KI-Infrastruktur für das Training großer Modelle in Deutschland

Eine Einordnung

KI BUNDESVERBAND

Executive Summary

1. Die Wirtschaft und Verwaltung in Deutschland braucht eigenständige, souveräne, mit individuellen Daten trainierte große KI-Modelle als Basis für künftige Innovation, Effizienzsteigerung und Wertschöpfung
2. Basis für solche großen Industrie-spezifische Modelle sind große wettbewerbsfähige Foundation-Modelle (Sprache, Multimodal)
3. Industriemodelle, die nicht auf LLM oder Multimodalen Modellen beruhen sind in der Regel in Umfang und Komplexität um Größenordnungen kleiner und benötigen nicht die Rechenkapazitäten großer Supercomputer-Cluster
4. Große Rechenzentren ermöglichen das Training von großen Modellen. Der Betrieb (Inferenz) kann auch verteilt in kleineren Rechenzentren und „at-the-Edge“ unterstützt werden.

Daraus folgt:

Die Planung für eine große KI-Supercomputer-Infrastruktur muss eingebettet sein in eine Gesamtstrategie, die auch die Erstellung von wettbewerbsfähigen europäischen Foundation-Modellen beinhaltet.

Um Chancen zu nutzen und Risiken zu begegnen, müssen wir unser Innovationsverständnis in Deutschland und Europa ändern. Das alte Modell – Grundlagenforschung in einen Silo, Entwicklung in einem anderen Silo, spätere Anwendung durch die Industrie – passt nicht mehr. Es ist zu langsam, zu fragmentiert und zu weit entfernt von den tatsächlichen Bedürfnissen.

Stattdessen brauchen wir integrierte Ökosysteme, in denen Forschung, Entwicklung und Anwendung Hand in Hand stattfinden, in einem Haus, zumindest in einem Dorf. KI-Forschende müssen nah dran sein an den realen Problemen. Und diejenigen, die KI-Lösungen entwickeln, brauchen frühzeitigen Zugang zu den Erkenntnissen aus der Forschung – um Systeme zu bauen, die nicht nur funktionieren, sondern auch ethisch, sicher und verlässlich sind. Und alle brauchen Zugang zu Hardware.

1) Einleitung

Die Bedeutung von KI für die wirtschaftliche und gesellschaftliche Entwicklung erhält breite Anerkennung in allen Bereichen. Durch den aktuellen „Hype“ um ChatGPT und andere Sprachmodelle sind vor allem diese Large Language Modelle (LLMs) in den Mittelpunkt der öffentlichen Wahrnehmung gerückt, u.a. unter folgenden Aspekten

- **LLMs benötigen sehr viele Daten** (und entsprechende Data-Pipelines) für das Training, z.B. benötigt ein 7B-Modell (Llama 3) 15 Billionen Token, das sind ca. 60 Terabyte (TB). Zuvor müssen jedoch deutlich mehr TB an Daten (Faktor 5 bis 10) beschafft werden, da nur wenige hochqualitative Daten für das Training eingesetzt werden. Der Großteil der Daten wird aus Mangel an Datenqualität nicht benutzt.[1]
- **Es sind enorme Rechenkapazitäten erforderlich, um LLMs zu trainieren**. Mehrere Millionen GPU-Stunden. Hierfür sind spezielle Rechenzentren notwendig, die mindestens 5000 GPUs lokal in einem Cluster vereinen. Z.B. wurde für das 8B LLAMA 3 eine Rechenzeit von 1,3 Mio GPU Stunden (400TFlops) auf einem H100 GPU Cluster benötigt. Auf einem 5000 GPU-Cluster sind das für einen Durchlauf ca. 10 Tage.[2] Das 70B LLAMA 3 Modell hat ca. 7,7 Mio GPU Stunden (65 Tage) benötigt.[3] In der Praxis kommen Pufferzeiten, Checkpoints, Daten-Streaming und andere Faktoren hinzu, diese Zahl zeigt das **optimale untere Limit**.
- Die **Investitionen** für den Aufbau von Infrastrukturen für das LLM-Training und die Beschaffung von Daten erfolgten hauptsächlich durch die Big-Tech Companies in den USA sowie in China.
- Die europäische Infrastruktur-Landschaft ist aktuell **nicht vergleichbar**. Große kommerzielle Rechenzentren mit mehreren tausend GPUs sind vorhanden und mit Jupiter (20.000 GPUs) wurde am Forschungszentrum Jülich der erste Exascale Rechner in Europa für die Forschung in Betrieb genommen. In Summe bleibt das aber Größenordnungen unter den US-Kapazitäten.
- **Europäische Unternehmen nutzen LLMs intensiv**, hauptsächlich auf Basis der Angebote von OpenAI/Microsoft, Google, Anthropic etc. Mistral aus Frankreich ist das einzige vergleichbare kommerzielle Angebot aus Europa. Für komplexe Aufgaben stehen hier nur die großen LLMs (>70 B) im Fokus.

[1] https://ai.meta.com/blog/meta-llama-3/
[2] https://github.com/meta-llama/llama3/issues/91
[3] https://huggingface.co/meta-llama/Meta-Llama-3-70B

- In Deutschland wurden mit Teuken/OpenGPTx, LLäMmlein, OpenEuroLLM und Occiglot bereits LLMs entwickelt, die aber aufgrund des Mangels an verfügbarer Rechenkapazität nicht in vergleichbare Leistungskategorien skaliert werden konnten.

Vor diesem Hintergrund und den aktuell, auf unterschiedlichen Ebenen stattfindenden Diskussionen möchten wir (die Unterzeichner) einige Sachverhalte und deren Einschätzungen im Folgenden transparent machen:

2) KI-Modelle und Rechenbedarf

a) Foundation Modelle

Ein Foundation Model („Grundlagenmodell“) ist ein großes, vortrainiertes KI-Modell, das auf umfangreichen, breit gefächerten Datensätzen trainiert wurde und sich anschließend für viele unterschiedliche Aufgaben feinjustieren lässt (z. B. durch Fine-Tuning).

Merkmale eines Foundation Models:

- Große Datenbasis: Training auf enormen Mengen unterschiedlicher Daten (Text, Bilder, Code, Audio, Video etc.).
- Selbstüberwachtes Lernen: Häufig ohne manuelle Labels trainiert (z. B. durch Vorhersage von fehlenden Wörtern).
- Skalierbarkeit: Kann auf leistungsstarken Infrastrukturen (z. B. mit Tausenden GPUs) trainiert werden.
- Multifunktionalität: Ein Modell kann viele Aufgaben übernehmen (z. B. Text generieren, übersetzen, zusammenfassen, Fragen beantworten).
- Anpassbarkeit: Nach dem initialen Vortraining kann es für spezifische Anwendungen angepasst werden (z. B. medizinische Texte, juristische Analysen).

Beispiele:

- LLMs (Large Language Models) und Multimodale Modelle: GPT-4, Claude, LLaMA, Gemini (Text + Bild + Audio), CLIP (Text-Bild-Verknüpfung)
- Bildgeneratoren: DALL·E, Stable Diffusion, Flux
- State-Space Models: Mamba, Hyeana, xLSTMs

Vor allem die LLMs und multimodale Modelle benötigen für das Training umfangreiche Datensätze und GPU-Rechenkapazitäten (in der Größenordnung von Millionen GPU Stunden).

Bildgeneratoren und weitere Modelle, die auf industriellen Daten (nicht sprach-basiert) trainiert wurden, sind in ihrem Rechenkapazitätsbedarf oft eine Größenordnung niedriger einzustufen (einige 100k GPU-Stunden).

Für das Training von wettbewerbsfähigen Modellen werden zudem hochqualitative Datensätze benötigt. Diese stehen zum Teil als Open Source zur Verfügung (z. B. (Occiglot-)Fineweb, Community OSCAR, LLäMmlein deutsche Daten oder Common Crawl). Dennoch ist bekannt, dass große Firmen darüber hinaus Daten für das Training bei kommerziellen Anbietern wie Scale AI lizensieren.

b) Rechenbedarf für Foundation-Modelle

Die Entwicklung eines offenen, leistungsfähigen deutschen Foundation-Modells mit 70 bis 100 Milliarden Parametern stellt eine zentrale strategische Investition dar, um die digitale Souveränität Europas im KI-Bereich zu sichern. Vergleichbare Modelle wie **LLaMA 3 70B** wurden mit rund **15 Billionen Token** (ca. **60 TB bereinigter Textdaten**) trainiert, wofür eine Ausgangsbasis von mehreren **hundert Terabyte Rohdaten** benötigt wird. Alleine für einen Trainings-Lauf beläuft sich der Rechenaufwand auf etwa **7,7 Millionen GPU-Stunden**, was rund **65 Tagen auf einem Cluster mit 5.000 H100-GPUs** entspricht. Dazu kommt noch die Herstellung von State-of-the-Art-Datensätzen für das Training des Modells, was ungefähr 2 Millionen GPU-Stunden entspricht.

Ein deutsches Basismodell dieser Größenordnung könnte anschließend in der Industrie **wiederverwendet und bei Bedarf effizient angepasst** werden. Industriespezifische KI Modelle basieren in der Regel auf bestehenden großen Foundation-Modellen wie LLaMA 3 oder DeepSeek und werden durch Methoden wie zum Beispiel **Finetuning, Adapter Lösungen** oder **Reinforcement Learning** an unternehmensspezifische Anforderungen angepasst. Diese Anpassungen sind technisch effizient möglich: Selbst bei großen Basismodellen wie LLaMA 3 70B kann ein Feintuning mit 100.000 Beispielen in wenigen Stunden auf einer einzelnen H100-GPU durchgeführt werden.

Large Reasoning Modelle (LRM). Die Anpassung eines vortrainierten Modells mit Industriespezifischen Daten durch Reasoning ist durch DeepSeek R1 oder O4 Mini Anfang 2025 bekannt geworden. Typischerweise werden mehrere Reasoning-Schritte durchgeführt. Die Anpassung eines R1 Modells für die Lösung von Matheaufgaben der Olympiade erfolgte in 27 h auf 8× A100.[4] Auf 4 A100 wurden nur 13 Stunden auf 7000 Trainingsdaten für einen ähnlichen Task veranschlagt.[5] Generell sind LRMs bei hochqualitativen Trainingsdaten und gut vorinitialisierten Basismodellen ähnlich kostenoptimal wie Finetuning -Ansätze für die Industrie. Dabei können diese Ansätze jedoch Modelle auf die Abarbeitung komplexerer Aufgaben, die Chain Of Thought Reasoning erfordern, vorbereiten.

c) Industriespezifische Modelle

In der Regel sind mehrere Aspekte von industriespezifischen Modellen zu unterscheiden

Modelle auf Basis von existierenden Foundation Modellen (LLM oder Multimodal). Hierfür werden diese Foundation-Modelle mit Industrie-, Unternehmens- oder Anwendungsfall-spezifischen Daten im Finetuning Verfahren „weitertrainiert". Die so entstandenen Modelle können dann auch durch unterschiedliche Verfahren auf Performance, Energiebedarf, Größe (Edge) weiter optimiert werden.

Anpassung mit Finetuning/Adapter/LoRA. LoRA-Finetuning auf großen LLMs (Text oder Multimodal) benötigt meist nur **einzelne GPU-Stunden** oder **wenige Stunden**, selbst bei großen Basismodellen. Das heißt, die Industrie kann Basismodelle wie LLAMA 7B[6] auf 100.000 Beispielen in wenigen Stunden auf einer einzelnen H100 finetunen. Die Voraussetzung dafür ist, dass diese Trainingsdaten und Zugriff auf eine GPU vorhanden sind.

Modelle, die komplett auf eigenen, spezifischen Daten trainiert werden. Hierfür gibt es viele Anwendungsfälle, z.B. optische Qualitätssicherung, Demand-Forecasting etc. Diesen ist in der Regel gemeinsam, dass die erforderliche Rechenleistung deutlich niedriger ist. Z.B. wurden für das Training von Yolo (Basis-Modell für Bilderkennung) lediglich 1000-2000 GPU genutzt.

[4] https://www.reddit.com/r/machinelearningnews/comments/1jhrlm6/sea_ai_lab_researchers_introduce_dr_grpo_a/

[5] https://arxiv.org/pdf/2505.09655

[6] https://medium.com/%40govindarajpriyanthan/parameter-elicient-fine-tuning-of-llama-3-1-a-comprehensive-guide-bed38d232285#

Modelle in der Robotik. Für die zum Training von Robotik-Modellen erforderlichen Rechenkapazitäten gibt es kaum veröffentlichte Informationen. Auf Basis von existierenden Publikationen und für ein realistisches Trainingsszenario auf einem **512-GPU-Cluster mit H100s** lässt sich z.B. der Rechenaufwand für Boston Dynamics' Modelle wie folgt abschätzen: Das Training von **Spot**, das auf etwa **1 Million Simulations-Episoden**[7] basiert, würde auf einem solchen Cluster etwa **1–3 Stunden** reine Simulationszeit beanspruchen, wobei unter realen Bedingungen (inkl. Policy-Updates, Logging, Checkpoints etc.) mit **12–24 Stunden** Trainingszeit zu rechnen ist. Für komplexere Modelle wie **Atlas**, bei denen laut Boston Dynamics bis zu **150 Millionen Simulationen** pro Aufgabe verwendet werden, ergibt sich auf demselben Cluster eine theoretische Simulationszeit von etwa **20 Stunden**, realistisch aber eher **2–4 Tage** inklusive Overhead. Diese Werte basieren auf typischen Durchsatzraten moderner Physiksimulatoren (z. B. Isaac Gym oder MuJoCo) bei H100-Leistung.

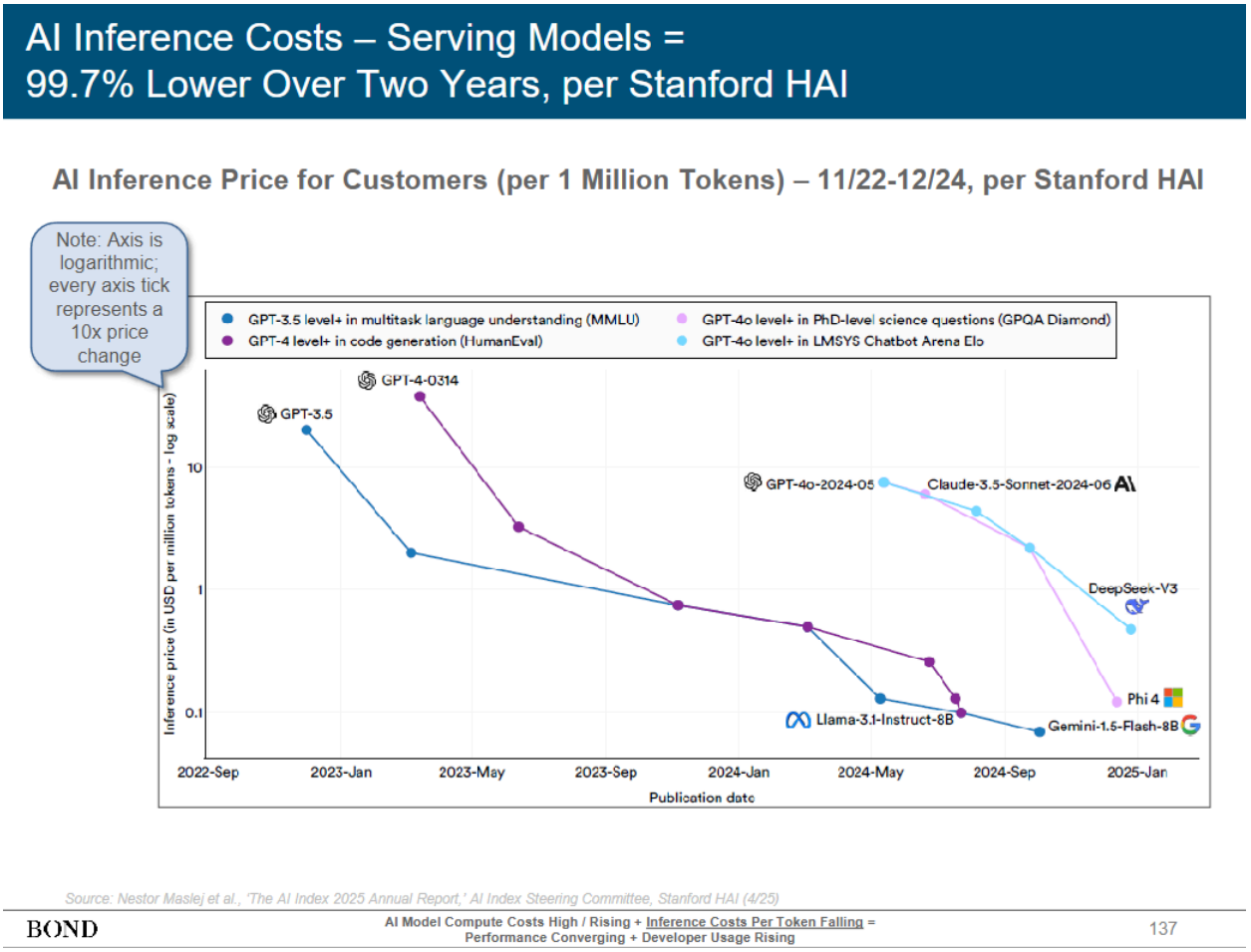
Der Bedarf in der Robotik ist daher sehr wahrscheinlich um eine bzw. zwei Größenordnungen kleiner als bei einem LLM.

3) KI-Rechenzentren

Wichtig bei der Diskussion um KI-Rechenzentren ist die Unterscheidung zwischen Model Training und Betrieb der Modelle (Inferenz).

Das Training von wettbewerbsfähigen Foundation-Modellen erfordert mehrere Millionen GPU Stunden, die auf einer großen Zahl von GPUs (> 10.000) parallel abgearbeitet werden und für den Verlauf des Trainings kontinuierlich verfügbar sein müssen. Hierfür sind speziell auf das Training von großen Modellen ausgelegte Compute-Cluster notwendig, die neben einer großen Anzahl GPUs mit großem Speicher vor allem eine schnelle Datenkommunikation zwischen den einzelnen GPUs benötigen.

[7] https://bostondynamics.com/blog/starting-on-the-right-foot-with-reinforcement-learning/
https://www.ctco.blog/posts/atlas-robot-rl-simulation-demo-reflection/



Der Betrieb eines Foundation Modells erfordert verglichen mit dem Training in der Regel oft weniger Ressourcen.Der Bedarf für die Inferenz steigt zwar durch z.B. den Einsatz von Reasoning-Modellen, jedoch können die erforderlichen GPU-Kapazitäten verteilt auf unterschiedliche Rechenzentren, nah am Nutzer oder „At-the-Edge“ installiert werden – sukzessive, dem steigenden Bedarf folgend. Ein zentrales Hochleistungsrechenzentrum inklusive der damit verbundenen hohen Investitionen ist hierfür nicht zwingend erforderlich.

4) Fazit und Empfehlung

Für den Aufbau von digitaler Souveränität in KI sind GPU-Rechenzentren ein wesentlicher Bestandteil, der sinnvoll in eine Gesamtplanung eingebettet sein muss.

Daneben benötigen wir u.a. qualitativ hochwertige Daten nach deutschen/europäischen Standards, eigene Open-Source Foundation Modelle (LLM, Multimodal etc.) sowie ein Ökosystem aus Frameworks, Plattformen und Komponenten zur automatisierten Erstellung von Anwendungen und KI-Agenten-Systemen.

In einem schlanken und fokussierten Aktionsplan sollten alle Aktivitäten zur Erstellung eines souveränen KI-Ökosystems koordiniert und mit den Beteiligten aus Wirtschaft, Wissenschaft, Verwaltung und Gesellschaft möglichst breit abgestimmt werden.

Über die Autoren:

Jörg Bienert
KI Bundesverband

Dr. Simon Ostermann
DFKI

Prof. Wolfgang Neidl
L3S Hannover

Dr. Mehdi Ali
Fraunhofer IAIS

Dr. Daniel Porta
DFKI

Prof. Alexander Löser
Berliner Hochschule für Technik

Prof. Antonio Krüger
DFKI

Dr. Nicolas Flores-Herr
Fraunhofer IAIS

Prof. Kristian Kersting
TU Darmstadt

Dr. Stefan Dietzel
Merantix Momentum

Prof. Josef van Genabith
DFKI

Prof. Andreas Hotho
Universität Würzburg

Daniel Abbou
KI Bundesverband

Die Autoren:

Jörg Bienert, KI Bundesverband
Dr. Nicolas Flores-Herr, Fraunhofer IAIS
Dr. Simon Ostermann, DFKI
Prof. Kristian Kersting, TU Darmstadt
Prof. Wolfgang Neidl, L3S Hannover
Dr. Stefan Dietzel, Merantix Momentum
Dr. Mehdi Ali, Fraunhofer IAIS
Prof. Josef van Genabith, DFKI
Dr. Daniel Porta, DFKI
Prof. Andreas Hotho, Universität Würzburg
Prof. Alexander Löser, Berliner Hochschule für Technik
Daniel Abbou, KI Bundesverband
Prof. Antonio Krüger, DFKI