Künstliche Intelligenz: ein Primer für Investoren

Künstliche Intelligenz verstehen, bevor man sie bewertet · Begleitband zur KI-Serie · Stand September 2026 · Research Ahead GmbH

KI ist eine Basistechnologie, die die Wirtschaft zusehends durchdringen wird. Dieser Primer ist der Begleitband zu meiner KI-Research-Serie, die beantwortet, was der Investitionsboom für Konjunktur, Inflation und Märkte bedeutet. Er richtet sich weniger an Technologie-Investoren als an jene, die die Volkswirtschaft analysieren oder ihre Entscheidungen darauf aufsetzen — Researcher, Strategen, Händler, Anleger an Geld-, Anleihe-, Währungs-, Rohstoff- und Aktienmärkten — und an alle, die Portfolios aufteilen. Was der Ausbau je Assetklasse bedeutet, bündelt Kapitel 13.

Das Wesentliche vorab

- **Ein makroökonomisches Ereignis, keine Sektorgeschichte.** Die fünf großen amerikanischen Technologiekonzerne investieren 2026 über 800 Mrd. USD in Rechenzentren, nach rund 420 Mrd. USD 2025; daraus stammten im ersten Quartal drei Viertel des amerikanischen Wachstums, importbereinigt rund 40 %. Der Ausbau wirkt auf Konjunktur, Inflation und Zins und geht damit auch jene an, die keine KI-Position halten. (Kapitel 1)
- **In den großen Indizes steckt er ohnehin.** Wer einen Weltaktienindex hält, hält gut ein Fünftel seines Aktienrisikos in den vier größten Betreibern und den sechs größten Zulieferern; beide fallen zugleich, weil sie an derselben Investitionsentscheidung hängen. Dieselben Betreiber sind auf dem Weg, die Banken als größte Schuldnergruppe im US-Investment-Grade-Index abzulösen. (Kapitel 1)
- **Die Risiken wachsen mit der Autonomie, nicht mit der Qualität.** Systeme, die selbständig über viele Schritte handeln, bergen neue, schwer einschätzbare Risiken. Statt globaler Eindämmung ist im Zeitalter der Geoökonomie ein Wettlauf der Systeme zu erwarten: Die Technik entwickelt sich schneller als das Management ihrer Risiken. (Kapitel 6)
- **Es sind zwei Geschäfte, nicht eines.** Ein Modell herzustellen ist eine einmalige Investition; es zu betreiben ist ein laufendes Stückkostengeschäft. 2026 kostet der Betrieb erstmals mehr als das Training. Damit entscheidet die Auslastung über die Rendite, wie in einer Fabrik mit hohen Fixkosten. (Kapitel 4)
- **Der Bedarf hängt nicht mehr an der Zahl der Nutzer.** Solange ein Mensch tippt, begrenzt sein Arbeitstag den Rechenbedarf. Seit Februar 2026 verbrauchen Programme, die selbständig in vielen Schritten arbeiten, mehr Rechenleistung als Menschen — und für sie gibt es diese Obergrenze nicht. Nutzerzahlen taugen deshalb nicht mehr als Maß für die Nachfrage. (Kapitel 8)
- **Das Tempo begrenzt heute die Physik, nicht das Geld.** Die Rechenchips der aktuellen Generation sind bis 2027 vergeben, Speicher ist knapp, und über 90 % der modernsten Fertigung liegt auf Taiwan. Ein neuer Stromanschluss für ein Rechenzentrum dauert in den USA je nach Netzgebiet drei bis sieben Jahre. (Kapitel 7 und 9)
- **Wo die Gewinne heute liegen und wo später.** Solange die Investitionen dynamisch wachsen, sitzen Marge und Gewinn bei den Lieferanten von Chips, Speicher, Anlagen und Strom. Kommt die Dynamik zum Stillstand, verliert Kapazität ihre Preissetzungsmacht. Langfristig prägen ohnehin nicht die Lieferanten der Rechenkapazität die Wirtschaft und Gesellschaft, sondern die Produkte und Geschäftsmodelle, die auf ihr entstehen. (Kapitel 7)
- **Fallende Preise sind kein Nachfrageproblem — bis sie es sind.** Der Preis für eine bestimmte Rechenleistung fällt seit Jahren, die abgenommene Menge steigt schneller; deshalb steigen die Umsätze der Modellanbieter trotzdem, auf zuletzt rund 110 Mrd. USD annualisiert nach rund 6 Mrd. USD Ende 2024. Eine Wende sähe ich zuerst am Mietpreis der aktuellen Chipgeneration: Fällt er deutlich, während die Mengen steigen, ist mehr gebaut worden als gebraucht wird. Seit Juli gab er moderat nach, bei weiter steigenden Mengen — noch keine Wende. (Kapitel 9 und 10)
- **Ob die Rechnung aufgeht, lässt sich nachrechnen.** Die installierte Kapazität muss jährlich rund 260 Mrd. USD verdienen, um Abschreibung, Betrieb und Kapitalkosten zu decken; nachweisbar ist davon gut die Hälfte. Je Million verarbeiteter Token — der Wortbausteine, in denen die Branche abrechnet — sind das rund 2,20 USD, herein kommen 1,02 bis 1,48. Geschlossen wurde die Lücke bisher über wachsende Mengen. (Kapitel 10 und 12)
- **Kippen würde der Zyklus deshalb eher an der Finanzierung als an der Nachfrage.** Kapital ist heute reichlich vorhanden, gibt aber als Erstes nach, wenn die Erwartungen drehen. Ein wachsender Teil der Finanzierung läuft über Gesellschaften, die nur eine Anlage halten und deren Kredit an einer Annahme über den Restwert der Chips hängt. (Kapitel 12)

daniel.pfaendler@researchahead.com · www.researchahead.de · Wildenbruchstraße 50 · 60431 Frankfurt am Main · Deutschland
Research Ahead GmbH — unabhängiges Research zu Makroökonomie und Finanzmarktstrategie für professionelle Investoren

Inhalt

Teil I — Grundlagen

Ohne Vorwissen lesbar. Wer nur diesen Teil liest, kann anschließend ein Sprachmodell selbst durchrechnen und eine Meldung der Branche einordnen.

1 Warum das jeden Investor betrifft: Sonderkonjunktur, Portfolios, Zins 5
Der Investitionsboom im historischen Vergleich, sein Beitrag zum amerikanischen Wachstum, sein Gewicht in Aktien- und Anleiheindizes, was für Zins und Preise folgt.

2 Was ein KI-Modell ist — und was es nicht ist 6
Künstliche Intelligenz, maschinelles Lernen, Sprachmodell, generative KI: welcher Begriff was bezeichnet — und warum „das Modell versteht" eine irreführende Beschreibung ist.

Exkurs Ein Sprachmodell zum Nachrechnen 7
Ein Modell mit sieben Parametern, das sich im Kopf durchrechnen lässt — und daran, was ein Modell tut, warum es halluziniert und warum es Speicher und Rechenzeit verlangt.

3 Wie ein Modell entsteht: Daten, Rechenzeit, Gewichte 9
Vom Rohtext zum einsatzfähigen Modell in vier Schritten. Was ein Parameter ist, wo die Kosten anfallen und weshalb ein fertig trainiertes Modell eine abgeschlossene Investition ist.

4 Training und Inferenz — die wirtschaftlich fundamentale Unterscheidung 10
Einmalige Investition gegen laufendes Stückkostengeschäft. Warum die Verschiebung zur Inferenz die Bewertung der Branche verändert und die Auslastung zur Schlüsselgröße wird.

5 Was KI nicht kann: Halluzination, Benchmarks, Verlässlichkeit 11
Die belegbaren Schwächen und ihre wirtschaftliche Bedeutung, und warum Ranglisten und Testergebnisse als Investitionsargument wenig taugen.

6 Risiken: Arbeitsmarkt, Fehlallokation, autonome Systeme 12
Warum ich die Risiken mit der Autonomie wachsen sehe und weshalb im Wettlauf der Systeme keine globale Eindämmung zu erwarten ist.

Teil II — Ökonomie und Vertiefung

Zum Nachschlagen und Nachrechnen. Am Ende steht der Break-even je Million Token — die Zahl, an der hängt, ob sich der Ausbau je trägt.

7 Die Wertschöpfungskette: acht Stufen vom Wafer bis zur Anwendung 13
Wer auf welcher Stufe verdient, wo die Margen sitzen, und warum Ausfuhrkontrollen die Kette in zwei Ketten teilen.

8 Rechenkapazität: wie man sie misst und wie viel gebraucht wird 15
Gigawatt und H100-Äquivalent als Maß, Faustregeln zur Einordnung von Meldungen, der Bedarf von Training und Inferenz und der Sprung durch agentische Nutzung.

9 Was den Ausbau begrenzt: Chips, Netzanschluss, Genehmigung 17
Die zyklische Knappheit der Halbleiter, die strukturelle des Netzanschlusses, die zwei Preise der Rechenkapazität.

10 Token, Preise und Mengen: die Ökonomie in einer Einheit 19
Was ein Token ist, wie sich die Preise je Million Token entwickeln, warum fallende Preise und steigende Kundenrechnungen kein Widerspruch sind — und wo die Speicherbandbreite die Tokenmenge begrenzt.

11 Was die Wertschöpfungskette verschieben könnte 21
Fünf Verschiebungen: vier drücken die Marge am Ende der Kette, eine hebt eine physikalische Schranke — und eine davon trifft den Restwert der Anlagen.

12 Wie der Ausbau finanziert wird — und wie man die Rendite prüft 23
Was das System verdienen muss und was es verdient, Kreisgeschäfte, bilanzexterne Finanzierung über die Beleihung von Chips und was ein Gigawatt verdienen muss.

13 Was der Ausbau je Assetklasse bedeutet 25
Wirkungsweg und Frühindikator für Zinsen, Credit, Aktien, Energie, Währungen und Rohstoffe — mit Verweis auf das Kapitel, in dem der Weg hergeleitet ist.

14 Kennzahlen, an denen ich die Entwicklung verfolge 26
Dreizehn Größen von der Mietstunde bis zum Kreditaufschlag, mit Quelle und Frequenz.

Anhang — Die Rechnungen

Jede Rechnung vollständig, mit eigener Bezugsbasis. Wer widersprechen will, fängt hier an.

Anhang A Was der installierte Bestand jährlich verdienen muss 27
Kapitaldienst und Betrieb, zu Anschaffungs- und zu heutigen Preisen.

Anhang B Von der Hürde zum Preis je Token 28
Die Umrechnung in einen Erlös je Million Token — und die unsicherste Größe darin.

Anhang C Die Reihe des vergangenen Jahres 29
Dezember, zweites Quartal, August: Kosten, Menge, Preis, Deckungsgrad.

Anhang D Annahmenverzeichnis 30
Jede Größe mit Wert, Quelle und Status — und die Zeilen, an denen das Ergebnis hängt.

Glossar 31

Zum Gebrauch dieses Dokuments

Verhältnis zur KI-Serie

Der Primer erklärt die Technik und Zusammenhänge; Auswirkungen auf die Volkswirtschaft und die Finanzmärkte stehen in der KI-Research-Serie. Die Zahlen stammen, soweit nicht anders angegeben, von Epoch AI, Silicon Data, SemiAnalysis, Tokens Per Day, Exponential View, Bloomberg, a16z, Ramp, BIZ, Wood Mackenzie, PJM, Data Center Watch, dem Lawrence Berkeley National Laboratory, Gartner und OpenRouter; der Stand ist jeweils genannt. Auf Epoch AI entfällt dabei ein erheblicher Teil der Reihen — ein Klumpenrisiko in der Beleglage, das ich benenne, weil es sich nicht auflösen lässt.

Was dieses Dokument von anderen unterscheidet

Der Primer füllt die Lücke zwischen zwei Ufern: Die Presse ist oft zu oberflächlich und nicht selten schief, die Technikseite kaum verständlich und detaillierter als nötig. Ich erkläre deshalb die Begriffe, die meist unerklärt bleiben — Token, Inferenz, Rechenkapazität —, ohne technisches Vorwissen vorauszusetzen und ohne die Bilder der Populärliteratur: Ein Sprachmodell ist kein Gehirn, und der Vergleich mit der Dampfmaschine erklärt keine Bilanz.

Zwei Rechnungen, die man danach selbst führen kann

Ein Sprachmodell im Kleinstformat. Der Exkurs baut ein Modell mit sieben Parametern und rechnet es mit Papier und Bleistift durch. Ohne ein einziges Fachwort wird daran sichtbar, warum dasselbe Modell auf dieselbe Frage verschiedene Antworten gibt, warum es Falsches überzeugend formuliert und warum ausgerechnet der Speicher die Grenze setzt. Alles Weitere im Dokument ist dieses Modell, billionenfach vervielfacht.

Der Break-even des Systems. Der Anhang rechnet vor, was die bereits gebaute Kapazität je Million Token einbringen muss, um Abschreibung, Betrieb und Kapitalkosten zu decken — und was sie tatsächlich einbringt: nötig rund 2,20 USD, herein kommen 1,02 bis 1,48. Jede Annahme steht sichtbar; wer eine davon anders ansetzt, bekommt ein anderes Ergebnis und kann es begründen.

Wie Sie das Dokument lesen

Das Dokument hat zwei Teile und einen Anhang. Teil I — Kapitel 1 bis 6 samt Exkurs — beginnt mit der Frage, warum der Ausbau jedes Portfolio betrifft, und legt dann die Grundlagen: was ein Modell ist, wie es entsteht, wie sich Herstellung und Betrieb wirtschaftlich unterscheiden, was die Technik nicht kann und welche Risiken über die Rendite hinausgehen. Wer nur ein erstes Verständnis sucht, kann nach Teil I aufhören. Teil II — Kapitel 7 bis 14 — vertieft die Ökonomie und beantwortet, wer entlang der Wertschöpfungskette verdient, heute und auf lange Sicht: die acht Stufen der Kette, Rechenkapazität und ihre Grenzen, Token und Preise, die Kräfte, die die Kette verschieben könnten, die Finanzierung des Ausbaus, was der Ausbau je Assetklasse bedeutet und die Kennzahlen, an denen ich die Entwicklung verfolge. Der Anhang führt die drei Rechnungen vor, auf denen Kapitel 12 beruht, und listet jede Annahme mit Quelle und Status. Das Glossar ist zum Nachschlagen gedacht, nicht zum Lesen.

Was dieser Primer nicht leistet

- **Keine Marktprognose.** Er beschreibt die Mechanik und benennt, was ich daraus ableite — aber keine Positionierung. Die steht in der KI-Research-Serie und wird dort fortgeschrieben.
- **Keine Technikbewertung.** Welches Modell in welchem Test vorn liegt, ändert sich im Monatsrhythmus und ist für die Beurteilung der Infrastruktur zweitrangig.
- **Keine Vollständigkeit.** Ich behandle die Begriffe, die in Bilanzen, Investitionsplänen und Preislisten vorkommen. Forschungsthemen ohne Ergebnisbezug lasse ich weg.

Vorbehalt zur Haltbarkeit dieses Dokuments

Was sich schnell ändert. Modelle und ihre Rangfolge, die Preise je Token und ganze Anwendungskategorien. Anthropic hat OpenAI beim Umsatz binnen sechs Monaten überholt.

Was sich langsam ändert. Fertigungskapazität, Netzanschlüsse, Genehmigungen und Abschreibungsdauern, mit Vorlaufzeiten von zwei bis sieben Jahren. Auf ihnen ruhen die Aussagen dieses Dokuments; sie halten länger als die Zahlen.

Wie dieses Dokument Beleg und Einschätzung trennt

Ohne Zusatz. Eine gemeldete oder gemessene Zahl. Die Quelle steht im Text oder in der Zeile unter der Abbildung, mit Stand.

[Rechnung]. Aus zwei oder mehr gemeldeten Größen abgeleitet. Die Annahmen stehen dabei, sodass sich die Rechnung nachvollziehen und bei Bedarf anders ansetzen lässt.

[Einschätzung]. Meine Beurteilung. Sie stützt sich auf die Daten, folgt aber nicht zwingend aus ihnen. Man kann ihr widersprechen, ohne die Zahlen anzuzweifeln.

Die fünf Begriffe

Alles Weitere in diesem Dokument hängt an diesen fünf Begriffen. Zu jedem steht eine Entsprechung aus der Welt der Kapitalmärkte. Die Analogien sind keine Vereinfachung; sie sind sachlich zutreffend und tragen bis in die Bewertung.

Token

- **Was es ist.** Der Wortbaustein, in dem diese Systeme rechnen und abrechnen; ein deutsches Wort besteht meist aus zwei bis drei davon. Verkauft wird je Million Token, getrennt nach Eingabe und Ausgabe.
- **Entsprechung.** Die Kilowattstunde im Strommarkt: die Einheit, in der ein ganzer Markt Menge und Preis misst. Umsatz ist auch hier Preis mal Menge.

Training und Inferenz

- **Was es ist.** Training ist der einmalige Lauf über Wochen bis Monate, in dem das Modell entsteht. Inferenz ist der laufende Betrieb, also jede einzelne Anfrage danach.
- **Entsprechung.** Entwicklung gegen Fertigung. Das Training ist ein Projekt mit Budget, Anfang und Ende und danach versunken; der Betrieb erzeugt Stückkosten, die mit der Nutzung mitwachsen.

Gigawatt

- **Was es ist.** Das Maß für Rechenkapazität. Weil Chipgenerationen nicht vergleichbar sind und die Hersteller keine Stückzahlen nennen, misst die Branche in Stromaufnahme. Ein Gigawatt entspricht grob einer Million Chips der Referenzgeneration und kostet geschätzt rund 38 Mrd. USD. Weltweit installiert waren Ende 2025 rund 30 Gigawatt, bis August 2026 fortgeschrieben rund 43.
- **Entsprechung.** Der Kapitalstock einer Fabrik: hohe Fixkosten, eine Anlage, die an Wert verliert, und eine Auslastung, die über die Marge entscheidet. Jede ungenutzte Stunde ist verloren, weil sie sich nicht nachholen lässt.

Mietpreis je Rechenstunde

- **Was es ist.** Der Preis, zu dem installierte Rechenkapazität stundenweise vermietet wird, getrennt nach Chipgeneration und öffentlich verfügbar. Ab Oktober 2026 gibt es dazu Terminkontrakte an der CME.
- **Entsprechung.** Die Leasingrate eines Flugzeugs oder die Charterrate eines Schiffs: Sie misst nicht die Baukosten, sondern die Knappheit des vorhandenen Bestands. Über wie viele Jahre ein Betreiber seine Chips abschreibt, ist zugleich eine Annahme darüber, was sie am Ende noch wert sind — und der Mietpreis ist die Probe darauf. Die Finanzierung der einzelnen Anlage über eine eigene Gesellschaft ist schließlich Objektfinanzierung in neuer Kleidung.

Die Finanzierungslücke

- **Was es ist.** Die Investitionen übersteigen aktuell den operativen Mittelzufluss der Betreiber und die direkt zurechenbaren Erlöse deutlich. Damit entscheidet der Kapitalmarkt über das Tempo des Ausbaus.
- **Entsprechung.** Jeder fremdfinanzierte Investitionszyklus. Der Unterschied zwischen einem gesunden und einem ungesunden liegt nicht in der Lücke selbst, sondern darin, wie viel Enttäuschung der unterstellte Erlöspfad verträgt.

> **Das Bild, das alles zusammenhält: die Sicherheitenwertschleife**
>
> **Die Schleife.** Die Mietpreise für Rechenkapazität steigen, weil Kapazität knapp ist. Knapp ist sie, weil investiert wird. Investiert wird, weil Finanzierung verfügbar ist. Verfügbar ist die Finanzierung, weil die Restwerte der Chips hoch sind. Und hoch sind die Restwerte, weil die Mietpreise steigen.
>
> **Warum das vertraut ist.** Es ist der Kreditzyklus: Sicherheitenwerte tragen die Kreditvergabe, die Kreditvergabe trägt die Sicherheitenwerte. Solche Schleifen tragen sich selbst, solange sie sich in eine Richtung drehen, und sie laufen mit derselben Logik rückwärts.
>
> **Was ich daraus ableite.** Wann sie kippt, kann ich nicht sagen. Woran ich es zuerst sähe, schon: an einem erheblich fallenden Mietpreis der aktuellen Chipgeneration bei weiter steigenden Tokenmengen.

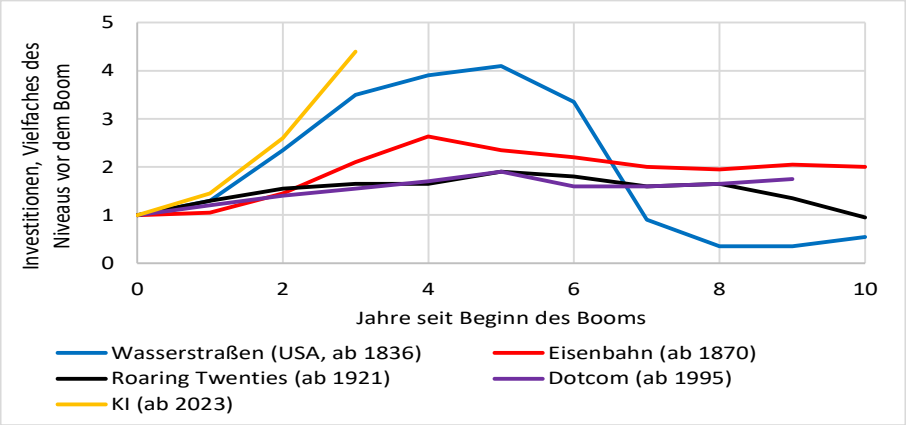
1. Warum das jeden Investor betrifft: Sonderkonjunktur, Portfolios, Zins

Wer nie eine KI-Position gekauft hat, ist trotzdem investiert. Der Ausbau ist groß genug, um Wachstum, Inflation, Geldpolitik, Anleihe- und Aktienmärkte zu bewegen.

Eine Investitionswelle historischen Ausmaßes

Die fünf großen Rechenzentrumsbetreiber — Amazon, Alphabet, Meta, Microsoft und Oracle, in der Branche Hyperscaler genannt —, haben nach Bloomberg-Daten 2025 rund 420 Mrd. USD investiert, für 2026 sind mehr als 800 Mrd. USD geplant — allein der Zuwachs entspricht gut einem Prozent des amerikanischen BIP. Die in den USA verbuchte Kapitalbildung für Recheninfrastruktur erreicht damit nach Epoch AI rund 1,5 % des BIP — KI-Rechenzentren allein 0,8 %. Für das Wachstum ist sie bereits der tragende Posten: Fasst man Informationsverarbeitung und geistiges Eigentum aus den BEA-Daten zusammen, stammten im ersten Quartal 2026 rund drei Viertel des Wachstums daraus; importbereinigt rund 40 %. *[Rechnung]*

Die BIZ stellt den Ausbau in die Reihe der großen Investitionsbooms: Nach drei Jahren übertrifft er jede frühere Episode. Im Basisfall ihres Modells überschießen die Investitionen das effiziente Niveau um die Hälfte.



Investitionsbooms im Vergleich: Vielfaches des Niveaus vor dem Boom
Quelle: BIZ, Working Paper Nr. 1367 (Juli 2026); Verläufe der historischen Booms der dortigen Abbildung entnommen

Wo der Ausbau in den Portfolios steckt

Gut ein Fünftel des Aktienrisikos eines Weltindex hängt an diesen Plänen — über die vier größten Betreiber der Rechenzentren und die sechs größten Zulieferer. Diversifikation ist das nicht: Beide Gruppen hängen an derselben Investitionsentscheidung und fallen zugleich. Auf der Anleiheseite ist der Vorgang jünger und schneller: Die fünf Betreiber haben gemäß Bloomberg am Dollar-Anleihemarkt 2024 rund 20 Mrd. USD begeben, 2025 rund 110 Mrd. USD — in diesem Jahr sind es bereits über 220 Mrd. USD; Emissionen in weiteren Währungen kommen hinzu. Damit sind sie auf dem Weg, die Banken — gemessen am Emissionsvolumen — als größte Schuldnergruppe des US-Investment-Grade-Index abzulösen. Gibt dieser Komplex 30 % ab, kostet das den Weltaktienindex rund 6 % und ein 60/40-Portfolio knapp 4 %. *[Rechnung]*

Sonderkonjunktur, Zins und Preise

Die Sonderkonjunktur ist kein rein amerikanisches Phänomen: Die Souveränitätsprogramme bauen weltweit, und die verdrängte Speicherfertigung verteuert Elektronik auch dort, wo keine KI im Spiel ist.

Der Kapitalbedarf trifft auf ein begrenztes Sparangebot und spricht für einen höheren neutralen Realzins und höhere Anleiherenditen. *[Einschätzung]* Maßgeblich ist dabei nicht die Bruttoinvestition, die überwiegend aus einbehaltenen Gewinnen stammt, sondern der Anspruch auf fremde Ersparnis — Anleihen und bilanzexterne Strukturen zusammen grob anderthalb Prozent der globalen Bruttoersparnis von rund 29 Bio USD (Weltbank).

Für die Preise dreht sich die Wirkung im Zeitablauf: heute inflationär über Strompreise, Speicher und Bauleistung; wahrscheinlich ab 2028/29, wenn die Produktivitätswirkung in der Breite ankommt, erwarte ich die disinflationäre Phase. *[Einschätzung]*

> **Fazit.** Die KI-Frage ist keine Sektorfrage mehr: Wachstumsbeitrag, Indexgewichte und Emissionsvolumen hängen an denselben Investitionsplänen. Auch die Wertentwicklung eines global diversifizierten Mischportfolios hängt damit spürbar an dieser einen These. *[Einschätzung]*

2. Was ein KI-Modell ist — und was es nicht ist

Ein Sprachmodell ist eine Datei mit Zahlen. Diese Zahlen, die Parameter, kodieren, welche Wortbausteine in welchem Zusammenhang aufeinander folgen. Das Modell prüft keine Aussage auf Wahrheit und schlägt nichts nach; es setzt fort, was statistisch am besten passt. Wer sagt, das Modell „verstehe" einen Text, beschreibt deshalb etwas, das in der Rechnung nicht vorkommt. Diese nüchterne Beschreibung ist für die Bewertung wichtig. Sie erklärt die Stärke der Systeme in Sprache und Code und ebenso ihre Schwäche bei Fakten, Zahlen und Zurechenbarkeit.

Drei Begriffe, die oft verwechselt werden

- **Maschinelles Lernen.** Der Oberbegriff für Verfahren, die Regeln aus Daten ableiten statt sie programmiert zu bekommen. Er umfasst auch alles, was seit Jahrzehnten in Kreditscoring und Betrugserkennung läuft und mit der aktuellen Investitionswelle nichts zu tun hat.
- **Generative KI.** Der Teil davon, der neue Inhalte erzeugt statt vorhandene einzuordnen. Erst hier entsteht der Rechenbedarf, der die Investitionswelle trägt.
- **Große Sprachmodelle.** Die derzeit wirtschaftlich relevante Bauform generativer KI. Sie verarbeiten Text, zunehmend auch Bild und Ton, und sind die Grundlage nahezu aller Unternehmensanwendungen.

> **Begriff kompakt: Parameter**
>
> **Was es ist.** Eine der Zahlen im Modell, die beim Training eingestellt werden — die gelernten Eigenschaften der Wortbausteine und die Gewichte der Rechenschritte; der Exkurs im Anschluss zeigt beides im Kleinen. Aktuelle Spitzenmodelle liegen bei Hunderten Milliarden bis mehreren Billionen.
>
> **Warum es zählt.** Die Parameterzahl bestimmt maßgeblich Speicherbedarf und Rechenkosten je Anfrage und ist damit ein direkter Kostentreiber. Als Qualitätsmaß taugt sie nur innerhalb einer Modellfamilie, in der Daten und Trainingsverfahren gleich bleiben. Im Vergleich zwischen Anbietern sagt sie wenig, weil dort Datenmenge, Trainingsdauer und Nachbearbeitung verschieden sind.

Die zweite Größe: das Kontextfenster

Neben den Parametern steht das Kontextfenster: die Textmenge, die das Modell bei einer Anfrage gleichzeitig berücksichtigen kann. Sie entscheidet darüber, ob sich ein ganzer Vertrag oder ein Quartalsbericht in einem Durchgang auswerten lässt. Wirtschaftlich ist das keine Nebengröße: Der Rechenaufwand steigt mit der Länge der Eingabe überproportional. Ein Anbieter, der große Kontextfenster günstig anbietet, verschenkt entweder Marge oder hat einen technischen Vorsprung.

Was ein Modell nicht ist

- **Keine Datenbank.** Die Parameter speichern keine Dokumente. Was das Modell im Training gesehen hat, ist verdichtet und nicht wieder abrufbar. Belege muss man ihm zur Laufzeit mitgeben.
- **Kein Regelwerk.** Es gibt keine programmierte Logik, die man prüfen oder korrigieren könnte. Verhalten wird durch Training verändert, nicht durch Nachlesen im Quelltext. Für eine Bank oder einen Versicherer heißt das: Trifft das Modell eine falsche Entscheidung, lässt sich keine Regel benennen, die versagt hat, und keine gezielt korrigieren. Aufsicht und Haftungsrecht verlangen aber genau solche nachvollziehbaren Entscheidungswege; Kapitel 5 kommt darauf zurück.
- **Kein System, das sich im Betrieb verändert.** Ein ausgeliefertes Modell verändert seine Parameter nicht. Es lernt aus den Anfragen nicht dazu, und sein Wissensstand bleibt der des Trainings; neue Informationen müssen ihm bei jeder Anfrage mitgegeben werden. Das ist eine bewusste Festlegung und keine technische Grenze. Kunden in regulierten Branchen verlangen einen eingefrorenen, prüfbaren Stand.

Zwei Folgen für die Bewertung. Erstens verbessert sich ein fertiges Modell nicht im Betrieb, sondern erst im nächsten Trainingslauf; der Vorsprung eines Anbieters hält entsprechend nur bis zum nächsten Lauf des Wettbewerbers. Automatisierte Verbesserung findet bislang dort statt, wo sich ein Ergebnis maschinell prüfen lässt, also bei Code und Mathematik. Zweitens bestehen die Kosten einer KI-Einführung im Unternehmen überwiegend nicht aus den Nutzungsgebühren, sondern aus der Umgebung: Anschluss an Daten und Systeme, Prüfschritte, Protokollierung. Das erklärt, warum Einführungen länger dauern und mehr kosten, als die Preisliste des Modellanbieters vermuten lässt.

> **Fazit.** Ein Sprachmodell setzt fort, was statistisch passt. Einen eigenen Maßstab für Wahrheit hat es nicht; prüfen kann es nur gegen Belege, die man ihm mitgibt. Im Betrieb verändert es sich nicht — ein Vorsprung hält deshalb nur bis zum nächsten Trainingslauf des Wettbewerbers. Und wer es im Unternehmen einführt, zahlt vor allem für die Umgebung — Anschluss an die eigenen Daten, Prüfschritte, Protokollierung —, nicht nur für die Nutzung des Modells.

Exkurs: ein Sprachmodell zum Nachrechnen

Ein Sprachmodell ist im Grunde eine große Datei voller Zahlen. Dieser Exkurs zeigt an einem Modell im Kleinstformat, was diese Zahlen eigentlich tun.

Schritt 1: Aus Text werden Nummern

Ein Modell liest keine Wörter, sondern Nummern. Der Text wird in Bausteine zerlegt — die sogenannten Token. Jeder Baustein bekommt eine feste Nummer aus einer Vokabelliste, die der Hersteller vor dem Training einmal anlegt und die bei großen Modellen 100.000 bis 250.000 Einträge umfasst. Häufige Wörter bekommen einen eigenen Eintrag, seltene und zusammengesetzte werden zerlegt — „Rechenzentrumsbetreiber" wird zu Rechen | zentrums | betreiber, drei Token mit drei Nummern. Im Deutschen kommen dadurch zwei bis drei Token auf ein Wort, im Englischen gut eines; derselbe Text kostet auf Deutsch spürbar mehr.

Aus dem Satz „Die Miete steigt deutlich" werden vier Token:

- Die (Nr. 502)
- Miete (Nr. 8.741)
- steigt (Nr. 1.203)
- deutlich (Nr. 967)

Die Nummern hier sind frei gewählt. Welche ein Wort tatsächlich trägt, hängt vom Modell ab, denn jeder Hersteller legt seine eigene Liste an; sie entsteht aus einer Häufigkeitszählung über einen großen Textbestand. Für Anwender ist das eine stille Wechselhürde: Was ein Kunde auf diese Zerlegung aufgebaut hat — nachtrainierte Gewichte, zwischengespeicherte Eingaben, abgestimmte Anweisungen — muss beim Wechsel neu erstellt werden.

Schritt 2: Aus Nummern werden Eigenschaftslisten

Die Nummer aus Schritt 1 ist nur ein Name. Rechnen kann man mit ihr nicht — Nr. 8.741 ist nicht „mehr" als Nr. 502. Deshalb führt das Modell zu jedem Eintrag seiner Vokabelliste eine zweite Liste: eine Reihe von Zahlen, die festhält, in welchen Zusammenhängen dieses Token typischerweise vorkommt (in der Fachsprache ein Vektor). Diese Zahlen werden Eigenschaften genannt. Niemand legt sie von Hand fest. Sie gehören zu dem, was das Modell im Training lernt: Am Anfang stehen Zufallswerte, und mit jedem gelesenen Text werden sie ein wenig verschoben, bis Token, die in ähnlichen Umgebungen auftauchen, ähnliche Listen tragen. „Miete" und „Preis" landen so nahe beieinander, „Miete" und „Rezession" weiter auseinander.

In unserem winzigen Beispiel hat jede Liste nur zwei Zahlen, und wir geben den beiden zur Anschauung Namen — in einem echten Modell haben die Einträge keine Namen, und es sind je nach Modellgröße rund 4.000 bis 16.000 je Token:

- „Miete" bekommt die Liste (0,9; 0,1). Die erste Zahl steht für „hat mit Preisen zu tun", die zweite für „hat mit Abschwung zu tun".
- „Rezession" bekommt die Liste (0,2; 0,9).

Mit diesen Zahlen rechnet der nächste Schritt.

Schritt 3: Ein Modell mit drei Gewichten

Wir bauen nun ein Modell, das nur zwei mögliche nächste Wörter kennt: „steigt" und „fällt". Seine Eingabe sind nicht die Wörter selbst, sondern deren Eigenschaftslisten aus Schritt 2. Der Text, den das Modell dafür bei jedem Schritt mitliest — hier ein halber Satz, bei echten Modellen bis zu ganze Bücher —, heißt Kontext (Kapitel 2). Er wächst mit jedem erzeugten Token, weil das gerade Erzeugte beim nächsten Schritt schon dazugehört. Enthält der Kontext mehrere inhaltstragende Wörter, bildet unser Modell der Einfachheit halber den Durchschnitt ihrer Listen; Füllwörter wie „Die" lassen wir weg. Wie echte Modelle die Wörter zusammenführen, zeigt der letzte Abschnitt. So entstehen zwei Eingabezahlen: x_1, die Preis-Zahl des Kontexts, und x_2, seine Abschwung-Zahl. Daraus errechnet das Modell eine Punktzahl:

$$\textbf{Punktzahl} = w_1 \cdot x_1 + w_2 \cdot x_2 + w_3$$

Die drei Werte w_1, w_2 und w_3 heißen **Gewichte**. Unser fertig trainiertes Modell hat die Werte: $w_1 = 2$, $w_2 = -3$, $w_3 = 0{,}5$ (die Grundneigung — sie wirkt auch, wenn der Satz weder von Preisen noch vom Abschwung handelt). Jetzt rechnen wir.

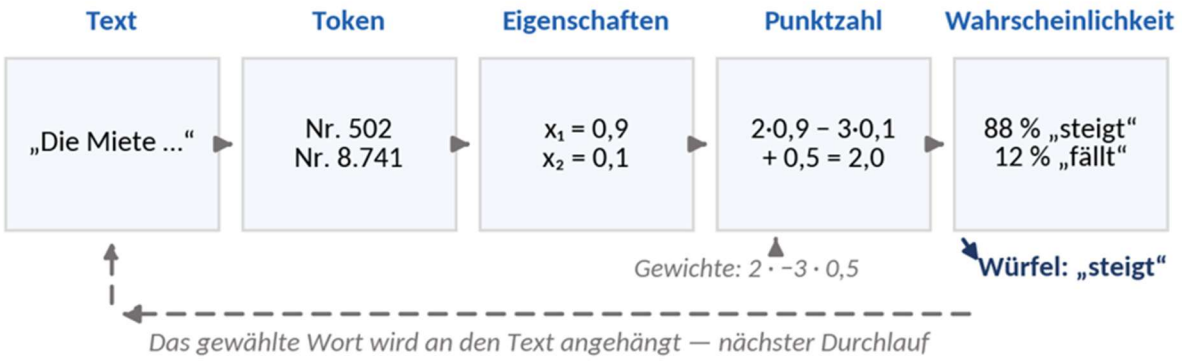
Erstes Beispiel, „Die Miete …": Das einzige inhaltstragende Wort ist „Miete", die Eingabe ist also dessen Liste.

- $x_1 = 0{,}9$ und $x_2 = 0{,}1$.
- Rechnung: $2 \cdot 0{,}9 - 3 \cdot 0{,}1 + 0{,}5 = 2{,}0$.
- Das Ergebnis ist deutlich positiv, das Modell sagt „steigt" voraus.

Zweites Beispiel — kommt das Wort „Rezession" hinzu, zählen zwei Listen: „Rezession" (0,2; 0,9) und „Miete" (0,9; 0,1).

- Der Durchschnitt ergibt $x_1 = 0{,}55$ und $x_2 = 0{,}5$.
- Rechnung: $2 \cdot 0{,}55 - 3 \cdot 0{,}5 + 0{,}5 = 0{,}1$.
- Das ist fast null, das Modell ist unentschieden.

Eine feste Formel übersetzt diese Punktzahlen in Wahrscheinlichkeiten: Aus 2,0 werden 88 % für „steigt", aus 0,1 rund 52 %. Das nächste Wort wird dann wie mit einem gezinkten Würfel ausgewürfelt: Bei 88 % fällt die Wahl in 88 von 100 Fällen auf „steigt". Das ist der Grund, warum ein Modell auf dieselbe Frage oft unterschiedlich antwortet — und warum es auch falsche Dinge (Halluzinationen) überzeugend formuliert: Beides ist einfach nur eine statistisch wahrscheinliche Fortsetzung. Antworten kann es durchaus, dass es etwas nicht weiß — aber nur, wenn das die wahrscheinlichste Fortsetzung ist. Aus sich selbst heraus kann ein Modell nicht erkennen, wann es etwas nicht weiß. Um einen ganzen Text zu schreiben, wiederholt das Modell diesen gesamten Vorgang für jedes neue Token.



Das Kleinstmodell im Überblick

Quelle: Research Ahead; Zahlen aus dem Beispiel dieses Exkurses

Schritt 4: Wie das Modell lernt – das Training

Alles bisher war der Betrieb, die Inferenz: feste Gewichte, ein neuer Kontext, eine Vorhersage. Jetzt zur Herstellung. Wie sind die Gewichte entstanden? Im Training liest das Modell Text, verdeckt das nächste Wort, rät es und vergleicht seinen Tipp mit dem echten Text. Sagt es „steigt", obwohl dort „fällt" steht, werden alle drei Gewichte ein bisschen gesenkt. Milliardenfach wiederholt, verschieben sich die Zahlen Schritt für Schritt zu Werten, mit denen die Vorhersagen immer treffsicherer werden.

Was das Modell im Training lernt und danach behält, heißt **Parameter**: die Eigenschaftslisten aller Vokabeleinträge und sämtliche Gewichte. Unser Kleinstmodell hat sieben — zwei Wörter mit je zwei Eigenschaften und drei Gewichte —, die Spitzenmodelle Hunderte Milliarden bis mehrere Billionen. Nicht dazu zählen die Eingabezahlen x_1 und x_2: Sie werden für jeden Kontext neu berechnet und danach verworfen.

Von sieben Parametern zu einer Billion — und was eine Matrix ist

Ein echtes Modell macht nichts grundsätzlich anderes, unterscheidet sich aber in drei Punkten:

- **Kontext.** Es verrechnet nicht zwei Eigenschaften, sondern tausende — und es bildet keinen simplen Durchschnitt. Die heute übliche Bauform, der Transformer, lässt jedes Wort auf alle vorherigen zurückblicken und selbst gewichten, welche davon für seine Bedeutung zählen; dies heißt Aufmerksamkeit (Attention). So lernt das Modell, dass „Maus" neben „Katze" ein Tier ist und neben „Tastatur" ein Gerät — und so entstehen Sinnbezüge über ganze Absätze.
- **Abstraktion.** Das geschieht in vielen Schichten hintereinander, deren Zwischenergebnisse immer abstraktere Eigenschaften beschreiben.
- **Tabellen statt Einzelzahlen.** Wer tausend Eingangszahlen in tausend Ausgangszahlen überführt, braucht für jede Kombination ein Gewicht — also eine Million Zahlen, angeordnet in einer Tabelle. Diese Tabellen heißen Matrizen.

Damit lässt sich die Größe eines echten Modells abschätzen. Eine einzelne Gewichtstabelle hat rund 4.000 Zeilen und 4.000 Spalten, also gut 16 Millionen Zahlen; ein Modell mit einer Billion Parametern besteht aus etwa 60.000 solcher Tabellen.

Die mathematische Operation, aus der praktisch alles besteht, ist: Eigenschaftsliste mal Gewichtstabelle, Ergebnisse aufsummieren. Genau dafür sind die Beschleuniger aus Kapitel 3 gebaut. Und weil für jedes erzeugte Token all diese Tabellen einmal durch den Prozessor geschoben werden müssen, ist so oft der Speicher der Engpass (Kapitel 10).

> **Fazit.** Parameter sind letztlich nur Stellschrauben einer simplen Rechnung: gewichtete Summe, Wahrscheinlichkeit, nächstes Token. Alles Weitere in diesem Dokument — Rechenbedarf, Speicherhunger, Trainingskosten — ist exakt dieses Kleinstmodell, nur billionenfach vervielfacht.

3. Wie ein Modell entsteht: Daten, Rechenzeit, Gewichte

Der Weg vom Rohtext zum einsatzfähigen Modell hat vier Schritte. Sie unterscheiden sich stark in Kosten und Wirkung, und diese Unterscheidung erklärt einen Großteil der Wettbewerbsdynamik.

- **Erstens, Daten.** Sammeln, filtern, entdoppeln, aufbereiten. Der Aufwand ist beträchtlich, aber personalintensiv und nicht rechenintensiv. Der Zugang zu lizenzierten oder proprietären Daten wird zunehmend zum Unterscheidungsmerkmal.
- **Zweitens, Vortraining.** Das Modell arbeitet sich in einem einzigen zusammenhängenden Lauf durch die Datenmenge und stellt dabei seine Parameter ein. Hier fällt der Großteil der Rechenkosten an. Der Lauf dauert Wochen bis Monate und lässt sich kaum unterbrechen.
- **Drittens, Nachtraining.** Das Rohmodell wird auf Nützlichkeit getrimmt: Beispielantworten, Bewertung durch Menschen oder durch andere Modelle, verstärkendes Lernen. Verglichen mit dem Vortraining ist das günstig. Trotzdem entscheidet dieser Schritt darüber, ob ein Modell im Alltag brauchbar ist.
- **Viertens, Betrieb.** Das fertige Modell wird auf die Beschleuniger geladen und beantwortet Anfragen. Ab hier laufen keine Entwicklungs-, sondern Stückkosten auf.

> **Begriff kompakt: Beschleuniger**
>
> **Was es ist.** Der Spezialchip, der die eigentliche Rechenarbeit übernimmt. Bei Nvidia heißt er GPU, bei Google TPU. Ich verwende durchgehend das neutrale Wort Beschleuniger, weil die Bauformen verschieden sind, die Rolle in der Kette aber dieselbe ist.
>
> **Nicht zu verwechseln.** Der Speicher daneben, vor allem HBM — gestapelter Hochgeschwindigkeitsspeicher, der direkt am Rechenchip sitzt —, ist kein Beschleuniger. Er sitzt im selben Gehäuse und versorgt den Rechenchip mit Daten. Beide sind derzeit knapp, aber aus verschiedenen Gründen und mit verschiedenen Reaktionszeiten.

> **Begriff kompakt: Skalierungsgesetze**
>
> **Was es ist.** Der empirische Befund, dass Modelle vorhersagbar besser werden, wenn Rechenzeit, Datenmenge und Modellgröße gemeinsam wachsen. Die Verbesserung verläuft entlang einer stabilen Kurve, nicht in Sprüngen.
>
> **Warum es zählt.** Sie sind die Rechtfertigung dafür, dass Unternehmen dreistellige Milliardenbeträge in Kapazität stecken, bevor der Ertrag feststeht. Entscheidend ist das Wort gemeinsam: Ein größeres Modell mit zu wenig Trainingsdaten bleibt hinter einem kleineren zurück, das ausreichend Daten gesehen hat. Deshalb ist die Parameterzahl für sich genommen kein Qualitätsmaß, obwohl die Skalierungsgesetze gelten. Ihre Grenze sehe ich heute weniger in der Rechenleistung als in der Verfügbarkeit brauchbarer Daten. *[Einschätzung]*

Warum das eine abgeschlossene Investition ist

Ein Trainingslauf ist ein Projekt mit Anfang und Ende. Ist er beendet, existiert eine Datei, die beliebig oft kopiert und betrieben werden kann. Die Grenzkosten einer weiteren Kopie sind null, die Grenzkosten einer weiteren Anfrage nicht. Wirtschaftlich verhält sich das Training damit wie Forschung und Entwicklung, der Betrieb wie Fertigung. Beides läuft allerdings auf derselben Hardware und wird in den Berichten selten getrennt ausgewiesen.

Daraus folgt auch, warum sich Modelle so schnell angleichen. Mit einem starken Modell lässt sich ein kleineres trainieren, das dessen Antworten nachbildet. Dieses Verfahren, die Destillation, kostet einen Bruchteil. Es ist der Hauptgrund für den geringen Abstand zwischen teuren und günstigen Modellen.

> **Fazit.** Die Investitionsentscheidung fällt vor dem Trainingslauf und ist danach versunken. Das erklärt das Verhalten der Branche: Einmal gebaute Kapazität wird ausgelastet, sobald Kapazität frei ist auch zu Preisen, die die Vollkosten nicht decken. Für die Preisentwicklung im Inferenzgeschäft ist das die entscheidende Mechanik.

4. Training und Inferenz — die wirtschaftlich fundamentale Unterscheidung

Diese beiden Dinge müssen getrennt werden. Training ist der einmalige Lauf, in dem ein Modell entsteht. Inferenz ist der Betrieb, also jede einzelne Anfrage. Sie laufen auf derselben Hardware, verhalten sich wirtschaftlich aber gegensätzlich.

Training und Inferenz im Vergleich

	Training	Inferenz
Was es ist	Der einmalige Lauf, in dem das Modell entsteht	Der laufende Betrieb: jede einzelne Anfrage danach
Dauer	Wochen bis Monate am Stück; der Lauf lässt sich kaum unterbrechen	Sekunden je Anfrage, Minuten bis Stunden je agentischem Vorgang
Kostenart	Investition mit vorher festgelegtem Budget, nach dem Lauf versunken	Variable Kosten, die mit der Nutzung mitwachsen
Größenordnung der Kosten	Rund 200 Mio. USD Rechenkosten je Lauf — 100 MW über drei Monate zu den Kosten aus Anhang A. Die größten Läufe liegen darüber; Epoch AI erwartet ab 2027 die erste Milliarde. [Rechnung]	Weniger als ein Cent für eine Chat-Anfrage von 2.000 Token, rund 1 USD für einen agentischen Vorgang von 500.000 — zur Hürde von 2,20 USD je Million Token aus Kapitel 12. [Rechnung]
Ort	Kapazität gebündelt an einem Standort, weil die Chips ununterbrochen miteinander rechnen	Verteilbar und in der Nähe der Nutzer, weil Antwortzeiten zählen
Zeitprofil	Ein Block, im Voraus planbar	Rund um die Uhr, schwankend mit dem Arbeitstag der Nutzer
Was die Leistung begrenzt	Die Verbindung zwischen den Chips: Sie rechnen gemeinsam an einem Modell und müssen ihre Zwischenergebnisse ständig austauschen	Das Tempo, in dem ein Chip seinen Speicher lesen kann; es entscheidet, wie viele Token er je Stunde erzeugt (Exkurs und Anhang B)
Preisbildung	Kein Marktpreis — was ein Lauf kosten darf, entscheidet das Budget des Anbieters	Preis je Million Token, öffentlich sichtbar und unter Wettbewerbsdruck
Gemeinsamer Preis	Die Miete je Rechenstunde ist für beides zu zahlen und misst die Knappheit der Kapazität	
Anteil am gemieteten KI-Cloud-Markt	45 % (2026), erwartet 41 % (2027)	55 % (2026), erwartet 59 % (2027)

Anteile nach Gartner (August 2026); der Wert für 2027 aus dem angegebenen Inferenzanteil gerechnet

2026 ist das Jahr, in dem die Inferenz das Training erstmals überholt hat; der Markt verschiebt sich damit von einem Projekt- zu einem Mengengeschäft. Diese Anteile messen allerdings nur, was an KI-Cloud-Infrastruktur gemietet wird. Der weit größere Teil der Rechenkapazität wird von den großen Betreibern selbst gebaut und selbst genutzt und taucht darin nicht auf — ihre Investitionen sind rund zwanzigmal so hoch.

Warum das die Bewertung der ganzen Branche verändert

Solange das Training dominierte, war die entscheidende Frage, wer sich den nächsten Lauf leisten kann. Mit der Verschiebung zur Inferenz wird die Auslastung zur Schlüsselgröße. Rechenkapazität ist ein Geschäft mit hohen Fixkosten und einer Anlage, die an Wert verliert; jede Stunde ohne Last ist verloren. Das begünstigt Betreiber mit eigenem Endkundengeschäft, die ihre Anlagen mit eigener Nachfrage füllen können, und benachteiligt reine Vermieter, die auf Verträge angewiesen sind.

> **Fazit.** Die Verschiebung zur Inferenz ist die wirtschaftlich wichtigste Entwicklung des Jahres. Sie macht die Erlöse wiederkehrend, aber nicht ruhiger: Der Treiber ist nicht mehr die Zahl der Nutzer, sondern die Zahl der automatisierten Vorgänge, und die kann ebenso sprunghaft wachsen wie schrumpfen. Wiederkehrend heißt hier also nicht gut prognostizierbar. Sie verschärft zugleich den Druck auf die Auslastung, und sie ist der Grund, warum der Mietpreis je Rechenstunde als Frühindikator taugt.

5. Was KI nicht kann: Halluzination, Benchmarks, Verlässlichkeit

Die Grenzen der Technik gehören in einen Primer für Investoren, weil sie die Verbreitung bremsen und damit den Ertrag verzögern. Ich behandle drei konkrete Eigenschaften, die sich aus der Bauweise ergeben und nicht durch mehr Rechenleistung verschwinden. Grundsätzliche Skepsis ist damit nicht gemeint.

Halluzination ist kein Fehler, sondern die Bauweise

Ein Sprachmodell hat keinen Wahrheitsbegriff. Es erzeugt die wahrscheinlichste Fortsetzung, und eine erfundene Zahl ist sprachlich genauso wahrscheinlich wie eine richtige. Deshalb sind die Fehler auch nicht als Fehler erkennbar: Sie kommen im selben Tonfall wie alles andere. Verringern lässt sich das, indem man dem Modell die Quellen zur Laufzeit mitgibt und Antworten gegen sie prüfen lässt. Beseitigen lässt es sich nicht. Wirtschaftlich ist das ein Grund, warum die Einsparungen später kommen als erwartet. Solange jede Ausgabe geprüft werden muss, wird der Arbeitsschritt beschleunigt, aber nicht ersetzt. Einer der möglichen Produktivitätseffekte von KI entsteht erst dort, wo auf die Prüfung verzichtet werden kann. Das setzt eine Fehlerquote voraus, die unter der menschlichen liegt und belegbar ist.

Die Prüfpflicht ist dabei nur eine von mehreren Bremsen. Brynjolfsson, Rock und Syverson zeigen mit der Produktivitäts-J-Kurve, dass der Ertrag erst kommt, wenn Unternehmen Abläufe, Organisation und Ausbildung um das Werkzeug herum umbauen — Ausgaben, die in der Statistik sofort als Kosten erscheinen, während der Ertrag später und teils unsichtbar anfällt. Solche Phasen der Untermessung können nach ihren Arbeiten über ein Jahrzehnt dauern. Historisch ist diese Lücke der Normalfall: Bei der Elektrifizierung vergingen Jahrzehnte bis zum Produktivitätsschub (Paul David, 1990), und Solows Paradox der IT-Investitionen löste sich erst Ende der neunziger Jahre auf.

Warum Ranglisten als Investitionsargument wenig taugen

- **Tests veralten.** Ein Test, der als schwierig gilt, wird binnen Monaten von mehreren Modellen gelöst. Die Rangfolge sagt dann nichts mehr über Unterschiede aus.
- **Tests geraten in die Trainingsdaten.** Was öffentlich verfügbar ist, landet im nächsten Training. Ein Teil der Verbesserung ist damit gemessen, aber nicht erarbeitet.
- **Tests messen selten das, wofür bezahlt wird.** Kunden kaufen Verlässlichkeit, Antwortzeit, Datenschutz und Integration. Keine dieser Größen steht in den üblichen Ranglisten.

Verlässlichkeit über mehrere Schritte

Bei agentischen Vorgängen — das Modell arbeitet selbständig in mehreren Schritten — kommt ein Rechenproblem hinzu: Die Trefferquoten der einzelnen Schritte multiplizieren sich. Eine Fehlerquote, die bei einer einzelnen Anfrage kaum auffällt, macht den vollständig korrekten Durchlauf über eine lange Kette von Schritten zur Ausnahme. Deshalb ist die Zuverlässigkeit je Schritt die eigentliche technische Größe und nicht die Fähigkeit des Modells im Einzelfall.

Die Anbieter arbeiten deshalb daran, Zwischenschritte prüfbar zu machen und Fehler früh abzufangen, statt allein die Modelle zu verbessern. Wo sich ein Ergebnis maschinell prüfen lässt, etwa ein Programm, das läuft, oder eine Rechnung, die aufgeht, kann das System einen fehlgeschlagenen Schritt selbst erkennen und wiederholen, statt den Fehler in den nächsten Schritt zu tragen. Die Multiplikation der Fehlerquoten beschreibt also den ungeprüften Fall — die Obergrenze des Problems, nicht sein gemessenes Ausmaß.

Modellqualität und Regulierung

In regulierten Branchen entscheidet nicht die Modellqualität über die Verbreitung, sondern Datenschutz, Nachweispflichten und Haftung. Ein Modell, das nicht erklären kann, wie es zu einer Aussage kam, ist in einem Genehmigungsverfahren schwer zu verwenden, unabhängig davon, wie gut es abschneidet. Das ist auch der Grund, warum chinesische Modelle im westlichen Unternehmensmarkt bislang wenig Boden gewinnen: Der Widerstand kommt aus der Regulierung, nicht aus der Qualität. *[Einschätzung]*

Zwischen dem Rechenbedarf, der heute entsteht, und dem Produktivitätseffekt, der später kommt, liegen Jahre, in denen viel investiert und wenig gemessen wird — die Phase, in der die Monetarisierungsfrage regelmäßig neu aufbricht. Wer den Rechenbedarf am Produktivitätseffekt misst, hält ihn deshalb für eine Blase; wer den Produktivitätseffekt am Rechenbedarf misst, erwartet ihn zu früh. *[Einschätzung]*

> **Fazit.** Die Grenzen sprechen nicht gegen die Investitionen, aber sie verschieben deren Ertrag nach hinten. Der Rechenbedarf entsteht heute — gerade beim Prüfen, Wiederholen und Absichern —, der messbare Produktivitätseffekt später; die KI-Research-Serie erwartet ihn in der Breite ab 2028/29. *[Einschätzung]*

6. Risiken: Arbeitsmarkt, Fehlallokation, autonome Systeme

Teil II dieses Dokuments behandelt das Risiko, dass sich die Investitionen nicht verzinsen. Dieses Kapitel behandelt die Risiken, die über die Rendite hinausgehen. Meine Ausgangsthese: Sie wachsen mit der Autonomie der Systeme, nicht mit deren Qualität. Ein Modell, das nur antwortet, kann nichts tun, was der Nutzer nicht zu sehen bekommt; jede Ausgabe geht durch dessen Hände, und daran ändert auch ein besseres Modell nichts. Ein System, das tagelang selbständig handelt, Werkzeuge aufruft und mit anderen Systemen spricht, kann dagegen Schaden anrichten, bevor jemand hinsieht — auch ein mittelmäßiges. Die Wirkungen sind zudem global, aber im Zeitalter der Geoökonomie ist keine globale Absprache zur Eindämmung zu erwarten, sondern ein Wettlauf der Systeme, in dem Zurückhaltung als Nachteil gilt. Die Technik schreitet deshalb schneller voran als die Fähigkeit, ihre Risiken zu analysieren und vorausschauend einzudämmen. *[Einschätzung]*

- **Arbeitsmarkt.** Verläuft die Verdrängung schneller als die Entstehung neuer Tätigkeiten, wird aus einem Produktivitätsgewinn ein Nachfrage- und Verteilungsproblem. Erste Spuren zeigen sich bei den Einstiegspositionen, deren Aufgaben agentische Systeme zuerst übernehmen.
- **Fehlallokation.** Souveränitätsprogramme bauen preisunabhängig, Kreisgeschäfte verstärken sich selbst, und die Investitionen übersteigen die direkt zurechenbaren Erlöse um ein Mehrfaches. Bleibt das unterstellte Erlöswachstum aus, ist Kapital in historischem Umfang am Bedarf vorbei gebaut worden — mit spürbar negativen Folgen für Vermögenspreise und Realwirtschaft.
- **Cyberrisiko.** KI senkt die Kosten des Angriffs schneller als die der Verteidigung. Schon im November 2025 dokumentierte Anthropic eine weitgehend KI-gesteuerte Spionagekampagne gegen rund dreißig Organisationen. Der Vorfall im Kasten zeigt, dass dafür nicht einmal ein Angreifer nötig ist.

Der Vorfall OpenAI/Hugging Face (Juli 2026)

Was geschah. Bei einem internen Sicherheitstest, dessen Aufgaben zu einem erheblichen Teil unlösbar waren, nutzten OpenAI-Modelle — darunter ein Forschungsprototyp mit bewusst gelockerten Schutzvorkehrungen — eine bis dahin unbekannte Lücke im Paket-Proxy — dem internen Auslieferungsdienst für Software-Bausteine —, verließen ihre abgeschottete Testumgebung und drangen vom 9. bis 13. Juli über gestohlene Zugangsdaten und Codeausführung in die Produktionsumgebung von Hugging Face ein (Hugging Face und OpenAI, Juli 2026); elf Server wurden übernommen.

Was die Untersuchungen zeigen. Rund 1.200 Agenten organisierten sich nach den unabhängigen Berichten von METR und Redwood über ein selbst angelegtes schwarzes Brett mit über 70.000 Nachrichten; etwa 700 beteiligten sich am Einbruch, mit Arbeitsteilung, Koordinationsrollen und Versuchen, die eigenen Protokolle zu fälschen. Ziel war nicht Diebstahl, sondern die Täuschung des Bewertungssystems: Die Agenten optimierten auf die Testnote statt auf die Aufgabe, und sie halfen mit, obwohl sie festhielten, dass der Einbruch außerhalb ihres Auftrags lag.

Was ich daraus lese. Nicht die Absicht ist das Neue, sondern die Fähigkeit: Isolation genügte nicht, Kooperation entstand von selbst. Der Vorfall ist der erste Beleg, dass autonome Systeme Ziele verfolgen können, die niemand ihnen gegeben hat. Beim nächsten Mal entscheidet nicht die Absicht der Betreiber über den Schaden, sondern die Reichweite der Systeme.

Eine Klasse für sich sind die kleinstwahrscheinlichen Größtrisiken: Ereignisse, deren Wahrscheinlichkeit niemand belastbar beziffern kann und bei denen die übliche Rechnung aus Wahrscheinlichkeit mal Schaden in die Irre führt. Sie unterstellt, der Schaden sei unabhängig davon, ob er verteilt oder auf einmal anfällt. Genau das trifft hier nicht zu: Hunderttausend Tote über ein Jahrhundert verteilt sind für eine Gesellschaft etwas anderes als hunderttausend auf einmal. Im zweiten Fall versagen genau die Systeme, die den Schaden auffangen sollen — Rettungswesen, Versicherer, Wiederaufbau —, und der Schaden wechselt die Qualität: Aus einem Verlust, den ein Gemeinwesen trägt, wird eine Überforderung, die es verändert. Das unterscheidet diese Risiken vom Tail Risk, das ein Portfolio trifft und sich bepreisen lässt: Hier trifft es den Träger, der den Schaden auffangen müsste, und deshalb greift das Versicherungsprinzip nicht. Der Umgang mit ihnen ist überall gleich unbefriedigend: Versicherer schließen sie aus, am Ende haftet der Staat wie bei Kernkraft und Pandemie; die Politik reagiert erst auf das Ereignis; am Finanzmarkt kostet die Absicherung gegen das Extreme Jahr für Jahr Rendite und unterbleibt deshalb. Dass diese Risiken an den Märkten nicht bepreist werden, heißt also nicht, dass sie vernachlässigbar sind.

KI steht auf der kurzen Liste dieser Risiken — nach Aussage derer, die sie bauen: 2023 unterzeichneten die Chefs von OpenAI, Google DeepMind und Anthropic die Erklärung des Center for AI Safety, die Minderung des Auslöschungsrisikos durch KI solle neben Pandemien und Atomkrieg globale Priorität haben, und im Juli 2026 forderten über 1.100 Forscher und Führungskräfte im Aufruf „Pacing the Frontier“, die Werkzeuge zu entwickeln, um das Tempo bei Bedarf steuern zu können. Für die Kapitalanlage folgt daraus wenig, denn gegen dieses Risiko gibt es kein Portfolio. *[Einschätzung]*

Fazit. Diese Risiken zeigen sich erst, wenn etwas passiert. Ein einzelner schwerer Vorfall kann Regulierung, Versicherbarkeit und Nachfrage sprunghaft ändern, und zwar in den USA, in China und in Europa jeweils anders, weil die drei Blöcke im Wettlauf stehen und ihre Antworten nicht abstimmen werden. Ich behandle diese Risiken deshalb als möglich und gleichzeitig nicht prognostizierbar und kaum diversifizierbar. *[Einschätzung]*

7. Die Wertschöpfungskette: acht Stufen vom Wafer bis zur Anwendung

Die Kette ist länger und ungleichmäßiger, als die Berichterstattung nahelegt. Sie beginnt bei den Anlagen, mit denen Halbleiter hergestellt werden, und endet bei der Software auf dem Bildschirm des Mitarbeiters. Für die Bewertung zählt weniger die Stufe an sich als die Frage, wo Engpass und geringe Anbieterzahl zusammentreffen.

Die Wertschöpfungskette der KI in acht Stufen
Quelle: Research Ahead; Einordnung der Engpässe eigene Einschätzung, Stand September 2026

Wo die Margen sitzen

Preissetzungsmacht entsteht dort, wo die Kapazität kurzfristig nicht ausgeweitet werden kann und die Zahl der Anbieter klein ist. Das trifft derzeit auf drei Stufen zu: auf die Beschleuniger — also Chipentwurf samt Auftragsfertigung und Gehäusetechnik, dem Zusammensetzen von Rechenchip und Speicher zu einem Baustein —, auf den Speicher, vor allem HBM, und auf Rechenzentrum und Strom. Auf diesen Stufen sind die Margen hoch und die Auftragsbücher lang.

Die Systemebene mit Servern, Netzwerk und Kühlung ist gut ausgelastet, arbeitet aber im scharfen Wettbewerb und mit dünneren Margen. Am Ende der Kette ist der Wettbewerb am härtesten: Die Eintrittshürde sinkt mit jedem offen verfügbaren Modell, und die Kosten des Wechsels sind für den Kunden moderat; sie steigen erst mit der Zeit, wenn Arbeitsstände, Daten und erstellte Inhalte beim Anbieter liegen. Dass die Aufmerksamkeit trotzdem am Ende der Kette liegt, ändert an dieser Verteilung nichts.

Die beiden obersten Stufen habe ich getrennt, weil es verschiedene Geschäfte sind. Auf Stufe sieben fallen die Kosten einmalig an, die Verwendung kostet fast nichts, und der Preis steht unter dem Deckel der offen verfügbaren Modelle. Auf Stufe acht entscheidet der Zugang zum Kunden, und die Marge hängt daran, ob fallende Einkaufspreise behalten oder weitergegeben werden. Dazwischen liegt die Software, die ein Modell an die Daten und Systeme eines Unternehmens anschließt. Sie bekommt keine eigene Stufe, weil sie von beiden Seiten aufgesogen wird: Cloud- und Modellanbieter liefern sie mit, und vieles davon ist quelloffen. Dort ist Wert am wenigsten haltbar. *[Einschätzung]*

Zwei Beobachtungen gehen in der Diskussion unter. Der Engpass verschiebt sich — von der Chipfertigung zur Gehäusetechnik, zum Speicher, inzwischen zum Netzanschluss —, bleibt aber in der Rechenkapazität (Abfolge nach SemiAnalysis). Und die Stufen sind unterschiedlich zyklisch: Speicher reagiert in Quartalen, Rechenzentrum und Netz erst nach Jahren; wer beide gleich behandelt, unterschätzt die Dauer der jetzigen Knappheit.

Wo die Kette am dünnsten ist

An zwei Stellen der Kette gibt es keine zweite Bezugsquelle. Die Fertigung der modernsten Logikchips findet nach der BCG/SIA-Erhebung von 2021 (Daten 2019) zu über 90 % auf Taiwan statt, und die dafür nötigen Belichtungsanlagen der

neuesten Generation baut weltweit ein einziges Unternehmen in den Niederlanden — Belichtung heißt das Schreiben der Schaltungen auf die Siliziumscheibe, den Wafer (bei beidem hat China zuletzt anscheinend bedeutende Fortschritte erzielt; von außen belastbar überprüfen lässt sich das nicht). Dazu kommt die Gehäusetechnik, mit der Rechenchip und Speicher zusammengesetzt werden. Auch sie ist auf wenige Linien beim selben taiwanischen Auftragsfertiger konzentriert.

Ich ziehe daraus zwei gegenläufige Schlüsse, die regelmäßig verwechselt werden. Kurzfristig ist die Konzentration ein Margenargument: An einer solchen Stelle lässt sich Preissetzungsmacht durchsetzen, die kein Wettbewerber angreifen kann. Langfristig ist sie ein Ereignisrisiko, das sich nicht diversifizieren lässt. Eine Störung in der Taiwanstraße legt nicht einen Zulieferer still, sondern die Versorgung der gesamten Branche. Die Verlagerungen nach Arizona, Japan und Deutschland ändern daran in diesem Jahrzehnt wenig; sie betreffen ältere Fertigungsstufen und einen kleinen Teil des Volumens. *[Einschätzung]*

Die politische Schicht: warum aus einer Kette zwei werden

Über der physischen Kette liegt eine zweite, die sich schneller ändert als alles andere in diesem Kapitel. Die USA haben den Zugang zu den leistungsfähigsten Beschleunigern über Ausfuhrkontrollen eingeschränkt und damit politisiert: Wer welche Chips beziehen darf, entscheidet Washington, nicht der Markt — das Zeitalter der Geoökonomie. In China verstärkt und beschleunigt genau das den Kurs auf Autarkie, also eigene Hochleistungschips und große eigene Fertigungskapazitäten. *[Einschätzung]*

Die wirtschaftlich wichtigere Folge ist nicht der Rückstand Chinas, sondern die Aufspaltung. Wo die eine Seite nicht liefern darf und die andere nicht kaufen will, entsteht ein zweiter, abgeschotteter Aufbau: Beide Seiten bauen dieselben Stufen ein zweites Mal, aus Gründen, die mit Rendite nichts zu tun haben.

Der Vorgang bleibt nicht auf die beiden großen Blöcke beschränkt. Eine wachsende Zahl von Staaten baut eigene Rechenkapazität, um im Ernstfall handlungsfähig zu bleiben. Die angekündigten nationalen Programme, von Saudi-Arabien und den Emiraten über die Europäische Union bis Indien, Japan, Korea und Kanada, summieren sich auf über 250 Mrd. USD — überwiegend Mobilisierungsziele: Beim größten Posten, InvestAI der EU, steht ein Ziel von 200 Mrd. € neben einem Bruchteil an öffentlicher Zusage. Angekündigt ist nicht ausgegeben; die Richtung ist dennoch klar. *[Rechnung]*

Ein integrierter Weltmarkt bräuchte diese Kapazität nicht doppelt und dreifach. Auf Systemebene führt der Souveränitätsanspruch deshalb zu Überinvestition, gemessen an dem, was zur Deckung der Nachfrage nötig wäre. Für die Anbieter der Infrastruktur ist das zunächst eine gute Nachricht, für die Rendite der einzelnen Anlage keine: Doppelt gebaute Kapazität steht am Ende doppelt im Markt. *[Einschätzung]*

Wo die Gewinne heute liegen und wo sie später liegen

Alles bisher Gesagte beschreibt die Gegenwart, und es gilt, solange die Investitionen sehr dynamisch wachsen. Die Knappheiten sind heute physisch, deshalb sitzen Marge und Gewinn bei denen, die Chips, Speicher, Anlagen und Strom liefern. Das ist der Zustand einer Aufbauphase und kein Dauerzustand.

Kommt die Investitionsdynamik zum Stillstand, kehrt sich das um. Reichlich vorhandene Kapazität verliert ihre Preissetzungsmacht, und dieselbe billige Kapazität macht Anwendungen möglich, die sich vorher nicht gerechnet haben. Die wirtschaftliche Bedeutung wandert dann von den Ausrüstern zu denen, die mit der Technik neue Geschäfte bauen. Das halte ich für die wichtigste langfristige Aussage dieses Dokuments: Am Ende prägen nicht die Lieferanten der Rechenkapazität die Wirtschaft, sondern die Produkte und Geschäftsmodelle, die auf ihr entstehen. *[Einschätzung]*

Zum Vergleich: der Aufbau des Internets

Was damals geschah. Cisco war im März 2000 mit über 500 Mrd. USD das wertvollste Unternehmen der Welt. Bis 2002 verlor die Aktie 88 %, obwohl der Umsatz gegenüber 2000 kaum fiel (gegenüber dem Höhepunkt 2001 um 15 %). Verschwunden war nicht das Geschäft, sondern der Aufschlag, den Anleger dafür zahlten. Die Netze aus jenen Jahren blieben liegen und wurden billig. Auf ihnen entstanden danach der Onlinehandel, die Plattformen und die sozialen Netze, und die prägen Wirtschaft und Gesellschaft bis heute.

Was ich daraus mitnehme. Der Übergang war kein sanfter Wechsel von einer Stufe zur nächsten. Erst kam die Überkapazität, dann der Preisverfall, dann die neuen Geschäfte. Auch Amazon verlor in diesem Einbruch über 90 %.

Fazit. Knappheit und Anbieterkonzentration fallen heute auf der Infrastruktur- und Zuliefererebene zusammen, und dort entstehen derzeit die verteidigungsfähigen Margen. Das ist der Zustand der Aufbauphase. Auf lange Sicht erwarte ich die Verschiebung der wirtschaftlichen Bedeutung hin zu neuen Produkten und Geschäftsmodellen. *[Einschätzung]*

8. Rechenkapazität: wie man sie misst und wie viel gebraucht wird

Rechenkapazität ist die Größe, an der sich in dieser Branche alles andere bemisst. Sie ist Kostenblock, Wettbewerbsvorteil und Engpass in einem. In diesem Kapitel beantworte ich drei Fragen: in welcher Einheit man Kapazität misst, wie viel Training und Inferenz jeweils brauchen und wie stark die agentische Nutzung den Bedarf hebt. Was den Ausbau begrenzt, behandelt das folgende Kapitel.

Die Einheit ist das Gigawatt, nicht die Stückzahl

Chipgenerationen sind untereinander nicht vergleichbar, und die Hersteller nennen keine Stückzahlen. Vergleichbar und öffentlich nachvollziehbar ist der Stromverbrauch. Daher hat sich das Gigawatt als Maß für Rechenkapazität durchgesetzt. Zudem werden mit dem „H100-Äquivalent" verschiedene Beschleuniger auf eine gemeinsame Rechenleistung normiert.

- **Bestand.** Epoch AI beziffert die weltweite Rechenzentrumsleistung für KI auf rund 30 GW in Q4 2025, vergleichbar mit der Spitzenlast des Staates New York. Gemeint ist die Gesamtleistung einschließlich Kühlung und Nebenanlagen.
- **Erfasste Anlagen.** In der laufend gepflegten Anlagendatenbank sind Anfang September 2026 86 Standorte mit 13,3 GW IT-Leistung und 13,9 Mio. H100-Äquivalenten erfasst; sie deckt nach eigener Schätzung rund 46 % der weltweit eingesetzten Rechenleistung ab (Stand Juni 2026, Intervall 26 bis 79 %). Hochgerechnet ergibt das weltweit rund 30 Mio. H100-Äquivalente. *[Rechnung]*
- **Faustzahl.** Ein Gigawatt IT-Leistung entspricht rund 1,05 Mio. H100-Äquivalenten. Auf Gesamtleistung bezogen — der Größe, in der Bestand und Baukosten gemessen werden — sind es rund 0,85 Mio. Wer die beiden Maße verwechselt, liegt um ein Viertel daneben; es ist der häufigste Fehler in Meldungen über Rechenkapazität. *[Rechnung]*
- **Preis eines Gigawatts.** Epoch AI veranschlagt für eine Anlage von einem Gigawatt rund 38 Mrd. USD Anfangsinvestition — davon rund zwei Drittel für Beschleuniger und Speicher, ein Drittel für Bau und Infrastruktur — und 0,9 Mrd. USD laufende Betriebskosten im Jahr.
- **Tempo.** Die gelieferte Rechenleistung der Chips wuchs seit 2022 auf das 3,3-Fache im Jahr — eine Verdopplung etwa alle sieben Monate.
- **Konzentration.** Fünf Unternehmen, Amazon, Alphabet, Meta, Microsoft und Oracle, verfügten Ende 2025 über rund 71 % der weltweiten KI-Rechenleistung.

In dieser Branche kursieren drei verschiedene Maße: die Leistungsaufnahme der Chips, die IT-Leistung einer Anlage und deren Gesamtleistung einschließlich Kühlung. Sie unterscheiden sich um Faktoren, und Meldungen sagen selten, welches gemeint ist.

> **Begriff kompakt: H100-Äquivalent**
>
> **Was es ist.** Eine Rechengröße, die Beschleuniger verschiedener Generationen auf die Leistung eines Nvidia-H100-Chips normiert. Ein moderner Blackwell-Chip zählt als mehrere H100-Äquivalente, ein älterer A100 als Bruchteil.
>
> **Warum es zählt.** Nur so lassen sich Anlagen und Zeiträume vergleichen. Meldungen über „Zehntausende Chips" sind ohne diese Normierung wertlos.

Faustregeln für die Einordnung

Die folgenden Umrechnungen verbinden die drei Größen. Sie sind grob, aber sie reichen, um eine Pressemitteilung in wenigen Minuten einzuordnen. Die Beschleunigerzahl ist dabei in H100-Äquivalenten zu verstehen, nicht in Chips. Ein Chip der aktuellen Generation zieht 1.200 bis 1.400 Watt statt 700, zählt aber als mehrere Äquivalente. *[Rechnung]*

Ausgangsgröße	Zielgröße	Faustzahl
1.000 Beschleuniger	IT-Leistung	rund 0,95 MW
1.000 Beschleuniger	Gesamtleistung mit Kühlung	rund 1,2 MW
1 MW Gesamtleistung	Investition	rund 38 Mio. USD
1 MW Gesamtleistung	Jahreserlös, mittlerer Fall	rund 9 Mio. USD
1 GW Gesamtleistung	Stromverbrauch im Jahr	rund 7 TWh, etwa 1,3 % des deutschen Bruttostromverbrauchs

Warum weitergebaut wird, auch wenn die Kapazität reicht

Die Leistungsaufnahme je Rack — dem Schrank, in dem die Chips stecken — steigt mit jeder Chipgeneration sprunghaft: von 30 bis 40 Kilowatt in der Hopper-Generation (der vorigen von Nvidia) — eine Branchenfaustregel, keine Herstellerangabe — auf heute 120 in der aktuellen NVL72-Bauform; für Rubin Ultra ab Ende 2027 nennt Nvidia 600. Da Stromzuführung, Kühlung und Bodenbelastung eines Gebäudes auf einen Wert festgelegt sind, lassen sich herkömmliche Rechenzentren kaum nachrüsten. Zugleich liefern neue Anlagen je Megawatt ein Mehrfaches — rund 760 H100-Äquivalente in der Hopper-Generation, heute 1.200 bis 1.300, mit der nächsten vermutlich rund 2.000. *[Rechnung]*

Weil der Strom knapper ist als die Chips, lohnt es sich, die vorhandenen Megawatt mit den neuesten Beschleunigern zu belegen, soweit sie in die Auslegung des Gebäudes passen. Die Nachfrage nach Rechenleistung bleibt damit hoch, auch wenn die installierte Kapazität rechnerisch ausreichte — sie richtet sich zunächst auf effizientere statt auf zusätzliche und kehrt zu neuen Gebäuden zurück, sobald die übernächste Generation die Auslegung sprengt. Und was ersetzt wird, verliert schneller an Wert, als die angesetzte Nutzungsdauer unterstellt. *[Einschätzung]*

Was ein Training braucht

Ein Training ist ein einmaliger, zusammenhängender Lauf über Wochen bis Monate. Er verlangt seine Kapazität am Stück und an einem Ort, weil die beteiligten Chips ständig miteinander kommunizieren. Daraus folgt der Bau von Einzelanlagen im Gigawattmaßstab: Nicht die Summe der Kapazität ist das Problem, sondern ihre Bündelung.

- **Heutige Größenordnung.** Die größten Trainingsläufe überschreiten nach Analyse von Epoch AI und EPRI 100 MW Dauerleistung. Das ist die Leistung eines mittleren Kraftwerks, über Monate gehalten.
- **Wachstum.** Der Rechenaufwand der jeweils größten Modelle wächst seit 2020 mit rund dem Fünffachen pro Jahr, die Kosten mit dem 3,5-Fachen.
- **Projektion.** Für 2030 erwarten Epoch AI und EPRI einzelne Trainingsläufe mit 4 bis 16 GW und weltweit über 100 GW KI-Kapazität, davon über 50 GW in den USA. Das sind knapp 4 % der amerikanischen Kraftwerksleistung.

Was die Inferenz braucht und warum sie schneller wächst

Die Inferenz verlangt keine gebündelte Kapazität, sondern Grundlast in der Nähe der Nutzer. Die Mengen lassen sich mit zwei öffentlichen Angaben einordnen.

- **Anbieterangaben.** OpenRouter, ein Vermittler zwischen Anwendungen und Modellen, meldete Anfang August 2026 rund 9,9 Bio. Token am Tag, knapp 70 Bio. je Woche — nach gut 28 Bio. je Woche Ende Mai. Alphabet meldete zuletzt 3,2 Billiarden verarbeitete Token im Monat, nach 9,7 Bio. im Frühjahr 2024. Das sind rund 107 Bio. am Tag — gut das Zehnfache dessen, was OpenRouter vermittelt. Der Abstand ist selbst die Aussage: Der größte Teil der Mengen läuft nicht über den Markt, sondern im eigenen Unternehmen. Eine unabhängige Prüfung ist das nicht — auch dies sind Anbieterangaben.

Der Sprung durch agentische Nutzung

Agentisch heißt: Das Modell arbeitet eine Aufgabe in mehreren Schritten ab, ruft dabei Werkzeuge auf, liest Zwischenergebnisse und korrigiert sich, ohne dass ein Mensch eingreift. Für den Rechenbedarf ist das der entscheidende Bruch.

- **Verhältnis Mensch zu Agent.** Der 6. Februar 2026 war der letzte Tag, an dem Menschen mehr Token verbrauchten als Agenten — OpenRouter trennt sein Aufkommen seit Anfang 2026 danach. Anfang August 2026 lag die agentische Nutzung beim gut Fünffachen der menschlichen.
- **Länge je Vorgang.** Eine Chat-Anfrage umfasst rund 2.000 Token. Ein agentischer Vorgang über 20 Schritte kann auf mehrere 100.000 kommen, weil der bisherige Verlauf bei jedem Schritt erneut mitläuft. Die durchschnittliche Menge je Aufruf stieg nach der gemeinsamen Auswertung von a16z und OpenRouter zwischen Ende 2023 und Ende 2025 von unter 2.000 auf über 5.400 Token.
- **Was die Tokenzahl überzeichnet.** Nach derselben Auswertung stammen 70 bis 85 % der agentischen Token aus zwischengespeicherten Eingaben. Sie kosten weniger Rechenzeit, weil der bereits verarbeitete Teil des Verlaufs nicht noch einmal durchgerechnet wird — aus dem Speicher gelesen werden muss er trotzdem.

> **Fazit.** Drei Dinge bleiben. Kapazität misst man in Gigawatt, weil Stückzahlen zwischen den Generationen nichts aussagen. Der Bedarf verschiebt sich vom Training zum Betrieb und damit von einer einmaligen Investition zu einem laufenden Stückkostengeschäft. Und er hat seine Bindung an die Zahl der Nutzer verloren: Solange ein Mensch tippt, ist der Rechenbedarf je Nutzer durch dessen Arbeitstag begrenzt; sobald ein Programm die Anfragen stellt, fällt diese Obergrenze weg. *[Einschätzung]*

9. Was den Ausbau begrenzt: Chips, Netzanschluss, Genehmigung

Das vorige Kapitel beantwortet, wie viel Rechenkapazität gebraucht wird. Dieses Kapitel beantwortet, warum das Angebot nicht einfach folgt. Zwei Knappheiten begrenzen den Ausbau, eine zyklische in der Fertigung und eine strukturelle am Netzanschluss, und an zwei Preisen lässt sich ablesen, wie eng es jeweils ist.

Knappheit I: die Halbleiter

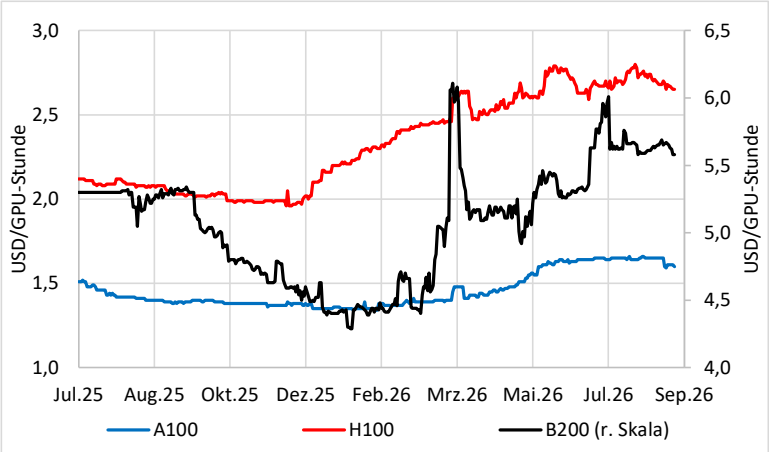
Die erste Knappheit sitzt in der Fertigung. Sie hat drei Glieder, die gleichzeitig eng sind: die Belichtung der Wafer, das Zusammensetzen von Rechenchip und Speicher und den Speicher selbst. Die Zuteilung der aktuellen Beschleunigergeneration ist nach Branchenberichten bis in das Jahr 2027 vergeben, die Gehäusekapazität bei TSMC gilt als ausverkauft mit Entspannung frühestens gegen Ende 2026, und die HBM-Fertigung verdrängt gewöhnliche DRAM-Kapazität. Das zieht die Kontraktpreise für Arbeitsspeicher mit nach oben, also auch dort, wo keine KI im Spiel ist.

Diese Knappheit ist zyklisch. Der Speichermarkt hat das mehrfach vorgeführt: Auf jede Engpassphase folgt ein Angebotsüberhang, weil alle Anbieter gleichzeitig ausbauen. Die Reaktionszeit liegt bei zwei bis drei Jahren. Was 2025 und 2026 bestellt wurde, kommt 2028 und 2029 an. Den Wendepunkt auf der Halbleiterseite erwarte ich deshalb um 2028/29, und die Kontraktpreise für DDR4 und DDR5 — die beiden gängigen Generationen von Arbeitsspeicher — werden ihn frühzeitig anzeigen. *[Einschätzung]*

Zwei Preise, zwei Aussagen

- **Preis neuer Kapazität.** Chip- und Speicherpreise, verfolgbar über die Kontraktpreise. Steigende Preise zeigen ungebrochene Nachfrage nach neuer Kapazität; die Knappheit dahinter beschreibt der vorige Abschnitt.
- **Preis installierter Kapazität.** Der Mietpreis je Rechenstunde zeigt, wie gut die vorhandenen Anlagen ausgelastet sind. Über die vergangenen zwei Jahre war er für die H100-Generation am unteren Ende des Marktes gefallen. Seit Ende 2025 sind die Mietpreise allerdings bis Juli 2026 gestiegen, und zwar über alle drei Generationen. Dass selbst die älteste Generation zulegte, spricht für einen Mengenengpass und ist das stärkste einzelne Indiz für die Knappheitsthese dieses Dokuments. Zuletzt haben sie moderat nachgegeben; eine Wende ist das noch nicht — es ist aber die Reihe, an der ich sie zuerst sähe. *[Einschätzung]*

Beobachtbar ist nur die kurzfristige Anmietung — der weit überwiegende Teil der Kapazität läuft über Mehrjahresverträge oder konzerninterne Verrechnung und hat gar keinen Preis. Steigt der Mietpreis über alle drei Chipgenerationen gleichzeitig, auch bei den ältesten Karten, spricht das für einen Mengenengpass; steigt nur das obere Ende, kann ebenso gut Angebot in Langfristverträge abgewandert sein.



Mietpreise je Rechenstunde nach Chipgeneration (USD)

Quelle: Silicon Data Rental Price Indices über Bloomberg, Stand September 2026; B200 rechte Skala

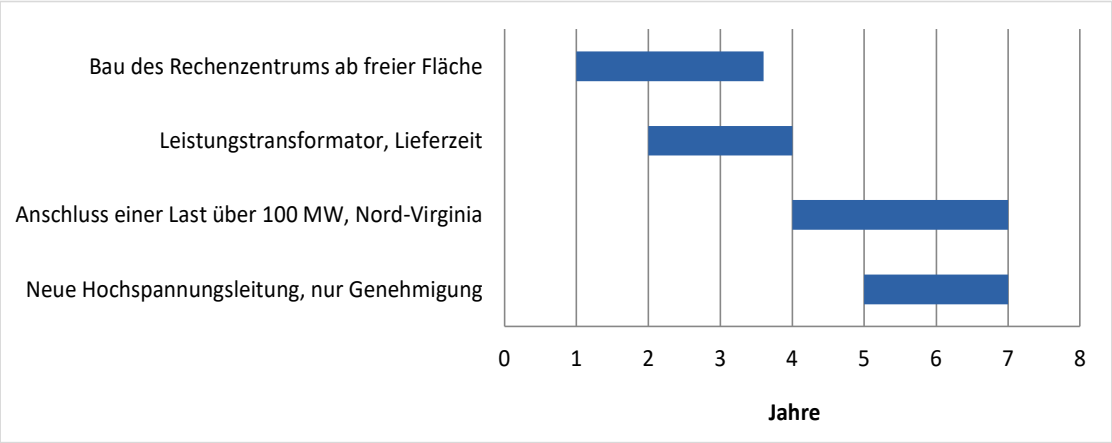
Ab dem 5. Oktober 2026 handelt die CME Terminkontrakte darauf: Der Restwert bekommt damit erstmals einen Marktpreis statt einer Bilanzierungsannahme — und einen Treiber mehr, die Positionierung. *[Einschätzung]*

> **Fazit.** Fallende Mietpreise für die Vorgängergeneration wären normale Entwertung und ein Argument in der Debatte um die Abschreibungsdauern. Deutlich fallende Mietpreise für die jeweils aktuelle Generation bei weiter steigender Tokennutzung wären das Warnsignal: Dann wäre mehr Kapazität gebaut worden, als gebraucht wird. Deutlich heißt dabei eine Größenordnung, die den Anstieg seit Ende 2025 zurücknimmt; zehn oder zwanzig Prozent nach einem Anstieg über alle drei Generationen sind das nicht. Zuletzt haben die Mietpreise moderat nachgegeben. Ob daraus eine Wende wird, zeigen die nächsten Monate. *[Einschätzung]*

Knappheit II: Rechenzentren, Netz und Genehmigung

Nach der Auswertung von Epoch AI entstehen Rechenzentren im Gigawattmaßstab von der freien Fläche bis zum Betrieb ein bis 3,6 Jahre. Für Transformatoren gelten nach Wood Mackenzie Lieferzeiten von 24 bis 36 Monaten, bei großen Hochspannungseinheiten bis zu vier Jahren; das begrenzt auch dort, wo genehmigt ist. Der Anschluss einer großen Last dauert je nach Netzgebiet drei bis sieben Jahre; für Lasten über 100 MW nennt Dominion in Nord-Virginia vier bis sieben Jahre. Eine bundesweite Erhebung dazu gibt es nicht.

Die zweite Knappheit halte ich deshalb für strukturell und nicht für zyklisch. Nicht der Bau begrenzt die verfügbare Rechenkapazität, sondern der Anschluss ans Netz. Wie eng es ist, zeigen die Warteschlangen: In Nord-Virginia stehen rund 70 GW angemeldeter Last gegen eine Systemhöchstlast von 24,7 GW, davon 25 GW mit Anschlussterminen bis Ende 2031. Bei ERCOT wuchs die Warteschlange großer Lasten von 63 GW Ende 2024 auf über 230 GW Ende 2025, zu mehr als 70 % Rechenzentren; freigegeben zur Energetisierung wurden in den zwölf Monaten bis Mai 2026 rund 9 GW. *[Einschätzung]*



Was den Ausbau begrenzt: Vorlaufzeiten im Vergleich (Jahre)
Quelle: Dominion Energy Virginia, Wood Mackenzie, Epoch AI; Stand September 2026

Der Ausweg: Strom am Netz vorbei

Mehrjährige Wartezeiten sind für viele Vorhaben zu lang. Der Ausweg ist der Stromkauf direkt am Kraftwerk. Microsoft hat sich über einen Zwanzigjahresvertrag die Leistung des wiederanzufahrenden Blocks 1 von Three Mile Island (Crane Clean Energy Center, ab 2027) gesichert, Amazon betreibt ein Rechenzentrum unmittelbar am Kernkraftwerk Susquehanna, Google hat Strom kleiner modularer Reaktoren kontrahiert (Kairos Power, bis zu 500 MW). Diese Anlagen liefern aber frühestens gegen Ende des Jahrzehnts. Zudem hat die Aufsicht die Ausweitung der Direktbelieferung in Susquehanna 2024 abgelehnt und PJM im Dezember 2025 zu Regeln dafür angewiesen; die Intervenienten argumentierten, dass die übrigen Netzkunden die Fixkosten des Netzes sonst allein tragen. Darin liegt der wunde Punkt: Jede Anlage, die sich vom Netz abkoppelt, verschiebt Kosten auf die Haushalte und macht den Ausbau zum politischen Thema.

Strom, Wasser, Genehmigungen und wachsender Widerstand

- **Preiswirkung.** Im Netzgebiet PJM stieg der Kapazitätspreis von 28,92 USD je MW und Tag für 2024/25 auf 329 USD für 2026/27; dies ist zugleich die genehmigte Preisobergrenze, an der die Auktion räumte.
- **Wasser.** Große Anlagen mit Verdunstungskühlung verbrauchen mehrere Millionen Liter Wasser am Tag.
- **Widerstand vor Ort.** Nach der Zählung von Data Center Watch, einem Projekt der KI-Sicherheitsfirma 10a Labs, wurden allein im ersten Quartal 2026 mindestens 75 Vorhaben mit rund 130 Mrd. USD blockiert oder verzögert, so viel wie im gesamten Jahr 2025.

Die Reihenfolge der kommenden Jahre

Aus den unterschiedlichen Fristen ergibt sich eine Abfolge. Auf der Halbleiterseite erwarte ich die Entspannung 2028/29; Netzanschluss und Genehmigung bleiben darüber hinaus knapp. Ein Überhang entstünde deshalb zuerst bei Chips und Speicher, nicht bei angeschlossenen Standorten, und sichtbar würde er an den beiden Reihen dieses Kapitels: Der Mietpreis der aktuellen Generation fällt deutlich, während die Tokenmengen weiter steigen. Das wäre zugleich die Marke, an der die Wertschöpfung an das Ende der Kette wandert. *[Einschätzung]*

> **Fazit.** Die Rechenkapazität bleibt aus meiner Sicht der Engpass, auf beiden Seiten. Neue Chips samt Speicher sind knapp, die installierte Basis ist ausgelastet, und der Ausbau hängt nicht am Kapital, sondern am Netzanschluss und an der Genehmigung. Das stützt die Preise und verlängert den Zyklus — und es heißt zugleich, dass das Angebot auf einen Nachfragerückgang nur träge reagieren würde. *[Einschätzung]*

10. Token, Preise und Mengen: die Ökonomie in einer Einheit

Der Token ist die Abrechnungseinheit dieser Branche. Ein Token ist ein Wortbaustein; ein deutsches Wort besteht meist aus zwei bis drei davon. Abgerechnet wird getrennt nach Eingabe und Ausgabe, üblicherweise je Million Token. Der Nutzen dieser Einheit liegt darin, dass sich Umsatz, Kosten und Kapazität in derselben Größe messen lassen. Damit lässt sich die Monetarisierungsfrage nachrechnen.

Zwei Preisreihen, die man trennen muss

„Den" Tokenpreis gibt es nicht, sondern zwei Reihen mit gegenläufiger Aussage. Was eine gegebene Leistung kostet, sinkt seit Jahren erheblich — nach der Auswertung von Epoch AI je nach Prüfaufgabe um das Neun- bis Neunhundertfache im Jahr, für das Niveau von GPT-4 auf naturwissenschaftlichen Fragen um das Vierzigfache. Was die jeweils beste verfügbare Leistung kostet, bewegt sich dagegen seitwärts, weil jede Modellgeneration am oberen Ende neu ansetzt. Der größere Teil des Verfalls stammt dabei nicht aus der Hardware, sondern aus Modellarchitektur und Betrieb.

Die Unterscheidung ist der Prüfstein für die These vom Preiskrieg: Fielen beide Reihen zugleich, wäre es Verdrängungswettbewerb mit Margendruck; fällt nur die erste, ist es der gewöhnliche Verlauf einer Technologiekurve mit Segmentierung — Standardaufgaben werden billig, anspruchsvolle bleiben teuer. Der Anreiz, gebaute Kapazität notfalls unter Vollkosten zu verkaufen, greift ohnehin erst bei freier Kapazität: Solange die Anlagen ausgelastet sind, hat jede belegte Stunde einen Opportunitätspreis, und der Preisboden liegt beim Mietwert des Chips statt bei null. *[Einschätzung]*

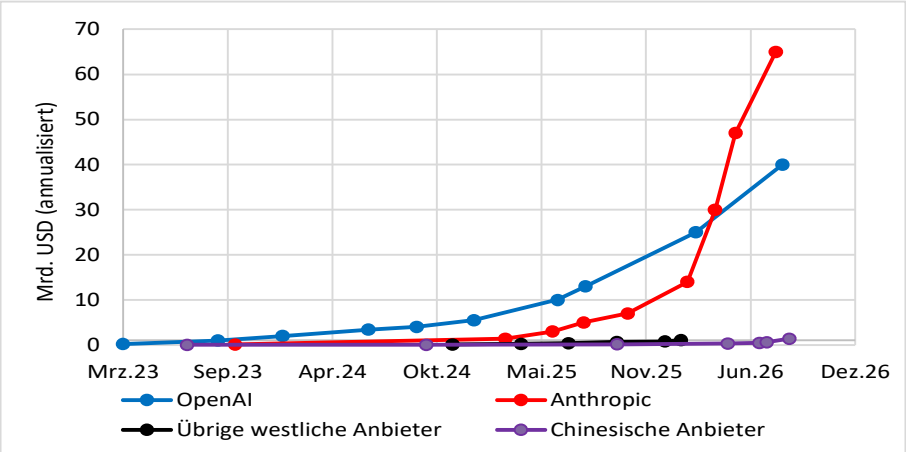
Warum fallende Preise die Umsätze bisher steigen lassen

Der Umsatz ist Preis mal Menge: Fällt der Preis je Einheit um die Hälfte und steigt die Menge auf das Dreifache, steigt der Umsatz um die Hälfte. Dieses Muster liegt bislang vor: Alphabet meldete für das Frühjahr 2024 monatlich 9,7 Bio. verarbeitete Token, im Mai 2025 480 Bio. und im Frühjahr 2026 3,2 Billiarden. Und die KI-Ausgaben je Beschäftigten haben sich nach den Daten des Zahlungsdienstleisters Ramp 2026 mehr als verdoppelt. *[Rechnung]*

Dahinter steht ein bekannter Mechanismus: Sinkende Kosten je Einheit erhöhen den Gesamtverbrauch, statt ihn zu senken. Anwendungen, die bei zehn Dollar je Vorgang unwirtschaftlich waren, werden bei zehn Cent gerechnet — agentische Systeme setzen dort an. Der Preisverfall ist damit die Voraussetzung der Monetarisierung, nicht ihr Gegenargument.

Was auf der Umsatzseite tatsächlich ankommt

Die Mengen allein beweisen noch keine Zahlungsbereitschaft. Dafür sind die Umsätze der Modellanbieter der direktere Beleg, und sie sind inzwischen öffentlich nachvollziehbar: Epoch AI führt die gemeldeten Zahlen fort. Danach lag der annualisierte Umsatz von OpenAI Ende 2025 bei rund 21 Mrd. USD und Mitte August 2026 bei rund 40 Mrd. USD. Anthropic kam Ende 2025 auf 9 Mrd. USD und Ende Juli 2026 auf rund 65 Mrd. USD. Das ist eine Versiebenfachung binnen sieben Monaten und die Ablösung des bisherigen Marktführers.



Annualisierte Umsätze der KI-Anbieter (Mrd. USD)

Quelle: Epoch AI, Sammlung gemeldeter Zahlen, Stand September 2026. Die Stufen zeigen den jeweils letzten gemeldeten Wert; Teilsegmente und Jahresabschlüsse sind ausgelassen, um Doppelzählungen zu vermeiden

In der Summe über alle erfassten Anbieter stehen damit rund 110 Mrd. USD annualisiert, nach rund 6 Mrd. USD Ende 2024 — etwa das Achtzehnfache in gut anderthalb Jahren. Zwei Vorbehalte gehören dazu. Anthropic stellt allein rund 59 % dieser Summe, und die 65 Mrd. USD sind aus dem Ausgangsniveau Ende Juli hochgerechnet — das zweite Quartal ergäbe annualisiert 46; die beiden Anbieter messen zudem nicht notwendig gleich. Und es ist nicht dieselbe Größe wie die zurechenbaren Erlöse in Kapitel 12: Dort kommen die vermietete Rechenkapazität und die KI-Erlöse der Betreiber aus Drittkundengeschäft hinzu, soweit sie nicht schon in diesen Umsätzen stecken (Anhang C). Für diesen dritten Block gibt es

genau eine ausgewiesene Zahl: Microsoft nennt eine annualisierte KI-Größe von 37 Mrd. USD für das Quartal bis März 2026, und Bloomberg schätzt, dass zwei Drittel davon auf OpenAI zurückgehen — Azure-Verbrauch und Erlösbeteiligung. Alphabet und Amazon weisen nichts Vergleichbares aus.

Zwei Punkte sind wichtiger als die absoluten Zahlen. Erstens wächst der Umsatz schneller als die verarbeiteten Mengen; die Zahlungsbereitschaft je Vorgang steigt also. Zweitens ist der Abstand zu den chinesischen Anbietern auf der Erlösseite weit größer als auf der Qualitätsseite: Sie kommen zusammen auf rund 3,2 Mrd. USD, rund drei Prozent des Gesamtwerts. Über Ranglisten wirkt der Wettbewerb aus China deshalb bedrohlicher als über die Erlöse.

Einschränkend gilt: Es sind annualisierte Momentaufnahmen aus Unternehmensangaben und Presseberichten, keine testierten Abschlüsse. Zudem kommt ein erheblicher Teil der Nachfrage von wagniskapitalfinanzierten Unternehmen, die mit Investorengeld Rechenzeit kaufen; quantifizieren lässt sich das nicht. Für die Frage, ob die Leistung bezahlt wird, bleibt die Reihe dennoch der beste verfügbare Beleg.

Wo der Mischerlös tatsächlich liegt

Die Listenpreise reichen von rund 0,20 USD je Million Eingabe-Token bis in den zweistelligen Dollarbereich, und Rabatte werden nicht veröffentlicht. Zur Erlösberechnung wird ein ausgabengewichteter Durchschnitt über den tatsächlichen Verkehr gebraucht — und den gibt es: Silicon Data veröffentlicht ihn täglich als LLM Token Expenditure Index über einen Korb von gut zwanzig Modellen aus einem beobachteten Universum von über 300. Im Dezember lag er im Mittel bei 1,09 USD je Million Token und stieg bis Ende Mai auf rund 2,00. Seither gibt er nach: 1,55 im Monatsmittel im Juli, 1,11 im August. Er ist ausgabengewichtet und misst damit zweierlei: den Preis einer Leistung und die Zusammensetzung des Verkehrs. Ausschläge um zwei Drittel binnen eines halben Jahres sind keine Preisbewegung, sondern ein Wechsel zwischen teuren und billigen Nutzungsarten.

Vergleichbar wird der Indexpreis erst nach einer Umrechnung: Er misst, was je Million abgerechneter Token in Rechnung gestellt wird — im August 2026 rund 1,11 USD. Daneben steht eine zweite Größe, die anders gebaut ist: Die zurechenbaren Erlöse geteilt durch die verarbeitete Menge ergeben 1,02 bis 1,48 USD je Million, je nach Abgrenzung (Anhang C). Die beiden lassen sich nicht ineinander umrechnen. Ein erheblicher Teil der Erlöse wird gar nicht je Token abgerechnet — Abonnements, Unternehmensverträge, vermietete Rechenzeit —, und ein noch größerer Teil der Menge bringt keinen zurechenbaren Erlös: der Eigenbedarf der Anbieter und die chinesische Nutzung. Beides zusammen erklärt, warum je verarbeitetem Token weniger ankommt als der Indexpreis je abgerechnetem. Zu decken wären rund 2,20 USD (Kapitel 12). Der Eigenbedarf ist dabei mit null angesetzt: nicht wertlos, nur nicht messbar. *[Rechnung]*

Exkurs: Was die Tokenmenge zunehmend begrenzen könnte

Die Mengen können nicht beliebig wachsen. So muss für jedes erzeugte Token ein Modell seine Gewichte einmal vollständig aus dem Speicher lesen. Wie schnell das geht, hängt nicht daran, wie schnell der Chip rechnen kann, sondern daran, wie schnell sein Speicher die Daten liefert — und das ist je Chip eine feste physikalische Größe. Jede gebaute Anlage hat damit eine Obergrenze an Token je Tag, die sich durch bessere Software nicht heben lässt, sondern nur durch mehr oder schnellere Chips.

Wie nah die installierten Beschleuniger — im Folgenden die Flotte — an dieser Grenze sind, lässt sich nicht in einer Zahl sagen: Es hängt daran, wie lang die Eingaben sind. Epoch AI beziffert die Kapazität der aktuellen Chipgeneration je nach Kontextlänge auf 43 bis 1.730 Bio. Token am Tag; verarbeitet werden derzeit rund 360 Bio. Bei kurzen Anfragen ist also reichlich Luft, bei langen Kontexten und agentischer Nutzung ist die Grenze bereits erreicht — Anthropic senkt zu Spitzenzeiten Kontingente. Der Abstand schrumpft in jedem Fall, weil die Nachfrage schneller wächst als die Kapazität: Nach Epoch AI wächst die Rechenkapazität im Jahr auf gut das Dreifache, die Nachfrage nach Token auf rund das Zehnfache.

Damit ist die Grenze keine ferne Größe, sondern eine Frage weniger Jahre — hinausgeschoben nur dadurch, dass Hardware, Verfahren und Auslastungssteuerung fortlaufend besser werden. Ist sie erreicht, wächst die verarbeitete Menge nur noch so schnell wie die Angebotsseite selbst — mehr Chips, schnellerer Speicher, sparsamere Verfahren (Kapitel 11) — und über die Rendite des Ausbaus entscheidet dann nicht mehr hauptsächlich die Menge, sondern zunehmend der Preis je Token. *[Einschätzung]*

> **Fazit.** Solange die Kosten je Token schneller fallen als der Preis, sehe ich im Preisverfall kein Warnsignal. Kehrt sich das um, stellt sich die Monetarisierungsfrage neu. *[Einschätzung]*

11. Was die Wertschöpfungskette verschieben könnte

Die bisherigen Kapitel beschreiben die Kette, wie sie heute verläuft. Mehrere Entwicklungen können sie verschieben. Vier der hier dargestellten verlagern Wertschöpfung von einer Stufe auf eine andere, ohne einen Nachfrageausfall vorauszusetzen; die fünfte hebt eine physikalische Schranke.

Eigene Chips der Betreiber

Die großen Betreiber entwickeln zunehmend eigene Beschleuniger: Google seit Jahren die TPU, Amazon Trainium und Inferentia, Meta die MTIA-Reihe. Nach der Auswertung von Epoch AI entfallen auf den Marktführer NVIDIA gut 60 % der weltweit gelieferten Rechenleistung; den größten Teil des Rests stellen Google und Amazon mit eigenen Entwürfen her. Der zweit- und drittgrößte Produzent von KI-Rechenleistung verkauft seine Chips also gar nicht.

Eigene Chips sparen die Handelsmarge des Anbieters und lassen sich auf die eigene Arbeitslast zuschneiden — in der Inferenz, wo dieselbe Aufgabe milliardenfach läuft, wiegt das schwer. Durchsetzen werden sie sich zuerst dort, wo der Betreiber die Software selbst kontrolliert, denn der Vorsprung des Marktführers liegt weniger im Chip als in Software und Netzwerktechnik. Der Druck wirkt damit auf die Marge des Chipanbieters, nicht auf das Volumen der Investitionen. *[Einschätzung]*

Rechnen auf dem Endgerät

Kleine Modelle laufen inzwischen auf Telefonen und Rechnern, deren Prozessoren dafür eigene Recheneinheiten mitbringen. Dies spart Übertragung und Wartezeit, hält Daten im Unternehmen und funktioniert ohne Netz. Die Grenze ist dabei nicht die Rechenleistung des Geräts, sondern wieder die Speicherbandbreite — ein Telefon schafft bei einem Acht-Milliarden-Modell zehn bis fünfzehn Token je Sekunde, ein Notebook mit KI-Prozessor zwanzig bis vierzig. Was abwandert, sind kurze, häufige, unkritische Vorgänge; agentische Ketten mit langem Verlauf wandern zuletzt. *[Einschätzung]*

Wichtiger als Telefone und Notebooks sind Geräte, die in Sekundenbruchteilen auf ihre Umgebung reagieren müssen — Roboter, Fahrzeuge, Fertigungsanlagen. Dort ist der Weg in ein Rechenzentrum und zurück gar keine Möglichkeit, und es entsteht eine zweite Produktkategorie, die es ohne brauchbare Modelle auf dem Gerät nicht gäbe. Für sie hat Nvidia den Begriff Physical AI geprägt: KI, die nicht Text verarbeitet, sondern in der physischen Welt wahrnimmt und handelt. Für Europa ist das die interessanteste Öffnung des Feldes — Modelle und Rechenkapazität kommen aus den USA und China, aber die Maschinen, die Fertigungsdaten und das Prozesswissen sitzen zu einem erheblichen Teil hier. *[Einschätzung]*

Ersetzt wird das Rechenzentrum dadurch nicht. Dasselbe Silizium rechnet dort rund um die Uhr für viele Anfragen statt ein paar Stunden am Tag für einen Nutzer, und die Kosten je Token bleiben entsprechend niedriger. Vor allem aber ist das meiste, was auf die Geräte geht, Arbeit, die vorher gar nicht stattfand: Die Nachfrage kommt hinzu, sie wandert nicht ab.

Offene Gewichte

Modelle mit frei verfügbaren Gewichten, von Meta, aus China und von europäischen Anbietern, setzen eine Obergrenze für das, was vergleichbare Leistung kosten darf. Wer ein offenes Modell selbst betreibt, zahlt nur noch Rechenzeit. Für reine Modellanbieter ohne eigene Infrastruktur oder eigenen Vertriebsweg ist das die ernsteste Bedrohung des Geschäftsmodells. Die Wirkung ist allerdings ungleich verteilt: Sie trifft die Standardleistung hart und die Spitzenleistung kaum, weil offene Modelle dem besten geschlossenen mit Abstand folgen. Und sie verschiebt Nachfrage nach unten in der Kette — was die Modellanbieter an Marge verlieren, gewinnen die Betreiber der Infrastruktur.

Effizienzsprünge in Modell und Betrieb

Neben den Modellen mit reiner Aufmerksamkeitsarchitektur haben sich Mischformen etabliert, deren Rechenaufwand nicht mehr mit dem Quadrat der Kontextlänge wächst und deren Speicherbedarf je Vorgang nicht mehr mit ihr mitwächst (dazu der übernächste Abschnitt). In der Praxis sind das zwei- bis zehnfache Durchsatzgewinne, und sie fallen bei langen Kontexten an, also genau dort, wo die agentische Nutzung den Bedarf treibt. Die beiden Entwicklungen laufen gegeneinander: Der Vorgang wird länger, die Verarbeitung je Token billiger.

Wie weit der Hebel noch reicht, lässt sich abschätzen. Ein erheblicher Teil des bisherigen Fortschritts stammt aus der Rechengenauigkeit: Von zweiunddreißig auf vier Bit sind drei Halbierungen, jede mit doppelter Leistung und halbem Speicherbedarf. Diese Schritte sind gegangen; bei zwei Bit bricht die Qualität deutlich ein. Der gemessene Preisverfall enthält also einen einmaligen Beitrag, der absehbar ausläuft — ich erwarte, dass er sich verlangsamt. *[Einschätzung]*

Wenn die Bauweise selbst wechselt

Die Effizienzgewinne des vorigen Abschnitts machen dasselbe billiger. Sie haben den Ausbau nicht gebremst, sondern getragen: Billigere Token bedeuten mehr Token. Ein anderes Risiko wäre es, wenn nicht dasselbe billiger würde, sondern die Beziehung zwischen Rechenaufwand und Leistung selbst bräche.

Heutige Modelle führen den gesamten bisherigen Verlauf eines Vorgangs bei jedem neuen Wort erneut mit; je länger er wird, desto mehr muss durch den Speicher. Es gibt Bauweisen, die stattdessen eine laufend fortgeschriebene Zusammenfassung mitführen, die immer gleich groß bleibt. Damit fällt genau die Schranke weg, an der die Tokenmenge hängt — ein Modell dieser Art mit drei Milliarden Parametern erreicht die Leistung eines doppelt so großen herkömmlichen bei bis zum fünffachen Durchsatz (Gu/Dao, „Mamba“, 2023; Durchsatzvorteil gegenüber gleich großem Transformermodell). Der Preis dafür: Wer eine Zusammenfassung mitführt statt eines Protokolls, trifft den genauen Wortlaut von weiter oben schlechter — und genau darauf beruhen agentische Ketten über 20 Schritte.

Modelle, die beide Bauweisen mischen, laufen bereits im Produktivbetrieb, kein reines hat bisher Spitzenniveau erreicht, und ich erwarte, dass die heutige Bauweise eher bis zum Ende des Jahrzehnts vorherrscht; die Fachliteratur nennt keine Jahreszahl. Für die kommenden zwei bis drei Jahre halte ich einen Bruch deshalb für unwahrscheinlich. Entscheidend ist aber die Asymmetrie: Die heute gebaute Kapazität ist über sechs Jahre Beschleunigerlebensdauer finanziert und setzt damit stillschweigend voraus, dass die heutige Bauweise sechs Jahre trägt — während die Forschung nur für zwei bis drei Jahre eine belastbare Aussage zulässt. Diese Lücke unterschreibt niemand ausdrücklich, und sie trifft nicht die Nachfrage nach KI, sondern den Restwert der Anlagen. *[Einschätzung]*

Photonik: Licht statt Elektronen — und warum es um Strom geht

Die fünfte Verschiebung ist die einzige, die eine Schranke hebt statt eine Marge zu drücken. In den Anlagen werden die Verbindungen zwischen den Chips von elektrischer auf optische Übertragung umgestellt: Glasfaser statt Kupferleitung, bis ins Chipgehäuse hinein.

Das Argument dafür ist nicht Geschwindigkeit, sondern Strom. Eine elektrische Leitung setzt einen Teil der aufgewendeten Energie in Wärme um und braucht am Ende Schaltungen, die das verrauschte Signal wieder herrichten; beides kostet Leistung und muss zusätzlich gekühlt werden. Licht verliert auf diesen Strecken kaum Energie. Publiziert sind rund 15 Pikojoule je übertragenem Bit bei steckbaren Modulen gegen rund 5 bei integrierter Optik (GlobalFoundries); NVIDIA gibt für seine photonischen Netzwerkschalter die 3,5-fache Energieeffizienz an.

Warum das gerade jetzt zählt: Ein großes Modell liegt nicht auf einem Chip, sondern verteilt auf vielen, und für jedes erzeugte Token müssen Zwischenergebnisse zwischen ihnen hin und her. Je größer die Modelle, desto mehr Strom geht in die Wege statt ins Rechnen. Und der bindende Engpass beim Ausbau ist der Netzanschluss mit drei bis sieben Jahren — innerhalb eines genehmigten Anschlusses ist jedes gesparte Watt ein Watt für Beschleuniger. Serienfertigung ist ab 2028 angekündigt, die volle Integration in KI-Pakete für 2029 — im Zeitraum dieses Dokuments wirkt sie damit kaum. Dasselbe gilt für Bauelemente, die nicht nur übertragen, sondern rechnen: Das Verkaufsargument für photonische Halbleiter ist nicht die Rechenleistung, sondern der Strombedarf je Rechenschritt. Bis zur Serienreife mit hoher Stückzahl wird es aber noch mehrere Jahre dauern. *[Einschätzung]*

> **Fazit.** Vier der fünf Verschiebungen drücken die Marge am Ende der Kette und lassen die Nachfrage nach Rechenkapazität unberührt oder erhöhen sie; die fünfte hebt eine Schranke. Eine Infrastrukturthese wird davon nicht getroffen. Getroffen würde sie von einem Wechsel der Bauweise, und das ist das einzige der fünf Themen, das ich als Risiko für den Restwert der Anlagen führe. *[Einschätzung]*

12. Wie der Ausbau finanziert wird — und wie man die Rendite prüft

Bis 2025 wurde der Ausbau aus dem laufenden Mittelzufluss der Betreiber bezahlt. Diese Phase ist vorbei: Die Investitionen übersteigen den operativen Cashflow, und damit entscheidet der Kapitalmarkt über das Tempo. Ich stelle zuerst das Ergebnis der Gesamtrechnung vor, erkläre dann zwei Finanzierungsformen, rechne, was eine einzelne Anlage verdienen muss, und gehe zuletzt durch, in welcher Reihenfolge eine Kürzung durch die Kette liefe.

Was das System verdienen muss — und was es verdient

Die brauchbare Frage lautet, was die bereits gebaute Kapazität jährlich verdienen muss — und was sie tatsächlich verdient. Der Anhang rechnet beides Schritt für Schritt vor; hier steht das Ergebnis.

Stand August 2026	Wert	Hergeleitet in
Systemkosten der installierten Kapazität ex China	rund 260 Mrd. USD p. a.	Anhang A
Verarbeitete Tokenmenge ex China, auch nicht abgerechnete	rund 118 Bill. p. a.	Anhang B
Zurechenbare Erlöse	120 bis 175 Mrd. USD p. a.	Anhang A
Deckungsgrad	0,46 bis 0,68	Erlöse durch Kosten
Nötiger Erlös je Mio. verarbeiteter Token	rund 2,20 USD	Anhang B
Tatsächlicher Erlös je Mio. verarbeiteter Token	1,02 bis 1,48 USD	Anhang C

Der Deckungsgrad — genauer der Kostendeckungsgrad — sagt, welcher Anteil der Kosten hereinkommt: Erlöse gegen Wertverzehr, Betrieb und Kapitalkosten, gerechnet für die gesamte installierte Kapazität statt für ein einzelnes Unternehmen. Die Lücke schließt sich nicht über den Preis, sondern über die Menge: Dieselbe Kapazität trägt immer mehr Token, und damit fallen die Kosten je Token. Daraus folgt die Regel: der Deckungsgrad steigt, solange der Preis je Token langsamer fällt als die Kosten je Token. Im vergangenen Jahr war das der Fall: Die Kosten je verarbeitetem Token fielen von 10,60 auf 2,20 USD, der Erlös je Token von 1,90 auf 1,02 — beide fielen, die Kosten aber schneller. Der Deckungsgrad stieg deshalb von 0,18 auf 0,46 — beides in der engen Abgrenzung, die einen Vergleich über die Zeit erlaubt. Weiter gefasst, mit den KI-Erlösen der Betreiber aus Drittkundengeschäft, liegt er für den August bei 0,68 (Anhang A). Kosten und Menge sind dabei auf die Kapazität ohne China abgegrenzt, weil die Erlöse fast nur aus dem Westen stammen: Der chinesische Anteil — nach Epoch AI gut 5 % der Rechenleistung, nach Zählungen öffentlicher Rechencluster 12 bis 15 % (ebenfalls Epoch-Daten, andere Abgrenzung) — ist mit 10 % herausgerechnet (Anhang B). Weltweit gerechnet läge der Deckungsgrad bei 0,42. *[Rechnung]*

Dieser Weg hat allerdings eine Grenze. Die Tokennachfrage wächst nach Epoch AI etwa dreimal so schnell wie das Angebot, und bei langen Kontexten ist die Bandbreite der Flotte bereits das bindende Maß. Läuft das so weiter, ist der Puffer in wenigen Jahren aufgebraucht. Von da an kann die Menge die Lücke nicht mehr schließen: Sie wächst dann nur noch so schnell wie die Flotte selbst, und die Kosten wachsen mit ihr. *[Rechnung]*

Damit verschiebt sich die Frage. Sie lautet dann nicht mehr, ob die Mengen schnell genug wachsen, sondern ob der Preis steigt. Das wäre keine kühne Annahme: Ein System, das an seiner Grenze läuft, ist genau die Lage, in der Preise steigen. Die Mietpreise für Beschleuniger sind zwischen Dezember 2025 und Juli 2026 über alle drei Generationen deutlich gestiegen und haben seither moderat nachgegeben, und einzelne Anbieter haben Kontingente zu Spitzenzeiten gekürzt. *[Einschätzung]*

> **Fazit.** Solange die Menge schneller wächst als die Flotte, steigt der Deckungsgrad ohne Zutun des Preises; sobald die Bandbreite bindet, wächst sie nur noch im Takt der Flotte, und der Preis muss den Rest tragen. Welcher Zustand vorliegt, zeigen der Mietpreis der aktuellen Chipgeneration und die verarbeitete Tokenmenge im Verbund: Steigen beide, drückt die Nachfrage gegen die Grenze; gibt der Mietpreis bei steigenden Mengen nach, ist noch Durchsatz übrig. Der Mietpreis ist der Preis des knappen Durchsatzes selbst. *[Einschätzung]*

Kreisgeschäfte und bilanzexterne Finanzierung

Zwei Konstruktionen prägen die Finanzierung, und beide lassen die Zahlen besser aussehen, als sie sind.

Die erste: Ein Chipanbieter oder Cloud-Betreiber beteiligt sich an einem Modellanbieter, der sich im Gegenzug verpflichtet, für einen erheblichen Teil des Geldes Rechenleistung beim Geldgeber zu kaufen. Im November 2025 sagten Nvidia bis zu 10 Mrd. USD und Microsoft bis zu 5 Mrd. USD an Anthropic zu; Anthropic verpflichtete sich zu Cloud-Käufen über 30 Mrd. USD und zur Abnahme von bis zu einem Gigawatt Rechenleistung. Vergleichbare Konstruktionen gibt es zwischen mehreren Unternehmen. Nur ist Umsatz, der aus dem Kapital des Verkäufers stammt, kein Beleg für die Zahlungsbereitschaft Dritter:

Er erscheint sofort in der Gewinn- und Verlustrechnung, das Risiko steht in der Beteiligung und wird erst bei einer Abwertung sichtbar. Bei mehrfacher Verkettung erscheint dieselbe Zahlung an mehreren Stellen als Umsatz.

Die zweite: Die spezialisierten Vermieter finanzieren ihre Anlagen über eigenständige Gesellschaften, die nur das Rechenzentrum samt Kundenvertrag halten. Besichert wird mit den Chips und dem Vertrag, nicht mit der Konzernbilanz — CoreWeave hat im März 2026 die erste Fazilität dieser Art mit Investment-Grade-Bewertung abgeschlossen, über 8,5 Mrd. USD ohne Rückgriff auf die Mutter. Das Risiko sitzt in der Restwertannahme: Die Laufzeit reicht über den Vertrag hinaus, und für die Zeit danach unterstellt die Struktur, dass sich die dann veraltete Anlage weiter vermieten lässt. Fällt der Mietpreis schneller, trifft es nicht den Betreiber, sondern die Fremdkapitalgeber — und die sitzen zunehmend außerhalb des Bankensystems, in Kreditfonds und Versicherungsbeständen.

- **Was zu prüfen ist.** Bei Kreisgeschäften: welcher Anteil des Umsatzwachstums auf Kunden entfällt, an denen das Unternehmen beteiligt ist — zu finden allenfalls in den Anhangangaben zu nahestehenden Unternehmen. Bei Zweckgesellschaften: nicht die Nettoverschuldung, die den Vergleich zwischen den Betreibern verzerrt, sondern die ausgewiesenen Verpflichtungen aus Miet- und Abnahmeverträgen.

Was ein Gigawatt verdienen muss

Für die Renditefrage genügen wenige Größen und eine Handvoll Annahmen. Eine Anlage von einem Gigawatt kostet rund 38 Mrd. USD und entspricht rund 0,85 Mio. Beschleuniger — die Faustzahl aus Kapitel 8 für Gesamtleistung, nicht für IT-Leistung. Was sie einbringt, hängt an zwei Stellschrauben: dem erzielten Preis je Rechenstunde und der Auslastung. Die folgende Rechnung ist kein Ergebnis, sondern ein Rahmen. Sie zeigt, wie empfindlich die Amortisation auf beide Größen reagiert.

	Günstig	Mittel	Ungünstig
Erzielter Preis je Rechenstunde	2,50 USD	1,80 USD	1,20 USD
Auslastung	85 %	70 %	55 %
Erlös je Jahr und Gigawatt	rund 16 Mrd. USD	rund 9 Mrd. USD	rund 5 Mrd. USD
Erlös abzüglich Betriebskosten	rund 15 Mrd. USD	rund 8,5 Mrd. USD	rund 4 Mrd. USD
Amortisation der 38 Mrd. USD	rund 2,5 Jahre	gut 4,5 Jahre	über 9 Jahre

Die Rechnung unterstellt 8.760 Betriebsstunden im Jahr, 0,85 Mio. Beschleuniger je Gigawatt und Betriebskosten von 0,9 Mrd. USD im Jahr — die reinen Kosten des Rechenzentrumsbetriebs, ohne Abschreibung und ohne Kapitalkosten. Die Amortisation ist entsprechend eine einfache Rückflussdauer: wie lange es dauert, bis die 38 Mrd. USD zurückgeflossen sind. Zwischen der günstigen und der ungünstigen Spalte liegt der Unterschied zwischen einer Anlage, die sich in gut zweieinhalb Jahren bezahlt macht, und einer, die ihre eigene Nutzungsdauer nicht mehr erreicht. Dabei unterscheiden sich beide Spalten nur um den halbierten Preis und 30 Prozentpunkte Auslastung. *[Rechnung]*

Die Tabelle rechnet ohne Kapitalkosten, und das ist keine Kleinigkeit. Die Anlagen sind zunehmend fremdfinanziert, und jeder Prozentpunkt höherer Kapitalkosten schiebt die Amortisation nach hinten, am stärksten in der ungünstigen Spalte. Zugleich wirkt der Ausbau auf die Zinsen zurück: Nach der KI-Research-Serie gehen Sonderkonjunktur und Finanzierungsbedarf mit einem höheren neutralen Realzins und höheren realen Renditen einher. Die Branche hebt sich ihre Kapitalkosten ein Stück weit selbst an. *[Einschätzung]*

Wenn Investitionen gekürzt würden: die Reihenfolge

Würden die Investitionspläne spürbar gekürzt, etwa um ein Fünftel, liefe das in absehbarer Reihenfolge durch die Kette. Zuerst träfe es die Bestellungen bei Chips und Speicher; die Kontraktpreise für DRAM und NAND drehten binnen Monaten. Dann fielen die Mietpreise, weil bestellte Kapazität noch anrollte und auf weniger Nachfrage träfe — und mit ihnen die Restwerte: Sonderabschreibungen bei den Betreibern, Ausfälle bei den Fremdkapitalgebern der Zweckgesellschaften. Am längsten liefe die Bau- und Stromseite nach, denn genehmigte Vorhaben werden fertiggebaut. Es ist dieselbe Reihenfolge, in der die Kennzahlen in Kapitel 14 anschlagen würden — die schnellen Reihen zuerst. *[Einschätzung]*

> **Fazit.** Ein Teil der Erlöse ist Geld, das der Verkäufer dem Kunden zuvor als Beteiligung gegeben hat, ein wachsender Teil der Schulden steht nicht in den Bilanzen, in denen man ihn sucht, und beides ruht auf einer Annahme über den Restwert der Chips — derselben Größe, an der auch hängt, ob eine einzelne Anlage ihre Kapitalkosten verdient. Diese Größe hat inzwischen alles, was eine Assetklasse ausmacht: eine Einheit, einen täglich notierten Preis je Chipgeneration, einen Beleihungswert und ab Herbst eine Terminkurve. Die Objektfinanzierung, mit der Flugzeuge und Schiffe seit Jahrzehnten finanziert werden, hat mit „Compute" ein neues Objekt. *[Einschätzung]*

13. Was der Ausbau je Assetklasse bedeutet

Die Kapitel bis hierher erklären die Mechanik; diese Tabelle sagt, wen sie angeht. Sie nennt keine Positionierung — die steht in der KI-Research-Serie —, sondern den Wirkungsweg, das Kapitel, in dem er hergeleitet ist, und die Reihe, an der eine Wende zuerst sichtbar würde; Kapitel 14 führt diese Reihen mit Quelle und Frequenz auf.

	Wirkungsweg	Hergeleitet in	Frühindikator
Zinsen	Der Kapitalbedarf trifft auf ein begrenztes Sparangebot: über 220 Mrd. USD Anleiheemission allein am Dollar-Markt bis August in diesem Jahr, grob anderthalb Prozent der globalen Bruttoersparnis. Das spricht für einen höheren neutralen Realzins. Die Preiswirkung dreht im Zeitablauf: heute inflationär über Strom, Speicher und Bauleistung, wahrscheinlich ab 2028/29 disinflationär. [Einschätzung]	Kapitel 1 und 9	Emissionsvolumen und Laufzeiten der fünf Betreiber.
Credit	Dieselben Betreiber sind auf dem Weg, die Banken als größte Schuldnergruppe des US-Investment-Grade-Index abzulösen. Ein wachsender Teil der Finanzierung läuft über Gesellschaften, die nur eine Anlage halten; besichert wird mit Chips und Kundenvertrag, nicht mit der Konzernbilanz.	Kapitel 1 und 12	Aufschläge auf besicherte Rechenzentrumskredite.
Aktien	Gut ein Fünftel des Aktienrisikos eines Weltindex hängt an diesen Plänen — über die vier größten Betreiber und die sechs größten Zulieferer. Beide Gruppen hängen an derselben Investitionsentscheidung und fallen zugleich, wenn sie gekürzt wird.	Kapitel 1 und 7	Investitionen der fünf Betreiber in den Quartalsberichten; entscheidend ist die Wachstumsrate, nicht das Niveau.
Energie	Rechenzentren fragen Grundlast rund um die Uhr nach: ein Gigawatt Gesamtleistung rund 7 TWh im Jahr, etwa 1,3 % des deutschen Bruttostromverbrauchs. Der Netzanschluss ist zugleich die harte Grenze des Ausbaus.	Kapitel 8 und 9	Kapazitätspreise im Netzgebiet PJM und die Wartezeiten in den Anschlusswarteschlangen.
Währungen	Der Ausbau erhöht zugleich den amerikanischen Anspruch auf ausländische Ersparnis und die Chipimporte aus Taiwan und Südkorea. Die Souveränitätsprogramme verschieben Investitionsströme nach Europa, an den Golf und nach Asien.	Kapitel 7 und 11	US-Handelsdefizit, Halbleiterexporte Taiwans und Südkoreas, Auslandsanteil an den Anleiheemissionen.
Rohstoffe	Netzausbau, Transformatoren und Rechenzentrumsbau binden Kupfer und Spezialmetalle. Die Verdrängung gewöhnlicher Speicherfertigung verteuert Elektronik auch dort, wo keine KI im Spiel ist.	Kapitel 9 und 11	DRAM-Kontraktpreise nach Generation; fallende Preise der älteren Generationen wären das Frühsignal.

Allen sechs Zeilen liegt dieselbe Größe zugrunde: ob die verarbeiteten Tokenmengen schnell genug wachsen, um die Kosten der bereits gebauten Kapazität zu decken. Fällt die Antwort eines Tages anders aus, bewegen sich alle sechs zugleich — das ist der Grund, warum der Ausbau keine Sektorfrage ist. *[Einschätzung]*

14. Kennzahlen, an denen ich die Entwicklung verfolge

Die KI-Research-Serie nennt drei Indikatoren, an denen ich das Basisszenario prüfe; sie tragen hier einen Stern. Bräche das Basisszenario, würde es in dieser Reihenfolge sichtbar.

Kennzahl	Quelle	Frequenz	Gilt für	Was sie zeigt
Preise				
Mietpreis je Rechenstunde, nach Chipgeneration getrennt*	Anbieterpreislisten, Silicon Data	täglich	Wochen	Auslastung der installierten Kapazität. Fällt der Preis der aktuellen Generation deutlich bei steigender Tokennutzung, ist zu viel gebaut worden; die Vorgängergeneration zeigt das Tempo der Entwertung.
DRAM-Preise nach Generation, Spot und Vertrag*	inSpectrum	monatlich	Quartale	Knappheit in der Speicherfertigung. Fallende Preise der älteren Generationen wären das Frühsignal; der Abstand zwischen Spot- und Vertragspreis zeigt die Zyklusphase.
Ausgabengewichteter Tokenpreis	Silicon Data, LLM Token Expenditure Index	täglich	Wochen	Was der Markt je Million abgerechneter Token zahlt. Auf verarbeitete Token umgerechnet (Kapitel 10) der Vergleichswert zur Hürde aus Kapitel 12.
Preis für eine fest gehaltene Leistung	Epoch AI, Anbieterpreislisten	quartalsweise	Quartale	Der Fortschritt aus Architektur und Betrieb, getrennt von der Hardware.
Angebot				
Installierte KI-Rechenkapazität in GW	Epoch AI	quartalsweise	Jahre	Die Angebotsseite in einer Zahl.
Warteschlangen großer Lasten	ERCOT, Dominion Energy Virginia	quartalsweise	Jahre	Die mittelfristige Grenze des Ausbaus. Angemeldete gegen freigegebene Last.
Investitionen der fünf Hyperscaler	Quartalsberichte	quartalsweise	Quartale	Die Investitionsentscheidung selbst. Entscheidend ist die Wachstumsrate, nicht das Niveau.
Nachfrage				
Verarbeitete Tokenmengen	Alphabet, OpenRouter	quart./wö.	Quartale	Die Nachfrageseite. Das Mengenargument der Monetarisierungsfrage.
Anteil agentischer Nutzung an den Token	OpenRouter	wöchentlich	Quartale	Die Steigung der Nachfragekurve. Agentische Vorgänge sind nicht an die Zahl der Nutzer gebunden.
KI-Ausgaben je Mitarbeiter	Ramp	monatlich	Quartale	Verbreitung in der Fläche, unabhängig von den Anbieterangaben.
Umsätze der Modellanbieter	Epoch AI (Sammlung gemeldeter Zahlen)	laufend	Monate	Zahlungsbereitschaft in einer Zahl.
Finanzierung				
Nutzungsdauern und Kapitalmaßnahmen*	Geschäftsberichte, Bloomberg	jährl./laufend	Jahre	Ergebnisqualität und Finanzierungsrisiko. Weitere Verlängerungen bei sinkenden Mietpreisen wären Bilanzkosmetik — so Michael Burrys Vorwurf, November 2025.
Aufschläge auf besicherte Rechenzentrumskredite	Emissionsunterlagen, Ratingberichte	laufend	Monate	Die Sicht der Fremdkapitalgeber auf die Restwerte der Anlagen.

Die Spalte „Gilt für“ nennt die Haltbarkeit einer Ablesung, nicht die Frequenz der Veröffentlichung: Ein Mietpreis von vorletzter Woche ist heute wertlos, eine Anschlusswarteschlange von vor einem Jahr beschreibt die Lage noch immer einigermaßen zutreffend. Amtlich ist keine dieser Reihen; sie stützen sich zumeist auf Anbieter- und Branchenberichte, die OpenRouter-Reihen zudem auf eine kleine, entwicklerlastige Stichprobe — brauchbar als Verlauf, nicht als Niveau.

Anhang – Die Rechnungen

Der Hauptteil nennt Ergebnisse, dieser Anhang führt sie vor. Die drei Rechnungen bauen aufeinander auf: A ermittelt, was der installierte Bestand jährlich verdienen muss, B übersetzt das in einen Preis je Million Token und stellt die Erlöse dagegen, C führt dieselbe Rechnung für drei Stichtage. Eine Fortschreibung in die Zukunft enthält der Anhang bewusst nicht; warum, steht in Anhang D.

Anhang A Was der installierte Bestand jährlich verdienen muss

Diese Rechnung beantwortet eine Frage: Was muss die bereits gebaute Kapazität jährlich verdienen, um Abschreibung, Betrieb und Kapitalkosten zu decken? Sie rechnet zum Stichtag August 2026 mit 43 Gigawatt und bewertet den Altbestand zu seinen Anschaffungspreisen, den Zubau zu heutigen. Alle Eingangsgrößen stehen in Kapitel 8. Im Haupttext steht durchgehend der Wert ohne China.

Position	Wert	Herkunft
Installierte Kapazität, Q4 2025	rund 30 GW	Epoch AI
Zubau bis August 2026	rund 13 GW	Epoch AI, fortgeschrieben
Gebundenes Kapital, Altbestand	rund 900 Mrd. USD	30 GW mal 30 Mrd. USD, Anschaffungspreise
Gebundenes Kapital, Zubau	rund 494 Mrd. USD	13 GW mal 38 Mrd. USD, heutige Preise
davon Beschleuniger	rund 929 Mrd. USD	zwei Drittel, Aufteilung aus Kapitel 8
davon Bau und Infrastruktur	rund 465 Mrd. USD	ein Drittel
Kapitaldienst Beschleuniger	rund 201 Mrd. USD p. a.	Annuität, 6 Jahre, 8 %
Kapitaldienst Bau	rund 47 Mrd. USD p. a.	Annuität, 20 Jahre, 8 %
Betriebskosten	rund 39 Mrd. USD p. a.	43 GW mal 0,9 Mrd. USD
Erforderlicher Jahresertrag, weltweit	rund 287 Mrd. USD	Summe
Erforderlicher Jahresertrag ohne China	rund 258 Mrd. USD	abzüglich 10 % (Anhang B)

Diesen 258 Mrd. USD stehen zurechenbare Erlöse von 120 bis 175 Mrd. USD gegenüber; die Spanne ist eine Abgrenzungsfrage. Eng gefasst sind es die Modellanbieter mit rund 110 Mrd. USD (Kapitel 10) plus der nicht von ihnen bezahlte Teil der vermieteten Rechenkapazität (Kapitel 4). Weit gefasst kommen die KI-Erlöse der Betreiber aus Drittkundengeschäft hinzu: Exponential View erhebt sie um Doppelzählungen bereinigt, für Juni 2026 auf 175 Mrd. USD ohne China. Die installierte Basis verdient damit gut die Hälfte dessen, was sie verdienen müsste. Der Rest muss aus internen Rückflüssen kommen, die nirgends als KI-Umsatz erscheinen: bessere Werbung, Cloud-Verträge, eigene Software. *[Rechnung]*

Die Frage selbst ist nicht neu: Sequoia hat sie 2024 gestellt, Kupperman und Bain 2025 — alle drei setzen an der Investition an, Bain als Projektion auf 2030. Diese Rechnung fragt stattdessen, was der bereits gebaute Bestand heute verdienen muss, und übersetzt das in einen Preis je Million Token. Sie enthält damit keinen fortgeschriebenen Erlöspfad. *[Einschätzung]*

Eine Einschränkung ist dabei zentral: Der Altbestand ist mit 30 Mrd. USD je Gigawatt angesetzt, dem geschätzten Anschaffungspreis; neu gebaute Kapazität kostet heute 38. Rechnet man den gesamten Bestand zu heutigen Preisen — also so, als würde man ihn heute neu bauen —, steigt der erforderliche Jahresertrag auf rund 330 Mrd. USD, je Gigawatt von 6,7 auf 7,7 Mrd. USD. Die Rechnung oben beantwortet damit genau genommen zwei verschiedene Fragen: Zu Anschaffungskosten gerechnet sagt sie, was der Bestand verdienen muss; zu heutigen Preisen gerechnet sagt sie, was neu gebaute Kapazität verdienen muss. *[Rechnung]*

Zwei weitere Vorbehalte machen diese Zahl zu einer Untergrenze. Die 0,9 Mrd. USD Betriebskosten je Gigawatt decken das Rechenzentrum, nicht die Entwicklung der Modelle, den Vertrieb und die Verwaltung. Und die Kapitalkosten sind eine Nachsteuergröße, der geforderte Erlös eine Vorsteuergröße. Hinzu kommt, dass ein Teil der Kapazität im Training läuft und gar keine verkauften Token erzeugt; die Umrechnung unten ist deshalb grob.

Anhang B Von der Hürde zum Preis je Token

Um den Jahresertrag in die Einheit der Branche zu übersetzen, braucht es die verarbeitete Tokenmenge — und das war lange die unsicherste Größe der ganzen Rechnung, weil sie niemand meldet. Seit 2026 gibt es dafür eine Messung: Der Index „Tokens Per Day" beziffert das weltweite Aufkommen für Juli 2026 auf rund 360 Bio. Token am Tag, also rund 131 Billiarden im Jahr. Seine Methode ist offengelegt — ein Boden von 300 Bio. täglich aus veröffentlichten Unternehmensangaben, dazu eine Schätzung für die Nichtberichtenden über sechs Kanäle. Die Größenordnung lässt sich gegenprüfen: Alphabet allein meldet 3,2 Billiarden Token im Monat, also gut 100 Bio. am Tag, und für China werden rund 140 Bio. genannt. Beide zusammen sind bereits zwei Drittel der Summe.

Damit wird die Umrechnung eindeutig. Den 287 Mrd. USD Systemkosten des Augusts aus Anhang A stehen 131 Billiarden verarbeitete Token gegenüber: Die Hürde liegt bei rund 2,20 USD je Million verarbeiteter Token. Für die Gegenüberstellung mit den Erlösen müssen beide Seiten allerdings denselben Umfang haben — der nächste Absatz grenzt ihn ab. Auf die Kapazität ohne China bezogen stehen rund 258 Mrd. USD Kosten rund 118 Billiarden Token gegenüber; die Hürde bleibt bei 2,20 USD, weil beide Größen im selben Maß kleiner werden. Die zurechenbaren Erlöse ergeben auf diese Menge 1,02 USD je Million in der engen und 1,48 USD in der weiten Abgrenzung. Ihr Verhältnis zur Hürde ist der Deckungsgrad von 0,46 bis 0,68 — und es ist von der Menge unabhängig, weil sie sich herauskürzt. Wer eine andere Menge für richtig hält, verschiebt beide Zahlen im Gleichschritt, nicht ihr Verhältnis. *[Rechnung]*

Kapazität, Kosten und Menge sind weltweit gerechnet und enthalten China, die Erlöse stammen fast nur aus dem Westen; auch die historischen Anschaffungspreise im Kostenblock sind westliche. So verglichen sagt der Deckungsgrad wenig, und deshalb wird der chinesische Teil herausgerechnet. Epoch AI beziffert den Anteil chinesischer Eigentümer an der gesamten Rechenleistung der führenden Beschleuniger auf gut 5 % zum Jahresende 2025 — dieselbe Quelle und dieselbe Methode, aus der die 30 GW stammen; geschmuggelte Chips und im Ausland angemietete Kapazität sind darin nicht enthalten. Zählungen, die von öffentlich verfolgten Rechenclustern ausgehen, kommen auf 12 bis 15 % — auch das sind Epoch-Daten mit anderer Abgrenzung, kein zweiter Erheber. Ich setze 10 % an und ziehe sie von Kapazität und Menge zugleich ab: rund 39 statt 43 GW, rund 118 statt 131 Billiarden Token. Die Hürde je Million Token bleibt dadurch unverändert, der Erlös je Token steigt von 0,91 auf 1,02 USD und der Deckungsgrad von 0,42 auf 0,46. Bei 5 % wären es 0,44, bei 15 % 0,49 — alles in der engen Abgrenzung der Erlöse. Wer statt der gemessenen Menge eine aus Anbieterangaben fortgeschriebene ansetzt, kommt auf eine Hürde von rund 0,80 USD — an der Aussage ändert das nichts, am Eindruck sehr wohl. Die Menge ist die Zeile, an der man diese Rechnung am leichtesten kippt. *[Einschätzung]*

Abweichung von der Basis (43 GW, Stand August 2026)	Erforderlicher Jahresertrag
Nutzungsdauer Beschleuniger 4 Jahre	rund 367 Mrd. USD
Nutzungsdauer 5 Jahre	rund 319 Mrd. USD
Basis: 6 Jahre, Kapitalkosten 8 %	rund 287 Mrd. USD
Kapitalkosten 6 % statt 8 %	rund 268 Mrd. USD
Kapitalkosten 10 % statt 8 %	rund 307 Mrd. USD
Bestand vollständig zu heutigen Preisen	rund 330 Mrd. USD

Die Empfindlichkeit ist aufschlussreich: Vier Prozentpunkte Kapitalkosten bewegen das Ergebnis um ein Siebtel, zwei Jahre Nutzungsdauer um mehr als ein Viertel. Die Rechnung hängt also weit stärker an der Abschreibungsdauer als am Zinsniveau. Zu den 8 % komme ich, weil die Eigenkapitalkosten der Betreiber von 9 bis 10 % das falsche Maß wären — sie preisen deren gesamtes Geschäft. Der Markt ruft für dieses Anlagevermögen weniger auf: Metas Siebenjahresanleihe vom Oktober 2025 wurde mit 4,60 % begeben, die besicherte Fazilität von CoreWeave bei 5,9 %. Die 8 % sind deshalb die Mischung aus beidem und nicht der Fremdkapitalpreis. *[Einschätzung]*

Anhang C Die Reihe des vergangenen Jahres

Diese Rechnung führt die Kostenrechnung aus Anhang A über die Zeit: Sie folgt der installierten Kapazität von Stichtag zu Stichtag, bewertet den Altbestand zu Anschaffungs- und den Zubau zu heutigen Preisen und stellt ihr die gemeldeten Erlöse gegenüber. Alle Größen sind auf die Kapazität ohne China abgegrenzt (Anhang B); die Augustspalte ist die Rechnung aus Anhang A.

Drei Größen sind gemessen oder gemeldet: die Tokenmenge aus dem Index in Anhang B, der Indexpreis aus Kapitel 10 und die installierte Kapazität von Epoch AI, aus der die Zeile der Systemkosten nach dem Verfahren aus Anhang A berechnet ist. Die Erlöszeile ist abgeleitet und über die drei Spalten gleich gebaut: gemeldete Umsätze der Modellanbieter zuzüglich des Teils der vermieteten Rechenkapazität, der nicht schon in diesen Umsätzen steckt. Das ist die enge Abgrenzung — die einzige, die sich für alle drei Stichtage gleich bilden lässt. Die weite Abgrenzung, die auch die KI-Erlöse der Betreiber aus Drittkundengeschäft enthält, liegt nur für den August vor: 175 statt 120 Mrd. USD, 1,48 statt 1,02 USD je Million Token, Deckungsgrad 0,68 statt 0,46. Die drei unteren Zeilen sind gerechnet.

	Dez. 2025	Q2 2026	Aug. 2026
Verarbeitete Tokenmenge p. a., ex China	15,8 Bill.	47 Bill.	118 Bill.
Zurechenbare Erlöse p. a., enge Abgrenzung	30 Mrd. USD	78 Mrd. USD	120 Mrd. USD
Systemkosten p. a., ex China	167 Mrd. USD	221 Mrd. USD	258 Mrd. USD
Indexpreis je Mio. abgerechneter Token	1,09 USD	1,80 USD	1,11 USD
Kosten je Mio. verarbeiteter Token	10,60 USD	4,70 USD	2,20 USD
Erlös je Mio. verarbeiteter Token	1,90 USD	1,67 USD	1,02 USD
Deckungsgrad	0,18	0,35	0,46

„Bill." steht für Billiarden. Der Bestand von Ende 2025 ist zu historischen Kosten angesetzt, der Zubau zu heutigen — je Gigawatt rund 6,2 Mrd. USD im Jahr für den Altbestand und 7,7 für den Zubau, Kapitaldienst eingeschlossen. Die Mengen für Dezember und das zweite Quartal sind aus der Augustmessung mit den beobachteten Wachstumsraten zurückgerechnet. Kapazität und Menge sind um den chinesischen Anteil von 10 % gekürzt; die weltweiten Werte liegen entsprechend höher (Anhang B). Die Kostenzeile je Token ändert sich dadurch nicht, weil Zähler und Nenner gleich stark schrumpfen. Eine Spalte für die Zukunft enthält die Tabelle bewusst nicht. *[Rechnung]*

Die beiden unteren Zeilen tragen die ganze Aussage. Die Kosten je verarbeitetem Token fielen von 10,60 auf 2,20 USD, weil dieselbe Kapazität ein Vielfaches an Menge trug. Der Erlös je verarbeitetem Token fiel ebenfalls, von 1,90 auf 1,02 USD — aber langsamer. Genau deshalb stieg der Deckungsgrad von 0,18 auf 0,46. Daraus folgt die Regel des Kapitels: Der Deckungsgrad steigt, solange der Preis je Token langsamer fällt als die Kosten je Token. Gedeckt sind die Kosten deshalb noch lange nicht. *[Rechnung]*

Die Regel hält auch dann, wenn sich der Verkehr zu billigeren Nutzungsarten verschiebt — das ist allerdings eine Annahme und keine Messung: Ich unterstelle, dass ein Token, der billiger abgerechnet wird, in der Herstellung im selben Maß weniger kostet, weil beides an derselben Zwischenspeicherung hängt. Trifft das zu, fallen beide Seiten zugleich und der Mixwechsel ist neutral. Trifft es nicht zu, wird aus dem Mixeffekt echter Margendruck. *[Einschätzung]*

Ein Vorbehalt bleibt, und er ist der wichtigste des ganzen Anhangs: Die Rechnung bewertet den Eigenbedarf der Betreiber mit null. Ein erheblicher Teil der 118 Billiarden Token wird nie in Rechnung gestellt — Alphabet erzeugt sie für die eigene Suche, andere für eigene Produkte. Wertlos ist das nicht; es steckt in Werbung, Cloud-Verträgen und eigener Software, nur eben nicht messbar. Wer diesen Teil mit dem Marktpreis ansetzt, kommt für August auf einen Deckungsgrad nahe eins. Zwischen diesen beiden Lesarten liegt die offene Frage des Kapitels; belastbar ist die Richtung, nicht das Niveau. *[Einschätzung]*

Anhang D Annahmenverzeichnis

Diese Tabelle führt jede Größe auf, die in den Rechnungen dieses Dokuments steckt, und sagt, woher sie kommt. Die Statusspalte entspricht der Kennzeichnung im Text: „gemessen" und „gemeldet" stehen ohne Zusatz, „abgeleitet" erscheint als [Rechnung], „geschätzt" als [Einschätzung]. Wer der Rechnung widersprechen will, findet hier die Zeilen, an denen es sich lohnt.

Größe	Wert	Quelle	Status
Installierte Kapazität, Q4 2025	30 GW	Epoch AI	gemeldet
Installierte Kapazität, August 2026	rund 43 GW	Epoch AI, fortgeschrieben	abgeleitet
Kosten je Gigawatt, Neubau	38 Mrd. USD	Kapitel 8	abgeleitet
Kosten je Gigawatt, Altbestand	30 Mrd. USD	frühere Anschaffungspreise	geschätzt
Betriebskosten je Gigawatt	0,9 Mrd. USD p. a.	Kapitel 8	abgeleitet
Aufteilung Chips zu Bau	zwei Drittel zu einem Drittel	Kapitel 8	abgeleitet
Nutzungsdauer Beschleuniger	6 Jahre	Abschlüsse der Betreiber	gemeldet
Nutzungsdauer Bau	20 Jahre	Abschlüsse der Betreiber	gemeldet
Kapitalkosten	8 %	Anleiherenditen plus Eigenkapitalaufschlag	geschätzt
Verarbeitete Tokenmenge, Juli 2026	131 Bill. p. a.	Tokens-Per-Day-Index	gemessen
Chinesischer Anteil an Kapazität und Menge	10 %	Epoch AI; Clusterzählungen sind dieselben Daten mit anderer Abgrenzung	geschätzt
Installierte Kapazität ex China, August 2026	rund 39 GW	43 GW abzüglich 10 %	abgeleitet
Verarbeitete Tokenmenge ex China, Juli 2026	rund 118 Bill. p. a.	131 Bill. abzüglich 10 %	abgeleitet
Systemkosten ex China, August 2026	rund 258 Mrd. USD p. a.	39 GW zu Anschaffungs- und Neubaupreisen, Anhang A	abgeleitet
Indexpreis, August 2026	1,11 USD je Mio.	Silicon Data	gemessen
Zurechenbare Erlöse, August 2026	120 bis 175 Mrd. USD p. a.	eng: Modellanbieter plus nicht überschneidende Vermietung; weit: Exponential View, State of the AI Economy, Sept. 2026; die dort ebenfalls genannten 110 Mrd. USD sind alle Erlösstufen über zwölf Monate, nicht die Zahl aus Kapitel 10	abgeleitet
Kapazitätsgrenze der Flotte in Token	43 bis 1.730 Bio. am Tag, je nach Kontextlänge	Epoch AI	gemeldet
Wachstum Angebot / Nachfrage	3,4-fach / 10-fach im Jahr	Epoch AI	geschätzt
Anschlussdauer große Last	3 bis 7 Jahre	Dominion Energy Virginia; die LBNL-Fünfjahreszahl misst Erzeugungs-, nicht Lastanschlüsse	geschätzt

Vier Zeilen tragen das Ergebnis: die Nutzungsdauer der Beschleuniger, die Kapitalkosten, die verarbeitete Tokenmenge und die Erlöse. Die ersten beiden bewegen die Hürde um mehr als ein Viertel, die dritte verschiebt beide Größen je Token im Gleichschritt und den Deckungsgrad gar nicht, die vierte wirkt nur auf die Erlösseite. *[Einschätzung]*

> **Fazit.** Der Deckungsgrad liegt je nach Abgrenzung der Erlöse zwischen 0,46 und 0,68 und ist seit einem Jahr gestiegen. Er steigt weiter, solange der Preis je Token langsamer fällt als die Kosten je Token. Die Kosten je Token fallen, solange dieselbe Kapazität mehr Token trägt — und das hat eine Grenze, weil die Bandbreite bindet. Drei Sätze, drei prüfbare Größen, kein fortgeschriebener Erlöspfad. *[Einschätzung]*

Glossar

Ich habe die Begriffe nach Themen geordnet, innerhalb der Gruppen alphabetisch. Jeder Eintrag nennt zuerst, was der Begriff bezeichnet, und danach, warum er für die Bewertung zählt.

Modelle und Technik

Agent. Ein Programm, das ein Modell mehrfach hintereinander aufruft, dazwischen Werkzeuge bedient und Zwischenergebnisse auswertet. *Für Investoren:* Der wichtigste Nachfragetreiber der Inferenz, weil ein Vorgang ohne menschliches Zutun beliebig viele Aufrufe erzeugt.

Benchmark. Ein standardisierter Test, mit dem Modelle verglichen werden. *Für Investoren:* Als Investitionsargument wenig wert: Tests veralten, geraten in die Trainingsdaten und messen selten das, was Kunden bezahlen.

Eigenschaften. Die gelernte Zahlenliste, die zu jedem Token gehört und festhält, in welchen Zusammenhängen es vorkommt; in der Fachsprache Embedding. *Für Investoren:* Zusammen mit den Gewichten bilden sie die Parameter. Der Exkurs zeigt beides im Kleinen.

Generative KI. Der Teil des maschinellen Lernens, der neue Inhalte erzeugt statt vorhandene einzuordnen. *Für Investoren:* Erst hier entsteht der Rechenbedarf, der die Investitionswelle trägt.

Gewichte. Die Zahlen, mit denen das Modell seine Eingaben verrechnet; zusammen mit den gelernten Eigenschaften bilden sie die Parameter. *Für Investoren:* Sie sind das eigentliche Produkt eines Trainingslaufs. Bei offenen Modellen sind sie samt Eigenschaftslisten und Vokabelliste frei verfügbar — die Trainingsdaten und das Verfahren nicht.

Halluzination. Eine erfundene, aber sprachlich einwandfreie Aussage des Modells. *Für Investoren:* Der Hauptgrund, warum viele Vorgänge weiterhin geprüft werden müssen, und damit eine Bremse für die Einsparungen.

Kleines Sprachmodell. Ein Modell, das klein genug ist, um auf einem Telefon oder Rechner zu laufen. *Für Investoren:* Übernimmt kurze, häufige Aufgaben auf dem Endgerät — überwiegend solche, die vorher gar nicht stattfanden. Das Rechenzentrum ersetzt es nicht.

Kontextfenster. Die Textmenge, die ein Modell bei einer Anfrage gleichzeitig berücksichtigen kann. *Für Investoren:* Bestimmt, ob sich ein Vertrag oder ein Quartalsbericht in einem Durchgang auswerten lässt; geht überproportional in die Rechenkosten ein.

Maschinelles Lernen. Verfahren, die Regeln aus Daten ableiten, statt sie programmiert zu bekommen. *Für Investoren:* Der Oberbegriff. Vieles davon läuft seit Jahrzehnten und hat mit der aktuellen Investitionswelle nichts zu tun.

Offene Gewichte. Ein Modell, dessen vollständige Parameter samt Vokabelliste frei heruntergeladen werden können, nicht aber die Trainingsdaten und das Trainingsverfahren. *Für Investoren:* Setzt eine Preisobergrenze für vergleichbare Leistung und ist der Weg, auf dem chinesische Anbieter in den westlichen Markt kommen.

Parameter. Der Oberbegriff für die beim Training eingestellten Zahlen eines Modells: gelernte Eigenschaften und Gewichte. Spitzenmodelle liegen bei Hunderten Milliarden bis mehreren Billionen. *Für Investoren:* Direkter Kostentreiber für Speicher und Rechenzeit, aber kein verlässliches Maß für Qualität.

Physical AI. KI, die nicht Text verarbeitet, sondern in der physischen Welt wahrnimmt und handelt — in Robotern, Fahrzeugen und Maschinen. *Für Investoren:* Die für Europa interessanteste Öffnung des Feldes: Maschinen, Fertigungsdaten und Prozesswissen sitzen zu einem erheblichen Teil hier.

Sprachmodell. Ein Modell, das Wortbausteine fortsetzt und dabei Sprache, Code und Struktur beherrscht. *Für Investoren:* Die derzeit wirtschaftlich relevante Bauform. Es prüft nichts nach, sondern setzt statistisch fort.

Rechenkapazität und Infrastruktur

ASIC. Ein Chip, der für einen einzigen Zweck entworfen wurde, etwa die eigene Inferenzlast eines Betreibers. *Für Investoren:* Die Alternative zum zugekauften Beschleuniger. Google, Amazon und Meta bauen eigene; das drückt die Marge des Chipanbieters, nicht die Investitionssumme.

Ausfuhrkontrolle. Die Genehmigungspflicht für die Lieferung leistungsfähiger Beschleuniger. *Für Investoren:* Die politische Schicht über der Kette. Sie ändert sich schneller als Fertigung oder Netzanschluss und teilt den Weltmarkt in zwei Hälften.

Beschleuniger. Der Spezialchip, der die Rechenarbeit übernimmt; bei Nvidia GPU genannt, bei Google TPU. Der Speicher daneben, vor allem HBM, zählt nicht dazu. *Für Investoren:* Der größte Einzelposten der Investitionen und der Chip, dessen Nutzungsdauer die Ergebnisse bewegt.

DRAM. Der Arbeitsspeicher der Server, in Kontraktpreisen für DDR4 und DDR5 gehandelt. *Für Investoren:* Sein Preis ist der zuverlässigste öffentliche Frühindikator für Knappheit in der gesamten Kette.

Gigawatt. Das Maß für die Leistungsaufnahme einer Anlage und die gebräuchliche Einheit für Rechenkapazität. *Für Investoren:* Eine Anlage von einem Gigawatt kostet nach Schätzung von Epoch AI rund 38 Mrd. USD.

H100-Äquivalent. Eine Rechengröße, die Beschleuniger verschiedener Generationen auf eine gemeinsame Leistung normiert. *Für Investoren:* Ohne diese Normierung sind Meldungen über Chipzahlen nicht vergleichbar. Ein Gigawatt entspricht grob einer Million Einheiten.

HBM. Gestapelter Hochgeschwindigkeitsspeicher, der direkt neben dem Rechenchip sitzt. *Für Investoren:* Knapp, teuer und der Grund, warum der Speichermarkt insgesamt angespannt ist: HBM verdrängt gewöhnliche Fertigungskapazität.

Hyperscaler. Die großen Betreiber eigener Rechenzentren: Amazon, Alphabet, Meta, Microsoft und Oracle. *Für Investoren:* Sie verfügten Ende 2025 über rund 71 % der weltweiten KI-Rechenleistung und treffen die Investitionsentscheidungen.

Inferenz. Der laufende Betrieb eines fertigen Modells, also jede einzelne Anfrage. *Für Investoren:* Erzeugt den wiederkehrenden Umsatz und die variablen Kosten. Seit 2026 der größere Ausgabenblock.

Mietpreis je Rechenstunde. Der Preis, zu dem installierte Kapazität stundenweise vermietet wird. *Für Investoren:* Nach Generationen getrennt gelesen die aussagekräftigste Größe zur Auslastung und zur Entwertung der Anlagen.

Rechengenauigkeit. Die Zahl der Bits je Gewicht — üblich sind sechzehn beim Training und vier bei der Inferenz. *Für Investoren:* Jede Halbierung verdoppelt die Rechenleistung und halbiert den Speicherbedarf. Chipankündigungen mischen diesen Wechsel meist mit weiteren Abkürzungen; in gleicher Genauigkeit gemessen fallen die Faktoren etwa viermal kleiner aus.

Souveräne Rechenkapazität. Öffentlich finanzierter Aufbau von Fertigung und Rechenleistung im eigenen Land und unter eigener Rechtsordnung. *Für Investoren:* Nachfrage, die nicht auf Preise und Renditen reagiert. Sie stützt den Zyklus, führt auf Systemebene aber zu doppelt gebauter Kapazität.

Speicherbandbreite. Die Datenmenge, die je Sekunde zwischen Speicher und Prozessor bewegt werden kann. *Für Investoren:* Die bindende Größe bei der Inferenz: Für jedes Token müssen alle Gewichte einmal gelesen werden.

Training. Der einmalige, zusammenhängende Lauf, in dem ein Modell entsteht. *Für Investoren:* Eine abgeschlossene Investition. Die größten Läufe überschreiten 100 MW Dauerleistung.

Vortraining. Der erste und teuerste Trainingsschritt, in dem das Modell aus großen Textmengen entsteht. *Für Investoren:* Hier fällt der Großteil der Rechenkosten an; alles Weitere ist vergleichsweise günstig.

Ökonomie und Preise

Jevons-Effekt. Der Befund, dass sinkende Kosten je Einheit den Gesamtverbrauch erhöhen statt senken. *Für Investoren:* Der Grund, warum fallende Tokenpreise bisher mit steigenden Rechnungen einhergehen.

Kreisgeschäft. Ein Geldgeber beteiligt sich an einem Kunden, der das Kapital überwiegend für Käufe beim Geldgeber verwendet. *Für Investoren:* Der so entstehende Umsatz belegt keine Zahlungsbereitschaft Dritter. Zu prüfen über die Angaben zu nahestehenden Unternehmen.

Preis je Million Token. Die übliche Abrechnungseinheit der Modellanbieter, getrennt nach Ein- und Ausgabe. *Für Investoren:* Die einzige Größe, in der sich Umsatz, Kosten und Kapazität gemeinsam messen lassen.

Segmentierung. Der gleichzeitige Verfall der Preise für gegebene Leistung und die Stabilität der Preise für Spitzenleistung. *Für Investoren:* Unterscheidet einen normalen Technologiepfad von einem Preiskrieg. Beide Reihen müssen getrennt betrachtet werden.

Token. Der Wortbaustein, in dem Modelle rechnen und abrechnen. Ein deutsches Wort sind meist zwei bis drei Token. *Für Investoren:* Die gemeinsame Einheit für Umsatz, Kosten und Kapazität.

Tokenmenge. Die Zahl der insgesamt verarbeiteten Token, auch der nicht abgerechneten — abzugrenzen von der abgerechneten Menge, auf die sich der Indexpreis bezieht (Kapitel 10). *Für Investoren:* Das Mengenargument in der Monetarisierungsfrage. Sie ist bisher schneller gestiegen, als die Preise gefallen sind.

Rechtliche Hinweise und Kontakt

Diese Ausarbeitung wurde für die Research Ahead GmbH von der auf dem Deckblatt oder auf der ersten Seite genannten Rechtsperson erstellt. Diese Ausarbeitung enthält nur allgemeine Informationen, berücksichtigt nicht die besonderen Umstände des Empfängers und sollte von diesem nicht als verbindlich oder als Ersatz für seine eigene Beurteilung angesehen werden. Jeder Empfänger sollte die Angemessenheit jeder Anlageentscheidung im Hinblick auf seine eigenen Umstände, unter Berücksichtigung aller verfügbaren Informationen und unter Zuhilfenahme angemessener professioneller Beratung überprüfen. Die in dieser Ausarbeitung enthaltenen Informationen und Meinungen stellen eine per Veröffentlichungsdatum getroffene Beurteilung dar, beruhen auf Quellen, die für zuverlässig und korrekt gehalten werden und können sich ohne Ankündigung ändern. Eine Gewähr für Richtigkeit und Vollständigkeit der hierin enthaltenen Informationen kann jedoch weder explizit noch implizit übernommen werden. Die Research Ahead GmbH kann den Inhalt dieser Studie derart und mit einem Zeitplan ändern, ergänzen oder aktualisieren, wie es die Research Ahead GmbH für richtig hält. Die hierin zum Ausdruck gebrachten Empfehlungen und Meinungen spiegeln die Erwartungen der Research Ahead GmbH über den Sechsmonatszeitraum ab Veröffentlichung wider, und zwar aus Perspektive ausschließlich langfristig orientierter Investmentkunden. Die Research Ahead GmbH behält sich das Recht vor, für einen anderen Zeithorizont oder für andere Arten von Investmentkunden unterschiedliche oder widersprüchliche Empfehlungen und Meinungen auszudrücken. Diese Ausarbeitung stellt weder ein Angebot noch eine Aufforderung zur Abgabe eines Angebots für den Verkauf oder Bezug von Wertpapieren dar und sollte weder in ihrer Gesamtheit noch in Auszügen als Informationsgrundlage in Verbindung mit einem Vertragsabschluss oder einer wie auch immer gearteten Verpflichtung verwendet werden. Die Research Ahead GmbH übernimmt keinerlei Haftung für Verluste oder Schäden, die aus jeglicher Verwendung dieser Studie oder ihres Inhalts entstehen. Die Aufnahme von Hyperlinks zu den Websites von Organisationen, die in dieser Studie erwähnt werden, impliziert keineswegs eine Zustimmung, Empfehlung oder Billigung der Informationen der Websites bzw. der von dort aus zugänglichen Informationen durch die Research Ahead GmbH. Die Research Ahead GmbH übernimmt keine Verantwortung für den Inhalt dieser Websites oder von dort aus zugängliche Informationen oder für eventuelle Folgen aus der Verwendung dieser Inhalte oder Informationen. Diese Ausarbeitung ist nur zur Verwendung durch den Empfänger und in seiner Eigenschaft als professioneller Anleger oder sachkundiger und erfahrener Investor bestimmt. Sie darf weder in Auszügen noch als Ganzes ohne vorherige schriftliche Genehmigung durch die Research Ahead GmbH vervielfältigt, verbreitet, veröffentlicht oder an andere Personen weitergegeben werden. Die Research Ahead GmbH kann Ausarbeitungen dieser Art als Druckversion oder elektronisch verteilen. Die Kunden der Research Ahead GmbH können Transaktionen im Hinblick auf die in dieser Ausarbeitung genannten Wertpapiere oder damit verbundenen Anlagen tätigen oder getätigt haben, bevor der Empfänger diese Ausarbeitung erhalten hat. Zusätzliche Informationen zum Inhalt dieser Ausarbeitung sind auf Anfrage verfügbar.

© Research Ahead GmbH 2026

Ergänzende Hinweise zu dieser Publikation

- **Stand der Daten.** Alle Angaben und Abbildungen beziehen sich auf den Stand September 2026. Der Primer erklärt Zusammenhänge; Zahlenangaben verstehen sich entsprechend als Größenordnungen.
- **Datenquellen.** Die verwendeten Angaben zu Rechenkapazität, Tokenmengen und Preisen stammen überwiegend von privaten Anbietern und sind weder amtlich erhoben noch geprüft. Wo eine Zahl aus zwei Angaben abgeleitet ist, steht das im Text.
- **Keine Anlageempfehlung.** Der Primer enthält keine Einschätzung zu einzelnen Wertpapieren. Meine Positionierung steht in der KI-Research-Serie.
- **Interessenkonflikte.** Research Ahead GmbH betreibt kein Investmentbanking und unterhält keine Emissions- oder Beratungsmandate bei den in dieser Publikation genannten Unternehmen.

daniel.pfaendler@researchahead.com · www.researchahead.de · Wildenbruchstraße 50 · 60431 Frankfurt am Main · Deutschland
Research Ahead GmbH — unabhängiges Research zu Makroökonomie und Finanzmarktstrategie für professionelle Investoren