokratisieren. Tausende Kunden (u.a. Amazon, Microsoft, Google, ...) nutzen dabei vor allem die Transformer Library. Eine über API verfügbare Library, die über 30 vorgefertigte NLP Modelle und über 100 Sprachen zusammenfasst.

Gemeinsam mit GENCI und IDRIS hat HuggingFace außerdem die Initiative BigScience ins Leben gerufen. Die Arbeitsgruppe bringt Forscher:innen aus den Feldern KI, NLP, Sozialwissenschaften, Ethik und Public Policy zusammen. Rund um den französischen Supercomputer Jean Zay soll so ein Netzwerk nach dem Vorbild CERN/LHC entstehen.

EleutherAI

EleutherAI ist eine Grassroot Bewegung aus ehrenamtlichen Entwicklern, Ingenieuren und Forschern. Ihr Ziel ist es, OpenAIs GPT-3 als Open-Source Software zu replizieren. Dafür arbeitet die Bewegung seit Juli 2020 an der GPT-Neo Familie. Ihr bisher größtes Modell GPT-J-6B haben sie mit 6 Milliarden Parametern trainiert und im Juni 2021 veröffentlicht.

Claire

Claire, die Confederation of Laboratories for Artificial Intelligence Research in Europe, ist ein Netzwerk aus mehr als 400 Forschungsinstituten und über 24.000 Forschern aus 37 europäischen Staaten. Das Ziel ist, die europäische KI-Forschung zu stärken. Ähnlich der Infrastruktur des CERN soll ein europäischer Hub entstehen, in dem KI-Forscher verschiedener Hintergründe zusammenkommen und forschen können.

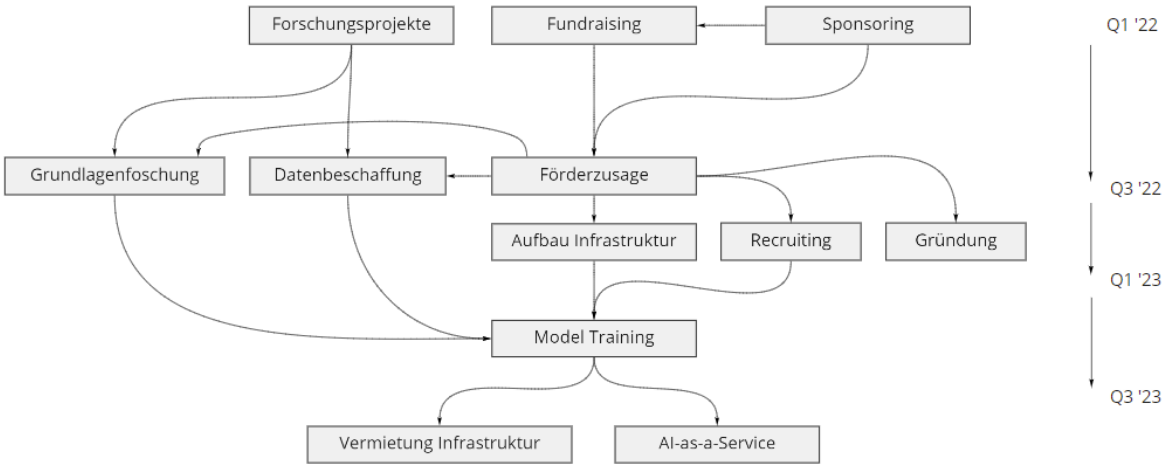
European Language Grid

Das European Language Grid ist eine Initiative aus Mitgliedern der europäischen Language Technology Community. Das EU-finanzierte Projekt sammelt Sprachtechnologien und Datensätze - auch von kleineren Sprach-Communities – und

trägt sie auf ihrer Cloud-Plattform zusammen. Das Ziel ist es, eine zentrale Plattform für Sprachtechnologien in Europa zu schaffen.

Zeitplan

Es ist sehr wichtig, dass die beabsichtigten Maßnahmen möglichst schnell umgesetzt werden, da sich ansonsten der einzuholende Rückstand weiter vergrößert. Wenn möglich sollten Wege abseits tradierter Forschungs- und Innovationsförderung gefunden werden, um Geschwindigkeit aufzunehmen. Nachfolgend ist dargestellt, welche Aktivitäten parallelisiert und mit welchen zeitlichen Abhängigkeiten durchgeführt werden können.



Zeitplan für die Realisierung von LEAM

Einzelne Forschungsprojekte sollten parallel zum Fundraising des eigentlichen LEAM-Projektes beantragt werden, um Grundlagenforschung und Datenbeschaffung frühzeitig anzugehen. Das Projekt OpenGPT-X ist bereits gestartet und bietet bereits einen ersten Erfahrungsschatz.

Ein Sponsoring aus der Privatwirtschaft soll aufgebaut werden, damit LEAM-Unterstützer ihre Zeit effektiver für LEAM allokieren können. Daneben soll eine öffentliche Förderung gewonnen werden. Der Zeitplan hierfür sind sechs bis 12 Monate.

Direkt danach kann der Aufbau und die Einrichtung der Infrastruktur beginnen. Mit der Fertigstellung kann nach etwa sechs bis zwölf Monaten gerechnet werden.

Das Training der Modelle kann erst erfolgen, nachdem die Infrastruktur vorhanden ist (auch wenn Vorläufermodelle bereits auf gemieteter, kleinerer Hardware gerechnet werden können). Für das erste LEAM-Models werden sechs Monate Entwicklungszeit angesetzt.

Ab diesem Zeitpunkt kann die Infrastruktur vermietet werden und das erste LEAM-Modell kann als AI-as-a-Service zur Verfügung gestellt werden.

Ausblick

Gemeinsam mit dem Konzeptpapier wird ein Teaser Dokument erstellt. Dieses wird der Politik sowie potenziellen Partnern zur Verfügung gestellt. Gleichzeitig ist geplant, mit anderen Initiativen, die sich mit einer ähnlichen Thematik befassen, ins Gespräch zu kommen.

Verantwortliche Landes- und Bundesministerien wurden bereits kontaktiert und es gab einen ersten Austausch. Schnellstmöglich sollen alle interessierten Stakeholder an einen Tisch gebracht werden.

Daneben soll ein Vorbereitungsprojekt gestartet werden. Dieses Vorbereitungsprojekt soll dazu dienen, die in diesem Papier getroffenen Annahmen zu überprüfen und zu präzisieren sowie noch offene Fragestellungen zu klären. Diese betreffen vor allem die Bereiche der Finanzierung und Organisationsform, mögliche Unterstützung der Partner sowie einen Budgetplan.

Anlage A: Große KI-Sprachmodelle

Überblick LEAM Sprachmodelle

Die folgende Tabelle gibt einen ersten Überblick über die geplanten drei Generationen der ersten LEAM Sprachmodelle. Wie auch in GPT wird es zu jeder Generation verschiedene Versionen der Modelle geben, die sich in der Auswahl der Lerndaten unterscheiden.

Generation	Type	Data	Pretraining	compliance measures	est. dev. time
LEAM-1	T5	GPT-3 data + European language data + business-relevant and society-relevant language data + KG data in NL	decoder pretrained autoregressive LM or mT5 encoder-decoder pretraining	care in balancing & weighting data resources supervised PALMS-like tuning	7 months
LEAM-2	T5	LEAM-1 data + selected non-European language data + text-sentiment pairs	decoder pretrained autoregressive LM or mT5 encoder-decoder pretraining	better weighting schemes, improved PALMS tuning exploitation of metadata schemes	5 months
LEAM-3	?	LEAM-2 data + multimedia content	?	?	6 months

Anwendungen der Sprachmodelle

Durch die zentrale Rolle der menschlichen Sprache in Wirtschaft, Bildung, Gesundheit, Kultur und allen anderen Bereichen des Lebens gibt es für Sprachtechnologien unzählige Anwendungen. Bisher musste die Sprachkompetenz für jeder dieser Anwendungen separat und aufgabenspezifisch mit großem Aufwand modelliert werden. Erst mit der Entwicklung der ersten großen neuronalen Sprachmodelle gelang es, eine grundlegende wiederverwendbare maschinelle Sprachkompetenz technologisch zu modellieren, die sich ähnlich der menschlichen Sprachfähigkeit durch mehr oder weniger aufwändige Anpassung für ganz verschiedene Zwecke nutzen lässt.

Die Erfahrung mit den bestehenden großen Sprachmodellen hat bereits gezeigt, dass sich solche Modelle in einer großen Bandbreite von sprachtechnologischen Anwendungen einsetzen lassen und dabei in nahezu jeder Anwendungsklasse neue Einsatzbereiche erschließen bzw. Funktionsumfang und/oder die Performanz bestehender Anwendungen signifikant verbessern.

Zu den Anwendungsklassen gehören:

- Informationszugriff (IR) und semantische Suche
- Textklassifikation (nach Genre, Themen, Herkunft, Hate Speech usw.)
- Sprachverstehen (Erkennung von Entitäten, Fakten, Ereignissen, Argumenten usw.)
- Analyse von Sentiment und Meinungen
- Textgenerierung
- Textzusammenfassung
- Textüberprüfung und -korrektur (Grammatik, Stil, Terminologie, Eignung, Compliance)
- Textübersetzung
- Textvereinfachung
- Konversationelle KI (Frage-Antwort Systeme, Chatbots, persönliche Assistenten)
- natürlichsprachliche Schnittstellen
- Tutoren- und andere Lernsysteme
- Virtuelle Realität und Computerspiele
- Sprache als ein Modus für komplexe Vorhersagesysteme, z.B. in der med. Diagnose
- Sprache in der Beschreibung von Systemen, z.B. Software-Systeme, Recht

Da es für jede dieser Anwendungsklassen eine Vielzahl von Einsatzfeldern in Wirtschaft und Gesellschaft gibt, ist das wirtschaftliche Potential einer Basistechnologie mit hinreichender Sprachkompetenz unermesslich.

In Deutschland gibt es bereits eine ständig wachsende Zahl von KI Unternehmen, die Anwendungen aus diesen Klassen in ihrem Produktportfolio haben. Durch die Verfügbarkeit der geplanten Sprachmodelle wird diese Zahl sicher noch signifikant zunehmen.

Die Aufgaben in den Benchmarks, die zur Evaluation der Modelle herangezogen werden, sind nicht immer bereits eigenständige Anwendungen, sie sind aber aussagekräftige

Indikatoren für die Eignung der Modelle in Bezug auf die Anwendungsklassen. LEAM:AI wird anfangs die bestehenden bewährten Benchmarks verwenden, wodurch auch eine Vergleichbarkeit mit den bekanntesten internationalen Sprachmodellen hergestellt wird. LEAM:AI wird darüber hinaus aber zumindest versuchsweise auch Benchmarks aufnehmen und neu entwickeln, die Kriterien wie Bias und Toxizität messen.

Die Forschungsabteilungen der Hyperscaler erhalten zusätzlich zu den Ergebnissen durch die Evaluation anhand von Benchmarks auch Rückmeldungen von den Unternehmensbereichen, in denen die Sprachmodelle für Anwendungen eingesetzt werden. In LEAM:AI werden neben der Evaluation durch Benchmarks auch Mechanismen für die Rückkopplung aus den deutschen KI Unternehmen geschaffen, welche die Modelle in ihrer Anwendungsentwicklung einsetzen.

Offene Punkte

Trotz intensiver Beratungen gibt es noch einige offene Punkte in Bezug auf die ersten zu entwickelnden Sprachmodelle. Dazu gehören bspw. die Auswahl der Daten, die Reduktion von Bias oder auch die Effizienzsteigerung beim Trainieren der KI-Modelle. Diese Punkte sollen in einem ersten Projekt aufgegriffen und geklärt werden.

Anlage B: Dimensionierung der Infrastruktur

Die LEAM Arbeitsgruppe Infrastruktur hat ein erstes Szenario für den Aufbau einer Infrastruktur entwickelt. Diese Zahlen sollen dabei nur eine erste Orientierung bieten und nicht als finale Infrastruktur verstanden werden.

CPU-Server (Pre-Processing)

CPU vCores	10.000
CPU-Server	200

GPU-Server (Training, Tuning, Inferenz)	
Anzahl GPU-Server pro Einheit	20
Anzahl Racks pro Einheit	5
Anzahl GPU-Server-Einheiten	23
Anzahl spezialisierter GPU-Server gesamt	**460**
Anzahl GPUs gesamt	3.680
Anzahl Racks gesamt	115
Leistung pro GPU-Server inkl. Netzwerk in kW	8,5
Gesamt-IT-Leistung in MW	**3,91**

Anlage C: Kosten Infrastruktur

Betrachtung der Investitionskosten / CAPEX

Basierend auf der in Anlage B entwickelten Infrastruktur stellt LEAM folgende Kostenschätzung für Investitionskosten in die Infrastruktur auf.

Kostenpunkt	**Preis pro Einheit**	**Gesamtpreis**	**Begründung**
GPU-Server-Einheit	10.000.000 €	230.000.000 €	Angabe GPU-Hersteller
Gebäude & Infrastruktur Invest in €/kW-IT	12.500 € bis 20.000 €	48.875.000 € bis 78.200.000 €	Annahme auf Basis von Branchenkennwerten
Gesamtpreis		**278.875.000 € bis 308.200.000 €**	

Betrachtung der Betriebskosten / OPEX

Zur Betrachtung der jährlichen Betriebskosten der Infrastruktur wurden verschiedene Annahmen getroffen. Diese basieren größtenteils auf den Erfahrungswerten der LEAM Unterstützer. Verschiedene Annahmen sind aber nur schwer einschätzbar (bspw. Auslastung der Infrastruktur) oder können sich ohne einen Einfluss von LEAM ändern (bspw. Strompreis). Daher sollen die Zahlen nicht als belastbare Zahlen dienen, sondern lediglich eine erste Orientierung bieten.

Annahmen

Wartung	2,50%
Strompreis in €/kWh	0,19 bis 0,25
Auslastung (der Leistung)	80%
Auslastung (kaufmännisch)	80%
PUE des Rechenzentrums	1,2
Benötigte Betriebsmitarbeiter (Anzahl; u.a. zur Abdeckung von 24/7 Support in IT Operations & Remote Hands)	10 bis 20
Personalkosten Betriebsmitarbeiter (pro Jahr)	150.000 €
CPU-Cloud IaaS-Kosten (pro Monat)	75.000 € bis 125.000 €

Kostenschätzung

Kostenpunkt	Gesamtkosten	Berechnung
Jährliche Wartung	5.750.000 €	Gesamtkosten GPU Server * Wartung
Jährliche Stromkosten	6.300.000 € bis 8.200.000	Gesamtleistung * PUE * Strompreis * Auslastung (gerundet)
Jährliche Personalkosten (Betrieb)	1.500.000 € bis 3.000.000 €	Anzahl Mitarbeiter * Jahresgehalt
Jährliche Internetkosten	120.000 € bis 240.000 €	Annahme
Jährliche Lizenzkosten	500.000 € bis 750.000 €	Annahme
Jährliche CPU-Cloud IaaS-Kosten	900.000 € bis 1.500.000 €	Annahme
OPEX jährlich	**15.070.000 € bis 19.440.000 €**	

Fünfjahresplan

Kostenpunkt	Gesamtkosten	Berechnung
Gesamte Kosten	**354.225.000 € bis 405.400.000 €**	5 Jahre OPEX plus CAPEX
Kosten pro Betriebsstunde	**8.100 € bis 9.300 €**	Gesamtkosten / Stunden in 5 Jahren (gerundet)

Vorgehen zur Spezifikation der Infrastruktur

Der nächste Schritt zur Spezifikation der Infrastruktur für LEAM sollte der Einstieg in die Erarbeitung eines Lastenhefts sein. Folgende Fragestellungen müssen im nächsten Schritt geklärt werden, zum Teil auch als Schnittstellenthemen mit den anderen LEAM-Arbeitsgruppen:

- Erstellung eines Betriebskonzeptes, z. B.
    - Definition des Software-Stack (Konzept zur Funktionalität, Lifecycle-Management und Security-Anforderungen der Cloud-Softwarearchitektur)
    - Definition von Hardwarearchitektur und Netzwerkschnittstellen (Definition von erwarteten Compute Jobs, Bucket-Größen durch das Data-Science Team und darauf aufbauend Entwicklung eines Compute-, Netzwerk- und Storage-Konzeptes)

  - Zuordnung von Anwendungsfällen zu Infrastruktur-Einheiten (Trennung vorwettbewerbliche und wettbewerbliche Teile, Ausgestaltung der „Tiered Architecture“)

- Definition der funktionalen Anforderungen und Ableitung der Implikationen auf die Infrastruktur, z. B.
  - Verfügbarkeit / Redundanzen
  - Sicherheitslevel / Gebäudeausstattung
  - Latenzen / Breitband-Anbindung
  - Erarbeitung szenarioabhängiger Service Level Objectives und Service Level Agreements im Infrastrukturbereich
  - Support auf Infrastruktur-Anwendungsebene und Troubleshooting

- Definition der nicht-funktionalen Anforderungen und Ableitung der Implikationen auf die Infrastruktur, z. B.
  - Nachhaltigkeit / Kühlungskonzepte
  - Prüfung modularer Konzepte und Nutzung verteilter Infrastruktur
  - Standortanforderungen

Anlage D: Bilanz Core Model Development

Kosten für das Core Model Development entstehen vor allem aus Personalkosten. Alle anderen Kosten sind bereits über die Organisationseinheit Infrastruktur abgerechnet. Eine mögliche interne Verrechnung der Infrastrukturkosten werden zu diesem Zeitpunkt nicht beachtet.

Aktuell gehen die LEAM Unterstützer von drei oder vier entwickelten Modellen pro Jahr aus.

Annahmen

Personalkosten Betriebsmitarbeiter (pro Jahr)	150.000 €
Anzahl Mitarbeiter (pro entwickeltem Modell)	10
Anzahl entwickelter Modelle (pro Jahr)	3 bzw. 4

Kosten

Anzahl Modelle	jährliche Kosten	Berechnung
3	**4.500.000 €**	Anzahl Modelle * Anzahl Mitarbeiter * Personalkosten
4	**6.000.000 €**	Anzahl Modelle * Anzahl Mitarbeiter * Personalkosten

Einnahmen

Einnahmen generiert die Organisationseinheit Core Model Development über das Bereitstellen der Infrastruktur für private Organisationen. Hierfür werden aktuell rund 10% der Leistung dieser Organisationseinheit eingeplant. Die LEAM Unterstützer arbeiten aktuell an einem belastbareren Szenario für die Einnahmen. Bis dahin soll die Berechnung des Nutzungspreises anhand der Kosten der Infrastrukturnutzung erfolgen. Da der Service den europäischen Unternehmen möglichst kostengünstig angeboten werden soll, wird von einem Minimalpreis ausgegangen.

Annahme		Berechnung
Kosten Betriebsstunde	**10.000 €**	Kosten Betriebsstunde Infrastruktur + Kosten Betriebsstunde Mitarbeiter
Nutzung Core Model Development	**85%**	
davon Nutzung für private Organisationen	**10%**	
Einnahmen pro Betriebsstunde	**850 €**	Kosten Betriebsstunde * Nutzung des Cure Model Development * Nutzung für private Organisationen (gerundet)
Einnahmen pro Jahr	**7.500.000 €**	Preis pro Betriebsstunde * Stunden pro Jahr (gerundet)

Anlage E: Bilanz Feintuning

Kosten für das Finetuning entstehen vor allem aus Personalkosten. Alle anderen Kosten sind bereits über die Organisationseinheit Infrastruktur abgerechnet. Eine mögliche interne Verrechnung der Infrastrukturkosten werden zu diesem Zeitpunkt nicht beachtet.

Einnahmen generiert die Organisationseinheit Finetuning über die Vergabe von Lizenzen und dem Beratungsgeschäft. Die LEAM UnterstützerInnen arbeiten aktuell an einem belastbaren Szenario für die Einnahmen.

Hier sollen mögliche Szenarien aufgezeigt werden, die zum aktuellen Zeitpunkt lediglich als Orientierung dienen sollen.

Annahme

Personalkosten Betriebsmitarbeiter (pro Jahr)	150.000 €

Kosten

Anzahl Mitarbeiter je nach Auslastung	**jährliche Kosten**	**Berechnung**
5	**750.000 €**	Anzahl Mitarbeiter * Personalkosten
8	**1.200.000 €**	Anzahl Mitarbeiter * Personalkosten
10	**1.500.000 €**	Anzahl Mitarbeiter * Personalkosten

Einnahmen

Einnahmen generiert die Organisationseinheit Finetuning über das Bereitstellen der Infrastruktur für private Organisationen. Hierfür werden aktuell rund 60% der Leistung dieser Organisationseinheit eingeplant. Darüber hinaus sollen Lizenzen vergeben und ein Beratungsgeschäft etabliert werden.

Die LEAM Unterstützer arbeiten aktuell an einem belastbareren Szenario für die Einnahmen. Bis dahin soll die Berechnung des Nutzungspreises anhand der Kosten der Infrastrukturnutzung erfolgen. Da der Service den europäischen Unternehmen möglichst kostengünstig angeboten werden soll, wird von einem Minimalpreis ausgegangen.

Annahme		**Berechnung**
Kosten Betriebsstunde	**10.000 €**	Kosten Betriebsstunde Infrastruktur + Kosten Betriebsstunde Mitarbeiter
Nutzung Finetuning	**10%**	
davon Nutzung für private Organisationen	**60%**	
Einnahmen pro Betriebsstunde	**600 €**	Kosten Betriebsstunde * Nutzung des Finetuning * Nutzung für private Organisationen
Einnahmen pro Jahr	**5.300.000 €**	Preis pro Betriebsstunde * Stunden pro Jahr

Anlage F: Bilanz Inference

Kosten für die Inference entstehen vor allem aus Personalkosten. Alle anderen Kosten sind bereits über die Organisationseinheit Infrastruktur abgerechnet. Eine mögliche interne Verrechnung der Infrastrukturkosten werden zu diesem Zeitpunkt nicht beachtet.

Einnahmen generiert die Organisationseinheit Finetuning über das Bereitstellen der KI Modelle über API.

Hier sollen mögliche Szenarien aufgezeigt werden, die zum aktuellen Zeitpunkt lediglich als Orientierung dienen sollen.

Annahme

Personalkosten Betriebsmitarbeiter (pro Jahr)	150.000 €
Preis pro 1000 Tokens	0,01 €
Wachstum Token-pro-Tag / Jahr	6.000.000

Kosten

Anzahl Mitarbeiter je nach Auslastung	**jährliche Kosten**	**Berechnung**
5	**750.000 €**	Anzahl Mitarbeiter * Personalkosten
8	**1.200.000 €**	Anzahl Mitarbeiter * Personalkosten
10	**1.500.000 €**	Anzahl Mitarbeiter * Personalkosten

Einnahmen

GPT-3 generierte nach neun Monaten bereits 4,5 Milliarden Tokens pro Tag. Diese Anzahl soll hier als Basis genutzt werden.

Anzahl Token pro Tag	Einnahmen pro Tag	Einnahmen pro Jahr
3 Mrd.	30.000 €	**10.950.000 €**
4,5 Mrd.	45.000 €	**16.425.000 €**
6 Mrd.	60.000 €	**21.900.000 €**

Eine andere mögliche Berechnung basiert auf einem konstanten Wachstum an Tokens pro Tag.

Anzahl Tokens	Einnahmen pro Jahr
6.000.000 (nach einem Jahr)	**10.950.000 €**
12.000.000 (nach zwei Jahren)	**31.850.000 €**
18.000.000 (nach drei Jahren)	**54.750.000 €**
24.000.000 (nach vier Jahren)	**76.650.000 €**
30.000.000 (nach fünf Jahren)	**98.550.000 €**

Anlage G: 5-Jahres-Bilanz

Basierend auf diesen Annahmen kann eine 5-Jahres Bilanz erstellt werden. Da diese Zahlen lediglich auf vagen Annahmen beruhen, sollen sie lediglich als Orientierung dienen.

Organisations-bereich	**Einnahmen (nach fünf Jahren)**	**Kosten (nach fünf Jahren)**	**Gesamt**
Infrastruktur	0 €	354.225.000 € bis 405.400.000 €	- 354.225.000 € bis - 405.400.000 €
Core Model Development	37.500.000 €	22.500.000 € bis 30.000.000 €	7.500.000 € bis 15.000.000 €
Modeltuning	26.500.000 €	3.750.000 € bis 7.500.000 €	19.000.000 € bis 22.750.000 €
Inference	54.750.000 € bis 273.750.000 €	3.750.000 € bis 7.500.000 €	47.250.000 € bis 270.000.000 €
Gesamt	**118.750.000 € bis 337.750.000 €**	**384.225.000 € bis 450.400.000 €**	**- 331.650.000 € bis - 46.475.000 €**