



# KI-Rechenzentren in Deutschland

Aktuelle Kapazität, künftiger Bedarf

Dr. Philip Fox (KIRA), Prof. Monika Schnitzer (LMU), Daniel Privitera (KIRA)

# Executive Summary

## Zentrale Befunde:

- Deutschlands Anteil an der globalen KI-Rechenkapazität droht in den nächsten Jahren **dramatisch zu sinken**.
- Eine mögliche KI-Gigafabrik – aktuell das größte in Deutschland vorgesehene Rechenzentrum – wäre mindestens **50x kleiner** als die weltweit größten angekündigten Cluster.
- Insgesamt würden die aktuell in Deutschland angekündigten Rechenzentren **bei weitem nicht ausreichen**, um die prognostizierte Bedarfslücke zu schließen.
- Ein ambitioniertes Umsteuern ist möglich – aber nur, wenn KI-Rechenkapazität **ressortübergreifend zur Priorität** erklärt wird.

**Ausreichende KI-Rechenkapazität ist künftig eine Grundvoraussetzung für ein wettbewerbsfähiges und souveränes Deutschland.** Allzweck-KI (*general purpose AI*) entwickelt sich rasant und durchdringt die Gesellschaft deutlich schneller als Technologien wie der Computer oder das Internet. Führende KI-Nationen werden in den kommenden Jahren immer mehr Prozesse in Wirtschaft und Verwaltung teilweise oder vollständig auf KI umstellen. Voraussetzung dafür sind Rechenzentren mit hochmodernen KI-Chips für das Training und den Betrieb von KI-Modellen. Ohne eigene KI-Rechenkapazität erhöht Deutschland seine Abhängigkeit von anderen Staaten und droht, wirtschaftlich zurückzufallen.

**Deutschlands Anteil an der globalen KI-Rechenkapazität ist gering und droht in den nächsten Jahren dramatisch zu sinken.** Aktuell befinden sich weniger als 2% der globalen KI-Rechenkapazität in Deutschland. In den nächsten Jahren werden nicht nur KI-Schergewichte wie die USA oder China ihre Kapazität deutlich stärker erweitern als Deutschland. Auch Länder wie Südkorea, Frankreich, das Vereinigte Königreich, Indien, Saudi-Arabien oder die Vereinigten Arabischen Emirate bauen ihre KI-Rechenkapazität viel ambitionierter aus. Eine mögliche KI-Gigafabrik, die in Deutschland teilweise als überdimensioniert kritisiert wird, wäre mindestens 50x kleiner als die weltweit größten angekündigten Cluster und bliebe weit unter dem prognostizierten Bedarf. Das Risiko eines Überangebots ist selbst bei einem stark beschleunigten Ausbau gering.

**Je nach Ambitionsniveau könnte die Bundesregierung verschiedene Strategien für den Ausbau von KI-Rechenkapazität verfolgen.** Deutschlands Bedarf an KI-Rechenkapazität hängt davon ab, wie stark KI in Wirtschaft und Verwaltung integriert werden soll und ob Deutschland eigene Spitzenmodelle trainieren will. Dieser Bericht beschreibt drei mögliche Strategien für den KI-Infrastrukturausbau bis Ende 2028:

1. KI in Teilbereichen anwenden: erfordert 850.000 H100-Äquivalente bzw. 0,8 GW IT-Anschlussleistung
2. KI in allen Bereichen anwenden: erfordert 3.400.000 H100-Äquivalente bzw. 3,4 GW IT-Anschlussleistung
3. KI in allen Bereichen anwenden und zusätzlich eigene Spitzenmodelle trainieren: erfordert 6.000.000 H100-Äquivalente bzw. 5,9 GW IT-Anschlussleistung

**Mit den richtigen Maßnahmen kann Deutschland verhindern, international abgehängt zu werden.**

Durch ein beherrztes Vorgehen kann die Bundesregierung ihr Ziel aus dem Koalitionsvertrag erreichen, Deutschland zur KI-Nation zu machen:

1. Kapazität für KI-Training und -Inferenz gemäß strategischer Prioritäten ausbauen
2. Politische Voraussetzungen schaffen und Kompetenzen bündeln
3. Günstige, verlässliche Energie für KI-Rechenzentren bereitstellen
4. Planungs- und Genehmigungsprozesse beschleunigen
5. Sicherheit von KI-Rechenzentren erhöhen
6. KI-Ökosystem umfassend stärken

# Autoren

**Dr. Philip Fox** ist Policy Specialist bei KIRA und Ko-Autor des International AI Safety Report. Er hat Philosophie und Volkswirtschaftslehre in Bayreuth, Oxford und Berlin studiert und wurde an der Humboldt-Universität zu Berlin in Philosophie promoviert.

**Prof. Dr. Dr. h.c. Monika Schnitzer** ist Professorin an der Ludwig-Maximilians-Universität in München und Vorsitzende des Sachverständigenrates zur Begutachtung der gesamtwirtschaftlichen Entwicklung.

**Daniel Privitera** ist Gründer und Executive Director bei KIRA. Er war Vice Chair für das Safety & Security Chapter des EU Code of Practice für Allzweck-KI und Lead Writer des International AI Safety Report.

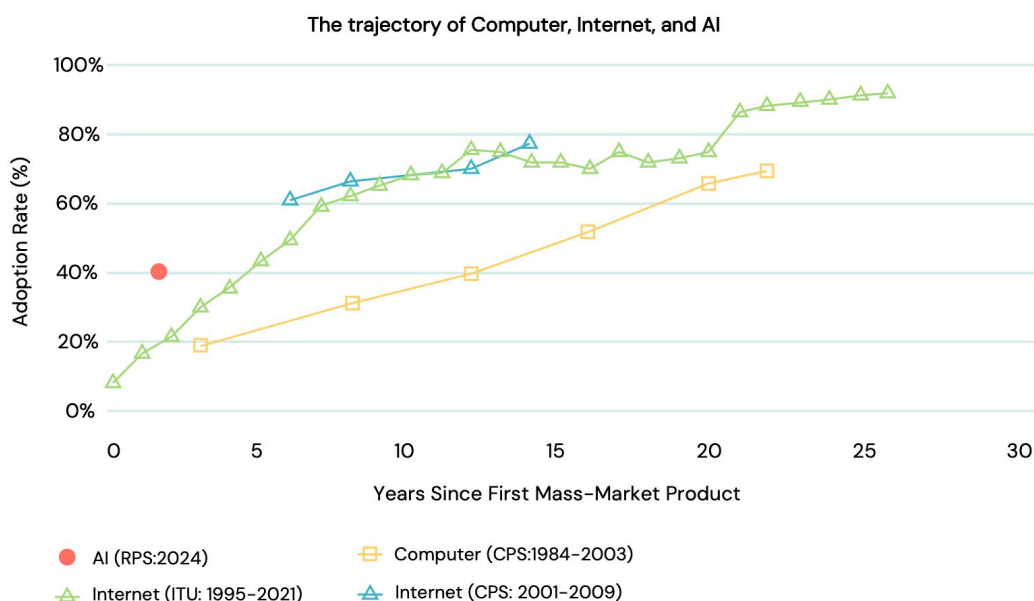
# Inhalt

Executive Summary.....	2
1. Status quo.....	5
2. Strategien.....	14
3. Empfehlungen.....	19
4. Fazit.....	23
Technischer Appendix .....	24
Literaturverzeichnis.....	27
Über diesen Bericht.....	31

# 1. Status quo

**Die Fähigkeiten von Allzweck-KI (fortan "KI") entwickeln sich rasant und werden sämtliche Bereiche in Wirtschaft und Verwaltung durchdringen.** Die Geschwindigkeit, mit der Künstliche Intelligenz (KI) stetig leistungsfähiger wird, überrascht selbst Fachleute. Besonders bemerkenswert ist die Entwicklung der sogenannten Allzweck-KI (*general-purpose AI*). Dazu gehören vielseitig einsetzbare, mit riesigen Datenmengen trainierte KI-Systeme wie ChatGPT, Gemini oder Claude sowie zahlreiche darauf aufbauende Anwendungen. **Wegen der überragenden Bedeutung von Allzweck-KI beschränkt sich dieser Report ausschließlich darauf; wo in diesem Report von "KI" die Rede ist, ist Allzweck-KI gemeint.** Während solche KI-Systeme noch vor fünf Jahren kaum einen kohärenten Text formulieren konnten, lösen sie heute in wenigen Sekunden komplexe Probleme aus den Naturwissenschaften oder der Informatik – teilweise auf einem ähnlichen Niveau wie Menschen mit Dokortitel. Gleichzeitig arbeiten die besten Forschungsteams mit immensen Budgets an der Lösung bestehender Probleme wie Halluzinationen oder der mangelnden Verlässlichkeit von KI-Agenten. Setzt sich der Trend der vergangenen Jahre fort – oder beschleunigt er sich gar, wie manche Fachleute erwarten – werden führende KI-Nationen absehbar immer mehr Prozesse in Wirtschaft und öffentlicher Verwaltung teilweise oder ganz auf KI umstellen. Wer nicht über die nötige Rechenkapazität für diesen Wandel verfügt, droht abgehängt zu werden.

**KI durchdringt die Gesellschaft schneller als frühere Technologien.** Wenige Jahre nach der Einführung liegt die Adoptionsrate von Allzweck-KI deutlich höher als bei Computern oder dem Internet zu einem vergleichbaren Zeitpunkt (Abbildung 1). Die schnelle Verbreitung ergibt sich unter anderem aus der einfachen individuellen Nutzung: Moderne KI ist über leicht zu bedienende, oft kostenlose Chatbots in einer vertrauten digitalen Umgebung verfügbar. Einige Fachleute gehen davon aus, dass bis 2030 praktisch alle Unternehmen kommerzielle KI nutzen werden (Abbildung 2). Aufgrund der schnellen Verbreitung und der langfristigen strategischen Bedeutung sollte Deutschland deshalb so schnell wie möglich die Voraussetzungen für einen starken KI-Standort schaffen.



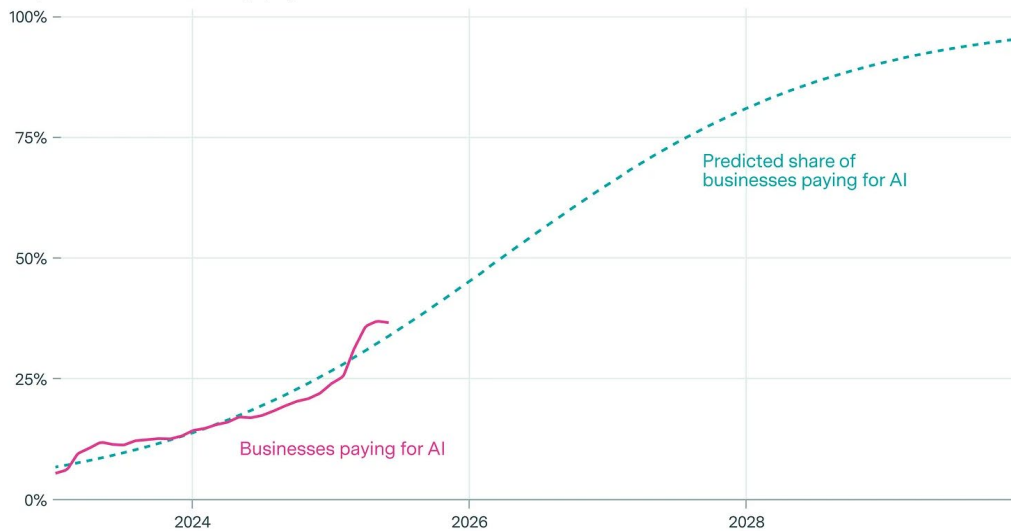
**Abbildung 1:** Daten zur beruflichen Nutzung weisen auf eine Adoptionsrate von generativer KI hin, die deutlich schneller wächst als bei früheren Technologien wie dem Computer oder Internet. Quelle: [Bick et al. \(2024\)](#)

# 1. Status quo

On current trends, most U.S. businesses will pay for AI by 2030

EPOCH AI

Proportion of U.S. businesses paying for AI



CC-BY

epoch.ai

**Abbildung 2:** Eine aktuelle Schätzung aus den USA geht davon aus, dass bis 2030 fast alle Unternehmen KI kommerziell nutzen werden. Quelle: [Berg & Ho \(2025\)](#)

**KI-Rechenkapazität ist im KI-Zeitalter eine strategische Kernressource.** Eine Grundvoraussetzung, um das immense Potential von KI auszuschöpfen, sind große Mengen hochkomplexer KI-Chips wie GPUs (Graphics Processing Units). Sie sind das Herzstück spezieller KI-Rechenzentren und werden sowohl für die Entwicklung ("Training") als auch den Betrieb ("Inferenz") von KI-Modellen benötigt (siehe Abschnitt 2). Ein Unternehmen, das KI-Modelle nutzen oder trainieren will, kann dazu entweder eigene Rechenzentren betreiben oder die nötige Rechenleistung bei externen Anbietern mieten (Cloud Computing). Handelt es sich beim Cloud-Anbieter um ein ausländisches Unternehmen, ist die auf diese Weise bezogene Rechenleistung nicht souverän (siehe Box 1).

## Box 1: Souveräne KI-Rechenkapazität: Vor- und Nachteile

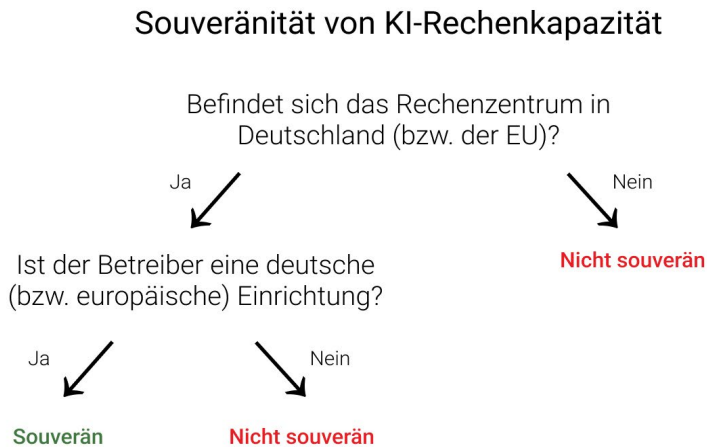
- Deutschland sollte dringend seine KI-Rechenkapazität erweitern, um international nicht den Anschluss zu verlieren. Dabei bieten KI-Rechenzentren auf deutschem Boden – unabhängig davon, wer sie betreibt – wichtige Vorteile gegenüber KI-Rechenleistung aus ausländischen Cloud-Rechenzentren:
  - **Ökosystem-Effekte.** Je mehr Rechenleistung hier zur Verfügung steht, desto attraktiver wird Deutschland als Forschungs- und Start-up-Standort für (inter-)nationales Spitzentalent.
  - **Latenz.** Wird KI in komplexe Industrieprozesse integriert, ist häufig eine niedrige Latenz wichtig – d. h. eine geringe Verzögerung zwischen einer Anfrage an ein KI-System und der berechneten Antwort. Viele KI-Anwendungen in der Robotik, beim autonomen Fahren und am Finanzmarkt erfordern ebenfalls niedrige Latenzen. KI-Rechenzentren, die sich in geografischer Nähe zu ihren Kunden befinden, minimieren Latenzen und ermöglichen entsprechende Anwendungen.
- Allerdings ist Souveränität nicht nur eine Frage des Standorts. **Dieser Bericht versteht KI-Rechenzentren nur dann als souverän, wenn sie sich (i) geografisch in Deutschland (bzw. der EU) befinden und (ii) von einer deutschen (bzw. europäischen) Einrichtung betrieben werden** (siehe Abbildung 3).

# 1. Status quo

## Fortsetzung Box 1: Souveräne KI-Rechenkapazität: Vor- und Nachteile

- Deutsche Unternehmen bzw. die öffentliche Verwaltung können souveräne Rechenkapazität daher auf zwei Arten beziehen: 1) Betrieb eines eigenen Rechenzentrums in Deutschland (bzw. der EU), 2) Nutzung eines deutschen (bzw. europäischen) Cloud-Dienstes, bei dem das relevante Rechenzentrum geografisch in Deutschland (bzw. der EU) liegt.
- Der Ursprung der Hardware-Komponenten ist für diese Definition unerheblich. Eine rein europäische Produktion konkurrenzfähiger KI-Chips ist in den nächsten Jahren nicht realistisch; ein handlungsrelevanter Souveränitätsbegriff sollte diese nicht zur Voraussetzung machen.
- Souveräne KI-Rechenzentren gemäß der obigen Definition bieten weitere entscheidende Vorteile:
  - **Geopolitischer Handlungsspielraum.** Die Nachfrage nach KI-Rechenkapazität ist in den vergangenen Jahren enorm gestiegen. Wer diese nur über ausländische Unternehmen beziehen kann, schwächt die eigene Verhandlungsposition und macht sich abhängiger von anderen Staaten, die den Zugang zu Rechenkapazität kappen oder an strenge Bedingungen knüpfen könnten. Wer KI-Chips hingegen einmal importiert und in souveränen Rechenzentren installiert hat, kann über diese wesentlich unabhängiger verfügen.
  - **Datensouveränität.** Souveräne KI-Rechenzentren erlauben eine vollständige Kontrolle über sensible Daten. Dazu zählen sensible Geschäftsdaten und geistiges Eigentum in Unternehmen sowie sensible personenbezogene Daten, etwa beim Einsatz von KI in der öffentlichen Verwaltung oder Medizin.
  - **Nationale Sicherheit.** In besonders sicherheitskritischen Bereichen – z. B. KI-Anwendungen in der kritischen Infrastruktur – sollten KI-Rechenzentren umfassend vor Sabotage und Spionage geschützt sein. Nur in souveränen KI-Rechenzentren können höchste Sicherheitsstandards unabhängig und gemäß nationaler Standards implementiert und jederzeit geprüft werden.
  - **Wertschöpfung.** Souveräne KI-Rechenzentren stimulieren die Wirtschaft und verhindern, dass ein beträchtlicher Teil der KI-Wertschöpfung an ausländische Hyperscaler abfließt.
- Gleichzeitig geht souveräne Recheninfrastruktur aktuell mit Nachteilen einher:
  - **Geringere Benutzerfreundlichkeit.** US-Hyperscaler wie AWS, Microsoft Azure oder Google Cloud bieten ein umfassendes Ökosystem aus Hardware und Software, das Unternehmen eine besonders leichte Entwicklung und Bereitstellung eigener KI-Anwendungen ermöglicht. Dieser "Platform-as-a-Service"-Ansatz nimmt vor allem Start-ups viel Aufwand ab und lässt sich durch kleinere Cloud-Anbieter nur schwer replizieren.
  - **Kapitalaufwand.** Neue KI-Rechenzentren und die dafür nötige Energieinfrastruktur erfordern Milliardeninvestitionen, die häufig die Finanzkraft einzelner Unternehmen übersteigen.
- Dieser Bericht schlägt vor, in Deutschland bis Ende 2028 eine souveräne Mindestkapazität von 200.000 H100-Äquivalenten aufzubauen (siehe Abschnitt 2).

# 1. Status quo



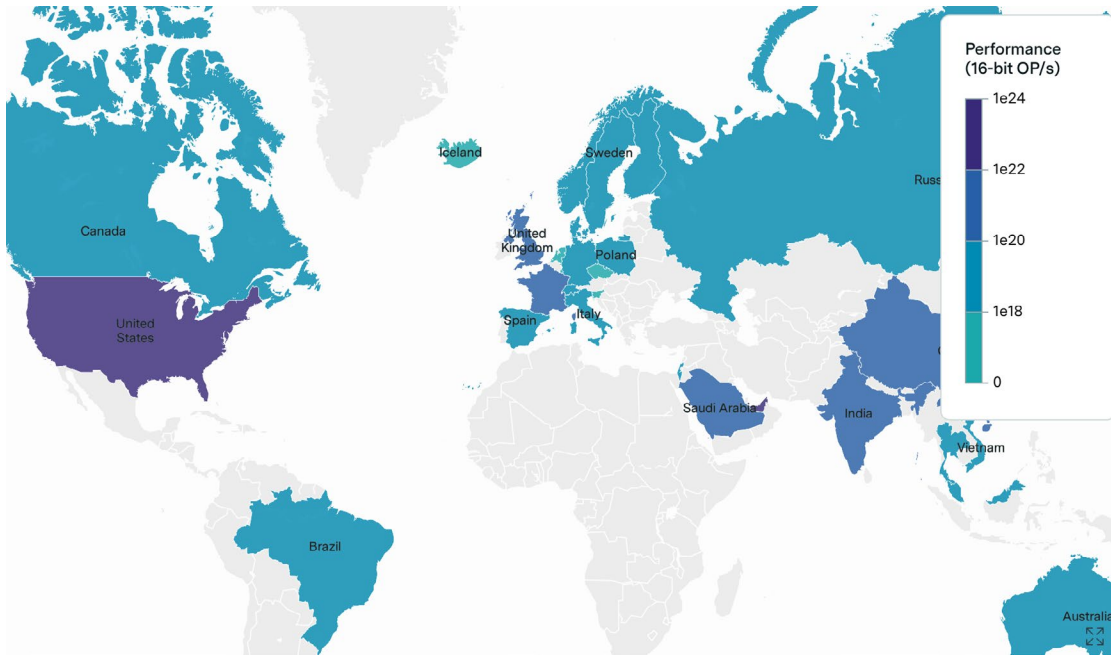
**Abbildung 3:** Ob Rechenkapazität im Sinne dieses Bericht als souverän gilt, hängt vom Standort des Rechenzentrums und Sitz des Betreibers ab.

**Wer über mehr Rechenkapazität verfügt, kann leistungsfähigere KI-Modelle trainieren und diese Modelle in mehr Prozesse integrieren.** Sogenannte „Scaling Laws“ beschreiben in der KI-Forschung einen empirischen Zusammenhang zwischen (u. a.) der Rechenleistung, die für das Training eines KI-Modells eingesetzt wird, und der Performance eines Modells – vereinfacht gesagt führt mehr Rechenleistung zu „intelligenterer“ KI. Neuere Erkenntnisse zeigen zudem, dass sich die Modell-Performance nicht nur durch mehr Rechenleistung während des Trainings steigern lässt. Die Performance von „Reasoning Modells“ wie GPT-5-Thinking (OpenAI) oder Gemini 2.5 Pro Deep Think (Google) erhöht sich auch, wenn dem Modell während des Betriebs zusätzliche Rechenleistung zur Verfügung gestellt wird. Diese Modelle können Rechenleistung in längere Denkprozesse umwandeln, um Probleme logisch stringent zu analysieren oder mehrere Lösungsansätze auszuprobieren – und damit besonders komplexe Aufgaben lösen.

**Deutschlands KI-Rechenkapazität fällt im internationalen Vergleich zurück.** Daten des Forschungsinstituts Epoch AI zeigen: Aktuell verfügt Deutschland über weniger als 2% der globalen KI-Rechenkapazität (Pilz et al. 2025). Zwar gehört Deutschland damit noch zu der kleinen Gruppe an Ländern, die relativ gesehen nach den KI-Schergewichten USA und China über die größte KI-Rechenkapazität verfügen. Dieses Bild verschlechtert sich aber deutlich, sobald man berücksichtigt, wie verschiedene Länder ihre Kapazität in den nächsten Jahren ausbauen wollen: Deutschland bleibt dann in der vorletzten von insgesamt vier Leistungsgruppen – zu der in diesem Fall auch Länder wie Brasilien, Russland, Kanada, Vietnam und Malaysia gehören –, während Frankreich und das Vereinigte Königreich in die zweithöchste Gruppe aufsteigen (Abbildung 4). Mit einer europäischen KI-Gigafabrik (siehe Box 2) könnte es Deutschland knapp in die zweithöchste Klasse schaffen – vorausgesetzt, Deutschland erhält einen Zuschlag der Europäischen Union, die Gigafabrik entsteht wie geplant und andere Ländern bauen ihre KI-Infrastruktur nicht stärker aus als aktuell erwartet. Selbst dann hätte Deutschland aber voraussichtlich eine geringere Gesamtkapazität als Südkorea, Frankreich, Indien, Saudi-Arabien und das Vereinigte Königreich – geschweige denn die USA, die Vereinigten Arabischen Emirate und China, deren Ausbaupläne mit Abstand am ambitioniertesten sind (Patel et al. 2025a, Patel et al. 2025b).<sup>1</sup>

<sup>1</sup> Für das Vereinigte Königreich ergibt sich dieser Vergleich aus der jüngsten Ankündigung, dort 120.000 GPUs der neuen Blackwell-Generation anzuschaffen (NVIDIA 2025b) – was mindestens 300.000 H100-Äquivalenten entspricht (siehe Fußnote 2). Der Vergleich mit den anderen Ländern basiert auf Daten von Epoch AI (Pilz et al. 2025).

# 1. Status quo



**Abbildung 4:** Deutschlands Pläne für den Ausbau von Rechenkapazität sind im internationalen Vergleich wenig ambitioniert. Eine mögliche europäische KI-Gigafabrik ist hier nicht eingerechnet. Auch mit einer Gigafabrik läge Deutschland im Ranking voraussichtlich hinter Ländern wie Südkorea, Frankreich, Indien, Saudi-Arabien oder den USA. Grundlage für den Vergleich sind Daten über die Rechenleistung von Supercomputern, gemessen in 16-Bit-Operationen pro Sekunde (FLOP/s) (Pilz et al. 2025). Der Datensatz deckt etwa 10-20% der globalen Supercomputer-Performance ab, ist die derzeit umfassendste Analyse dieser Art und liefert eine repräsentative Übersicht der globalen Verteilung von KI-Rechenkapazität.

## Box 2: Europäische KI-Gigafabriken

- Die Europäische Union plant im Zuge ihres AI Continent Action Plan bis zu 5 europäische KI-Gigafabriken: große Rechenzentren für KI-Training und -Inferenz, ausgestattet mit jeweils etwa 100.000 leistungsfähigen KI-Chips (European Commission 2025a, EuroHPC 2025).
- Dafür stellt die EU 20 Milliarden € im Rahmen der InvestAI-Initiative zur Verfügung und rechnet je Gigafabrik mit Kosten von 3-5 Milliarden € (EuroHPC 2025). Andere halten hingegen 6-8 Milliarden € für realistischer (Hess 2025). Nicht eingerechnet sind Ausgaben für den laufenden Betrieb, z. B. Energiekosten. Die EU bzw. ihre Mitgliedsstaaten unterstützen den Bau einer Gigafabrik mit bis zu 35% des Kapitalaufwands (CAPEX). Weitere Kosten, einschließlich der gesamten Betriebskosten (OPEX), werden vom Privatsektor getragen (EuroHPC 2025).
- Die Bundesregierung will mindestens eine Gigafabrik in Deutschland ansiedeln. Nachdem die Verhandlungen führender deutscher Unternehmen über eine gemeinsame Bewerbung scheiterten, haben verschiedene Unternehmen eigene Interessenbekundungen eingereicht – darunter die Schwarz Gruppe, die Deutsche Telekom, Ionos (gemeinsam mit Hochtief), der Freistaat Bayern, das Start-up Lyceum sowie ein Konsortium unter der Führung von Silicon Saxony (Handelsblatt 2025, FAZ 2025).
- Es ist derzeit unklar, ob ein deutsches Unternehmen bzw. Konsortium den Zuschlag erhält und, falls ja, ob in der Folge eine KI-Gigafabrik in Deutschland entsteht – Lyceum plant diese beispielsweise aufgrund günstigerer Energie und schnellerer Genehmigungsverfahren in Dänemark (FAZ 2025).
- Das offizielle Bewerbungsverfahren soll 2026 beginnen und bis zum Ende des 2. Quartals abgeschlossen sein. Selbst bei einer erfolgreichen Bewerbung würde eine Gigafabrik deshalb wohl frühestens 2027 in Betrieb gehen. Verzögerungen im Ausschreibungsverfahren, bei der Finanzierung, im Planungs- und Genehmigungsprozess oder bei der Netzanbindung könnten auch zu einem späteren Betriebsbeginn führen.

# 1. Status quo

**Deutschland ist mit seinen KI-Supercomputern international nicht konkurrenzfähig und wird in den nächsten Jahren weiter zurückfallen.** Deutschland droht, aufgrund mangelnder Ambition beim Ausbau von Rechenkapazität als KI-Standort weiter abgehängt zu werden. Ein Blick auf deutsche KI-Supercomputer im internationalen Vergleich stützt diesen Befund (Tabelle 1). Auch der oft als Leuchtturmprojekt bezeichnete Supercomputer "Jupiter" am Forschungszentrum Jülich liegt mit seinen ca. 24.000 H100-Äquivalenten ([FZ Jülich 2025](#)) deutlich hinter den größten US-Supercomputern, die mit bis zu 275.000 H100-Äquivalenten ausgestattet sind (siehe Tabelle 1). Ein Blick auf geplante Supercomputer zeigt, dass der Abstand in Zukunft eher noch weiter wachsen wird: Megaprojekten wie einem Cluster mit 20 Millionen H100-Äquivalenten in den VAE oder 5 Millionen H100-Äquivalenten in Südkorea steht die im Koalitionsvertrag für Deutschland vorgesehene KI-Gigafabrik mit 100.000 H100-Chips gegenüber ([Pilz et al. 2025](#)). Diese wäre mindestens 50x kleiner als die weltweit größten angekündigten Cluster (Tabelle 1). Auch ein Bericht des BMWK aus der vergangenen Legislaturperiode kommt zu dem Ergebnis, dass Deutschland in Sachen Rechenkapazität trotz der aktuellen Pläne weiter hinter Länder wie die USA zurückfällt ([BMWK 2025](#)).

# 1. Status quo

## KI-Supercomputer in Deutschland und international

Deutschland						International					
Super-computer	Standort	Eröffnung	# GPUs	Typ	H100-Äquivalente <sup>2</sup>	Super-computer	Standort	Eröffnung	# GPUs	Typ	H100-Äquivalente
<b>Vorhanden</b>						<b>Vorhanden</b>					
Jupiter AI Factory	Jülich	2025	23.752	NVIDIA GH200	<b>23.752</b>	xAI Colossus	Memphis, USA	2025	230.000	NVIDIA H100/200 /GB200	<b>ca. 275.000<sup>3</sup></b>
JUWELS-Booster	Jülich	2020	3.744	NVIDIA A100	<b>1181</b>	Meta 100k Cluster	USA	2024	100.000	NVIDIA H100	<b>100.000</b>
Hunter	Stuttgart	2025	752	AMD Instinct MI300A	<b>745</b>	OpenAI/ Micro-soft Cluster	Goodyear, USA	2024	100.000	NVIDIA H100	<b>100.000</b>
Capella	Dresden	2024	624	NVIDIA H100	<b>624</b>	Oracle OCI Supercluster	USA	2025	65.536	NVIDIA H200	<b>65.536</b>
hessian.AI fortythree	Darmstadt	2024	912	NVIDIA A100/H100	<b>479</b>	Tesla Cortex Phase 1	USA	2024	50.000	NVIDIA H100	<b>50.000</b>
Helma	Nürnberg	2024	384	NVIDIA H100	<b>384</b>	El Capitan	Livermore, USA	2024	44.544	AMD Instinct MI300A	<b>44.143</b>
<b>Geplant</b>						<b>Geplant</b>					
KI-Gigafabrik	Deutschland	?	100.000	?	<b>100.000</b>	Stargate UAE 5GW Campus	Abu Dhabi	2030	?	?	<b>20.262.759<sup>4</sup></b>
Telekom/ Nvidia Industrial Cloud <sup>5</sup>	Deutschland	2026	10.000	? <sup>6</sup>	<b>max. 25.265</b>	3GW Cluster	Südkorea	2028	?	?	<b>5.103.588</b>
ELBJUWEL	Dresden	?	?	?	<b>25.265</b>	Meta Louisiana Datacenter	Holly Ridge, USA	2030	?	?	<b>5.103.588</b>
HammerAI AI Factory	Stuttgart	?	?	?	<b>?<sup>7</sup></b>	Stargate USA Supercluster 3	Abilene, USA	2027	?	?	<b>5.103.588</b>
Herder	Stuttgart	2027	?	?	<b>?<sup>8</sup></b>	DataVolt Neom Phase 2	Saudi-Arabien	2031	?	?	<b>5.103.588</b>
Blue Lion	Garching	2027	~4.400	?	<b>?<sup>9</sup></b>	OpenAI/ Micro-soft Wisconsin	Mt. Pleasant, USA	2026	700.000	NVIDIA GB 200	<b>1.768.570</b>
OpenAI for Germany	?	2026	4.000	?	<b>?</b>	HUMAIN Phase 2	Saudi-Arabien	2030	?	?	<b>1.520.970</b>

**Tabelle 1:** Eine Auswahl bedeutender KI-Supercomputer in Deutschland und international zeigt, dass die deutsche Rechenkapazität sowohl bei vorhandenen als auch geplanten Clustern hinter anderen Ländern zurückfällt. Die Daten für geplante (und teilweise auch für vorhandene) Cluster basieren teilweise auf inoffiziellen Berichten und sind entsprechend unsicher, liefern aber einen hinreichend repräsentativen Überblick über die globale Verteilung von KI-Rechenkapazität. Quelle: [Pilz et al. 2025](#), eigene Recherche

<sup>2</sup> H100-Äquivalente sind ein gängiges Maß, um die Rechenleistung unterschiedlicher GPU-Typen miteinander zu vergleichen (siehe technischer Appendix).

<sup>3</sup> Laut xAI-Geschäftsführer Elon Musk verfügte Colossus im Juni 2025 über 150.000 H100-, 50.000 H200- und 30.000 GB200-GPUs ([Y Combinator 2025](#)).

<sup>4</sup> Bei einigen angekündigten Projekten sind Anzahl und Typ der GPUs unbekannt. Epoch AI schätzt in diesen Fällen die Anzahl an H100-Äquivalenten anhand weiterer verfügbarer Informationen wie der angekündigten IT-Anschlussleistung.

<sup>5</sup> Die Deutsche Telekom sieht die gemeinsam mit Nvidia geplante Industrial Cloud als Vorstufe einer europäischen KI-Gigafabrik ([Telekom 2025](#)). Sollte der Zuschlag für eine Gigafabrik an die Telekom gehen, könnten die ca. 25.000 H100-Äquivalente der Industrial Cloud daher Teil einer Gigafabrik mit insgesamt 100.000 GPUs werden.

<sup>6</sup> Laut offizieller Ankündigung wird die Industrial Cloud mit 10.000 GPUs ausgestattet sein, darunter Nvidia-Chips der neuen Blackwell-Generation ([NVIDIA 2025a](#)). Unklar ist, ob neben GPUs des Typs B200 auch ältere GPUs eingesetzt werden (und, falls ja, wie viele).

<sup>7</sup> Im Rahmen der EuroHPC-Initiative plant das HLRS Stuttgart eine europäische AI Factory. Das Projektbudget beträgt 85 Millionen € ([HLRS 2024](#)).

<sup>8</sup> Der Herder-Supercomputer am HLRS Stuttgart folgt auf das Übergangssystem Hunter (siehe Tabelle) aus dem Jahr 2025 und ist als Exascale-System konzipiert. Typ und Anzahl der Chips werden laut HLRS bis Ende 2025 festgelegt ([HLRS 2023](#)).

<sup>9</sup> Blue Lion entsteht am LRZ in Garching bei München mit Nvidia-GPUs der Rubin-Architektur, die noch nicht auf dem Markt ist ([LRZ 2025](#)). Schätzungen zufolge könnten dabei 4.400 Rubin-GPUs zum Einsatz kommen, was ca. 26.000 H100-Äquivalenten entsprechen könnte ([The Next Platform 2024](#)). Das Projektvolumen beträgt 250 Millionen €.

# 1. Status quo

**Ein Mangel an Rechenkapazität gefährdet Deutschlands Wettbewerbsfähigkeit und geopolitische Souveränität.** Mit der geringen Ambition beim Ausbau von KI-Infrastruktur gehen für die Zukunft gravierende Risiken einher:

- **Abgehängte Wirtschaft.** Das deutsche BIP-Wachstum könnte langfristig hinter Staaten zurückfallen, die über mehr Rechenkapazität für die Entwicklung und Nutzung von KI verfügen und damit einen größeren Teil der Wertschöpfung vereinnahmen können.
- **Abgehängte Verwaltung.** Die staatliche Verwaltung könnte im internationalen Vergleich zunehmend rückschrittlich und dysfunktional werden, wenn durch KI ermöglichte Effizienzgewinne ungenutzt bleiben.
- **Datenabfluss.** Deutschland könnte gezwungen sein, sensible Industrie- oder personenbezogene Daten an ausländische Cloud-Anbieter preiszugeben.
- **Talentabfluss.** Deutschland könnte als Standort für heimisches oder ausländisches KI-Talent noch unattraktiver werden, weil andere Länder eine bessere KI-Infrastruktur bieten.

Angesichts der herausragenden strategischen Bedeutung von KI sollte die Bundesregierung dringend den Ausbau von KI-Rechenkapazität vorantreiben. Die folgenden Abschnitte stellen dazu konkrete Strategien und Empfehlungen vor.

## Box 3: DeepSeek – (k)ein Gamechanger?

- Im Januar 2025 veröffentlichte das chinesische KI-Unternehmen DeepSeek das Reasoning-Modell R1, dessen Performance nur knapp hinter führenden US-Modellen lag. Laut eigenen Angaben erreichte DeepSeek diese Performance mit deutlich weniger Rechenleistung – und damit wesentlich günstiger – als die amerikanische Konkurrenz. Für das zugrunde liegende Modell V3 bezifferte DeepSeek die Kosten für den finalen Trainingslauf – die nur einen Teil der gesamten Entwicklungskosten abbilden – auf bloß 5,6 Millionen Dollar ([DeepSeek-AI 2024](#)). Dieser Wert liegt einerseits deutlich unter den 100 Millionen Dollar, die beispielsweise OpenAI im Jahr 2023 für die Entwicklung ihres Modells GPT-4 aufgewendet hat. Andererseits bleiben darin viele Kostenfaktoren unberücksichtigt, was einen Vergleich mit der Zahl von OpenAI erschwert.
- Auf die Veröffentlichung folgte eine heftige Reaktion an den Finanzmärkten. Die Börsenwerte von Chipunternehmen wie Nvidia brachen ein – manche Marktteilnehmer gingen möglicherweise davon aus, dass Spitzenmodelle in Zukunft keine so hohen Mengen an Rechenleistung mehr benötigen würden. Parallel dazu wuchs in den USA die Angst, im globalen KI-Wettlauf gegen das womöglich effizientere China zu verlieren. In Europa ließ der Erfolg DeepSeeks hingegen die Hoffnung aufkommen, bald zu vergleichbar niedrigen Kosten eigene Spitzenmodelle trainieren zu können. Diesen Reaktionen liegen aber zum Teil irreführende Annahmen zugrunde:
  - Die öffentlich häufig genannte Summe von 5,6 Millionen Dollar bezieht sich allein auf den finalen Trainingslauf von V3. Weitere Kosten, wie z. B. für GPUs, Strom oder Gehälter, sind nicht eingerechnet. Experten schätzen, dass die tatsächlichen Entwicklungskosten deutlich höher liegen ([Patel et al. 2025b](#)).
  - DeepSeek hat gezeigt, dass sich die Effizienz von KI-Modellen durch kluge Ideen in der Tat beachtlich steigern lässt. Es war allerdings schon vorher bekannt, dass sich ein gegebenes Performance-Niveau nach einigen Monaten mit deutlich weniger Ressourcen erreichen lässt ([Pilz, Heim & Brown 2023](#)). Neu waren deshalb nicht die Effizienzsteigerungen selbst, sondern dass diese von einem chinesischen Unternehmen realisiert wurden. Damit bleibt aber der grundlegende Trend unberührt, der ein entscheidender Faktor für die rasant wachsende Nachfrage nach KI-Recheninfrastruktur ist: Wer mehr Rechenkapazität besitzt, kann bessere Modelle trainieren – was sich auch daran zeigt, dass über ein halbes Jahr nach der Veröffentlichung von DeepSeeks R1 die führenden Modelle weiterhin aus den USA stammen.

# 1. Status quo

## Fortsetzung Box 3: DeepSeek – (k)ein Gamechanger?

- Rechenkapazität braucht es nicht nur für KI-Training, sondern auch für KI-Inferenz (siehe Box 4). Selbst wenn China ähnlich gute Modelle zu geringeren Kosten trainieren könnte, könnte es diese weniger intensiv nutzen als andere Länder. Zum Beispiel kann DeepSeek Berichten zufolge die eigenen Modelle nicht wie gewünscht in der Breite anbieten, weil dafür aufgrund US-amerikanischer Exportkontrollen die Chips fehlen ([The Information 2025](#)).
- Für die deutsche KI-Rechenzentrenstrategie ergeben sich zwei wichtige Implikationen:
  - Rechenkapazität ist weiterhin ein entscheidender Faktor für leistungsfähige KI. Das Training von Spitzenmodellen erfordert in der Zukunft wahrscheinlich Milliarden-Investitionen. Auch etwas weniger leistungsfähige Modelle, deren Performance einige Monate hinter den besten US-Modellen liegt, werden sich in den nächsten Jahren voraussichtlich nicht mit wenigen Millionen Euro entwickeln lassen.
  - Selbst wenn sich Spitzenmodelle in der Zukunft günstiger trainieren ließen – oder Deutschland ganz auf das Training solcher Modelle verzichten würde –, bräuchte Deutschland zusätzliche Rechenkapazität für KI-Inferenz, um KI flächendeckend in Wirtschaft und Verwaltung einzusetzen.

## 2. Strategien

**Die passende Compute-Strategie hängt davon ab, wie die Bundesregierung die strategische Relevanz von KI bewertet und welche Priorität sie dem Thema beimisst.** Die Bundesregierung sollte Klarheit darüber schaffen, wie ambitioniert die deutsche KI-Infrastruktur bis zum Ende der Legislaturperiode ausgebaut werden soll. Die Entscheidung darüber hängt von zwei zentralen Faktoren ab:

1. Wie hoch schätzt die Bundesregierung die zukünftige strategische Bedeutung von KI ein?
2. Wie hoch ist die Bereitschaft der Bundesregierung, die notwendigen politischen und wirtschaftlichen Rahmenbedingungen für einen erfolgreichen KI-Standort zu schaffen?

Je nach Antwort auf diese Fragen könnte sich Deutschland primär als Technologienehmer sehen, der ausländische KI-Modelle in die eigene Wirtschaft und Verwaltung integriert, oder primär als Technologieführer, der zusätzlich die Entwicklung eigener Spitzen-KI vorantreibt. Diese Entscheidung bestimmt, wie viel Kapazität Deutschland jeweils für KI-Inferenz bzw. KI-Training benötigt (siehe Box 4). Schätzt die Bundesregierung die strategische Bedeutung von KI als gering ein oder misst dem Thema aus anderen Gründen keine Priorität bei, würde sich eine passende Compute-Strategie darauf konzentrieren, genügend Rechenkapazität für die Nutzung ausländischer KI in Schlüsselbereichen zu schaffen. Schreibt die Bundesregierung KI eine mittlere strategische Relevanz zu und priorisiert das Thema entsprechend, würde eine passende Compute-Strategie darauf abzielen, genügend Rechenkapazität für die Nutzung ausländischer KI in allen Bereichen der Wirtschaft und des öffentlichen Sektors schaffen. Bewertet die Bundesregierung die strategische Relevanz hingegen als hoch und ist bereit, durch außergewöhnliche Maßnahmen die notwendigen Bedingungen zu schaffen, um Deutschland zu einem wirklich führenden KI-Standort zu machen, bietet sich eine ambitioniertere Compute-Strategie an. Eine solche Strategie würde z. B. auf das Training eigener Spitzenmodelle setzen und voraussichtlich u. a. ein Sondervermögen im niedrigen dreistelligen Milliardenbereich erfordern.

### Box 4: Training vs. Inferenz

- Rechenkapazität wird in verschiedenen Phasen des KI-Lebenszyklus benötigt:
  - Während des **Trainings** werden die Parameter – die internen Stellschrauben eines Modells – anhand großer Datenmengen optimiert. Dabei lernt das Modell, in den Daten identifizierte Muster auf neue Fälle anzuwenden. Unterschieden wird zwischen dem *Pre-Training*, bei dem ein Modell vor allem auf der Grundlage von Internetdaten allgemeines Wissen erlangt, und dem *Post-Training*, bei dem kuratierte Datensätze und menschliches bzw. KI-generiertes Feedback zum Einsatz kommen, um das Modell für spezifische Aufgaben oder Verhaltensweisen zu optimieren (z. B. mathematische Probleme zu lösen). Während das Pre-Training traditionell der rechenintensivste Schritt ist, wird bei neueren Modellen immer mehr Rechenleistung für das Post-Training aufgewendet ([You 2025](#)).
  - Während der **Inferenz** generiert ein zuvor trainiertes Modell für einen gegebenen Input den entsprechenden Output und wendet dabei das im Training erlangte Wissen an. Kommerzielle KI-Unternehmen bieten ihre Modelle dafür teils Hunderten Millionen Menschen an ([Chatterji et al. 2025](#)). Seit Ende 2024 zeigt sich ein Trend zu zunehmend rechenintensiver Inferenz: Reasoning-Modelle liefern bessere Ergebnisse, wenn sie während der Inferenzphase mehr Rechenleistung zur Verfügung haben und ausführlicher über ein Problem nachdenken können.
- Grundsätzlich kann jedes KI-Rechenzentrum für Training und Inferenz genutzt werden. Allerdings ist nicht jede Nutzungsart gleich sinnvoll, weil beide Phasen mit jeweils unterschiedlichen Anforderungen einhergehen ([Fist & Datta 2024](#)):
  - **Für Inferenz genügen kleinere Rechenzentren.** Das Training von Spitzenmodellen erfordert aktuell eine hohe Anzahl an Chips in einem großen, zentralen Cluster.<sup>10</sup> Während der Inferenzphase werden hingegen kleinere Datenmengen verarbeitet, so dass auch kleinere Rechenzentren in Frage kommen. Diese benötigen weniger Energie und können geografisch über ein Land verteilt sein.

<sup>10</sup> Distributed-Training-Ansätze versuchen, KI-Modelle in mehreren kleineren Rechenzentren zu trainieren. Aktuell sind sie zentralisierten Ansätzen unterlegen. Die Bedeutung von Distributed Training könnte in der Zukunft allerdings wachsen.

## 2. Strategien

### Fortsetzung Box 4: Training vs. Inferenz

- **Für Inferenz eignen sich auch ältere Chips.** Chips der jeweils neuesten Generation haben meist sowohl für Training als auch für Inferenz die beste Performance. Chips älterer Generationen, die für das Training kompetitiver Modelle nicht mehr geeignet sind, lassen sich aber in der Regel noch wirtschaftlich für Inferenz nutzen. (Bei unklarem strategischen Fokus kann der Kauf neuester Chips trotzdem sinnvoll sein, weil diese vielseitiger sind.)
- **Für Inferenz spielt mitunter die geografische Nähe zum Nutzer eine Rolle.** Im Gegensatz zum Training spielt bei KI-Inferenz für manche Anwendungen – z. B. am Finanzmarkt oder in der Industrieproduktion – die Latenz eine wichtige Rolle (d. h. die Verzögerung zwischen der Eingabe eines Nutzers und der Antwort des Modells). Diese lässt sich minimieren, indem KI-Rechenzentren in geografischer Nähe zu ihren Nutzern gebaut werden.

### Je nach Ambitionslevel ergeben sich für Deutschland drei mögliche Compute-Strategien:

#### Strategie 1 (niedrige Ambition)

- Diese Strategie sieht vor, sich auf die Nutzung (Inferenz) ausländischer Spitzenmodelle in bestimmten Sektoren zu konzentrieren und dafür in begrenztem Maße eigene Kapazität zu schaffen, z. B. um KI-Agenten in zentrale Bereiche der Wirtschaft und Verwaltung zu integrieren. Darüber hinaus würde Deutschland nur branchenspezifische KI-Modelle und -Anwendungen entwickeln, deren Training vergleichsweise wenig Rechenleistung benötigt. Wir nehmen an, dass der Gesamtbedarf für Inferenz und branchenspezifisches Training in Deutschland Ende 2028 bei ca. 4,3 Mio. H100-Äquivalenten liegt (siehe Appendix). Weiterhin gehen wir davon aus, dass Deutschland bei niedriger Ambition nur 20% dieses Bedarfs durch KI-Rechenzentren im eigenen Land abdecken würde. Dafür bräuchte Deutschland ca. 850.000 H100-Äquivalente bzw. 0,4% der globalen Rechenkapazität bis Ende 2028 gemäß unserer Schätzung. Das entspricht einer IT-Anschlussleistung von ca. 0,8 GW. (Ob Deutschland die restliche Nachfrage durch ausländische Cloud-Dienste bedienen könnte, hängt davon ab, wie schnell die global verfügbare KI-Rechenkapazität wächst und wie gewillt andere Staaten sind, diese mit Deutschland zu teilen.)

#### Strategie 2 (mittlere Ambition)

- Diese Strategie konzentriert sich ebenfalls auf Inferenz und branchenspezifisches Training und würde einen hohen Teil des Gesamtbedarfs dafür – 80% – durch KI-Rechenzentren im eigenen Land abdecken. Dafür bräuchte Deutschland bis Ende 2028 ca. 3,4 Mio. H100-Äquivalente bzw. 1,5% der globalen Rechenkapazität gemäß unserer Schätzung. Das entspricht einer IT-Anschlussleistung von ca. 3,4 GW. Damit wäre Deutschland deutlich besser positioniert, KI-Agenten in allen Bereichen der Wirtschaft und Verwaltung einzusetzen und würde seine Abhängigkeit von ausländischen Cloud-Anbietern reduzieren.

#### Strategie 3 (hohe Ambition)

- Diese Strategie sieht vor, zusätzlich Trainings- und Inferenz-Kapazität für eigene, kompetitive Spitzenmodelle mit allgemeiner Anwendung zu schaffen. Deutschland würde diese Modelle gemäß eigener Sicherheits- und Wertmaßstäbe entwickeln; ausländische Akteure könnten den Zugang zu diesen Modellen nicht unmittelbar einschränken. Dafür bräuchte Deutschland eine deutlich höhere Rechenkapazität von ca. 6 Mio. H100-Äquivalenten, was Ende 2028 gemäß unserer Schätzung ca 2,7% der globalen Rechenkapazität entsprechen würde – ein immer noch deutlich geringerer Anteil als Deutschlands Anteil am globalen BIP (nominal) von derzeit ca. 4,2% (IMF 2025). Eine solche Rechenkapazität würde eine IT-Anschlussleistung von 5,9 GW erfordern.

Tabelle 2 fasst diese Strategien und die dafür notwendigen politischen Erfolgsbedingungen zusammen. Ein technischer Appendix am Ende dieses Berichts erläutert die Berechnung der genannten GPU-Kapazitäten bzw. IT-Anschlussleistungen.

# 2. Strategien

## Mögliche Compute-Strategien für Deutschland

Ambition	Strategie bis Ende 2028	Beispielhaft: politische Erfolgsbedingungen
Niedrig	<p>Compute-Ausbau:</p> <ul style="list-style-type: none"> <li>• Deutschland schafft eigene Inferenz-Kapazität für sicherheits- und latenzkritische Anwendungen und die Nutzung von KI-Agenten in Schlüsselsektoren.</li> <li>• Zudem schafft Deutschland Trainings-Kapazität für branchenspezifische KI-Modelle und -Anwendungen (z. B. Industrial Foundation Models).</li> </ul> <p>Benötigte GPUs: 850.000 H100-Äquivalente (= 0,4% der globalen Kapazität gemäß unserer Schätzung)</p> <ul style="list-style-type: none"> <li>• davon z. B. 250.000 H100-Äquivalente in einem zentralen Cluster primär für KI-Training</li> <li>• der Rest in kleineren, dezentralen Cluster für Inferenz</li> </ul>	<ul style="list-style-type: none"> <li>• Die Bundesregierung sieht KI als ein wichtiges Digitalthema, vergleichbar mit der elektronischen Patientenakte.</li> <li>• "Business-as-usual"-Mentalität: Die Bundesregierung behandelt KI mit derselben Priorität wie viele andere Themen im politischen Tagesgeschäft.</li> <li>• Die Bundesregierung betont in ihrer KI-Strategie die Bedeutung schneller Planungs- und Genehmigungsprozesse für KI-Rechenzentren und beruft dazu eine Bund-Länder-Arbeitsgruppe ein.</li> <li>• Verfügbare IT-Anschlussleistung für KI-Rechenzentren von ca. 0,8 GW.</li> <li>• Die Bundesregierung fördert KI-Rechenzentren und das breitere KI-Ökosystem mit einem niedrigen zweistelligen Milliardenbetrag.</li> <li>• Ein Referat im zuständigen Ministerium koordiniert den Infrastruktur-Ausbau und stimmt sich mit anderen relevanten Ressorts ab.</li> </ul>
Mittel	<p>Compute-Ausbau:</p> <ul style="list-style-type: none"> <li>• Wie bei Ambitionslevel "niedrig", aber höhere Inferenz-Kapazität, um KI-Agenten in allen Bereichen der Wirtschaft und Verwaltung zu nutzen und damit die Abhängigkeit von ausländischen Cloud-Diensten zu reduzieren</li> </ul> <p>Benötigte GPUs: 3.400.000 H100-Äquivalente (= 1,5% der globalen Kapazität gemäß unserer Schätzung)</p> <ul style="list-style-type: none"> <li>• davon z. B. 500.000 H100-Äquivalente in einem zentralen Cluster primär für KI-Training</li> <li>• der Rest in kleineren, dezentralen Clustern für Inferenz</li> </ul>	<ul style="list-style-type: none"> <li>• Die Bundesregierung sieht KI als zentrale Zukunftstechnologie, die die Welt in etwa so stark verändern wird wie das Internet.</li> <li>• "Fast-track"-Mentalität: Die Bundesregierung schreitet bei KI schneller voran als bei anderen Technologien und schafft staatliche Kapazität, die das Thema vorrangig behandelt.</li> <li>• Die Bundesregierung novelliert binnen sechs Monaten relevante Gesetze (z. B. BauGB, BImSchG) und verkürzt Planungs- und Genehmigungsprozesse für KI-Rechenzentren gemeinsam mit Ländern und Kommunen auf 1-1,5 Jahre (statt wie bisher oft zwei Jahre oder länger).</li> <li>• Verfügbare IT-Anschlussleistung für KI-Rechenzentren von ca. 3,4 GW.</li> <li>• Die Bundesregierung fördert KI-Rechenzentren und das breitere KI-Ökosystem mit einem mittleren zweistelligen Milliardenbetrag.</li> <li>• Die Bundesregierung setzt eine Taskforce mit zuständigen Staatssekretären/-sekretärinnen unter Leitung des Kanzleramtsministers ein, die den Infrastruktur-Ausbau über die Ressorts hinweg koordiniert.</li> </ul>
Hoch	<p>Compute-Ausbau</p> <ul style="list-style-type: none"> <li>• Wie bei Ambitionslevel "mittel", aber zusätzlich Kapazität für das Training und die Inferenz eigener Spitzenmodelle mit allgemeiner Verwendung</li> </ul> <p>Benötigte GPUs: 6.000.000 H100-Äquivalente (= 2,7% der globalen Kapazität gemäß unserer Schätzung)</p> <ul style="list-style-type: none"> <li>• davon z. B. 3.000.000 H100-Äquivalente in einem zentralen Supercluster für das Training von Spitzenmodellen</li> <li>• der Rest in kleineren und mittleren Clustern für Inferenz und das Training branchenspezifischer Modelle</li> </ul>	<ul style="list-style-type: none"> <li>• Die Bundesregierung sieht KI als Technologie mit tiefgreifenden geopolitischen Auswirkungen, vergleichbar mit einer "Industriellen Revolution im Zeitraffer". Sie hält eine Allgemeine Künstliche Intelligenz (AGI) vor Ende des Jahrzehnts für möglich.</li> <li>• "Whatever it takes"-Mentalität: Die Bundesregierung erklärt KI-Souveränität zur obersten Priorität, schwört alle relevanten Akteure darauf ein und setzt alle Hebel in Bewegung, um dieses Ziel mit "Warp Speed" zu erreichen.</li> <li>• Die Bundesregierung bringt in wenigen Wochen ein "KI-Beschleunigungsgesetz" durch den Bundestag, das ausgewählte KI-Rechenzentren als im "überragenden öffentlichen Interesse" einstuft, geeignete Flächen ausweist und Genehmigungsprozesse auf max. sechs Monate verkürzt.</li> <li>• Verfügbare IT-Anschlussleistung für KI-Rechenzentren von ca. 5,9 GW.</li> <li>• Die Bundesregierung fördert KI-Rechenzentren und das breitere KI-Ökosystem mit einem Sondervermögen im niedrigen dreistelligen Milliardenbereich.</li> <li>• Die Bundesregierung setzt eine Taskforce mit allen relevanten Bundesministern/-ministerinnen unter Leitung des Kanzleramtsministers ein, die den Infrastruktur-Ausbau über die Ressorts hinweg koordiniert (angelehnt an die "Bunkerrunde" zum Bau von LNG-Terminals während der Energiekrise 2022).</li> </ul>

**Tabelle 2:** Je mehr KI-Souveränität Deutschland anstrebt, desto ambitionierter sollte die nationale Compute-Strategie sein. Damit diese nicht ins Leere läuft, sollten auf politischer Ebene verschiedene Mindestbedingungen gegeben sein.

## 2. Strategien

**Selbst bei einem niedrigen Ambitionslevel gemäß Tabelle 2 müsste Deutschland seine KI-Rechenkapazität wesentlich stärker ausbauen als aktuell vorgesehen.** Werden alle derzeit bekannten Projekte realisiert, würde Deutschland seine KI-Rechenkapazität in den nächsten zwei bis drei Jahren auf höchstens 225.000 H100-Äquivalente steigern. Das ist nur etwa ein Viertel der 850.000 H100-Äquivalente, die Strategie 1 als Ziel vorsieht. Unabhängig vom geplanten Ambitionslevel besteht daher dringender Handlungsbedarf.

**Deutschland sollte unabhängig von seinem Ambitionslevel über mindestens 200.000 H100-Äquivalente in souveränen KI-Rechenzentren verfügen.** Souveräne Rechenzentren (siehe Box 1) sind essentiell, um KI in sensiblen oder sicherheitskritischen Sektoren zu nutzen. Dazu gehören der Einsatz von KI in der kritischen Infrastruktur, der ein Höchstmaß an Souveränität und Ausfallsicherheit verlangt, oder die Verarbeitung sensibler Daten, wie z. B. sensible Geschäftsdaten in der Industrie. Für diese Bereiche braucht Deutschland eigene Trainings- und Inferenz-Kapazität, die höchsten Sicherheitsstandards entspricht. Der genaue Bedarf lässt sich aufgrund von Unsicherheiten über technische Faktoren und das Diffusionstempo von KI nur schwer präzise ermitteln. Es wäre allerdings ein strategisches Versäumnis, deshalb gar nicht zu handeln. Angesichts des rasanten KI-Fortschritts und der wachsenden Bedeutung von KI für staatliche und wirtschaftliche Schlüsselfunktionen liegt das größere Risiko darin, in Kernbereichen gefährliche Abhängigkeiten einzugehen. Deutschland sollte deshalb bis spätestens 2028 eine souveräne Mindestkapazität von 200.000 H100-Äquivalenten schaffen.

**Je ambitionierter die Compute-Strategie, desto mehr politische Voraussetzungen müssen erfüllt sein.** Der massive Ausbau von KI-Recheninfrastruktur gemäß Strategie 3 ist ein Vorhaben von nationaler Tragweite, das mit einem hohen finanziellen und logistischen Aufwand einhergeht. Mit einer "Business-as-usual"-Mentalität ist so ein Vorhaben zum Scheitern verurteilt. Erfolgreich wäre es nur, wenn die Bundesregierung KI als geopolitisches Schlüsselthema behandeln und alle relevanten Akteure aus Politik und Wirtschaft darauf einschwören würde, KI-Infrastruktur über die nächsten Jahre mit der höchsten Priorität auszubauen. Unabhängig vom Ambitionslevel trägt der Ausbau außerdem nur dann Früchte, wenn die Bundesregierung das gesamte KI-Ökosystem umfassend stärkt. Nur im Zusammenspiel mit weiteren Faktoren wie Kapital und Talent wird Rechenkapazität zum Katalysator eines starken KI-Standorts. Sind diese politischen Voraussetzungen nicht erfüllt, droht der Ausbau von Rechenkapazität zu versanden oder im schlimmsten Fall teure KI-Rechenzentren zu hinterlassen, deren Potential mangels eines starken KI-Ökosystems nicht ausgeschöpft wird.

**Eigene Spitzenmodelle sind auch als europäisches Verbundprojekt mit geteilter Infrastruktur denkbar.** Der Kapitalaufwand für das Training eigener Spitzenmodelle (Strategie 3) ist sehr hoch. Als drittgrößte Volkswirtschaft ist Deutschland grundsätzlich fähig, die dafür nötigen Ressourcen bereitzustellen. Die dafür notwendige Kraftanstrengung ließe sich allerdings reduzieren, indem Deutschland souveräne Spitzenmodelle gemeinsam mit anderen Staaten entwickelt. Dafür kommen primär andere europäische Staaten in Frage – entweder als bilaterale Partnerschaft, z. B. mit Frankreich oder Großbritannien, oder als EU-weites Verbundprojekt –, aber auch Staaten wie Kanada, Australien, Japan oder Südkorea. Eine Kooperation würde Ressourcen bündeln und könnte von unterschiedlichen Standortfaktoren profitieren – zum Beispiel von verlässlicher Nuklearenergie in Frankreich oder günstiger Wasserkraft in Norwegen. Um in den nächsten Jahren international konkurrenzfähige Spitzenmodelle zu trainieren, müsste ein solches Verbundprojekt jedoch deutlich über die fünf KI-Gigafabriken hinausgehen, die laut "Apply AI Strategy" der Europäischen Kommission für Fortschritte im Bereich Spitzen-KI vorgesehen sind ([European Commission 2025c](#)).

**Deutschland muss sich angesichts einer unsicheren Zukunft entscheiden, für welche KI-Szenarien es sich wappnen will.** Es ist nicht auszuschließen, dass sich der KI-Fortschritt in den nächsten Jahren verlangsamt oder sogar ein Plateau erreicht. Selbst dann könnten aber Hilfssysteme (*scaffolding*) und eine effizientere Einbettung in bestehende Prozesse dafür sorgen, dass KI einen vergleichbaren Einfluss auf die Weltwirtschaft hat wie das Internet. Aktuell entwickeln sich KI-Fähigkeiten allerdings rasant – in einigen Bereichen womöglich sogar exponentiell ([METR 2025](#)).

# 2. Strategien

Viele Fachleute – darunter der meistzitierte KI-Forscher der Welt, Yoshua Bengio, und der Nobelpreisträger und KI-„Urvater“ Geoffrey Hinton – halten es für möglich, dass eine Allgemeine Künstliche Intelligenz (AGI)<sup>11</sup> vor Ende des Jahrzehnts erreicht wird. Damit würde Intelligenz in einem nie dagewesenen Maß kopierbar und damit skalierbar – ein manchmal als “Millionen Genies in einem Datenzentrum” beschriebenes Szenario. Der Einfluss von KI könnte dann mit der Industriellen Revolution vergleichbar sein: eine systemverändernde Disruption wirtschaftlicher und gesellschaftlicher Gleichgewichte, die sich noch dazu im Zeitraffer vollziehen würde. Eine AGI könnte zudem in der Lage sein, sich selbst stetig zu verbessern – und dadurch eine Superintelligenz (ASI) schaffen, die Menschen kognitiv in allen Hinsichten übertreffen würde (Kokotajlo et al. 2025). Die damit verbundenen Veränderungen könnten präzedenzlos sein. Angesichts dieser Bandbreite an möglichen Szenarien ergibt sich eine strategische Unsicherheit: Für welches Maß an Veränderung sollte sich Deutschland wappnen?

**Unterschiedliche Compute-Strategien sind gegenüber verschiedenen Zukunftsszenarien unterschiedlich robust.** Strategie 1 und 2 könnten insbesondere dann erfolgreich sein, wenn der Einfluss von KI auf die Welt disruptiv, aber nicht transformativ ist – in etwa vergleichbar mit dem Internet. In diesem Fall ermöglichen sie es, zu vertretbaren Kosten an der globalen KI-Wertschöpfung zu partizipieren. Sie gehen aber auch mit einem hohen Risiko einher: Ist der Einfluss von KI transformativer als frühere Technologien, könnte Deutschland ohne ausreichende Inferenz-Kapazität und eigene Grundlagenmodelle irreversibel abgehängt werden – vergleichbar mit einer Gesellschaft, die sich der Dampfmaschine oder dem Ackerbau verweigert hätte. Strategie 3 ist robuster: Ist der Einfluss von KI mit dem Internet vergleichbar, könnte sich Deutschland als führende KI-Nation Wohlstand und globalen Einfluss sichern, wenn auch zu potentiell höheren Opportunitätskosten. Führt KI hingegen zu noch tiefgreifenderen Veränderungen, erscheint ein ambitionierter Compute-Ausbau gemäß Strategie 3 als der einzige Weg, um in Zukunft souverän und wettbewerbsfähig zu sein (siehe Tabelle 3).

	Einfluss von KI auf die Welt		
	1. disruptiv (≈Internet)	2. transformativ (≈Industrielle Revolution, im Zeitraffer)	3. zivilisationsverändernd (≈Neolithische Revolution, im Zeitraffer)
Strategie 1 (niedrige Ambition)	Wertschöpfung primär im Ausland, aber Partizipation an wirtschaftlicher Entwicklung	Abhängigkeit von ausländischen Technologieführern	Abhängigkeit von ausländischen Technologieführern
Strategie 2 (mittlere Ambition)	Führend in Schlüsselsektoren	Teilweise Abhängigkeit	Weitgehende Abhängigkeit
Strategie 3 (hohe Ambition)	Wachstum und Souveränität, ggf. hohe Opportunitätskosten	Wachstum und Souveränität	Wachstum und Souveränität

**Tabelle 3:** Die verschiedenen Compute-Strategien beeinflussen die deutsche Wettbewerbsfähigkeit und geopolitische Souveränität unterschiedlich – abhängig davon, welchen Einfluss KI auf die Gesellschaft, Wirtschaft, Verwaltung und internationale Ordnung haben wird.

<sup>11</sup> Der Begriff AGI hat keine kanonische Definition, meint im Allgemeinen aber eine KI, die dem Menschen in kognitiver Hinsicht ebenbürtig ist.

# 3. Empfehlungen

Beim Ausbau von KI-Rechenkapazität kommt dem Staat eine besondere Verantwortung zu: Angesichts der hohen Kosten für neue Infrastruktur und regulatorischer bzw. technologischer Unsicherheiten sind die Investitionsanreize für einzelne Unternehmen derzeit limitiert. Die Bundesregierung sollte den Privatsektor aufgrund der langfristigen strategischen Bedeutung von KI daher gezielt unterstützen. Sechs Prioritäten sollten dabei im Zentrum stehen.

## 1. Kapazität für KI-Training und -Inferenz gemäß strategischer Prioritäten ausbauen

Deutschlands Pläne zum Ausbau von KI-Infrastruktur sind bisher nicht ambitioniert genug. Selbst wenn alle angekündigten Projekte realisiert würden, würde die KI-Rechenkapazität in den nächsten 3 Jahren nur auf etwa 175.000-250.000 H100-Äquivalente steigen. Im internationalen Vergleich läge Deutschland damit voraussichtlich hinter Staaten wie den USA, China, den Vereinigten Arabischen Emiraten, Südkorea, Saudi-Arabien, Frankreich, dem Vereinigten Königreich und Indien. Der Bau einer europäischen KI-Gigafabrik in Deutschland wäre ein erster Schritt und sollte bei einer positiven Bewerbung schnellstmöglich umgesetzt werden, bliebe aber deutlich hinter dem prognostizierten Bedarf zurück.

### Empfehlungen:

- Die Bundesregierung sollte angesichts der oft mehrjährigen Planungs- und Bauphase von KI-Rechenzentren jetzt damit beginnen, deren Ausbau gezielt zu fördern. Angesichts der langfristigen strategischen Bedeutung von KI sollten dabei auch Public-Private-Partnerships genutzt werden, um finanzielle Risiken für Privatunternehmen zu begrenzen und Investitionen zu stimulieren.
- Parallel dazu sollte sie strategische Klarheit darüber schaffen, wie ambitioniert und mit welchem Schwerpunkt (Inferenz bzw. Training) die KI-Rechenkapazität bis zum Ende der Legislaturperiode ausgebaut werden soll (siehe Abschnitt 2).

## 2. Politische Voraussetzungen schaffen und Kompetenzen bündeln

Der Ausbau von KI-Infrastruktur ist ein über Jahre andauerndes Milliarden-Projekt, dessen Erfolg von politischer Ambition, einer klaren Strategie und gebündelter Koordinierung abhängt. Aktuell mangelt es an diesen Faktoren. Zwar herrscht über die strategische Relevanz von KI im politischen Diskurs weitgehend Konsens und die Bedeutung eigener Grundlagenmodelle wird explizit benannt ([Welt 2025](#)). Aber die infrastrukturellen und politischen Voraussetzungen für ein derart ambitioniertes Ziel sind aktuell bei Weitem nicht gegeben. Der Ausbau von Inferenz-Kapazität, die es auch für weniger ambitionierte Strategien braucht, wird ebenfalls noch nicht systematisch unterstützt. Zudem wird KI-Infrastruktur noch nicht ausreichend als ein ressortübergreifendes, zentral und mit hoher Priorität zu steuerndes Thema behandelt. Das Beispiel der LNG-Terminals während der Energiekrise 2022 zeigt, dass sich strategisch wichtige Infrastrukturprojekte auch in kurzer Zeit realisieren lassen, wenn dies aus dem Bundeskanzleramt heraus koordiniert wird.

### Empfehlungen:

- Je nach Strategie sollte die Bundesregierung die in Tabelle 2 genannten politischen Voraussetzungen schaffen. Dazu gehört u. a. ein in der gesamten Regierung herrschendes Bewusstsein über die strategische Relevanz von KI, ein Ausbau der für KI-Rechenzentren notwendigen Energieinfrastruktur, vereinfachte Regulatorik sowie eine gezielte Förderung des KI-Ökosystems.
- Eine zentrale Instanz – z. B. ein KI-Infrastruktur-Board im Bundeskanzleramt ([KI Bundesverband 2024](#), [Bitkom 2025](#)) – sollte den Ausbau von Rechen- und Energieinfrastruktur steuern. Sie sollte Zuständigkeiten über die relevanten Ressorts (Digitales, Energie, Wirtschaft, Inneres, Forschung) hinweg koordinieren, bürokratische Hürden abbauen und eng mit europäischen Initiativen wie EuroHPC zusammenarbeiten.
- Das Bundeskanzleramt sollte eine verlässliche Versorgung mit modernsten US-amerikanischen KI-Chips sicherstellen und sich dafür einsetzen, US-Chips weiterhin für souveräne KI-Rechenzentren erwerben zu können – statt ausnahmslos in Rechenzentren, die von US-Firmen betrieben werden, wie zuletzt einige Berichte nahelegten ([Export Compliance Daily 2025](#)).

# 3. Empfehlungen

## 3. Günstige, verlässliche Energie für KI-Rechenzentren bereitstellen

Je nach Strategie benötigt Deutschland in der Zukunft insgesamt eine IT-Anschlussleistung für KI-Rechenzentren von 0,8 GW, 3,4 GW bzw. 5,9 GW (siehe Abschnitt 2). Um diesen Bedarf zu decken, sollte Deutschland so schnell wie möglich die nötige Energieinfrastruktur ausbauen – ein Unterfangen, das selbst unter den besten Voraussetzungen Jahre dauern wird. Darüber hinaus sollte Strom nicht nur verfügbar, sondern auch wettbewerbsfähig bepreist sein.

### Empfehlungen:

- Der Ausbau der KI-relevanten Energieinfrastruktur sollte als nationale Priorität zentral vom Bundeskanzleramt aus gesteuert werden.
- KI-Rechenzentren sollten (wie im Koalitionsvertrag angekündigt) in die Strompreiskompensation einbezogen werden. Zudem sollten sie über einen Industriestrompreis gefördert werden ([eco 2025](#)). Die Bundesregierung sollte dafür die beihilferechtlichen Voraussetzungen schaffen, z. B. durch die Aufnahme der KI-Rechenzentrenbranche in die KUEBLL-Liste der EU ([eco 2023](#)).
- Darüber hinaus sind Steuervergünstigungen für KI-Rechenzentren denkbar, wie sie z. B. in den USA und einigen europäischen Staaten üblich sind ([Lorentz et al. 2025](#)).
- Im Rahmen möglicher Verbundprojekte sollte die Bundesregierung überprüfen, einen Teil der in Deutschland benötigten Rechenkapazität in anderen EU-Ländern anzusiedeln – vorzugsweise in Ländern mit mehr bzw. günstigerer Energie.

## 4. Planungs- und Genehmigungsprozesse beschleunigen

Langwierige Planungs- und Genehmigungsprozesse sind eine große Hürde beim Bau von KI-Rechenzentren in Deutschland. Wer Milliardeninvestitionen in KI-Infrastruktur erwägt, muss sich darauf verlassen können, dass geplante Vorhaben nicht jahrelang durch bürokratische Prozesse verzögert werden. In anderen EU-Ländern wie Dänemark werden Genehmigungen teils in nur wenigen Monaten erteilt, statt wie in Deutschland erst nach zwei Jahren oder mehr ([FAZ 2025](#), [German Datacenter Association 2024](#)). Der Anschluss eines neuen Rechenzentrums an das Stromnetz dauert in Deutschland ebenfalls lange: laut der Internationalen Energieagentur bis zu 7 Jahre, verglichen mit 3-5 Jahren in Spanien und 1-3 Jahren in den USA ([IEA 2025](#)).

### Empfehlungen:

- Die Bundesregierung sollte "Special Compute Zones" (SCZs) einrichten, in denen eine vereinfachte Regulatorik für neue KI-Rechenzentren und die dafür nötige Energieinfrastruktur gilt ([Wiseman, McClements & Horsley 2024](#), [Juijn et al. 2025](#), [Datta & First 2025](#)). Die Bundesregierung würde dazu gemeinsam mit Ländern und Kommunen besonders geeignete Flächen identifizieren (z. B. stillgelegte Industriestandorte mit vorhandener Netzkapazität). Für Genehmigungen in diesen Flächen wäre eine einzelne Behörde zuständig und es würde eine "presumption of conformity" gelten: Legt die zuständige Behörde nicht innerhalb einer festen Frist (z. B. 3-6 Monate) einen begründeten Widerspruch ein, gilt das Projekt als genehmigt.
- Auch jenseits von SCZs sollten Planungs- und Genehmigungsprozesse vereinfacht und beschleunigt werden, z. B. indem KI-Rechenzentren als im "überragenden öffentlichen Interesse" eingestuft, verschiedene Verfahren (wie Umweltverträglichkeitsprüfung und Baugenehmigung) parallelisiert und KI-Rechenzentren in die Liste bauplanungsrechtlich privilegierter Außenbereichsvorhaben aufgenommen werden ([BMWK 2024](#)).
- Die Bundesregierung sollte die Eignung bundeseigener Grundstücke für den Bau von KI-Rechenzentren überprüfen.

# 3. Empfehlungen

- Vergabeverfahren für die Netzanbindung sollten ebenfalls vereinfacht werden und eine prioritäre Behandlung strategisch wichtiger KI-Rechenzentren erlauben. Die Netzanbindung könnte insbesondere für KI-Rechenzentren beschleunigt werden, deren Betreiber sich verpflichten, bei Spitzenlast im Stromnetz ihren Energiebedarf zu reduzieren. Daten aus den USA zeigen, dass dies nur an wenigen Tagen im Jahr nötig wäre, um eine signifikante Menge zusätzlicher Rechenkapazität ans Netz zu bringen ([Norris et al. 2025](#)).
- Die Bundesregierung sollte zur Entlastung des bestehenden Netzes "Behind-the-meter"-Projekte unterstützen, bei denen KI-Rechenzentren durch eigene, vom Betreiber selbst errichtete Kraftwerke versorgt werden.

## 5. Sicherheit von KI-Rechenzentren erhöhen

Je leistungsfähiger KI-Modelle werden und je stärker sie in Wirtschaft und Verwaltung integriert werden, desto attraktiver werden sie als Ziel von Cyberattacken. Ein Bericht des Beratungsunternehmens Gladstone kommt zu dem Ergebnis, dass aktuell kein KI-Rechenzentrum ausreichend gegen Angriffe geschützt ist – insbesondere, wenn diese von bestens ausgestatteten, staatlich finanzierten Gruppen ausgehen ([Harris & Harris 2025](#)). Neben Sabotage droht der Diebstahl sensibler Daten bis hin zu den Modellparametern selbst. Wer diese durch einen physischen oder Cyberangriff in seinen Besitz bringt, könnte anschließend auf die Fähigkeiten des Modells zugreifen – und diese z. B. für militärische Zwecke missbrauchen. Weil es einem Bericht der RAND Corporation zufolge mehrere Jahre dauern wird, KI-Rechenzentren effektiv vor staatlichen Cyberangriffen zu schützen, sollte die Sicherheit hochsensibler Daten von Anfang an mitgedacht werden ([Nevo et al. 2024](#)). Tut Deutschland das nicht, könnten andere Staaten den Betrieb ihrer Spitzenmodelle auf deutschen KI-Rechenzentren untersagen, um nicht den Verlust wertvollen geistigen Eigentums zu riskieren.

### Empfehlungen:

- KI-Rechenzentren, die Spitzenmodelle betreiben oder für Anwendungen in der kritischen Infrastruktur relevant sind (z. B. Energieversorgung, Gesundheitswesen, Finanzmärkte), sollten absehbar mit einer ausreichend hohen Sicherheitsstufe ausgestattet sein – z. B. RAND SL4 oder, in hochsensiblen Fällen, SL5 (vgl. [Nevo et al. 2024](#)).
- Weil bisher kein Rechenzentrum auf der Welt diese hohen Sicherheitsstufen erfüllt, sollte die Bundesregierung relevante Forschung und Pilotprojekte im Bereich IT-Sicherheit unterstützen, um Deutschland als Vorreiter für hochsichere Rechenzentren zu etablieren.
- Staatliche Fachleute für IT-Sicherheit sollten eng mit privaten Betreibern zusammenarbeiten, um KI-Rechenzentren (auch auf Basis nicht-öffentlicher Informationen) vor möglichen Angriffen zu schützen.
- Deutsche Fachbehörden (BSI, Cyberagentur) und Forschungseinrichtungen sollten sich auf internationaler Ebene aktiv an der Erarbeitung, Umsetzung und Prüfung von Sicherheitsstandards für KI-Rechenzentren beteiligen.

## 6. KI-Ökosystem umfassend stärken

Der Ausbau von KI-Recheninfrastruktur ist nur dann sinnvoll, wenn es parallel dazu ein wachsendes Ökosystem an Organisationen gibt, die das Potential neuer Rechenkapazität tatsächlich nutzen. Zwar gibt es in Deutschland einige Positivbeispiele wie die Start-ups DeepL oder Parloa, und auch in der Industrie spielen Industrial Foundation Models eine zunehmende Rolle. Dennoch ist die viel beschworene "KI made in Germany" gerade eher ein Buzzword als eine ökonomische Realität.

### Empfehlungen:

- *Wirtschaft:*
  - Pensionskassen, Versicherungen und Förderbanken sollten leichter in Tech-Startups investieren können, um dem akuten Mangel an Risikokapital entgegenzuwirken.
  - Die Bundesregierung sollte die Startup- und Scaleup-Strategie der EU unterstützen, die den europäischen Kapitalmarkt stärker vereinheitlichen, Investitionsflüsse erleichtern und regulatorische Unsicherheit beseitigen soll ([European Commission 2025b](#)).

# 3. Empfehlungen

- KI-Adoption in Unternehmen sollte gezielt gefördert werden, z. B. durch Sonderabschreibungen für KI-Investitionen oder eine bürokratiearme Umsetzung der europäischen KI-Verordnung.
- Vereinfachte Visa-Prozesse sollten genutzt werden, um den deutschen KI-Standort für internationales Top-Talent attraktiv zu machen.
- *öffentliche Verwaltung:*
  - Die öffentliche Hand sollte bei technischer Gleichwertigkeit als Ankerkunde für heimische KI-Lösungen auftreten, um für vielversprechende Unternehmen verlässliche Umsätze in der Wachstumsphase zu generieren. Öffentliche Beschaffungsregeln sollten möglichst einfach ausgestaltet sein.
  - Die Bundesregierung könnte zudem den Bau von KI-Rechenzentren attraktiver machen, indem sie Abnahmegarantien für einen Teil der Rechenleistung übernimmt und diese nutzt, um die öffentliche Verwaltung effizienter und zukunftsfähiger zu machen.
- *Forschung:*
  - Der Transfer zwischen universitärer Forschung und wirtschaftlicher Anwendung sollte vereinfacht werden, z. B. durch leichtere Ausgründungen, einen schnellen und unbürokratischen Zugang zu öffentlicher Rechenleistung für Start-ups, einen attraktiven Rechtsrahmen für Mitarbeiterbeteiligungen sowie eine vereinfachte Regulatorik, insbesondere für kleine Unternehmen.
  - Die öffentliche Forschungsförderung sollte ambitionierte Moonshot-Projekte mit europäischen Partnerländern in den Blick nehmen, damit der nächste technische Durchbruch für leistungsstarke sowie sichere und verlässliche KI aus Europa stammt. Auch bilaterale Kooperationen mit Ländern wie Großbritannien, Kanada, Japan oder Australien sind denkbar.

# 4. Fazit

Deutschland steht vor einer kritischen Weichenstellung. KI-Rechenkapazität ist eine strategische Kernressource, die die deutsche Wettbewerbsfähigkeit und geopolitische Souveränität über Jahrzehnte hinweg bestimmen wird. Aktuell droht Deutschland im internationalen Vergleich zurückzufallen.

Aber noch ist ein Umsteuern möglich. Entscheidend sind dabei sechs Hebel: der Ausbau von Trainings- und Inferenz-Kapazität gemäß strategischer Prioritäten, eine Bündelung politischer Kompetenzen, vereinfachte Regulatorik, verlässliche und günstige Energie, der Schutz von KI-Rechenzentren vor Sabotage und Diebstahl sowie die umfassende Stärkung des KI-Ökosystems.

# Technischer Appendix

Dieser Appendix erklärt, wie die im Abschnitt 2 genannten GPU-Kapazitäten bzw. IT-Anschlussleistungen für die drei Compute-Strategien berechnet wurden.

## A.1 Berechnung von H100-Äquivalenten

Dieser Bericht gibt KI-Rechenkapazität primär in H100-Äquivalenten an. Dabei handelt es sich um ein gängiges Maß, um die Rechenleistung verschiedener GPU-Typen zu vergleichen. Wir nutzen eine gängige Umrechnungsformel (Pilz, Sanders et al. 2025, S. 18). Diese teilt die Leistung einer beliebigen GPU in der niedrigsten unterstützten Präzision (aus 32-, 16- oder 8-Bit) durch die Leistung einer Nvidia H100-GPU in 8-Bit-Präzision. Die Umrechnung liefert einen groben Referenzwert, um die Kapazität von KI-Rechenzentren miteinander ins Verhältnis zu setzen. Vor allem für jüngere Rechenzentren, die 2021 und später entstanden sind, liefert diese Umrechnungsformel sinnvolle Ergebnisse (Pilz et al. 2025).

## A.2 Berechnung der nötigen IT-Anschlussleistung

Wir verwenden folgendes Produkt, um die für den Betrieb einer bestimmten Anzahl an GPUs benötigte Menge an elektrischer Leistung zu berechnen (Pilz, Sanders et al. 2025, S. 25):

$$\text{Thermal Design Power} \times \text{Anzahl der GPUs} \times \text{Overhead-Faktor} \times \text{Power Usage Effectiveness}$$

Die Thermal Design Power (TDP) bezieht sich auf die maximale Menge an Wärme, die eine GPU unter Betriebsbedingungen erzeugen kann und variiert je nach GPU-Typ. Bei einer Nvidia H100 liegt sie bei 700 Watt (Patel & Nishball 2024). Ein Overhead-Faktor berücksichtigt weitere Hardware neben GPUs und wird hier pauschal als 2,03 angenommen. Die Power Usage Effectiveness (PUE) ist eine Kennzahl für die Energieeffizienz eines Rechenzentrums und wird hier pauschal als 1,1 angenommen.

## A.3 Global verfügbare KI-Rechenkapazität

Die global verfügbare KI-Rechenkapazität wächst jedes Jahr aufgrund von zwei Faktoren: 1) Chips werden stetig effizienter (d. h. sie können pro Zeiteinheit mehr Berechnungen durchführen), 2) Chip-Hersteller wie TSMC weiten ihre Produktion stetig aus. Fachleute nehmen an, dass die Chip-Effizienz um den Faktor 1,35x/Jahr wächst und die Chip-Produktion um den Faktor 1,65x/Jahr (Epoch AI 2025b, Kokotajlo et al. 2025). Aktuell beträgt die globale GPU-Kapazität etwa 10 Mio. H100-Äquivalente (Kokotajlo et al. 2025). Setzt sich der aktuelle Trend fort, ergeben sich bei einer Gesamtwachstumsrate von 2,25x/Jahr bis Ende 2028 225 Mio. H100-Äquivalente. Diese Schätzung ist unsicher, weil sie nur auf öffentlich verfügbaren Informationen basiert und sich die zugrunde liegenden Hardware-Trends verlangsamen oder beschleunigen könnten. Für die Berechnung von GPU-Kapazitäten/IT-Anschleisungen in diesem Bericht spielt sie keine Rolle. Ihr Zweck ist allein zu illustrieren, welchen Anteil Deutschland je nach Strategie an der globalen Rechenkapazität hätte (siehe Tabelle 2).

## A.4 Benötigte GPU-Kapazität unter Strategie 1 & 2

Das britische Wissenschaftsministerium (DSIT) hat McKinsey beauftragt, die Nachfrage nach KI-Rechenleistung im Vereinigten Königreich bis 2035 auf Grundlage eines volkswirtschaftlichen Modells zu prognostizieren (DSIT 2025). Weil das McKinsey-Modell proprietär ist, sind nicht alle Details öffentlich einsehbar. Es fußt aber – wie auch Strategie 1 und 2 im hier vorliegenden Bericht – auf der Annahme, dass Rechenleistung vor allem für Inferenz sowie für das Training branchenspezifischer Modelle benötigt wird, nicht aber für das Training allgemeiner Spitzenmodelle (DSIT 2025, S. 6, S. 14).

Wir wenden die Ergebnisse des McKinsey-Modells auf den deutschen Kontext an und betrachten dazu das "Medium"-Szenario, in dem die Nachfrage nach KI-Rechenkapazität jährlich um 27% wächst. Zu einem ähnlichen Ergebnis kommt eine Studie von Deloitte über KI-Infrastruktur in Deutschland: Dort ergibt ein S-Kurven-Modell zur Diffusion neuer Technologien, dass Deutschlands Bedarf an KI-Rechenkapazität bis 2030 um jährlich 24,5% wachsen wird (Lorentz et al. 2025).

# Technischer Appendix

Wir verwenden hier die etwas höhere Wachstumsrate aus dem McKinsey-Modell, wodurch sich für Deutschland – ausgehend von einer IT-Anschlussleistung von 1,6 GW im Jahr 2025<sup>12</sup> – bis Ende 2028 ein Bedarf von 4,2 GW ergibt. Wir nehmen an, dass dieser Gesamtbedarf durch ca. 1,9 Mio. B200-GPUs gedeckt wird.<sup>13</sup> Bei einer Thermal Design Power von 1000 W (Patel & Nishball 2024) ergibt sich unter Verwendung der in A.2 genannten Formel die o. g. Gesamtleistung von 4,2 GW.

Für die Berechnung der H100-Äquivalente ergibt sich ein Umrechnungsfaktor von 2,27: Die 8-Bit-Rechenleistung eines H100 entspricht 1.979 TFLOP/s; die eines B200 4.500 TFLOP/s (Patel & Nishball 2024). Der für Deutschland relevante Gesamtbedarf liegt in diesem Fall bei ca. 4,3 Mio. H100-Äquivalenten.

Strategie 1 deckt ca. 20% der deutschen Nachfrage (0,84 GW) durch eigene KI-Rechenzentren ab. (Die restliche Nachfrage würde, sofern möglich, durch ausländische Cloud-Dienste gedeckt werden und andernfalls ungedeckt bleiben.) Dafür würde Deutschland bis Ende 2028 etwa 850.000 H100-Äquivalente benötigen. Dies entspricht im Jahr 2028 ca. 0,4% der globalen KI-Rechenkapazität (bei 225 Mio. H100-Äquivalenten insgesamt – siehe A.3).

Strategie 2 deckt ca. 80% der deutschen Nachfrage (3,36 GW) durch eigene KI-Rechenzentren ab. Dafür würde Deutschland bis Ende 2028 ca. 1,5 Mio. B200-GPUs benötigen, d. h. 3,4 Mio. H100-Äquivalente. Dies entspricht zu diesem Zeitpunkt ca. 1,5% der globalen KI-Rechenkapazität gemäß unserer Schätzung.

## A.5 Benötigte GPU-Kapazität unter Strategie 3

Grundlage für Strategie 3 sind eigene Berechnungen basierend auf einem umfangreichen Datensatz des KI-Forschungsinstituts Epoch AI (Epoch AI 2025a). In einem ersten Schritt schätzen wir die benötigte Rechenleistung, um Ende 2028 ein Spitzenmodell mit allgemeiner Anwendung (*frontier model*) zu trainieren. Diese Rechenleistung (*training compute*) ist seit 2010 etwa um den Faktor 4x/Jahr gewachsen. Es ist Fachleuten zufolge nicht nur technisch möglich, sondern auch wahrscheinlich, dass sich dieser Trend bis zum Ende des Jahrzehnts fortsetzt (Sevilla et al. 2024, Epoch AI 2025c).

Allerdings verlangt dieses enorm schnelle Wachstum immer größere Investitionen in KI-Rechenzentren und Energieversorgung. Allein OpenAI plant, in den nächsten vier Jahren 500 Mrd. Dollar zu investieren (OpenAI 2025). Es ist ungewiss, ob diese Ambition wirtschaftlich tragfähig ist oder technische Hürden die Skalierung von Rechenleistung stärker ausbremsen als erwartet. So erwarten manche Fachleute, dass sich das Wachstum auf etwa 3,5x im Jahr verlangsamen wird (Greenblatt 2025). Sie könnte aber auch deutlich niedriger liegen.

Bei einem hohen Ambitionslevel gemäß unseres Vorschlags würde die Bundesregierung – wie auch die derzeit führenden KI-Unternehmen – davon ausgehen, dass sich der historische Wachstumstrend in den nächsten Jahren in etwa fortsetzt. Wir nehmen für Strategie 3 deshalb an, dass die jährliche Wachstumsrate der Trainings-Rechenleistung zwischen 3,5x und 4x liegt. Bei einer 16-Bit-Präzision im Training (Fist & Datta 2024) und einer GPU-Auslastung von 40%<sup>14</sup> impliziert das, dass Ende 2028 der Trainingslauf eines Spitzenmodells eine Rechenleistung von  $1,5\text{-}2,3 \times 10^{28}$  FLOP erfordert. Bei einer Trainingsdauer von 258 Tagen<sup>15</sup> und einer Leistung von 989 TFLOP/s einer einzelnen H100-GPU entspricht das ca. 1,7-2,6 Mio. H100-Äquivalenten. Ein KI-Rechenzentrum dieser Größenordnung, in dem auf B200-GPUs trainiert wird, benötigt eine IT-Anschlussleistung von 1,7-2,5 GW.

12 Es handelt sich hierbei um Hyperscale- und Co-Location-Rechenzentren, die aufgrund ihrer Bauweise grundsätzlich für KI-Workloads geeignet sind (Lorentz et al. 2025).

13 Der Nvidia B200 ist der Nachfolge-Chip des Nvidia H100, der sowohl für Training als auch Inferenz eine bessere Performance liefert (NVIDIAc 2025). Wir gehen davon aus, dass in Deutschland in neuen KI-Rechenzentren bis Ende 2028 vor allem B200-Chips zum Einsatz kommen werden. Weil der B200 in Sachen Stromverbrauch und Performance zwischen dem H100 und noch stärkeren Chips wie dem B300 liegt, die ebenfalls zum Einsatz kommen werden, liefert eine Berechnung auf Grundlage von B200-Chips eine sinnvolle Approximation der benötigten GPU- bzw. Stromkapazitäten.

14 Mit "Auslastung" ist der Anteil der theoretisch möglichen Rechenleistung einer GPU gemeint, der tatsächlich für das Training genutzt wird. Faktoren wie eine langsame Datenübertragung zwischen Speicher- und Recheneinheiten führen zu einer Auslastung unterhalb von 100%. Wir nehmen für das Training von Spitzenmodellen eine Auslastung von 40% an (Fist & Datta 2024).

15 In den vergangenen Jahren hat sich bei Spitzenmodellen ein Trend zu immer längeren Trainingsläufen gezeigt (Epoch AI 2025a). Dieser Trend wird sich in den nächsten Jahren wahrscheinlich fortsetzen, weil längere Trainingsläufe Energie sparen (Fist & Datta 2024). Aufgrund schneller Iterationszyklen ist der Trend aber nach oben hin gedeckelt. Fachleute gehen davon aus, dass die Trainingsdauer bis 2027 ihr Maximum von 8,6 Monaten (258 Tage) erreichen wird (Emberson & Edelman 2025).

# Technischer Appendix

In einem zweiten Schritt berechnen wir die benötigte Anzahl an GPUs für Inferenz. Weil sich der gesamtwirtschaftliche Bedarf an Inferenz-Rechenleistung schwer prognostizieren lässt – unsichere Variablen sind u. a. das Diffusionstempo von KI-Technologie und die zukünftige Bedeutung von “inference-time scaling” (Ma et al. 2025) –, nehmen wir an, dass der Inferenz-Anteil an der gesamten Rechenleistung 15%, 25% oder 50% betragen könnte.<sup>16</sup> Berücksichtigt man außerdem ein, dass GPUs während der Inferenzphase eine geringere Auslastung von etwa 15% haben<sup>17</sup>, ergibt sich:

H100-Äquivalente für Frontier-Training	Inferenz-Anteil	H100-Äquivalente für Inferenz	H100-Äquivalente insgesamt
1,7 bis 2,6m	15%	0,8 bis 1,2m	2,5 bis 3,8m
	25%	1,5 bis 2,3m	3,3 bis 4,9m
	50%	4,6 bis 6,9m	6,3 bis 9,5m

**Tabelle A:** Unter Strategie 3 hat Deutschland einen GPU-Bedarf von 2,5 bis 9,5 Mio. H100-Äquivalenten, abhängig von verschiedenen technischen Annahmen über das Wachstum von Trainings-Rechenleistung und die relativen Anteile von Training und Inferenz. Bei der Summenbildung ergeben sich rundungsbedingte Abweichungen.

Der GPU-Wert für Strategie 3 – 6 Mio. H100-Äquivalente – entspricht dem Mittelwert aus den in Tabelle A genannten Mindest- bzw. Höchstwerten von 2,5 bzw. 9,5 Mio. H100-Äquivalenten. 6 Mio. H100-Äquivalente machen Ende 2028 gemäß der Schätzung aus A.3 ca. 2,7% der globalen KI-Rechenkapazität aus. Dies liegt unter dem deutschen Anteil am globalen BIP (nominal) von derzeit ca. 4,2% (IMF 2025).

Unter der Annahme, dass für Training und Inferenz ca. 2,6 Mio. B200-GPUs genutzt werden, ergibt sich daraus eine benötigte IT-Anschlussleistung von insgesamt 5,9 GW (2,5 bis 9,3 GW).

<sup>16</sup> Höhere Inferenz-Anteile erscheinen für diesen Zeitrahmen unplausibel, weil ein in Deutschland trainiertes Spitzenmodell im Vergleich zu den USA anfangs geringere Nutzerzahlen hätte.

<sup>17</sup> Diese Schätzung basiert auf Gesprächen mit Hardware-Experten und deckt sich in etwa mit dem Wert von 17%, den Erdil (2025) für das Modell Llama 3 70B berechnet.

# Literaturverzeichnis

**Berg & Ho (2025).** "After the ChatGPT Moment: Measuring AI's Adoption", Epoch AI Gradient Updates, <https://epochai.substack.com/p/after-the-chatgpt-moment-measuring>

**Bick et al. (2024).** "The Rapid Adoption of Generative AI", National Bureau of Economic Research, Cambridge, MA, [https://www.nber.org/system/files/working\\_papers/w32966/w32966.pdf](https://www.nber.org/system/files/working_papers/w32966/w32966.pdf)

**Bitkom (2025).** "Deutschland zum KI-Hotspot machen: 10 Empfehlungen für eine erfolgreiche KI-Zukunft", Bitkom e. V. <https://www.bitkom.org/sites/main/files/2025-05/bitkom-publikation-deutschland-zum-ki-hotspot-machen.pdf>

**BMWK (2025).** "Stand und Entwicklung des Rechenzentrumsstandorts Deutschland", Gutachten im Auftrag des Bundesministeriums für Wirtschaft und Klimaschutz, [https://www.bundeswirtschaftsministerium.de/Redaktion/DE/Publikationen/Technologie/stand-und-entwicklung-des-rechenzentrumsstandorts-deutschland.pdf?\\_\\_blob=publicationFile&v=10](https://www.bundeswirtschaftsministerium.de/Redaktion/DE/Publikationen/Technologie/stand-und-entwicklung-des-rechenzentrumsstandorts-deutschland.pdf?__blob=publicationFile&v=10)

**Chatterji et al. (2025).** "How People Use ChatGPT", OpenAI, <https://cdn.openai.com/pdf/a253471f-8260-40c6-a2cc-aa93fe9f142e/economic-research-chatgpt-usage-paper.pdf>

**Datta & First (2025).** "Compute in America: A Policy Playbook", Institute for Progress, <https://ifp.org/special-compute-zones/>

**DeepSeek-AI (2024).** "DeepSeek-V3 Technical Report", arXiv, <https://arxiv.org/abs/2412.19437>

**DSIT (2025).** "Compute Evidence Annex", Research Report 2025/025, Department for Science, Innovation & Technology, <https://assets.publishing.service.gov.uk/media/687f74f4fdc190fb6b846868/compute-evidence-annex-final.pdf>

**eco (2023).** "Einordnung des BMWK-Konzeptes für einen Industriestrompreis in das EU-Beihilferecht", eco - Verband der Internetwirtschaft e.V., [https://www.eco.de/wp-content/uploads/2023/09/20230810\\_eco-hintergrundpapier\\_industriestrompreis.pdf](https://www.eco.de/wp-content/uploads/2023/09/20230810_eco-hintergrundpapier_industriestrompreis.pdf)

**eco (2025).** "eco Allianz fordert Industriestrompreis für Rechenzentren", eco - Verband der Internetwirtschaft e.V., <https://www.eco.de/presse/eco-allianz-fordert-industriestrompreis-fuer-rechenzentren/>

**Emberson & Edelman (2025).** "Frontier training runs will likely stop getting longer by around 2027", Epoch AI, <https://epoch.ai/data-insights/longest-training-run>

**Epoch AI (2025a).** "Data on AI Models", Online-Zugriff am 03.09.2025, <https://epoch.ai/data/ai-models>

**Epoch AI (2025b).** "Machine Learning Trends", Online-Zugriff am 03.09.2025, <https://epoch.ai/trends#hardware>

**Epoch AI (2025c).** "AI in 2030: Extrapolating current trends", Online-Zugriff am 18.09.2025, [https://epoch.ai/files/AI\\_2030.pdf](https://epoch.ai/files/AI_2030.pdf)

**Erdil (2025).** "Inference economics of language models", arXiv, <https://arxiv.org/abs/2506.04645>

**EuroHPC (2025).** "Call for expression of interest in AI Gigafactories (AIGFs)", European High Performance Computing Joint Undertaking, [https://www.eurohpc-ju.europa.eu/document/download/47492db7-592e-4ad8-b672-9c822f94afa0\\_en?filename=AI%20GIGAFACTORIES%20CONSULTATION.pdf](https://www.eurohpc-ju.europa.eu/document/download/47492db7-592e-4ad8-b672-9c822f94afa0_en?filename=AI%20GIGAFACTORIES%20CONSULTATION.pdf)

# Literaturverzeichnis

**European Commission (2025a).** "AI Continent Action Plan", Online-Zugriff am 03.09.2025, <https://digital-strategy.ec.europa.eu/en/factpages/ai-continent-action-plan>

**European Commission (2025b).** "EU Startup and Scaleup Strategy", Online-Zugriff am 03.09.2025, [https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/jobs-and-economy/eu-startup-and-scaleup-strategy\\_en](https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/jobs-and-economy/eu-startup-and-scaleup-strategy_en)

**European Commission (2025c).** "Apply AI Strategy", Online-Zugriff am 08.10.2025, <https://digital-strategy.ec.europa.eu/en/policies/apply-ai>

**Export Compliance Daily (2025).** "New AI Diffusion Rule Will Let Allies Buy US Chips With Conditions, Commerce Head Says", [https://exportcompliancedaily.com/article/2025/06/05/new-ai-diffusion-rule-will-let-allies-buy-us-chips-with-conditions-commerce-head-says-2506040042?BC=bc\\_685e6c5ba1f0a](https://exportcompliancedaily.com/article/2025/06/05/new-ai-diffusion-rule-will-let-allies-buy-us-chips-with-conditions-commerce-head-says-2506040042?BC=bc_685e6c5ba1f0a)

**Fist & Datta (2024).** "How to Build the Future of AI in the United States", Institute for Progress, <https://ifp.org/future-of-ai-compute>

**FAZ (2025).** "Das elfköpfige Start-up, das eine AI Gigafactory bauen will", Frankfurter Allgemeine Zeitung, <https://www.faz.net/pro/digitalwirtschaft/kuenstliche-intelligenz/lyceum-start-up-aus-berlin-will-ai-gigafactory-bauen-110586207.html>

**FZ Jülich (2025).** "Supercomputer JUPITER erreicht Rekord-Rechenleistung in Europa", Forschungszentrum Jülich, <https://www.fz-juelich.de/de/aktuelles/news/pressemitteilungen/2025/supercomputer-jupiter-erreicht-rekord-rechenleistung-in-europa>

**German Datacenter Association (2024).** "GDA-Positionspapier zum Konsultationsverfahren BNetzA über ein Verfahren zur Zuteilung von Entnahmeleistungen aus Netzebenen oberhalb der Niederspannung", German Datacenter Association e. V., [https://www.germandatacenters.com/fileadmin/documents/publications/2024-12-30\\_GDA\\_Position\\_Konsultation\\_BK\\_6-24-245.pdf](https://www.germandatacenters.com/fileadmin/documents/publications/2024-12-30_GDA_Position_Konsultation_BK_6-24-245.pdf)

**Greenblatt (2025).** "What's going on with AI progress and trends? (As of 5/2025)", Redwood Research, <https://blog.redwoodresearch.org/p/whats-going-on-with-ai-progress-and>

**Harris & Harris (2025).** "America's Superintelligence Project", Gladstone AI, <https://superintelligence.gladstone.ai/>

**Handelsblatt (2025).** "Das sind die deutschen Bewerber für eine AI Gigafactory", <https://www.handelsblatt.com/technik/ki/ki-rechenzentren-das-sind-die-deutschen-bewerber-fuer-eine-ai-gigafactory-01/100137764.html>

**Hess (2025).** "Eine Gigafactory reicht vorerst: Besonnen statt groß denken!", Süddeutsche Zeitung Dossier, <https://www.sz-dossier.de/gastbeitraege/2025-05-02-julia-hess-eine-gigafactory-reicht-voerst-besonnen-statt-gross-denken-8e094bdb>

**HLRS (2023).** "Exascale-Supercomputing kommt nach Stuttgart", Höchstleistungsrechenzentrum Stuttgart, <https://www.hlrs.de/de/news/detail/exascale-supercomputing-kommt-nach-stuttgart>

**HLRS (2024).** "HammerHAI wird eine AI Factory für Wissenschaft und Industrie aufbauen", Höchstleistungsrechenzentrum Stuttgart, <https://www.hlrs.de/de/presse/detail/hammerhai-wird-eine-ai-factory-fuer-wissenschaft-und-industrie-aufbauen>

**IEA (2025).** "Energy and AI", World Energy Outlook Special Report, International Energy Agency, <https://iea.blob.core.windows.net/assets/40a4db21-2225-42f0-8a07-addcc2ea86b3/EnergyandAI.pdf>

# Literaturverzeichnis

**IMF (2025).** "World Economic Outlook Database", International Monetary Fund, Online-Zugriff am 03.09.2025, <https://www.imf.org/en/Publications/WEO/weo-database/2025/April/weo-report>

**Juijn et al. (2025).** "Tripling the EU's data centre stock with special AI compute zones", Center for Future Generations, <https://cfg.eu/special-compute-zones-submission/>

**KI Bundesverband (2024).** "Für ein starkes KI-Deutschland: Impulspapier zur Bundestagswahl 2025", Bundesverband der Unternehmen der Künstlichen Intelligenz in Deutschland e.V., [https://ki-verband.de/wp-content/uploads/2025/05/Impulspapier\\_Bundestagswahl2025\\_KI-Bundesverband\\_2024.12-1.pdf](https://ki-verband.de/wp-content/uploads/2025/05/Impulspapier_Bundestagswahl2025_KI-Bundesverband_2024.12-1.pdf)

**Kokotajlo et al. (2025).** "AI 2027", AI Futures Project, <https://ai-2027.com/>

**Lorentz et al. (2025).** "KI-Infrastruktur: Wie Deutschland im globalen KI-Rennen aufholen kann", Deloitte, <https://www.deloitte.com/content/dam/assets-zone2/de/de/docs/issues/sustainability-climate/2025/Deloitte-KI-Infrastruktur-Studie.pdf>

**LRZ (2025).** "Blue Lion mit Vera Rubin-Architektur", Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften, <https://www.lrz.de/news/detail/blue-lion-mit-vera-rubin-architektur>

**Ma et al. (2025).** "Inference-Time Scaling for Diffusion Models beyond Scaling Denoising Steps", arXiv, <https://arxiv.org/abs/2501.09732>

**METR (2025).** "How Does Time Horizon Vary Across Domains?", Model Evaluation & Threat Research, <https://metr.org/blog/2025-07-14-how-does-time-horizon-vary-across-domains/>

**Nevo et al. (2024).** "Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models", RAND Corporation, [https://www.rand.org/content/dam/rand/pubs/research\\_reports/RRA2800/RRA2849-1/RAND\\_RRA2849-1.pdf](https://www.rand.org/content/dam/rand/pubs/research_reports/RRA2800/RRA2849-1/RAND_RRA2849-1.pdf)

**Norris et al. (2025).** "Rethinking Load Growth: Assessing the Potential for Integration of Large Flexible Loads in US Power Systems", Nicholas Institute for Energy, Environment & Sustainability, Duke University, <https://hdl.handle.net/10161/32077>

**NVIDIA (2025a).** "NVIDIA Builds World's First Industrial AI Cloud to Advance European Manufacturing", <https://nvidianews.nvidia.com/news/nvidia-builds-worlds-first-industrial-ai-cloud-to-advance-european-manufacturing>

**NVIDIA (2025b).** "NVIDIA and United Kingdom Build Nation's AI Infrastructure and Ecosystem to Fuel Innovation, Economic Growth and Jobs", <https://nvidianews.nvidia.com/news/nvidia-and-united-kingdom-build-nations-ai-infrastructure-and-ecosystem-to-fuel-innovation-economic-growth-and-jobs>

**NVIDIA (2025c).** "NVIDIA Blackwell Datasheet", <https://resources.nvidia.com/en-us-gpu-resources/datasheet?lx=CPwSfP>

**OpenAI (2025).** "Announcing The Stargate Project", <https://openai.com/index/announcing-the-stargate-project/>

**Patel & Nishball (2024).** "Nvidia Blackwell Perf TCO Analysis – B100 vs B200 vs GB200 NVL72", SemiAnalysis, <https://semianalysis.com/2024/04/10/nvidia-blackwell-perf-tco-analysis/>

# Literaturverzeichnis

**Patel et al. (2025a).** "DeepSeek Debates: Chinese Leadership On Cost, True Training Cost, Closed Model Margin Impacts", SemiAnalysis, <https://semianalysis.com/2025/01/31/deepseek-debates/>

**Patel et al. (2025b).** "AI Arrives In The Middle East: US Strikes A Deal with UAE and KSA", SemiAnalysis, <https://semianalysis.com/2025/05/16/ai-arrives-in-the-middle-east-us-strikes-a-deal-with-uae-and-ksa/>

**Pilz, Heim & Brown (2023).** "Increased Compute Efficiency and the Diffusion of AI Capabilities", arXiv, <https://arxiv.org/abs/2311.15377>

**Pilz et al. (2025).** "Data on GPU Clusters", Epoch AI, Online-Zugriff am 03.09.2025, <https://epoch.ai/data/gpu-clusters>

**Pilz, Sanders et al. (2025).** "Trends in AI Supercomputers", arXiv, <https://arxiv.org/abs/2504.16026>

**Sevilla et al. (2024).** "Can AI Scaling Continue Through 2030?", Epoch AI, <https://epoch.ai/blog/can-ai-scaling-continue-through-2030>

**Telekom (2025).** "KI-Turbo für Gigafactories: Telekom bildet europäische KI-Infrastruktur mit NVIDIA für Industrie", Deutsche Telekom, <https://www.telekom.com/de/medien/medieninformationen/detail/ki-turbo-nvidia-und-deutsche-telekom-1093524>

**The Information (2025).** "DeepSeek's Progress Stalled by U.S. Export Controls", <https://www.theinformation.com/articles/deepseeks-progress-stalled-u-s-export-controls>

**The Next Platform (2023).** "LRZ Adopts Nvidia Engines For €250 Million "Blue Lion" Supercomputer In 2027", <https://www.nextplatform.com/2024/12/16/lrz-adopts-nvidia-engines-for-e250-million-blue-lion-supercomputer-in-2027/>

**Welt (2025).** "'Werte und Wissen widerspiegeln' – Digitalminister fordert europäische KI-Modelle", <https://www.welt.de/wirtschaft/article256390626/Karsten-Wildberger-Digitalminister-fordert-europaeische-KI-Modelle.html>

**Wiseman, McClements & Horsley (2024).** "Getting AI datacentres in the UK", Inference Magazine, <https://inferencemagazine.substack.com/p/getting-ai-datacentres-in-the-uk>

**Y Combinator (2025).** "Elon Musk: Digital Superintelligence, Multiplanetary Life, How to Be Useful", YouTube-Video, Online-Zugriff am 03.09.2025, <https://www.youtube.com/watch?v=cFllta1GkiE&t=1894s>

**You (2025).** "How far can reasoning models scale?", Epoch AI Gradient Updates, <https://epoch.ai/gradient-updates/how-far-can-reasoning-models-scale>

# Über diesen Bericht

**Autoren:** Philip Fox, Monika Schnitzer, Daniel Privitera

**Kontakt:** philip@kira.eu

**Zitierempfehlung:** Fox, P., Schnitzer, M. & Privitera, D. (2025). "KI-Rechenzentren in Deutschland: Aktuelle Kapazität, zukünftiger Bedarf", KIRA Center, Berlin.