BEITRAG ZU EINER KI-STRATEGIE FÜR DEUTSCHLAND

Version 1.4 vom 02. November 2023

Künstliche Intelligenz (KI), engl. Artificial Intelligence (AI), bündelt mehrere Teildisziplinen der Informatik, die Maschinen intelligenter machen sollen. Eine dieser Teildisziplinen befasst sich mit sogenannten Large-Language-Modellen (LLMs). LLMs verarbeiten natürliche Sprache. Andere Modelle können hingegen Bilder, Filme, Musik, Proteinfaltungen usw. verarbeiten. Neu ist, dass die LLMs zunehmend verlässlich in Domänen arbeiten, in denen Maschinen bisher nicht ausreichende Ergebnisse lieferten.

Diese Verlässlichkeit basiert darauf, dass die Modelle mit riesigen Datenmengen trainiert werden und auf Basis ihrer Programmierung selbstständig teilweise mehrere hundert Milliarden Parameter für ihren Einsatzzweck definieren. Im Falle von LLMs ist dieser Zweck bspw. das Zusammenfassen von Texten oder Beantworten von Fragen.[1] Ein prominentes Beispiel für eine LLM-Anwendung ist ChatGPT. Die Entwicklung von LLMs benötigt nicht nur deren Programmierung und eine umfangreiche Datenbasis, sondern auch spezialisierte Rechenkapazitäten, um sie zu trainieren. Im Jahr 2020 lagen die Kosten für das Training eines umfangreichen LLMs bei ca. 10 Mio. US-Dollar[2] – eine relativ hohe Markteintrittsbarriere für neue Marktteilnehmer.

Eine Aufholjagd Deutschlands im Bereich dieser generischen LLMs, bspw. wie von LEAM[3] angezielt, scheint dringend erforderlich. Etablierte Unternehmen, wie bspw. OpenAI, Anthropic, Microsoft, Google, aber auch Baidu und Huawei geben in diesem Bereich ein sehr hohes Entwicklungstempo vor[4] und investierten große Summen. Die Nachahmung dieser Vorgehensweise ergäbe kein direktes Differenzierungsmerkmal für deutsche Unternehmen. Auch wenn ‚digitale Souveränität' häufig als Feature genannt wird, so reichte dies in der Vergangenheit als Wettbewerbsmerkmal für die Durchsetzung europäischer Unternehmen nicht aus, z. B. zur Etablierung europäischer Hyperscaler. Daher muss die Antwort ein anderes Geschäfts- und Entwicklungsmodell sein, um Aussicht auf Erfolg zu haben.

Für die Wettbewerbsfähigkeit von LLMs sind mehrere Aspekte ausschlaggebend: Einerseits definieren die Menge und der Gehalt der Daten, mit denen die Modelle trainiert werden, die Wertigkeit der Modelle. Die deutsche Industrie verfügt in verschiedenen Bereichen über hochwertige Spezialdaten, bspw. aus Herstellungs- und Automatisierungsprozessen, die zum Training eigener LLMs verwandt werden können. Darüber hinaus besteht in Deutschland (und der EU) noch großes Potenzial hinsichtlich der Verfügbarmachung öffentlicher Daten für Forschung und Entwicklung. Andererseits ist ausreichend zugängliche Rechenkapazität für das Training der Modelle essentiell. Hier besteht strategischer Investitionsbedarf, denn die verfügbaren Rechenkapazität ist in Deutschland beschränkt. Die Kapazitäten können auch nicht ohne weiteres eingekauft werden, denn die dafür unter anderem erforderlichen Grafikprozessoren von NVIDIA müssen aktuell vornehmlich aus den USA beschafft werden. Außerdem werden zukünftig die Nachvollziehbarkeit, Transparenz und

[1] https://www.cloudcomputing-insider.de/was-ist-ein-large-language-model-llm-a-9b7bdd0c3766b5a9c0ee1e0c909790a3/
[2] Sharir, Peleg und Shoham (2020): The Cost of Training NLP Models, URL: https://arxiv.org/pdf/2004.08900.pdf
[3] Die Initiative Large European AI Models (LEAM) fordert, dass ein Hochleistungsrechenzentrum aufgebaut wird, speziell für die Entwicklung von Anwendungen und die Erforschung dieser im Bereich der Künstlichen Intelligenz (s. https://ki-verband.de/leam-grosse-ki-modelle-in-europa/).
[4] Ratnaparkhi (30. Mai 2023): Evolution of NLP. URL: https://medium.com/@Ratnaparkhi/evolution-of-nlp-unleashing-the-potential-of-large-language-models-and-prompts-fee7ba02f72b

Vertrauenswürdigkeit der Modelle eine bedeutende Rolle einnehmen. Eine Chance ergibt sich daraus, dass diese Kriterien aktuell oft im Widerspruch zu den Geschäftsmodellen der etablierten KI-Unternehmen stehen.

Deutschland sollte sich in der Entwicklung von KI auf seine Kompetenzen konzentrieren und die Bedürfnisse, aber auch die Assets der hiesigen technologisch führenden Industrien, der Gesellschaft und Wissenschaft berücksichtigen. Nur dann und nur mit zügigem Handeln können trotz des derzeitigen Rückstandes bald einige entscheidende Wettbewerbsvorteile entstehen.

Für die Erreichung dieses Ziels schlägt SPRIND vier parallele Handlungsstränge vor:

1. **SPRIND Challenges zur Entwicklung anwendungsspezifischer KI** für die konkrete Nutzung und Weiterentwicklung der KI in den Verticals,
2. **SPRIND Challenge zur Entwicklung von Datenpools** zur Bereitstellung und Kuration von hochwertigen, einzigartigen Datenpools aus der Wirtschaft, Verwaltung, Forschung und Gesellschaft,
3. **Förderung von Open-Source-LLMs**, analog zum Modell des Sovereign Tech Funds, für einen Aufbau hiesiger, zukünftiger Champions,
4. **Bereitstellung von Rechenkapazität** mit paralleler **Entwicklung von spezieller Hardware.**

Dieses Dokument beschreibt eine missionsorientierte Vorgehensweise, die schnell in Aktion gebracht werden kann. SPRIND kann als Instrument für die Durchführung von Challenges und zur Projektfinanzierung genutzt werden. SPRIND kann per In-Haus-Vergabe beauftragt werden.

1. SPRIND Challenges zur Entwicklung anwendungsspezifischer KI

SPRIND Challenges haben sich als schlagkräftiges Instrument zur Entwicklung von Innovationen in komplexen Themenfeldern bewährt. Fünf SPRIND Challenges sind derzeit aktiv.[5] Challenges adressieren Akteure aus Wissenschaft, Wirtschaft und Zivilgesellschaft, leuchten somit ein Feld umfassend aus. Das Verfahren ist schnell, denn es erfordert nur wenige Wochen zwischen Einreichungsfrist-Ende bis zur Mittelvergabe. Es ist effizient, denn es nutzt die vorkommerzielle Auftragsvergabe, ohne Zuwendungs- oder Projektmittel-Logik mit aufwändigem Controlling und ohne kleinere Akteuren a-priori auszuschließen. Innerhalb der SPRIND Challenges entwickelt sich eine große Dynamik durch das mehrstufige Wettbewerbsformat.

SPRIND schlägt daher eine Serie von Challenges vor, die mit dem Ziel der Anwendung von KI im Feld vielversprechende Kompetenzen der deutschen Wissenschaft, Wirtschaft und Zivilgesellschaft adressieren und die Entwicklungen im Wettbewerbsformat vorantreiben. Diese Liste ist jederzeit erweiterbar und kann an den Interessenfokus der Partner angepasst werden, also bspw.:

- **AI Engineer** (z. B. Entwicklungsunterstützung im Maschinenbau, Robotik),
- **AI Scientist** (z. B. Generierung von Daten, Ableitung und Test von Hypothesen in der Pharmaforschung, Molekularbiologie oder der Materialforschung),
- **AI Lawyer** (z. B. Optimierung von Gesetzgebung, Antragsstellung und Bearbeitung, Vertragswesen und Steuerwesen),
- **AI Doctor** (z. B. Unterstützung in der Anamnese, Pathogenese, Salutogenese),

[5] https://www.sprind.org/de/challenges/

- AI Public Officer (Optimierung und Begleitung von Arbeitsprozessen der Verwaltung),
- AI Reflector (Überprüfung und Test von LLMs oder der Authentizität und Identität von Dokumenten),
- AI Developer (z. B. Automatisierte Softwareentwicklung, Qualitätssicherung).
- AI Teacher (Bildung neu denken im AI-Zeitalter)

2. SPRIND Challenge zur Entwicklung von Datenpools

Daten bestimmen die Qualität der Modelle und somit ihre Leistungsfähigkeit in den entsprechenden Anwendungsfällen. Trotz einiger Initiativen des Bundes ist es bisher nicht gelungen, relevante Daten in ausreichend großer Menge bereitzustellen. Die wesentlichen Gründe dafür sind eine geringe Bereitschaft der Akteure zum Datenaustausch und auch eine Unsicherheit in Bezug auf die Datenschutzgrundverordnung. Darüber hinaus fehlt es an einer gezielten Förderung datenorientierter Projekte und dem damit verbundenen Erwerb von Kompetenzen im Umgang mit den Daten.

Deutsche Unternehmen verfügen in einer Vielzahl ihrer Systeme über große Mengen und verschiedene Datenklassen. Die allermeisten der Unternehmen sind nicht dazu in der Lage, ihre eigenen Daten aufzubereiten oder gar zusammenführen. Aus diesem Grund wissen die Unternehmen auch häufig wenig über das Potenzial, das in ihren Daten ruht. Eine weitere SPRIND Challenge könnte auf die Befähigung der Unternehmen abzielen, genau dieses Potenzial zu erkennen und es zu heben.

Eine SPRIND Challenge könnte die Entwicklung von Datenpools in Deutschland beschleunigen, indem vorhandene Daten aktiviert und unterschiedliche Quellen zusammengeführt werden.[6] Um eine möglichst große Innovationkraft zu entwickeln, würde die Challenge bestimmte Rahmenbedingungen definieren, wie bspw., dass die Daten trotz gewisser Exklusivität offen erreichbar sein müssen und dass ein Geschäftsmodell für die spätere Monetarisierung mitentwickelt werden muss.

Etablierte Unternehmen, wie bspw. Google, sind bereits sehr weit mit der Entwicklung ihrer horizontalen Datenpools. In Anbetracht der tiefgehenden Daten deutscher Industrieunternehmen erscheint ein Nachahmen aber ohnehin nicht sinnvoll. Ein Aufbau vertikaler Datenpools über mehrere Stufen einer Wertschöpfungskette ist bei einer hohen Kooperationsbereitschaft schneller möglich und eröffnet zudem ein deutlich größeres Innovationspotenzial als bei horizontalen Datenpools. Das Ziel der Challenge könnte daher eine Art vertikaler Crawler zur Erstellung von Datenpools sein, wobei zunächst nicht der fertige Datenpool, sondern lediglich die Entwicklung eines Crawler-Demonstrators als Instrument im Fokus steht. Dieser Demonstrator könnte dann in einer zweiten Stufe der Challenge einem umfangreicheren Praxistest unterzogen werden. Dabei können sowohl die Entwickler eines Crawlers als auch+ die Datenlieferanten an der Challenge teilnehmen.

3. Förderung von Open-Source-LLMs

Nachdem die bis dahin offenen Modelle von Unternehmen wie OpenAI sichtbar erfolgreich wurden, begann eine rasante Kommerzialisierung. Diese fand ihren bisherigen Höhepunkt in einer Investition von Microsoft in Höhe von 10 Mrd. US-Dollar in die vormals noch als

[6] Es können auch bereits begonnene Projekte beschleunigt werden (z. B. „Idealist“ vom Helmholtz Zentrum München).

Non-Profit-Organisation geführte Open-Source-Firma OpenAI.[7] Diese Investition bedeutete das Ende der Open-Source-Ambitionen dieser Firma.[8] Daraufhin entwickelte sich die KI-Open-Source-Gemeinschaft rasant. Sie begann eigene Modelle zu entwickeln, um diese Machtzentralisierung zu stoppen und die Nutzung von LLMs für die Allgemeinheit zu ermöglichen.[9] Im Ergebnis nahm die Zahl von Open-Source-LLMs zuletzt stark zu.[10] Diese Zunahme lässt vermuten, dass es einfacher für neue Marktteilnehmer wird, Dienstleistungen auf Basis von Open-Source-LLMs anzubieten, weil eine große Hürde entfällt.[11] Ein Markteintritt mit Open-Source-LLMs erfordert keine Investitionen in die Entwicklung und das Training gänzlich neuer, eigener LLMs. Stattdessen können Open-Source-LLMs als Basis verwendet und bspw. für konkrete Einsatzbereiche angepasst werden. Diese Feinabstimmung ist günstig und erlaubt eine wesentlich schnellere Optimierung der LLMs im Vergleich zu einem LLM, das zentral und proprietär entwickelt wird.[12] Nicht zuletzt lassen sich diese Open-Source-LLMs auch auf privaten Rechnern ausführen und können so in den unterschiedlichsten Szenarien souverän eingesetzt werden. Im Vergleich zu Cloud-Lösungen sind sie dadurch relativ geschützt und sicher vor dem ungewollten Abfluss von Daten.

SPRIND schlägt vor, Open-Source-LLMs und andere Modellarten zu unterstützen. Eine solche Förderung könnte z. B. analog zum Sovereign Tech Fund[13] erfolgen. Zusätzlich sollten gezielt hiesige Firmen gefördert werden, die die Kommerzialisierung der Modelle betreiben, bspw. im sehr erfolgreichen Enterprise-Open-Source-Modell, wie es u. a. bei den Firmen SUSE Linux und RedHat seit Jahrzehnten zur Anwendung kommt. Ebenso können die durch Gaia-X angestoßenen Projekte, wie Catana-X und demnächst Manufacturing-X, sowie die weiteren durch das BMWK angeschobenen Leuchtturmprojekte, einen deutlichen KI-Schub erfahren (z. B. im Health-Bereich TEAM-X und Health-X dataLOFT).

4. Bereitstellung von KI-spezifischer Hardware

Die aktuell für KI-Aktivitäten bereitstehende Rechenleistung wird als zu gering eingeschätzt und vorhandene, leistungsstarke Systeme werden überwiegend von Forschungsinstitutionen in Anspruch genommen.[14] SPRIND sieht daher kurz- und mittelfristig drei besonders wichtige Notwendigkeiten:

Erstens sollte Mittelstand, Industrie und digitale Gesellschaft zur Nutzung von KI ermächtigt werden, indem spezielle KI-Hardware für sie entwickelt wird. Wichtig ist dabei die generelle Verfügbarmachung von Hardware, die Förderung des Wissens um die Notwendigkeit der Nutzung von KI bis zur Bereitstellung personeller und finanzieller Ressourcen sowie der Aufbau von Expertise und eines Sicherheitsbewusstseins der Anwender:innen.

Zweitens sollte sog. Middleware entwickelt werden, die entsprechende Software-Schnittstellen und -Dienste für komplexe Anwendungen bereitstellt.

Drittens sollten fortgeschrittene, erfolgsversprechende und schnell verfügbare Hardware-Lösungen für Hochleistungsrechnen unterstützt werden.

---

[7] https://www.golem.de/sonstiges/zustimmung/auswahl.html?from=https%3A%2F%2Fwww.golem.de%2Fnews%2Fopenai-microsoft-investiert-in-chat-gpt-unternehmen-2301-171383.html&referer=https%3A%2F%2Fwww.google.com%2F
[8] https://de.wikipedia.org/wiki/OpenAI
[9] Eine Liste aktueller Open-LLMs befindet sich auf https://github.com/eugeneyan/open-llms.
[10] https://medium.datadriveninvestor.com/list-of-open-source-large-language-models-llms-4eac551bda2e
[11] https://cobusgreyling.medium.com/large-language-models-are-being-open-sourced-537dcd9c2714
[12] https://bdtechtalks.com/2023/05/08/open-source-llms-moats/
[13] https://sovereigntechfund.de/de/
[14] https://background.tagesspiegel.de/digitalisierung/wo-stecken-die-ki-milliarden

Diesen Notwendigkeiten kann aus Sicht der SPRIND wie folgt konkret begegnet werden:

a) Ausbau der bestehenden Supercomputer-Zentren mit KI-spezifischer, ‚klassischer' Hardware, ggf. in Zusammenarbeit mit der Europäischen Union.
b) Unterstützung jener Hardware-Entwicklungsprojekte, die absehbar den größten Erfolg versprechen und deren Ergebnisse schnell für eine Anwendung verfügbar sein werden. Das Umfeld der SPRIND bietet bereits einige Ansätze, die aus der Arbeit der letzten drei Jahre, sowie aus der SPRIND Challenge „New Computing Concepts"[15] entstanden sind.
c) Entwicklung einer Softwarelösung, bestehend aus allen notwendigen Software-Komponenten (sog. Full-Stack), um die unterliegende Hardware leichter nutzbar zu machen.
d) KI-unterstützte Chip- und Softwareentwicklung, um bspw. FPGA-Designs für den Einsatz in KI-Anwendungen zu optimieren. Hier sollten Referenzfälle geschaffen werden, mit denen die Vorteile der KI-gestützten Entwicklung nachgewiesen werden können.

Auf mittlere Sicht sollte Deutschland eigene KI-spezifische Hardware entwickeln, sowohl für bisher unbesetzte Nischenanwendungen als auch für solche, die einen Effizienzvorteil um mindestens Faktor 10 gegenüber dem aktuellen Stand der Technik bieten. Parallel sollten Algorithmen und Software entwickelt werden, die auf die neue Hardware abgestimmt sind, um so den größtmöglichen Nutzen zu erreichen. SPRIND kann hierzu aus der aktuellen Arbeit zwei Beispiele beitragen:

- **Memristoren und Kondensatoren** zählen zu den wichtigsten Bauelementen der zukünftigen Hardware für KI.[16] Daher sollte ein Ökosystem für Entwicklung von Memristoren aufgebaut und entwickelt werden. Die SPRIND-Tochter Memlog GmbH kann dabei unterstützen. Auch können die Teams der SPRIND Challenge „New Computing Concepts" bei der Initiierung eines Ökosystems helfen. Dies könnte bspw. durch die bereits in der Entwicklung befindlichen speziellen Compiler und auf Memristoren abgestimmten arithmetisch-logischen Rechenwerke (Arithmetic Logic Units) für Prozessoren erfolgen.
- Beschleunigung der Projekte aus der SPRIND Challenge „New Computing Concepts" zur **Erstellung alternativer Hardware** zur KI-Beschleunigung und Effizienzsteigerung mit dem Ziel, jeweils mindestens Faktor 10 zu erreichen.

5. Zusammenfassung

Die 4 empfohlenen Handlungsstränge

1. Challenges zur Entwicklung anwendungsspezifischer KI,
2. Challenge zur Entwicklung von Datenpools,
3. Förderung von Open-Source-LLMs und
4. Bereitstellung von Rechenkapazität mit paralleler Entwicklung von spezieller Hardware

können und sollten parallel aufgesetzt werden, um dem rasanten Entwicklungstempo dieser Technologien gerecht zu werden.

[15] https://www.sprind.org/de/challenges/newcomputing
[16] https://www.enas.fraunhofer.de/de/geschaeftsfelder/micro_and_nanoelectronics/Beyond-CMOS-und_HF-Bauelemente/Memristoren_fuer_die_Rechner_von_morgen.html