Die Rendite von KI-Chips hängt an der Rechenleistung für Kunden

Ein Gigawatt, das Kunden bedient, bringt rund 88 Mrd. USD Umsatz im Jahr bei 8,5–15 Mrd. USD Kosten, doch das braucht heute nur ein Zehntel der Flotte. Die Kapazität je Gigawatt verdreifacht sich bis 2028; die Rendite setzt voraus, dass die Nachfrage jährlich auf das Vier- bis Sechsfache wächst.

STAND
14. November 2026

UMFANG
KI-Flotte ohne China bis 2028: Token je Gigawatt, Kundenanteil, Umsatz und Kosten je Gigawatt

KERNPUNKTE

- **Ein Gigawatt verarbeitet bis 2028 rund dreimal so viele Token.** Neuere Chips und bessere Software heben ihn von 0,21 auf 0,65 Billiarden (Brd.) Token pro Tag (getestet: 0,15–1,49); die Hälfte aus Software, abgeleitet aus dem gemessenen Trend.
- **Der Betrieb für Kunden braucht rund ein Zehntel der Flotte, und bei diesem Anteil hält die Kapazität mit der Nachfrage Schritt.** Die Anfragen der Kunden beanspruchen 8 %–11 % der heutigen Flotte. Bleibt es dabei, wächst die Kapazität im Kundenbetrieb im nächsten Jahr auf das 4,8-Fache, die berichtete Nutzung auf das 4,0- bis 6,2-Fache.
- **Ein Gigawatt im Kundenbetrieb erlöst 88 Mrd. USD im Jahr, doch nur ein Viertel von Anthropics Rechenleistung dient Kunden.** Das ist das Sechs- bis Zehnfache der jährlichen Eigentumskosten von 8,5–15 Mrd. USD. Über die gesamte Rechenleistung erlöst Anthropic 22 Mrd. USD je Gigawatt und liegt knapp unter der Gewinnschwelle.

Der Basisfall für die Kapazität

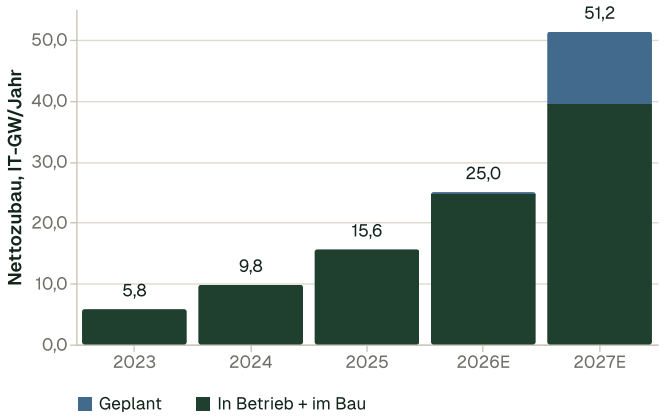
Ausgangspunkt ist die gesamte Rechenzentrumskapazität, die bereits in Betrieb ist oder noch ans Netz gehen soll. Auf dieser Basis verteilen wir Rechenleistung auf verschiedene Verwendungen und schätzen, wie viel Umsatz ein GW erzielt. Die beste öffentliche Quelle dafür ist SemiAnalysis.

Die Bauprognose vom 19. Oktober beziffert den jährlichen Nettozubau an IT-Leistung von Rechenzentren außerhalb Chinas auf 25,0 GW im Jahr 2026 und 51,2 GW im Jahr 2027. Darin sind Projekte im Bau ebenso enthalten wie geplante, sodass ein Fertigstellungsrisiko bleibt.

In diesem Papier beziffern wir, wie sich die Zusammensetzung dieser Flotte verändert und wie Fortschritte bei Hardware und Software die Token-Produktion beeinflussen. Das vergleichen wir mit der heutigen Nachfrage und dem Wachstum, das nötig ist, um das zusätzliche Angebot aufzunehmen. Danach untersuchen wir, wie sich diese Rechenleistung auf Kundenbetrieb und Forschung verteilt und was ein GW erlöst und kostet, gestützt auf Anthropics Börsenprospekt.

SemiAnalysis erwartet einen schnelleren Ausbau

Jährlicher IT-Zubau in GW ohne China, nach Baustatus



E = Prognose. 2026 geplant = 25,0 – 24,6 = 0,4 GW. Veröffentlichte Summe 2027: 51,2 GW; gerundete Segmente: 51,3 GW.

Im August bezifferte Dylan Patel von SemiAnalysis den weltweiten Zubau an KI-Rechenleistung für 2028 auf rund 70 GW und stimmte zu, dass der Bestand zum Jahresende über 200 GW liegen werde. Wir verwenden eine niedrigere Schätzung der außerhalb Chinas installierten KI-Flotte, die wir bis dahin bei 127 GW sehen. SemiAnalysis erwartet zudem, dass 2028 wieder rund 35 % des Zubaus ohne China außerhalb der USA entstehen.

Bislang haben Moratorien und Auflagen für Rechenzentren nur wenig Kapazität verzögert. Das Ergebnis der Zwischenwahlen könnte den Ausbau erschweren, trifft aber vor allem Projekte, die 2028 ans Netz gehen. Die Projekte für 2026 und 2027 sind weit fortgeschritten und vollständig genehmigt. Der Ausbau könnte sich auch beschleunigen, wenn modulare Bauweisen sich durchsetzen. Vollständig modulare Konzepte könnten die Bauzeit von 18–24 auf 12–18 Monate senken, Genehmigungen nicht eingerechnet.

Wie sich die Hardware der Flotte verändert

Mit jeder neuen Hardwaregeneration ändert sich die Zusammensetzung der weltweiten GPU-Flotte. Die starke Nachfrage nach KI-Diensten hält ältere Generationen länger in Betrieb.

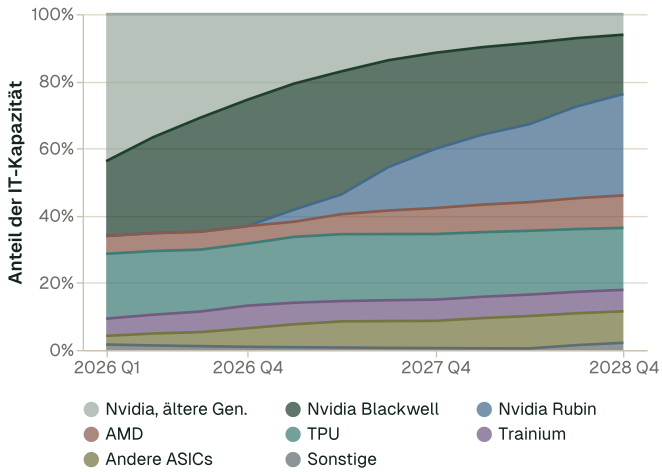
Neue Hardware ist leistungsfähiger und effizienter und besser auf heutige Workloads und Modelle abgestimmt. Im Mittelpunkt steht zunehmend das gesamte Rack. GB200 NVL72 etwa verbindet 72 GPUs über NVLink und verringert so Kommunikationsengpässe beim Betrieb großer Modelle.

Training und Inferenz lassen sich jeweils eigens optimierten Hardwarepools zuweisen. Inferenz teilt sich weiter in Prefill (verarbeitet die Eingabe) und Decode (erzeugt die Antwort). Beim Reinforcement Learning nutzt die Erzeugung von Rollouts Vorwärtsdurchläufe der Inferenz, während die Aktualisierung der Gewichte Vorwärts- und Rückwärtsdurchläufe erfordert.

Die Effizienzgewinne aus Hardware, Software und Modelldesign können sich verstärken und die Token-Produktion je GW erhöhen. Der gemeinsame Gewinn hängt davon ab, welche Engpässe jede Verbesserung beseitigt.

Die Zusammensetzung der KI-Flotte ändert sich

Ohne China; geschätzter Anteil an der installierten IT-Leistung



Eigene Schätzung des Hardwaremix; keine Prognose von SemiAnalysis. Anteile an der installierten IT-Kapazität ohne China. Die GW-Basis nutzt veröffentlichte SemiAnalysis-Zahlen.

Auch Frontier-Labore und Hyperscaler bringen eigene Hardware ans Netz. Googles TPU und AWS Trainium sind die bekanntesten Beispiele. Hinzu kommen Microsofts Maia für die Inferenz im Produktivbetrieb und Metas MTIA für Ranking und Empfehlungen, mit Inferenzchips für generative KI auf der Roadmap. OpenAI plant mit Broadcom erste Jalapeño-Installationen für Ende 2026. Googles TPU zeigt, was die gemeinsame Entwicklung von Hardware und Software leisten kann: Chips, Verbindungen, Compiler und Modelle werden zusammen optimiert.

Auch spezialisierte Chipfirmen zielen auf schnelle Inferenz. Cerebras nutzt Prozessoren in Wafergröße, SambaNova rekonfigurierbare Dataflow-Chips und Groq eigene Inferenzprozessoren. Start-ups wie Etched bauen Inferenzsysteme für Frontier-Modelle und optimieren Chips, Speicher und Verbindungen für Prefill und Decode.

Output bei vorgegebenem Antworttempo

Die Grafik unten vergleicht den Agentenbetrieb von DeepSeek V4 Pro auf sieben Systemen, gemessen in Ausgabe-Token je IT-MW bei einem Mindesttempo von 60 Token/s/Nutzer im Median und 80 oder 100 bei P90.

In Nvidias Vorabwerten liefert Rubin das 1,16- und das 1,27-Fache des Outputs von GB300 je IT-MW bei den Schwellen Median 60 und P90 100; das 1,90-Fache bei P90 80 beruht auf einem einzigen Lauf knapp an der Schwelle. Diese Verhältnisse berücksichtigen bereits die angenommene Systemleistung einschließlich des höheren Overheads.

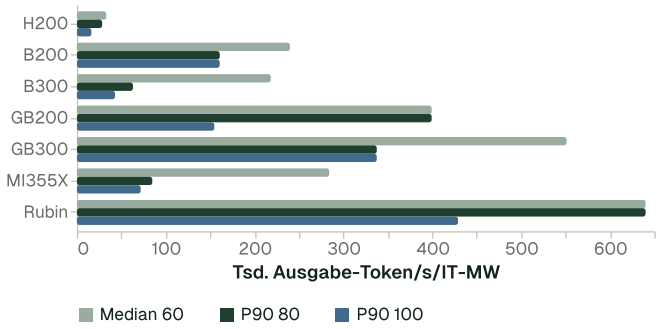
Der Übergang zu Rack-Systemen erweitert die Scale-up-Domäne von acht GPUs in einem B300-Server auf 72, die in GB300 NVL72 über NVLink verbunden sind. Das kann Engpässe beim Austausch zwischen MoE-Experten verringern.

Im selben Agenten-Workload mit DeepSeek V4 Pro liefert GB300 NVL72 bei den drei Schwellen das 2,53-, 5,35- und 7,86-Fache des Outputs von B300 je IT-MW. Die gewählten Setups nutzen 24–28 GB300-GPUs gegenüber 4–8 B300-GPUs, bei angenommenen 2,12 und 2,10 kW je GPU einschließlich System-Overhead. Die B300-Läufe sind Ein-Knoten-Setups ohne getrennte Pools für Prefill und Decode, sodass die höheren Verhältnisse teils eine nicht optimierte Vergleichsbasis widerspiegeln; die ältere B200 übertrifft B300 bei allen drei Schwellen.

Jeder Chip bringt zudem Fortschritte bei HBM und Rechenleistung, die verschiedene Teile des Workloads betreffen. Wie Nvidia beschreibt, hilft mehr Speicherbandbreite der speichergebundenen Erzeugung, mehr Speicherkapazität fasst Modellgewichte und zwischengespeicherten Kontext, und mehr Rechenleistung beschleunigt rechengebundene Schritte.

Output bei vorgegebenem Mindesttempo

DeepSeek V4 Pro · Agenten-Workload · konfigurierte IT-Leistung



Mindesttempo: Median 60, P90 80 oder P90 100 Token/s/Nutzer; keine Latenzgrenze für das erste Token. Engines/Präzision variieren; Systemleistung angenommen. Rubin: Vorabwerte.

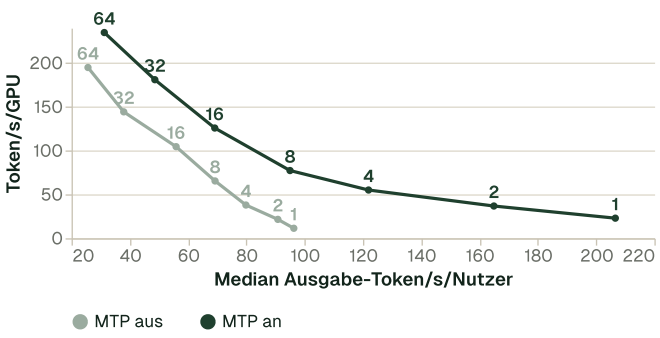
Schnellere Antworten kosten Durchsatz

Interaktivität misst die erzeugten Token pro Sekunde und Nutzer, Durchsatz den Output über alle Anfragen je GPU. Werden mehr Anfragen gemeinsam bedient, steigt in der Regel der Durchsatz, während jeder Nutzer langsamer bedient wird; das kann schnelle Modi für Coding-Agenten teurer machen. Fortschritte bei Hardware und Software verschieben diesen Zielkonflikt: SemiAnalysis berichtete, dass Vorabsysteme von Nvidia Rubin NVL72 am 27. Oktober mit Kimi K3 mehr als 850 Token/s/Nutzer und am 1. November mit GLM-5.3 mehr als 1.000 erreichten. Gemessen ist jeweils das Antworttempo bei P90; beide Kurven zeigen weiterhin weniger Token je Dollar, je höher das Antworttempo ist.

Ein Softwarebeispiel bei gleicher Konfiguration: In einem InferenceX-Lauf mit DeepSeek V4 Pro, FP4, acht B200 und vLLM 0.25.0 erlaubt Multi-Token-Prediction (MTP) 16 statt acht parallele Anfragen bei weiterhin rund 69 Token/s/Nutzer im Median. Das ergibt das 1,92-Fache an Ausgabe-Token je GPU. Die Grafik unten zeigt, wie sich DeepSeek V4 entlang der Kurve aus Durchsatz und Interaktivität betreiben lässt.

Tuning verschiebt die Tempo-Durchsatz-Kurve

DeepSeek V4 Pro · B200 · FP4 · 8K Eingabe / 1K Ausgabe



12. Juli 2026; derselbe Knoten mit acht GPUs und vLLM 0.25.0. Beschriftung: parallele Anfragen. Nur Ausgabe-Token; volle Knotenzeit. Latenz bis zum ersten Token separat.

Software-Engineering erhöht den Token-Output

Neben der Hardware selbst optimieren Hardwarehersteller und die Teams der Serving-Engines gemeinsam einzelne Kombinationen aus Modell und Hardware. Nvidia und SGLang dokumentieren Gewinne aus Kernel-Fusion, Speicherverwaltung, Scheduling und Fehlerbehebungen. Diese Verbesserungen wirken, sobald Teams sie umsetzen und ausrollen.

Unsere Studie für Januar bis September 2026 umfasst 279 Beobachtungen, 47 vergleichbare Konfigurationen und 6 Modelle. Wir halten Modellname, Hardware, Präzision, Engine, Decoding-Verfahren und GPU-Aufteilung konstant. Die Zahl der gemeinsam bedienten Anfragen darf sich ändern, solange das mediane Antworttempo mindestens 50 Token/s/Nutzer beträgt und die mediane Zeit bis zum ersten Token höchstens zwei Sekunden.

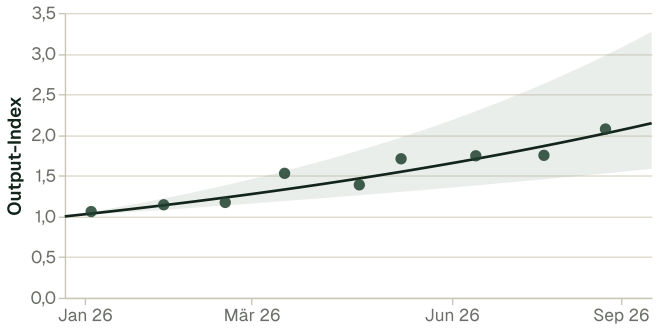
$$\log q_{c,t} = \alpha_c + \beta t + \varepsilon_{c,t}.$$

Der geschätzte Trend zeigt 9,1 % mehr Output je GPU pro Monat. Die nominale 95-%-Bootstrap-Spanne beträgt 5,4 %–14,5 %, wobei jedes der sechs Modelle gleich gewichtet ist. Lässt man jeweils ein Modell weg, liegen die monatlichen Gewinne bei 7,3 %–11,0 %. Nach den ersten Wochen eines Modells beträgt der Trend 7,4 % pro Monat; bei diesem Tempo dauert die Verdopplung, die wir im weiteren Verlauf unterstellen, etwa 10 Monate.

Wir fassen hier alle Arten von Softwareverbesserung zusammen. Für ein festes Modell und GPU-Setup erwarten wir, dass der Fortschritt nachlässt, je näher er an die Grenze dessen kommt, wie schnell die Chips rechnen und Daten bewegen können. Unsere Formel misst das bisherige Wachstum, ohne diese Grenze zu schätzen. Inference-Engineering ist noch ein junges Feld, daher erwarten wir vorerst kein Plateau.

Geschätzter Softwaretrend bei gleicher Konfiguration

6 Modelle · 47 Konfigurationen · Jan.–Sep. 2026



Punkte: bereinigte Monatsmittel. Linie: geschätzter Trend, 5. Jan. = 1. Band: 95-%-Intervall über Modelle. Rückblickend, wechselnde Stichprobe; kein Flottenindex.

Modelle und Serving-Software entstehen gemeinsam

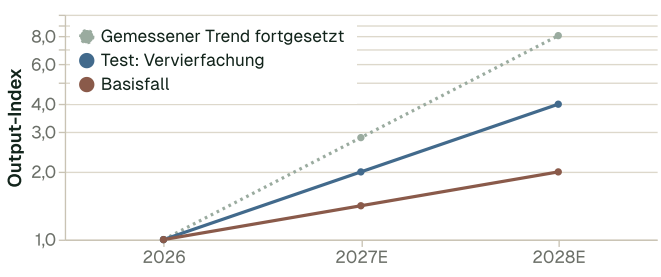
DeepSeek hat kürzlich das offene Inferenz-Framework DeepInfer veröffentlicht, das ursprünglich für V4.1 Flash auf vLLM aufbaute und eigenen GPU-Code, Drafting und Scheduling ergänzt. Auch Moonshots Kimi Linear verbindet spezialisierten Attention-Code mit vLLM. Post-Training-Firmen könnten abgestimmte Einstellungen oder Draft-Modelle liefern. Erscheinen diese Teile mit dem Modell, könnte mehr Optimierung schon vor dem Start erfolgen.

Modellspezifische Software macht Allzweck-GPUs für ein Modell effizienter und hält sie für andere nutzbar. Wir sehen darin einen Zwischenschritt vor modellspezifischer Hardware, an der DeepSeek Berichten zufolge arbeitet.

Inferenzanbietern, die Token verkaufen, könnte das den Aufwand für neue Modelle und die Betriebskosten senken. Nutzen Wettbewerber denselben Code, dürfte ein Teil der Ersparnis über niedrigere Preise weitergegeben werden. Die Anbieter würden weiter über Tempo, Zuverlässigkeit und GPU-Auslastung konkurrieren. Die Grafik vergleicht den gemessenen Softwaretrend mit dem Gewinn, den wir später unterstellen.

Unsere Annahme liegt unter dem Softwaretrend

Output je GPU allein durch Software · Ende 2026 = 1 · log. Skala



Gemessener Trend: 9,1 % pro Monat aus unserer Panelstudie, zwei Jahre fortgeschrieben. Basisfall und Vervierfachung: Annahmen im weiteren Verlauf.

Modelltypen und Modellgrößen

Die Brücke von GW zu Token-Output beginnt bei den Modellen und ihrer Parameterzahl. Dichte Modelle nutzen für jedes Token alle Gewichte des Sprachmodells, Mixture-of-Experts-Modelle (MoEs) nur einen Teil. MoE ist unter führenden offenen Modellen wegen der günstigen Kosten verbreitet, auch wenn sich Expertenzahl und Routing stark unterscheiden.

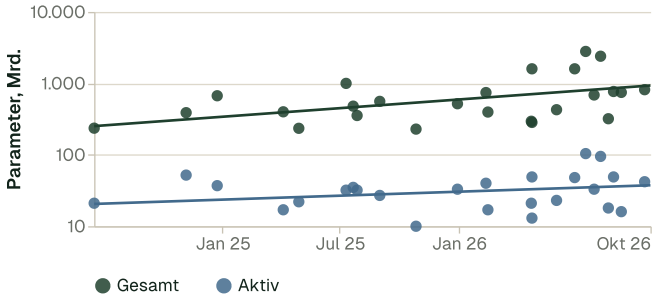
Für große offene MoEs mit mindestens 150 Mrd. Parametern schätzen wir getrennte Trends für gesamte und aktive Parameter. Der Vergleich von Architekturen desselben Labors ergibt ein jährliches Wachstum von 1,74× (gesamt) und 1,29× (aktiv), über 25 Beobachtungen und 9 Labore. Die Schätzung für die aktiven Parameter ist deutlich unsicherer, mit einer nominalen 95-%-Spanne von 0,94× bis 2,08× pro Jahr.

Für die Frontier insgesamt schätzt Epoch, dass die Trainingsrechenleistung seit 2020 jährlich auf rund das 5-Fache wächst, während sein Capabilities Index (ECI) seit September 2024 rund 14 Punkte pro Jahr zulegt.

Die Trainingsrechenleistung scheint schneller zu wachsen als die Parameterzahl, auch wenn beide Schätzungen verschiedene Modelle und Zeiträume abdecken. Fähigkeiten und Aufgabenhorizonte verbessern sich weiter. Mehr Rechenleistung kann in längeres Training und RL fließen, ohne dass die Parameterzahl steigt; bessere Daten und besseres Modelldesign erhöhen die Leistung bei gegebener Größe ebenfalls. Womöglich erreichen wir eine praktische Obergrenze der Modellgröße und erhalten dennoch deutlich leistungsfähigere Modelle.

Gesamte und aktive Parameter im Zeitverlauf

Große offene MoEs ≥150 Mrd. · 2024–2026 · logarithmische Skala



25 Architekturen / 9 Labore. Punkte: veröffentlichte Werte; Linien: Trends je Labor. Gesamt 1,74× pro Jahr, aktiv 1,29×. Intervall der aktiven Rate: 0,94–2,08×.

Eine kombinierte Schätzung der Modellgrößen

Um zu sehen, welche Modelle auf der Hardware laufen, kombinieren wir Quellen zur Token-Nutzung ohne China; in China entwickelte, anderswo betriebene Modelle zählen mit. OpenRouter und Vercel zeigen die Modellverteilung gut, überzeichnen aber Modelle mit offenen Gewichten. Mit Anbieterangaben und Kanalannahmen schätzen wir die Modellanteile insgesamt.

Unsere beispielhaften MoE-Größenklassen, mit aktiven Parametern je erzeugtem Token:

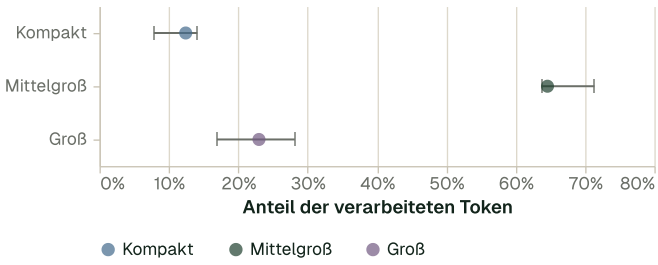
- **Kompakt:** 20 Mrd.–150 Mrd. gesamt, 3 Mrd.–15 Mrd. aktiv.
- **Mittelgroß:** 150 Mrd.–1 Bio. gesamt, 10 Mrd.–70 Mrd. aktiv.
- **Groß/Frontier:** 1 Bio.–6 Bio. gesamt, 40 Mrd.–300 Mrd. aktiv.

Für geschlossene Modelle verankern öffentliche Schätzungen die Spannen. Eine Community-Rekonstruktion von Astra nennt 2 Bio.–6 Bio. residente Parameter; ein benchmarkbasiertes Modell sieht Claude Opus 5 bei rund 3 Bio. (50-%-Spanne 1,0–9,1 Bio.). Epoch schätzt 50 Mrd.–300 Mrd. aktive Parameter für GPT-5. Die Schätzungen betreffen verschiedene Modellversionen.

Damit liegen die modellierten Token-Anteile ohne China bei rund 12 % kompakt, 65 % mittelgroß und 23 % groß. Die Abrechnungs- und Gateway-Fälle ergeben bei festen Größen und nativem Routing Spannen von etwa 8 %–14 %, 64 %–71 % und 17 %–28 %. Sie erfassen nicht die gesamte Unsicherheit.

Verteilung der Modellgrößen (Arbeitsannahme)

Token ohne China · feste Größen und natives Routing



Punkte: Referenz; Linien variieren Abrechnungs-, Traffic- und Gateway-Fälle. Natives Routing und Modellgrößen bleiben fest; weitere Unsicherheit ist nicht erfasst.

Wie die Quellen kombiniert werden

Wir schätzen, wie viele Token jeder Anbieter verarbeitet, und ordnen sie nach Modellgröße. Googles letzte Gesamtzahl vom Mai skalieren wir mit dem Wachstum seines API-Volumens (von 19 auf knapp 32 Mrd. Token pro Minute); API- und interne Token zählen nur einmal. Mit einem Zuschlag für andere Anbieter und private Systeme ergeben sich ohne China rund 0,56 Brd. verarbeitete Token pro Tag, oder 0,49 mit Google auf dem Mai-Niveau.

Bei OpenAI und Anthropic gehen wir vom für September berichteten annualisierten Umsatz von rund 50 Mrd. und 70 Mrd. USD aus und trennen Nutzungsentgelte von Abonnements; Anthropic weist die Aufteilung im S-1/A aus, für OpenAI schätzen wir sie.

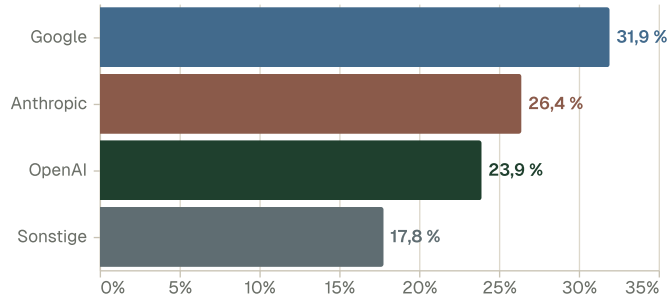
Den Umsatz aus bezahlter Nutzung teilen wir durch einen geschätzten Durchschnittspreis je verarbeitetem Token aus veröffentlichten API-Tarifen, Modellwahl, Caching und dem Eingabe-Ausgabe-Verhältnis. Der native Nutzungsmix ist unveröffentlicht, daher nutzen wir Gateway-Daten. Bei OpenAI gewichten wir die günstigsten Modelle herab, bis das bezahlte Volumen zur API-Angabe vom April passt, fortgeschrieben mit früherem Wachstum.

Bei Abonnements setzen wir die durchschnittliche Nutzung zu API-Preisen der Gebühr gleich, testen leichtere und stärkere Nutzung und behandeln Gratis- und interne Nutzung getrennt.

Gateway-Ranglisten bestimmen die Verteilung auf Modellgrößen; demnach werden bei OpenAI häufiger günstige Modelle genutzt als bei Anthropic, zu einem niedrigeren Preis je Token. Natives Routing und proprietäre Modellgrößen bleiben unsicher.

Traffic-Gewichte hinter der Schätzung

Anteil der modellierten verarbeiteten Token ohne China



Umsatz von 50 Mrd. USD (OpenAI) und 70 Mrd. USD (Anthropic) mit Annahmen zu Abos, Gratis- und interner Nutzung. Google: Mai-Summe mit API-Wachstum skaliert. Sonstige: Pauschale.

Wie sich der Mix bis 2028 ändern könnte

Der Modellmix 2028 ist schwer vorherzusagen, die Szenarien hier haben daher große Fehlerspannen. Kleinere Modelle könnten mehr der heutigen Routinearbeit übernehmen, während Nutzer KI zugleich schwierigere Arbeit übertragen könnten, die weiterhin große Modelle erfordert.

Wir testen drei Möglichkeiten:

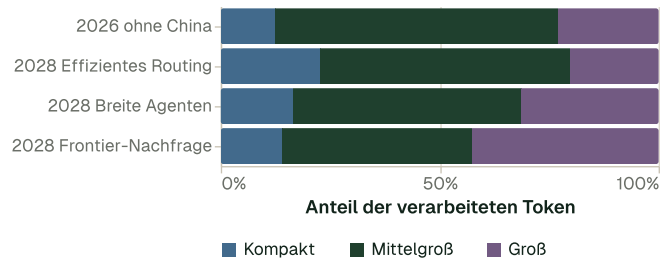
- **Effizientes Routing:** Kleinere Modelle übernehmen mehr der heutigen Aufgaben.
- **Breite Agenten:** KI erledigt mehr Aufgaben mit mehreren Schritten, etwa Dateien lesen, Code ändern und das Ergebnis testen. Das Modell wird dabei wiederholt aufgerufen.
- **Frontier-Nachfrage:** Nutzer setzen KI häufiger für schwierigere Aufgaben ein, die große Modelle erfordern.

Wir schätzen, wie viele Aufgaben Nutzer der KI geben, wie oft ein Modell je Aufgabe aufgerufen wird und wie viel Text es liest und schreibt. Aus diesen Annahmen ergeben sich die Größenanteile in der Grafik.

Hinzu kommen viele Unbekannte bei künftigen Neuerungen in Modellarchitektur, Destillation, Speicher, Reasoning und Inferenzeffizienz; sie könnten verändern, welche Modellgrößen sich für welche Aufgaben am besten eignen, und den gesamten Modellmix verschieben.

Modellmix nach Token: heute und 2028

Verarbeitete Token ohne China · abgeleitete Workflow-Szenarien



Anteile 2028 folgen Annahmen zu Workflow-Wachstum, Aufrufen, Token-Budgets und Routing. Die Ausgangskohorte ist kalibriert, nicht beobachtet. Keine Wahrscheinlichkeiten.

Was ein Gigawatt leisten kann

Wir fragen nun, wie viele Token ein GW im Kundenbetrieb verarbeitet.

Das hängt von vier Größen ab: den installierten Chips, den Modellen darauf, der Serverauslastung und der Qualität der Serving-Software. Die Chips entsprechen der oben gezeigten Flotte zum Jahresende. Der Modellmix bewegt sich vom heutigen zu den Fällen für 2028; 2027 liegt auf halbem Weg.

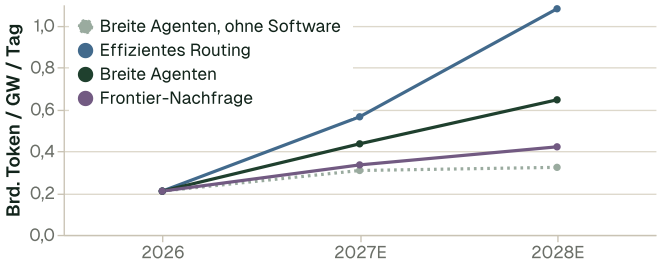
Wir nehmen an, dass Server über den Monat 45 % ihres Benchmark-Outputs liefern, weil die Nachfrage ruhige Stunden hat und Betreiber Reserven für Spitzen vorhalten; den Wert leiten wir aus Azures Produktivclustern ab, deren Spitzenlast das 1,7-Fache des Durchschnitts beträgt. Den gemessenen Softwaretrend schreiben wir langsamer fort: Der Output verdoppelt sich bis Ende 2028, was rund 8 Monaten der oben gemessenen Gewinne entspricht.

Damit verarbeitet ein GW im Kundenbetrieb Ende 2028 rund dreimal so viele Token wie heute: 0,21 Brd. pro Tag Ende 2026 und 0,65 im Fall Breite Agenten, oder 0,42–1,08 je nach Modellmix. Das sind verarbeitete Token, umgerechnet aus den erzeugten Token der Benchmarks zum heutigen Verhältnis von rund 15 zu eins.

Neuere Chips und bessere Software tragen etwa gleich viel bei. Chips allein heben den Output je GW bis 2027 auf das 1,8-Fache und bis 2028 auf das 2,3-Fache, Software verdoppelt ihn; ohne Softwaregewinn läge der Wert für Breite Agenten nur bei 0,32. Mit effizientem Routing verarbeitet ein GW 2,6-mal so viele Token wie bei Frontier-Nachfrage.

Token je Gigawatt, 2026 bis 2028E

Billiarden Token pro Tag je GW im Kundenbetrieb · Flotten zum Jahresende



45 % Auslastung; Software verdoppelt den Output bis 2028. Gepunktet: Breite Agenten ohne diesen Gewinn. Mix 2027E auf halbem Weg. Raten meist geschätzt.

Warum der Modellmix wichtig ist

Große Modelle brauchen je Token weit mehr Strom: Jedes Token durchläuft mehr Parameter, und die Gewichte verteilen sich auf mehr GPUs, die Daten austauschen müssen.

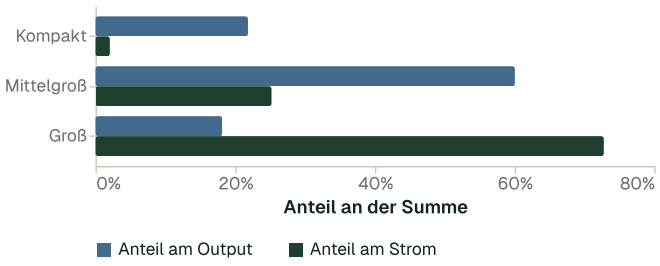
Benchmarks zeigen den Unterschied. Auf GB300-Racks mit InferenceX' Agenten-Traces erzeugt ein GW mit DeepSeek V4.1 Flash, dem meistgenutzten OpenRouter-Modell, rund 55-mal so viele Token wie mit Kimi K3.1 Pro, dem Nachfolger von Kimi K3 (2,8 Bio. Parameter) und dem jüngsten offenen Frontier-Modell mit Benchmark-Werten. DeepSeek V4 Pro, das große Modell hinter unseren Raten, liegt dazwischen: bei einem Zehntel von V4.1 Flash und dem 5,5-Fachen von Kimi. Alle Läufe halten mindestens 80 Ausgabe-Token/s/Nutzer bei P90; Engines, Caching und Testdatum variieren.

Geschlossene Frontier-Modelle könnten größer sein als V4 Pro, und Kimi zeigt, dass größere Modelle ein Mehrfaches an Strom brauchen können. Wir behalten die V4-Pro-Rate, weil sie nahe an Anthropics berichteten Kosten im Kundenbetrieb liegt, und testen sie von Kimis Tempo bis zum Doppelten.

Der Token-Anteil eines Modells kann stark vom Stromanteil abweichen. Auf der Flotte 2028 braucht ein Token des großen Modells rund das 10-Fache an Strom des mittelgroßen. Im Fall Breite Agenten liefern große Modelle 18 % des erzeugten Outputs, brauchen aber 73 % des Inferenzstroms. Wandert etwas Nachfrage zu Frontier-Modellen, bleibt deutlich weniger Strom für alles andere.

Output-Anteil ist nicht gleich Stromanteil

Breite Agenten · Flotte 2028 · einheitliches 8K/1K-Setup



Stromanteil: Outputanteil jedes Profils geteilt durch seine Flottenrate, auf 100 % skaliert. Die Raten sind meist geschätzte Übertragungen.

Der Vergleich mit heute

Wir prüfen die Methode an den Angaben der beiden größten Labore zu ihrer eigenen Rechenleistung.

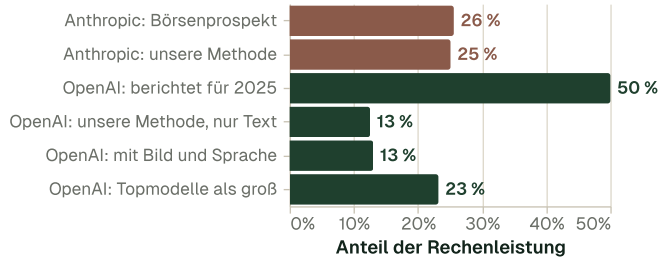
Anthropics S-1/A erlaubt den direktesten Vergleich: In den ersten neun Monaten 2026 kostete die Rechenleistung für Kunden 14,5 % des Umsatzes und machte 26 % der Rechenkosten aus. Unsere Methode benötigt für Anthropics bezahlten Traffic und Modellmix bei 45 % Auslastung rund 0,7 GW, also 25 % der rund 2,8 GW, die Anthropic nach unserer Schätzung Ende September hatte (aus weiter unten genannten Berichtszahlen).

Die Werte stimmen nur überein, weil wir eine Annahme am Prospekt ausgerichtet haben: Von drei getesteten Geschwindigkeiten großer Modelle behielten wir die, die Anthropics Wert traf; die beiden langsameren benötigten deutlich mehr Rechenleistung. Der Vergleich ist zudem ungenau, weil ein Ausgabenanteil kein Stromanteil ist und unsere Eingaben Schätzungen sind.

Bei OpenAI passen die Werte nicht zusammen. Laut Epoch AI bediente 2025 etwa die Hälfte von OpenAIs Rechenleistung Nutzer (Text, Bild, Video und Sprache); unsere Methode benötigt für den heutigen Traffic nur 13 %, Bild und Sprache eingerechnet. Hauptgrund ist die Modellgröße: Wir ordnen nur 8 % von OpenAIs Token großen Modellen zu, gegenüber 60 % bei Anthropic, und große Modelle brauchen heute rund das 6-Fache an Strom je Token. Stuft man OpenAIs Topmodelle als groß ein, sind es 23 %. Der Rest liegt vermutlich am Zeitpunkt: Anthropics Kundenanteil fiel seit 2025 von 38 % auf 26 %. Stimmt die berichtete Hälfte, braucht OpenAIs Kundenbetrieb rund das 3,8-Fache des modellierten Stroms, und unsere Flottenanteile sind zu niedrig.

Anteil der Labor-Rechenleistung im Kundenbetrieb

2025–26 · berichtet vs. unsere Methode bei 45 % Auslastung



Anthropic: Umsatzkosten an allen Rechenkosten, neun Monate bis September 2026. OpenAI: Epochs etwa hälftige Aufteilung, alle Modalitäten. Methode: Anteil an der Rechenleistung.

Was das Ergebnis bewegt

Im Fall Breite Agenten ergeben sich 2028 0,65 Brd. Token pro Tag je GW. Dahinter stehen fünf Annahmen, die wir nicht beobachten können; die Grafik variiert sie einzeln.

Auslastung. Wir nehmen an, dass Server 45 % ihres Benchmark-Outputs liefern. Das Ergebnis beträgt 0,29 bei 20 %, nahe der GPU-Auslastung eines nachgestellten Produktiv-Traces, und 0,86 bei 60 %.

Tempo neuerer Chips. In diesem Workload gemessene Chips tragen nur 13 %–30 % des Stroms der Flotte 2028; Rubin (21 %) ist in einem anderen Workload gemessen, 31 % entfallen auf noch nicht ausgelieferte Chips, der Rest ist von anderer Hardware abgeleitet. Das Ergebnis beträgt 0,36, falls unsere Schätzungen zu hoch sind, und 1,49, falls sie zu niedrig sind.

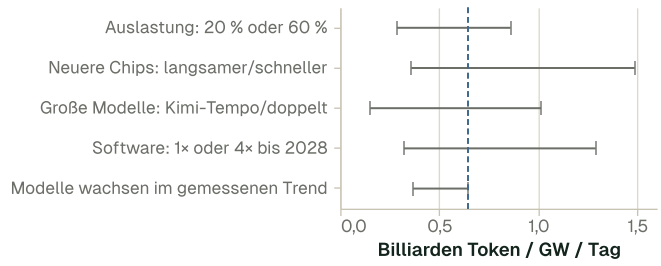
Tempo großer Modelle. Das Ergebnis beträgt 0,37 beim halben angenommenen Tempo und 1,01 beim doppelten. Beim Tempo von Kimi K3.1 Pro, um den Faktor 5,5 unter V4 Pro, wären es 0,15, unter den heutigen 0,21.

Softwaregewinn. Wir unterstellen bis 2028 eine Verdopplung des Outputs, weniger als gemessen (9,1 % pro Monat). Ohne Gewinn sind es 0,32, bei einer Vervierfachung 1,29 (rund 16 Monate im gemessenen Tempo).

Größere Modelle. Wir halten die Modellgrößen auf heutigem Niveau. Stiegen die Betriebskosten mit dem oben gemessenen Trend der aktiven Parameter, rund 1,29× pro Jahr, fiele das Ergebnis auf 0,37.

Was das Ergebnis für Breite Agenten bewegt

Brd. Token pro Tag je GW, wenn sich eine Annahme ändert · Linie = Basisfall



Je eine Änderung; gestrichelt: Basisfall (45 % Auslastung, Software verdoppelt). Auslastung 20 %–60 %; Software ohne Gewinn bis vierfach; große Modelle: Kimi K3.1 Pro bis doppelt.

Wohin der Strom heute fließt

Nun zur Frage, wie viel der Flotte Kunden bedient, zunächst je Labor.

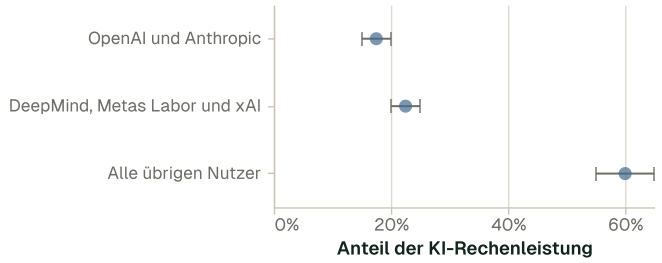
Die einzigen Zahlen aus einem Unternehmensabschluss liefert Anthropics S-1/A: Der Betrieb für zahlende Kunden machte 2024 30 % der Rechenkosten aus, 2025 38 % und in den ungeprüften ersten neun Monaten 2026 26 %. SemiAnalysis schätzt, dass OpenAI und Anthropic 2026 zusammen 34 %–38 % ihrer Kapazität für Inferenz nutzen. Laut Epoch AI verteilte OpenAI seine Rechenleistung 2025 etwa hälftig auf Forschung und Inferenz.

Der Großteil der übrigen Rechenleistung der Labore fließt nicht in Training im üblichen Sinn. Nach Epochs Schätzung entfiel 2024 rund ein Zehntel von OpenAIs Forschungsrechenleistung auf finale Trainingsläufe, der Rest auf Experimente, synthetische Daten und andere Forschung.

Die Grafik ordnet die Labore mit Epochs Schätzungen für Ende 2025 in die Flotte ein. OpenAI und Anthropic nutzten 15 %–20 % der weltweit betriebenen KI-Rechenleistung, ihr Kundenbetrieb also nur 6 %–10 %. Google DeepMind, Metas Labor und xAI nutzten nach unserer Lesart weitere 20 %–25 %. Die übrigen 55 %–65 % dienen Cloud-Kunden, kleineren Laboren, offenen Modellen und Systemen ohne Sprachbezug wie Empfehlungssystemen; Epoch schlüsselt sie nicht auf. Die Spannen sind grob, und der Anteil beider Labore steigt schnell: Patel sieht sie 2026 bei rund 30 % des Zubaus und 2027 bei bis zu 40–50 %.

Wer die KI-Rechenleistung der Welt nutzt

Grobe Schätzung · Anteil an der KI-Rechenleistung in Betrieb · Ende 2025



Spannen von Epoch AI oder daraus abgeleitet; Punkte sind Mittelwerte. Gemietete Kapazität zählt beim Labor, das sie nutzt.

Warum die Aufteilung eine Entscheidung ist

Die Zuteilung von GPUs auf Aufgaben ist zunehmend eine wirtschaftliche Entscheidung. Anthropic steuert seine Rechenleistung nach eigenen Angaben als „einen einzigen, fungiblen Pool“ und lenkt sie dorthin, „wo die zusätzliche Einheit voraussichtlich am produktivsten ist“.

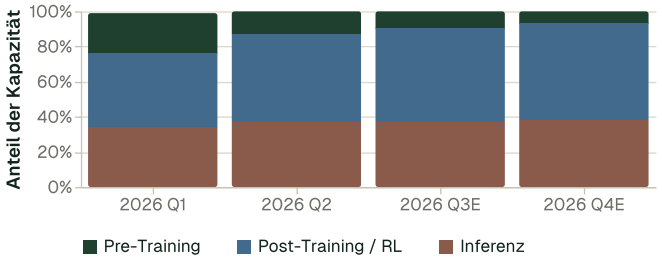
Ein GW hat also zwei konkurrierende Verwendungen: Der Kundenbetrieb bringt heute Umsatz, Forschung ein besseres Modell, das später Umsatz bringt. Laut Anthropic überstieg die Nachfrage stets die verfügbare Kapazität, sodass Strom für die eine der anderen fehlt. Das Narrativ rekursiver Selbstverbesserung und das Rennen der Labore erschweren die Wahl noch.

Bei Anthropic stieg der Kundenbetrieb 2025 auf 38 % der Rechenkosten und fiel 2026 auf 26 %, während der Umsatz auf das 13-Fache wuchs. Die Grafik zeigt die Schätzung von SemiAnalysis für beide Labore.

Der Rückgang hat zwei Lesarten. Laut Patel fließt ein steigender Anteil jedes neuen Megawatts in die Forschung; er sieht ein Frontier-Labor heute bei 40 % Inferenz, 50 % Forschung und 10 % Entwicklung. Anthropic verweist auf den Zeitpunkt: Kapazität kommt in großen Schritten, bevor der Umsatz folgt, den sie ermöglicht, und im dritten Quartal entfiel etwa die Hälfte der Rechenkosten auf kurzfristig mit Aufschlag gekaufte Kapazität.

Ein Drittel der Labor-Rechenleistung ist Inferenz

OpenAI + Anthropic; Kapazitätsanteile laut SemiAnalysis



Schätzungen vom 13. September; E = Prognose. Q1 ergibt gerundet 99 %. Anteile an der Kapazität der Labore, nicht an der IT-Energie der gesamten Flotte. Die Quelle endet 2026.

Heutige Nutzung gegenüber dem Angebot 2027

Außerhalb Chinas verarbeiten KI-Dienste für ihre Nutzer rund 0,6 Brd. Token pro Tag; nach unserer Methode sind das 8 %–11 % der heutigen Flotte.

Im nächsten Jahr wächst die installierte Flotte von 29 auf 66 GW, und jedes GW im Kundenbetrieb verarbeitet dank neuerer Chips und besserer Software 0,44 statt 0,21 Brd. Token pro Tag. Zusammen steigt die Kapazität auf das 4,8-Fache; derselbe Flottenanteil könnte 2,7 Brd. Token pro Tag verarbeiten. Ohne den unterstellten Softwaregewinn wäre es nur das 3,4-Fache, und das 3,7-Fache, wenn die Modelle wie oben gemessen weiterwachsen.

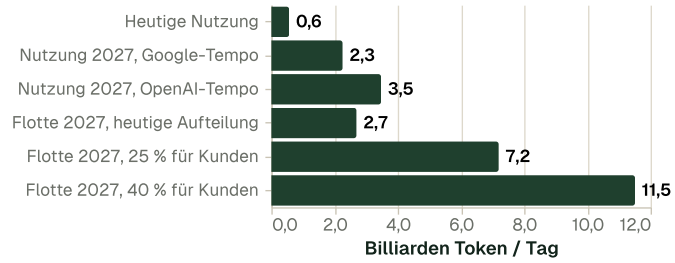
Weil die gesamte Token-Nutzung schwer zu messen ist, orientieren wir uns am berichteten API-Wachstum von Google und OpenAI. Googles API-Volumen hat sich im vergangenen Jahr mehr als vervierfacht; das von OpenAI stieg in den sechs Monaten bis April auf das 2,5-Fache, also rund das 6,2-Fache pro Jahr. Mit Googles Wachstumstempo steigt die gesamte Nutzung von 0,6 auf 2,3 Brd. Token pro Tag, mit dem von OpenAI auf 3,5. Die 2,7, die die Flotte 2027 bei heutiger Aufteilung verarbeiten könnte, liegen dazwischen; Angebot und Nutzung bleiben also etwa im Gleichgewicht. Token-Zahlen überzeichnen den Rechenbedarf, weil zwischengespeicherte Eingaben weit weniger Rechenaufwand kosten und ihr Anteil an den Agenten-Token auf OpenRouter dieses Jahr von rund 60 % auf knapp 90 % stieg.

Weit mehr ließe sich bedienen, wenn sich die Aufteilung änderte: Ein Viertel der Flotte von Ende 2027 könnte, wenn es nur Text bediente, 7,2 Brd. Token pro Tag verarbeiten, rund das 13-Fache der heutigen Nutzung.

Dieser Strom müsste aus der Forschung oder von anderen Eigentümern kommen. Nach unserer Schätzung halten OpenAI und Anthropic Ende 2026 27 %–31 % der Flotte ohne China (unten Patels Anteil am Zubau, oben die Ziele der Labore) und ein Jahr später 34 %–42 %, wenn sie 40–50 % des Zubaus erhalten, das obere Ende der SemiAnalysis-Erwartung.

Heutige Nutzung gegenüber dem Angebot 2027

Billiarden Token pro Tag für Nutzer · ohne China



Angebot: installierte Flotte Ende 2027 × Kundenanteil × unsere Token je GW. Nutzung: heute, ein Jahr im Tempo der API-Angaben fortgeschrieben. Heutiges Eingabe-Ausgabe-Verhältnis.

Wie Nachfrage und Mix die Aufteilung verschieben

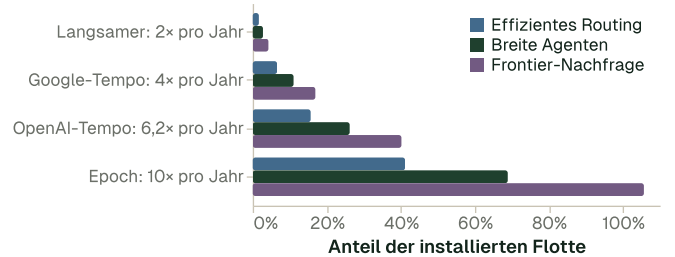
Wie viel der Flotte 2028 der Textbetrieb braucht, hängt zuerst vom Nachfragewachstum der nächsten zwei Jahre ab; wir betrachten vier Pfade. Die Nutzung könnte auf eine jährliche Verdopplung abflauen, Googles API-Wachstum von mehr als dem Vierfachen pro Jahr oder OpenAIs 6,2-Faches halten oder das rund Zehnfache pro Jahr erreichen, das Epoch AI für die Nachfrage bei heutigen Preisen schätzt. Im Mix Breite Agenten erfordert das 2 %–3 %, 10 %–12 %, 23 %–30 % und 60 %–78 % der Flotte von 127 GW.

Der Modellmix verschiebt jeden Wert, weil dieselben Token bei Frontier-Nachfrage das 2,6-Fache an Strom brauchen wie bei effizientem Routing. Beim schnellsten Pfad bräuchte Frontier-Nachfrage 92 %–119 % der Flotte und ließe nichts für andere Arbeit.

Nachfrage und Forschung konkurrieren somit um denselben Strom: Je schneller die Nachfrage wächst und je größere Modelle die Nutzer wählen, desto weniger bleibt für Forschung und andere Workloads; langsamere Nachfrage setzt Strom frei. Schnelleres Wachstum halten wir für wahrscheinlicher, weil Google und Anthropic zuletzt so stark wuchsen, obwohl die Nachfrage nach eigenen Angaben ihre Kapazität überstieg. Die Pfade beruhen auf wenigen Firmenangaben; der langsame Fall ist unsere Annahme.

Schnellere Nachfrage, weniger Flotte für Forschung

Anteil der Flotte 2028 für Text · nach Nachfragetempo und Modellmix



Heutige Nutzung, zwei Jahre im jeweiligen Tempo fortgeschrieben, bei unseren Token je GW; Mittelwerte. Langsam: unsere Annahme; 10×: Epoch AIs Schätzung bei heutigen Preisen.

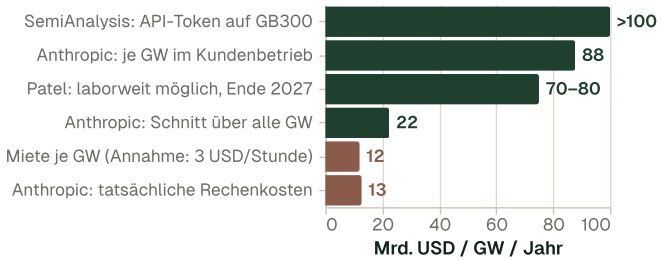
Was ein Gigawatt erlöst

An einem GW verdienen zwei Ebenen: Ein Infrastrukturanbieter vermietet die Rechenleistung, ein Modellanbieter verkauft die Arbeit seiner Modelle darauf. SemiAnalysis' August-Szenario für einen GB300-Cluster sieht den API-Umsatz bei über 100 Mrd. USD je GW und Jahr, bei rund 12 Mrd. USD Miete zu 3 USD je GPU-Stunde; Patel hält bis Ende 2027 für ein Labor insgesamt 70–80 Mrd. USD je GW für möglich, wenn es das beste Modell behält.

Anthropics Prospekt erlaubt die Prüfung an einem realen Unternehmen, dessen Umsatz in den ersten neun Monaten 2026 im Schnitt rund 44 Mrd. USD pro Jahr betrug. Nach unserer Schätzung verfügte es in dieser Zeit im Schnitt über rund 2,0 GW, zwischen berichteten 1,4 GW Ende 2025 und dem Ziel von 3,5 GW für Ende 2026; der ausgewiesene Kundenanteil ergibt rund 0,5 GW im Kundenbetrieb. Das sind rund 88 Mrd. USD je GW im Kundenbetrieb, weniger als im Szenario von SemiAnalysis. Über die gesamte Rechenleistung sind es nur 22 Mrd. USD je GW, deutlich unter Patels Wert, weil der Großteil keine zahlenden Kunden bedient. Die Rechenkosten lagen bei rund 13 Mrd. USD je GW, nahe der von SemiAnalysis angenommenen Miete.

Umsatz und Rechenkosten je Gigawatt

Mrd. USD je Gigawatt und Jahr · Umsatz grün, Rechenkosten braun



Anthropic: neun Monate bis September 2026, annualisiert, über unsere Schätzung seiner GW (Kundenbetrieb, gesamt). Szenarien: SemiAnalysis und Patel, August 2026. Kein Gewinn.

Was ein Gigawatt erlösen muss

Der Eigentümer eines GW muss die Baukosten wieder einspielen, die Epoch AI für Nvidia-GB200-Systeme auf rund 38 Mrd. USD und BCG auf rund 42 Mrd. USD beziffert. Über die Lebensdauer der Anlagen zu branchenüblichen Kapitalkosten annualisiert, samt Strom und Instandhaltung, kommt Epoch auf rund 8,5 Mrd. USD pro Jahr. Das BCG-Modell, das auch Schulden und eine Eigenkapitalrendite abdeckt, setzt den nötigen Ertrag auf 15 Mrd. USD an.

Der Eigentümer erhält die Miete, nicht den Umsatz des Labors; die 12–13 Mrd. USD, die Labore zahlen, decken Epochs Wert, aber nicht den von BCG. Auf Basis von Epochs Kosten, mit Chips und Netzwerk über fünf und dem Gebäude über vierzehn Jahre abgeschrieben, bleiben dem Eigentümer bei voller Vermietung nach Abschreibung und Betriebskosten 5,1–5,8 Mrd. USD pro Jahr. Das entspricht einer operativen Marge von 42 %–45 % und einer Rendite von 13 %–15 % pro Jahr, vor Steuern und Zinsen, auf die investierten 38 Mrd. USD.

Anthropics 22 Mrd. USD je GW liegen über beiden Werten, obwohl nur ein Viertel seiner Rechenkosten Kunden dient; sie entsprächen den Eigentumskosten, wenn 10 %–17 % eines GW zu Preisen von 2026 zahlende Kunden bedienten.

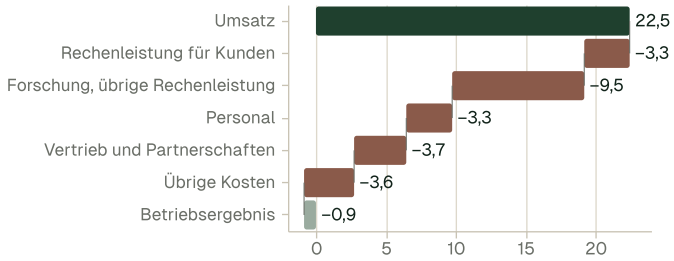
Wohin der Umsatz eines Labors fließt

Für das Labor, das das GW mietet, ist der Umsatz kein Gewinn, weil neben der Rechenleistung auch Personal, Vertrieb und Finanzierung zu bezahlen sind. Die Rechenleistung beansprucht 12,7 Mrd. USD der 22,5 Mrd. USD, die jedes GW erlöst; davon dienen 9,5 Mrd. USD nicht zahlenden Kunden. Personal, Vertrieb und übrige Kosten beanspruchen weitere 10,6 Mrd. USD, und das Unternehmen landet knapp unter der Gewinnschwelle.

Diese Werte halten die Preise auf dem Niveau von 2026; da sich die Token je GW nach unserer Schätzung im nächsten Jahr etwa verdoppeln, steigt der Umsatz je GW nur, wenn der Preis je Token um weniger als die Hälfte fällt.

Wohin der Umsatz eines Anthropic-Gigawatts fließt

Mrd. USD je Gigawatt seiner gesamten Rechenleistung · 2026, annualisiert



Betriebsaufwand laut S-1/A, annualisiert, über unsere Schätzung von Anthropics gesamten GW. Rechenleistung für Kunden: näherungsweise Umsatzkosten. Ergebnis vor Zinsen.

Fazit

Der Ausbau rentiert sich bei der Rechenleistung, die Kunden bedient; die Gesamtrendite hängt stärker von Nachfrage und Aufteilung der Rechenleistung ab als vom Chipangebot. Die heutigen Nutzer beanspruchen rund ein Zehntel der Flotte. Bleibt es dabei, wächst die Kapazität für Text im nächsten Jahr auf das 4,8-Fache und die berichtete Nutzung auf das 4,0- bis 6,2-Fache; Angebot und Nutzung bleiben nur etwa im Gleichgewicht, solange die Nachfrage ihr Tempo hält und die Serving-Software besser wird.

Ein GW im Dienst von Anthropics Kunden erlöst ein Mehrfaches seiner Eigentumskosten, doch der Großteil der Rechenleistung der Labore bedient keine Kunden, und Anthropic insgesamt liegt knapp unter der Gewinnschwelle. Der Rest muss sich über Forschung für bessere Modelle auszahlen, über nicht erfasste Workloads oder über schneller wachsende Nachfrage.

Der Modellmix zählt zu den größeren Effekten: Mit den größten Modellen brauchen dieselben Token das 2,6-Fache an Strom, zulasten der Forschung.

Die Ergebnisse beruhen auf Größen, die wir nicht beobachten oder konstant halten: Flottenauslastung und künftige Preise sind nicht veröffentlicht, Modellgrößen und Anfrageprofile bleiben auf heutigem Niveau, und unsere Rate für große Modelle ist an einem einzigen Prospekt kalibriert.

Quellen

1. Diffusion Capital - lab research and serving budgets
2. Diffusion Capital - lab compute allocation model
3. Diffusion Capital - installed AI fleet estimate
4. Diffusion Capital - compute demand and revenue model
5. Alphabet - CFO Anat Ashkenazi at Morgan Stanley TMT conference
6. HuggingNews - Anthropic's AI compute goal for 2026 drops to 3.5 GW
7. Anthropic S-1/A revenue channels
8. Lyx - GPT6 Astra architecture hypothesis
9. BCG - Who pays for AI infrastructure
10. Diffusion Capital - token traffic and model-size estimate
11. Diffusion Capital - fleet use and returns model
12. Diffusion Capital - data-centre site database
13. Diffusion Capital - data-centre capacity and operator dataset
14. Diffusion Capital - tokens per gigawatt model
15. Diffusion Capital - 2028 model-mix scenarios
16. Dwarkesh Podcast - Dylan Patel (SemiAnalysis) on AI compute
17. Epoch AI - Introducing the AI chip users explorer
18. Epoch AI - AI data center cost breakdown
19. Epoch AI - How much AI compute do frontier labs use?
20. Epoch AI - published general training-compute and ECI trends
21. Epoch AI - Is a compute crunch coming
22. Epoch AI - How many digital workers could OpenAI deploy
23. Epoch final training runs versus R&D compute spending
24. Etched - Frontier Inference Clusters
25. Exponential View - The State of the AI Economy
26. Anpaure - Frontier Parameter Model
27. Investing.com (FT report) - OpenAI annualized revenue at $50bn far below reports
28. Google - Q3 2026 model API usage
29. Google Cloud - Inside the Ironwood AI stack
30. Alphabet Q2 2026 CEO remarks
31. Groq - NVIDIA inference technology licence
32. InferenceX — DeepSeek V4 Pro B200 vLLM0.25 paired MTP sweep
33. InferenceX - GB300 NVL72 versus B300 specifications
34. SemiAnalysis InferenceX - agent-trace histories for recent models
35. SemiAnalysis InferenceX - common8K/1K three-profile histories
36. Diffusion Capital — matched InferenceX software panel study
37. Chatterji et al. - How People Use ChatGPT (NBER w34255)
38. NVIDIA - Blackwell MoE serving improvements
39. OpenAI - the next phase of enterprise AI
40. OpenAI DevDay2025
41. OpenRouter consumption snapshot through15October2026
42. Patel et al. - Characterizing power management opportunities for LLMs in the cloud
43. ROSE - reinforcement learning rollouts on serving accelerators
44. SemiAnalysis frontier lab compute mix - September 2026
45. SemiAnalysis - Datacenters 2027: The Year of Acceleration, +50GW in a Single Year
46. SemiAnalysis - The great GPU shortage: rental capacity
47. SemiAnalysis InferenceX serving benchmark
48. SemiAnalysis - The Wild Wild West Of LEGO Datacenters
49. SemiAnalysis - Datacenter moratoriums mapped
50. SemiAnalysis - RL systems, matching generator and trainer pools
51. SemiAnalysis - Rubin GLM response-speed preview
52. SemiAnalysis - Rubin Kimi K3 response-speed preview
53. SemiAnalysis: SpaceX terrestrial capacity
54. SemiAnalysis Weekly 039: 2027 forecast
55. Original publisher model card: deepseek-ai/DeepSeek-V4.1-Flash
56. Original publisher model card: moonshotai/Kimi-K3
57. Original publisher model card: openai/gpt-oss-120b

Research zu Diskussionszwecken und keine Anlageberatung. Nichts hiervon ist eine Empfehlung, ein Angebot oder eine Aufforderung zum Kauf oder Verkauf eines Wertpapiers, und die Verhältnisse einzelner Leser bleiben unberücksichtigt. Die Quellen werden als verlässlich erachtet, eine Gewähr wird nicht übernommen. Die Einschätzungen gelten zum angegebenen Datum und können sich ohne Ankündigung ändern. Szenarioschätzungen sind keine Prognosen, sofern sie nicht ausdrücklich als solche gekennzeichnet sind.