AI chip returns depend on how much compute serves customers

A gigawatt serving customers earns about $88B a year against $8.5–15B to own, but serving takes about a tenth of the fleet today. Capacity per gigawatt triples by 2028, so the return rests on demand growing four to six times a year.

AS OF
14 November 2026

SCOPE
Non-China AI fleet to 2028: tokens per gigawatt, serving share, revenue and cost per gigawatt

KEY POINTS

- **One gigawatt processes about three times as many tokens by 2028.** Newer chips and better software lift it from 0.21 to 0.65 quadrillion tokens a day (tested range 0.15–1.49), half of it from software gains projected from the measured trend.
- **Serving customers takes about a tenth of the fleet, and at that share capacity keeps pace with demand.** Customer traffic needs 8%–11% of today's fleet. Held there, serving capacity grows 4.8 times next year, against disclosed use growing 4.0–6.2 times.
- **A gigawatt serving customers earns $88B a year, but only a quarter of Anthropic's compute does so.** That is six to ten times the $8.5–15B it costs to own. Across all its compute it earns $22B per gigawatt, slightly below break-even.

The capacity base case

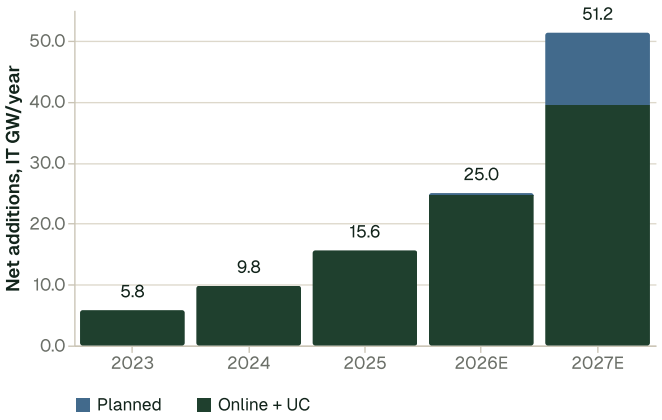
We start this analysis with the total datacentre capacity available and expected to come online. This capacity is the baseline for allocating compute to different use cases and assessing how much revenue each GW can generate. The best public source for this is SemiAnalysis.

The 19 October construction forecast puts annual non-China datacentre IT net additions at 25.0 GW in 2026 and 51.2 GW in 2027. This includes projects under construction as well as planned projects, so some delivery risk remains.

In this note, we aim to quantify how the composition of this fleet will change over time and how hardware and software improvements will affect token production. We will compare this with current demand and the trajectory needed to absorb the additional supply. We then examine how this compute is split between serving customers and research, and what a gigawatt earns and costs, using Anthropic's filing.

SemiAnalysis expects the buildout to accelerate

Annual non-China IT GW added, by building status



UC = under construction; E = forecast. 2026 planned = 25.0 – 24.6 = 0.4 GW. 2027 published total: 51.2 GW; rounded segments: 51.3 GW.

In August, SemiAnalysis's Dylan Patel put global AI-compute additions at roughly 70 GW in 2028 and agreed that the stock would exceed 200 GW by year-end. For this note we use a lower estimate of the AI fleet installed outside China, which we expect to reach 127 GW by then. SemiAnalysis also expects the share of non-China additions built outside the US to recover to around 35% in 2028.

So far, moratoriums and datacentre restrictions have delayed a relatively small amount of capacity. The midterm results could be challenging for the datacentre buildout but are primarily impacting projects coming online in 2028. Projects in 2026 and 2027 are already too far ahead and have all necessary permits. This buildout could also accelerate as modular designs become more prevalent. Fully modular designs could reduce construction times from 18–24 months to 12–18 months, excluding permitting.

How the fleet's hardware is changing

As new hardware generations come online, they change the composition of the global GPU fleet. Strong demand for AI services is keeping older generations online for longer.

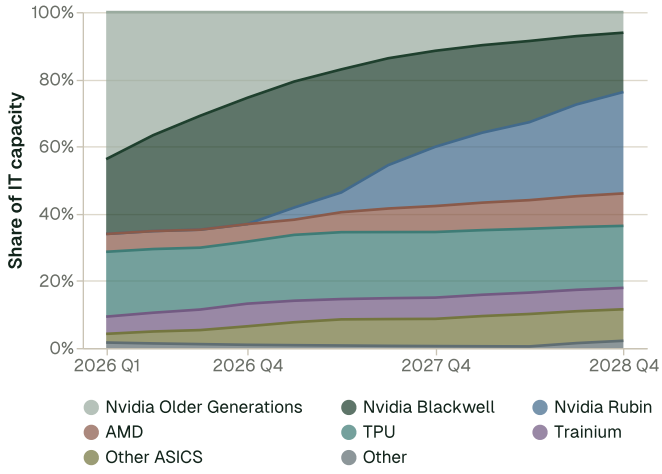
New hardware is more powerful and efficient, and also more closely aligned with current workloads and models. The focus is increasingly on the entire rack. GB200 NVL72, for example, connects 72 GPUs over NVLink, reducing communication bottlenecks when serving large models.

Training and inference can be allocated to separately optimised hardware pools. Inference can be further split into prefill, which processes the prompt, and decode, which generates the response. In reinforcement learning, rollout generation uses inference forward passes, while weight updates require both forward and backward passes.

These efficiency gains can compound across hardware, software and model design, increasing token production per GW. The combined gain depends on which bottlenecks each improvement removes.

The installed AI fleet changes composition

Non-China; estimated share of installed provisioned IT



Internal hardware-mix estimate; not a SemiAnalysis forecast. Shares of non-China installed IT capacity. The GW base uses published SemiAnalysis figures.

Frontier labs and hyperscalers are also bringing their own hardware online. Google's TPU and AWS Trainium are the most prominent examples. Other examples are Microsoft's Maia, which serves production inference, and Meta's MTIA, which runs ranking and recommendation workloads, with generative-AI inference chips on its roadmap. OpenAI plans initial Jalapeño deployments with Broadcom in late 2026. Google's TPU shows the potential of hardware/software co-design: chips, interconnects, compilers and models optimised together.

Specialist chip companies are also targeting fast inference. Cerebras uses wafer-scale processors, SambaNova reconfigurable dataflow chips and Groq dedicated inference processors. Startups such as Etched are building frontier-model inference systems, optimising chips, memory and interconnects for prefill and decode.

Output at required generation speeds

The chart below compares DeepSeek V4 Pro agentic serving across seven systems, measuring output tokens per IT-MW at minimum response speeds of 60 tokens/s/user at the median and 80 or 100 at P90.

In NVIDIA's preview submissions, Rubin delivers 1.16x and 1.27x GB300's output per IT-MW at the median-60 and P90-100 thresholds; the 1.90x at P90 80 rests on a single run at the threshold's edge. These ratios already account for assumed system power, including the increased overhead.

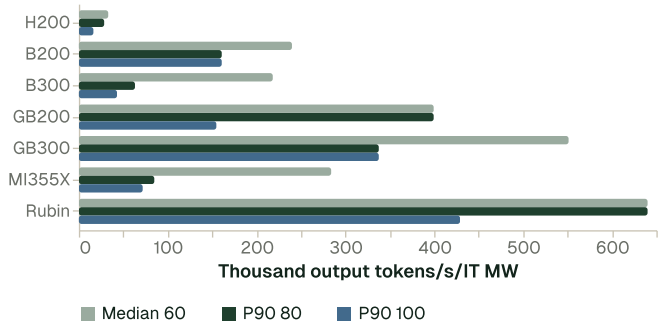
The shift towards rack-scale systems expands the scale-up domain from eight GPUs in a B300 server to 72 linked by NVLink in GB300 NVL72. This can reduce bottlenecks during MoE expert exchanges.

On the same DeepSeek V4 Pro agentic workload, GB300 NVL72 delivers 2.53x, 5.35x and 7.86x B300's output per IT-MW at the three thresholds, respectively. The selected setups use 24–28 GB300 GPUs versus 4–8 B300 GPUs, assuming 2.12 and 2.10 kW per GPU including system overhead. The B300 runs are single-node recipes without separate prefill/decode pools, so the larger ratios partly reflect an untuned baseline; the older B200 outperforms B300 at all three thresholds.

Each chip also brings HBM and compute improvements that affect different parts of the workload. As NVIDIA describes, more memory bandwidth helps memory-bound generation, more memory capacity holds model weights and cached context, and more compute accelerates compute-bound stages.

Output at minimum generation speeds

DeepSeek V4 Pro · agentic workload · configured IT power



Minimum speeds: median 60, P90 80 or P90 100 tokens/s/user; no first-token latency limit. Engines/precision differ; system power assumed. Rubin: NVIDIA preview submissions.

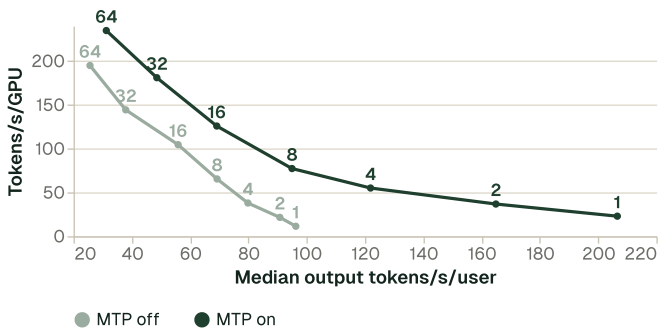
Faster responses compete with aggregate output

Interactivity measures generated tokens per second per user; throughput measures output across all requests per GPU. Serving more requests together generally raises throughput while slowing each user, which can make fast coding-agent modes more expensive. Hardware and software improvements shift this trade-off: SemiAnalysis reported NVIDIA Rubin NVL72 previews exceeding 850 tokens/s/user on Kimi K3 on 27 October and 1,000 on GLM-5.3 on 1 November. These are P90 response speeds; both curves still show fewer total tokens per dollar as response speed rises.

For a matched software example, in an InferenceX DeepSeek V4 Pro run using FP4, eight B200s and vLLM 0.25.0, multi-token prediction (MTP) supports 16 concurrent requests rather than eight while maintaining roughly 69 median tokens/s/user. This produces 1.92x as many output tokens per GPU. The chart below shows how DeepSeek V4 can be delivered along the throughput-interactivity curve.

Tuning changes the speed–throughput curve

DeepSeek V4 Pro · B200 · FP4 · 8K input / 1K output



12 July 2026; same eight-GPU node and vLLM 0.25.0 build. Labels: concurrent requests. Output-only tokens; full shared-node time. First-token latency is separate.

Software engineering increases token output

Apart from hardware improvements themselves, hardware vendors and serving-engine teams are working together to optimise particular model–hardware combinations. NVIDIA and SGLang document gains from kernel fusion, memory management, scheduling and bug fixes. These improvements arrive as teams implement and deploy them.

Our January–September 2026 study covers 279 observations, 47 matched configurations and 6 models. We hold the model name, hardware, precision, engine, decoding method and GPU layout fixed. The number of requests served together can change, provided median response speed stays at least 50 tokens/s/user and median time to the first token stays within two seconds.

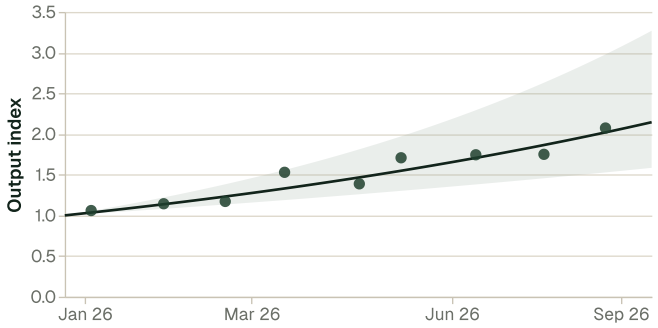
$$\log q_{c,t} = \alpha_c + \beta t + \varepsilon_{c,t}.$$

The fitted trend shows 9.1% more output per GPU per month. The nominal 95% bootstrap range is 5.4%–14.5%, with each of the six models given equal weight. Leaving out any one model gives monthly gains of 7.3%–11.0%. After a model's first weeks the trend is 7.4% a month, at which the doubling we project later in the note takes about 10 months.

We group all kinds of software improvement together here. For a fixed model and GPU setup, we would expect progress to slow as it approaches the limits of how quickly the chips can calculate and move data. Our formula measures past growth, without estimating that limit. Inference engineering is still a young field, so we do not expect this to plateau soon.

Estimated software trend across matched stacks

6 models · 47 configurations · Jan–Sep 2026



Dots: adjusted monthly means. Line: fitted trend, rebased to5Jan=1. Band:95% model-cluster interval. Retrospective, changing sample; not a fleet index.

Models and serving software are developed together

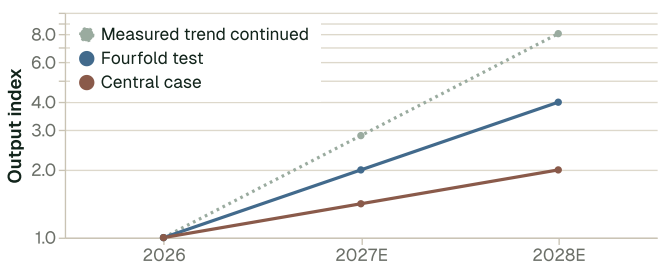
DeepSeek has recently released an open inference framework, DeepInfer, originally built on vLLM for V4.1 Flash, adding custom GPU code, drafting and scheduling. Moonshot's Kimi Linear also pairs specialised attention code with vLLM. Post-training companies could also supply tuned settings or draft models. Releasing these parts with a model could bring more optimisation before launch.

Model-specific software makes general-purpose GPUs more efficient for a particular model while keeping them usable for other models. We see this as an intermediate step before model-specific hardware, which DeepSeek is reported to be working on.

For inference providers selling tokens, this could reduce the engineering needed to support new models and lower serving costs. If competitors reuse the same code, some savings may be passed on through lower prices. Providers would still compete on speed, reliability and how well they keep GPUs busy. The chart compares the measured software trend with the gain we project later in the note.

We project less than the measured software trend

Output per GPU from software alone · end-2026 = 1 · log scale



Measured trend: 9.1% a month from our panel study, continued for two years. The central case and the fourfold test are the projections used later in the note.

Model types and sizes

To build a reliable bridge between GW and token output, we start with the models themselves and their parameter counts. Dense models use their full language-model weights for each token. Mixture-of-experts models, or MoEs, activate a subset. MoE is common among leading open models because of its favourable economics, although expert counts and routing techniques differ widely.

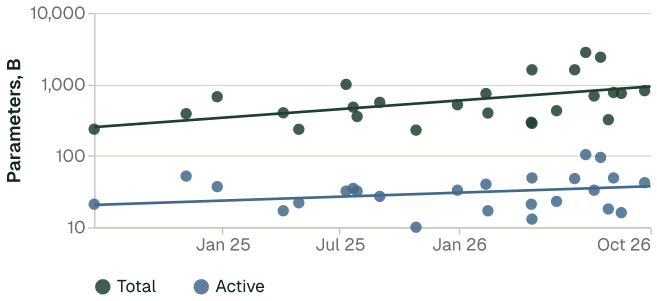
For large open MoEs with at least 150B total parameters, we fit separate trends to total and active counts. Comparing architectures within the same lab gives 1.74× annual growth in total parameters and 1.29× in active parameters, across 25 observations and 9 labs. The active-count estimate is much less certain: its nominal 95% range is 0.94–2.08× per year.

Across the broader frontier, Epoch estimates that training compute has grown around 5× a year since 2020, while its Capabilities Index (ECI) has gained about 14 points a year since September 2024.

Training compute appears to be growing faster than parameter counts, although the two estimates cover different models and periods. Capabilities and task horizons continue to improve. More compute can go into longer training and RL without increasing parameter counts. Better data and model design can also improve performance at a given size. We may eventually approach a practical upper limit on model size and still get much more powerful models.

Total and active parameters over time

Large open MoEs ≥150B · 2024–2026 · logarithmic scale



25 architectures / 9 labs. Dots: disclosed counts; lines: same-lab fits. Total 1.74×/yr, active 1.29×. Active rate interval: 0.94–2.08×.

A combined model-size estimate

To assess which models are served on the hardware, we combine token-usage sources on an ex-China basis. Chinese-developed models deployed elsewhere are still part of the estimates. OpenRouter and Vercel show model distributions well, but they over-represent open-weight models. We combine this with provider disclosures and channel assumptions to get to overall model share estimates.

Our illustrative MoE workload bands, with active counts per generated token, are:

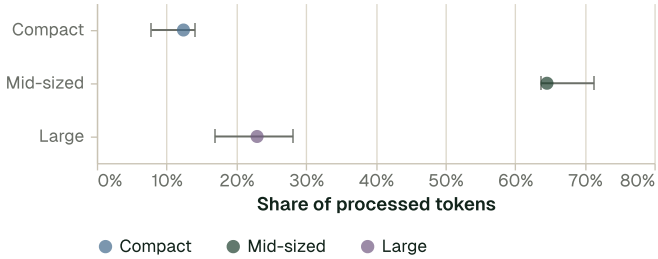
- **Compact:** 20B–150B total, 3B–15B active.
- **Mid-sized:** 150B–1T total, 10B–70B active.
- **Large/frontier:** 1T–6T total, 40B–300B active.

For closed models, public estimates help anchor the ranges. A community Astra reconstruction proposes 2T–6T resident parameters; a benchmark-based model estimates Claude Opus 5 near 3T, within a 50% range of 1.0–9.1T. Epoch estimates 50B–300B active parameters for GPT-5. These remain estimates with distinct model versions.

With these reference profiles, modelled ex-China token shares are about 12% compact, 65% mid-sized and 23% large. The billing and gateway cases give working ranges of approximately 8%–14%, 64%–71% and 17%–28%, holding sizes and native routing fixed. These spans do not capture all uncertainty.

Working model-size distribution

Ex-China tokens · fixed sizes and native routing



Dots show the reference; lines vary billing, traffic and gateway cases. Native routing and model sizes stay fixed; wider uncertainty is not captured.

How the sources are combined

We estimate how many tokens each provider processes and then sort them by model size. Google last published a total in May, which we scale by the growth of its API volume since then (19 to nearly 32 billion tokens a minute), counting API and internal tokens once. An allowance for other providers and private deployments brings the total to about 0.56 quadrillion processed tokens a day outside China, or 0.49 with Google left at its May level.

For OpenAI and Anthropic, we start from annualised revenue of about $50B and $70B, the levels reported for September, and separate usage fees from subscriptions. Anthropic discloses this split in its S-1/A; for OpenAI we estimate it.

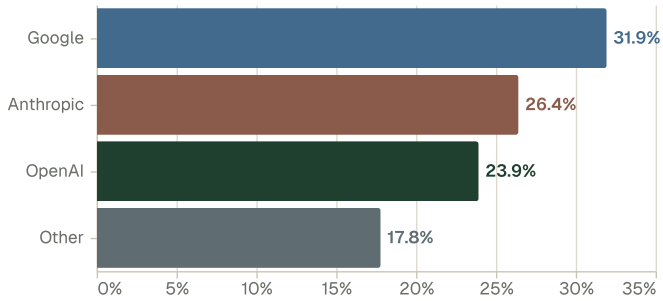
We divide paid usage revenue by an estimated average price per processed token, which reflects published API tariffs, model choice, caching and the balance of input and output. Native usage mixes are undisclosed, so we use gateway data. For OpenAI we down-weight its cheapest models until paid volume matches its April API disclosure carried forward at its earlier growth.

For subscriptions, we assume that average usage, valued at API prices, equals the fee. We test lighter and heavier use, and treat free users and internal use separately.

Gateway rankings then guide the allocation across model sizes, which gives OpenAI more exposure to inexpensive models than Anthropic and a lower price per token. Native routing and proprietary model sizes remain uncertain.

Traffic weights behind the estimate

Share of modelled ex-China processed tokens



Revenue of $50B (OpenAI) and $70B (Anthropic) with assumed subscription, free and internal use. Google: May total scaled by API growth. Other: allowance.

How the mix could change by 2028

The model mix in 2028 is hard to forecast, so the scenarios here carry wide error bars. Smaller models could take over more of today's routine work, while people may also ask AI to do harder work that still needs large models.

We test three possibilities:

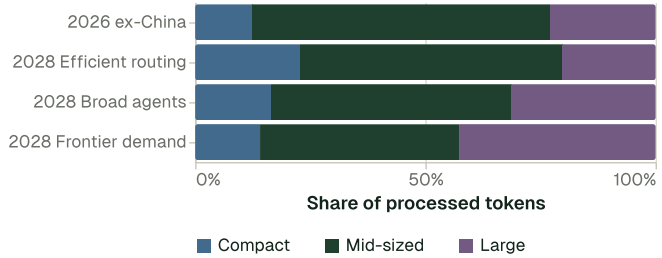
- **Efficient routing:** Smaller models take over more existing tasks.
- **Broad agents:** AI handles more tasks with several steps, such as reading files, changing code and testing the result. The model is used repeatedly during that work.
- **Frontier demand:** People use AI for harder tasks that need large models more often.

We estimate how many tasks people give AI, how often a model is used during each task, and how much text it reads and writes. These assumptions produce the chart's size shares.

There are also many unknowns around future innovations in model architecture, distillation, memory, reasoning and inference efficiency, which could change which model sizes are best suited to different tasks and shift the overall model mix.

Model-token mix: now and 2028

Ex-China processed tokens · derived workflow scenarios



2028 shares follow assumed workflow growth, calls, token budgets and routing. Baseline cohort split is calibrated, not observed. No scenario probabilities.

What one gigawatt can produce

We now look at how many tokens a gigawatt can process when it serves customers.

That number depends on four things: the chips installed, the models they run, how busy the servers are and how good the serving software is. The chips follow the year-end fleet composition from earlier in the note. The model mix moves from today's towards the different 2028 cases, and we place 2027 halfway between.

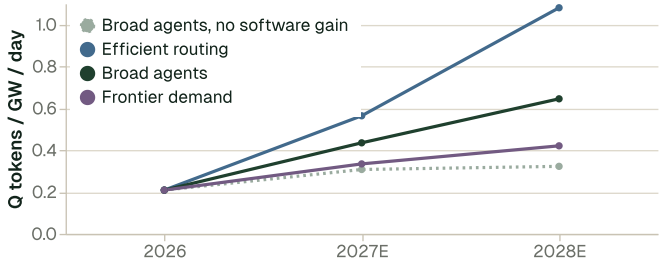
We assume that servers deliver 45% of their benchmark output over the month, because demand has quiet hours and operators keep spare capacity for peaks, a level we derive from Azure's production clusters, where peak load is 1.7 times the average. For software we project the measured trend at a slower pace: output doubles by the end of 2028, which equals about 8 months of the gains measured earlier in the note.

On these inputs, a gigawatt serving customers processes about three times as many tokens by the end of 2028 as today: 0.21 quadrillion a day at the end of 2026 and 0.65 in the broad-agent case, or 0.42–1.08 depending on the model mix. These are processed tokens, converted from the benchmarks' generated tokens at today's ratio of about 15 to one.

Newer chips and better software contribute about equally. Chips alone raise output per gigawatt 1.8 times by 2027 and 2.3 times by 2028, and software doubles it, so the broad-agent figure is only 0.32 without the software gain. A gigawatt processes 2.6 times as many tokens under efficient routing as under frontier demand.

Tokens per gigawatt, 2026 to 2028E

Quadrillion tokens a day per GW serving customers · year-end fleets



45% load; software doubles output by 2028. Dotted: broad agents without that gain. 2027E mix halfway to each 2028 case. Current prompt-to-output ratio. Mostly estimated rates.

Why the mix matters

A large model uses much more power for each token. Every token passes through more parameters, and the weights sit on more GPUs that have to exchange data.

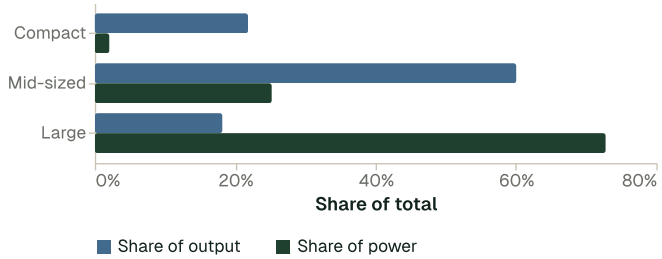
The benchmarks show how large the difference is. On GB300 racks running InferenceX's agent traces, one gigawatt generates about 55 times as many tokens with DeepSeek V4.1 Flash, the most used model on OpenRouter, as with Kimi K3.1 Pro, the successor to the 2.8T-parameter Kimi K3 and the most recent frontier open model benchmarked. DeepSeek V4 Pro, the large model on which we base our rates, lies between them, at a tenth of V4.1 Flash's output and 5.5 times Kimi's. Each result holds at least 80 output tokens/s/user at P90, though engines, caching and test dates differ.

The closed frontier models may be larger than DeepSeek V4 Pro, and the Kimi result shows that a larger model can need several times the power. We keep V4 Pro's measured rate because it comes close to what Anthropic reports spending on serving, and we test it from Kimi's speed to double.

A model's share of tokens can therefore be very different from its share of power. On the 2028 fleet, a token from the large model takes about 10 times the power of a token from the mid-sized model. In the broad-agent case, large models produce 18% of generated output but need 73% of inference power. If a little more demand moves to frontier models, much less power is left for everything else.

Output share is not power share

Broad agents · 2028 fleet · common 8K/1K serving recipe



Power share is each profile's output share divided by its fleet serving rate, rescaled to 100%. Rates are mostly estimated transfers.

How that compares with today

We check the method against what the two largest labs report about their own compute.

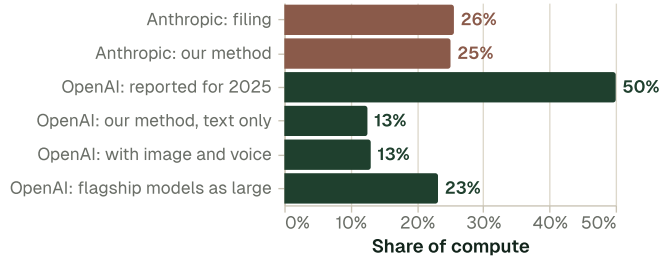
Anthropic's S-1/A allows the most direct comparison: in the first nine months of 2026, compute to serve customers cost 14.5% of revenue and made up 26% of its compute spending. Our method, applied to Anthropic's paid traffic and its own model mix at 45% load, needs about 0.7 GW, which is 25% of the roughly 2.8 GW we estimate it had at the end of September, from the reported figures given later in the note.

The two figures agree only because we used the filing to set one assumption: of three speeds we tried for large models, we kept the one that reproduced Anthropic's figure, and the two slower ones needed much more compute. The comparison is also inexact, because a share of spending is not a share of power and our inputs are estimates.

For OpenAI the two figures do not match. Epoch AI reports that about half of OpenAI's compute served users in 2025, across text, images, video and voice, while our method needs only 13% for this year's traffic, images and voice included. The main reason is model size: we class only 8% of OpenAI's tokens as large-model traffic, against 60% of Anthropic's, and a large model needs about 6 times the power per token today. With its flagship models classed as large the figure is 23%. Timing probably explains the rest, because Anthropic's own serving share fell from 38% to 26% between 2025 and this year. If the reported half is right, OpenAI's serving needs about 3.8 times the power we model, and the fleet shares in this note are too low.

Share of lab compute serving customers

2025–26 · reported against our method at 45% load



Anthropic: cost of revenue over all compute spending, nine months to September 2026. OpenAI: Epoch's roughly even split, all modalities. Method: share of lab compute.

What moves the result

The broad-agent case gives 0.65 quadrillion tokens a day per gigawatt in 2028. It rests on five assumptions we cannot observe, which the chart changes one at a time as a what-if test.

Server load. We assume servers deliver 45% of their benchmark output. The result is 0.29 at 20%, near the GPU utilisation of a replayed production serving trace, and 0.86 at 60%.

Speed of newer chips. Chips measured on this workload carry only 13%–30% of the 2028 fleet's power, while Rubin (21%) is measured on a different workload, 31% sits on chips not yet shipped and the rest is rated from other hardware. The result is 0.36 if our estimates prove too high and 1.49 if too low.

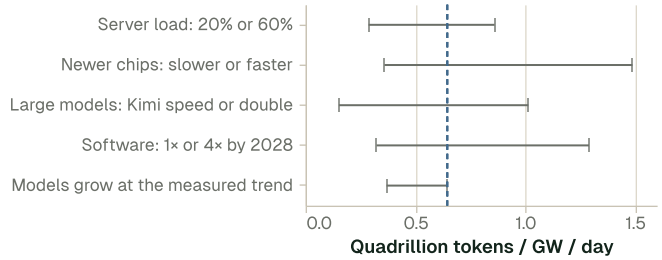
Speed of large models. The result is 0.37 at half the speed we assume and 1.01 at double. At Kimi K3.1 Pro's speed, 5.5 times below DeepSeek V4 Pro, it would be 0.15, below today's 0.21.

Software gain. We project a doubling of output by 2028, less than the 9.1% a month measured earlier in the note. The result is 0.32 with no further gain and 1.29 with a fourfold gain, about 16 months at the measured pace.

Larger models. We hold model sizes at today's levels. If serving cost rose with the trend in active parameters measured earlier, about 1.29 times a year, the result would fall to 0.37.

What moves the broad-agent result

Q tokens a day per GW if one assumption changes · line = central case



One change at a time; dashed line: central case (45% load, software doubling). Load 20%–60%; software none to fourfold; large models from Kimi K3.1 Pro's speed to double.

Where the power goes today

The next question is how much of the fleet serves customers, starting with the individual labs.

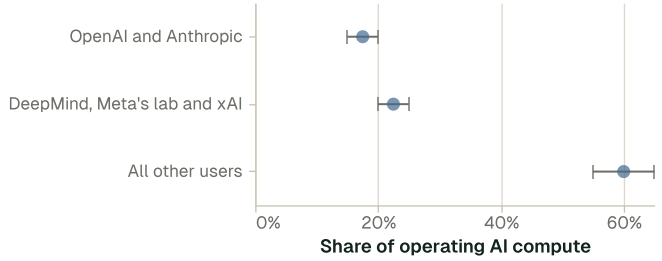
The only figures from company accounts come from Anthropic's S-1/A filing, in which serving paying customers took 30% of its compute spending in 2024, 38% in 2025 and 26% in the unaudited first nine months of 2026. SemiAnalysis estimates that OpenAI and Anthropic together use 34%–38% of their capacity for inference in 2026. Epoch AI reports that OpenAI split its 2025 compute roughly evenly between research and inference.

Most of the remaining lab compute does not go to training in the usual sense. By Epoch's estimate, final training runs took about a tenth of OpenAI's research compute in 2024, while experiments, synthetic data and other research took the rest.

The chart places the labs in the whole fleet, using Epoch's estimates for the end of 2025. OpenAI and Anthropic used 15%–20% of the world's operating AI compute, so serving their customers took only 6%–10% of it. Google DeepMind, Meta's lab and xAI used a further 20%–25% on our reading of Epoch's figures. The remaining 55%–65% serves cloud customers, smaller labs, open-weight models and non-language systems such as recommenders, and Epoch does not break it down further. These are rough ranges, and the two labs' share is rising fast, since Patel puts them at about 30% of new compute in 2026 and as much as 40–50% in 2027.

Who uses the world's AI compute, a rough estimate

Share of operating AI compute · end-2025



Ranges from or derived from Epoch AI; dots are midpoints. Rented capacity counts with the lab using it.

Why the split is a choice

Allocating GPUs to different tasks is increasingly an economic decision. Anthropic says it manages compute as "a single, fungible pool" and directs it to "where the marginal unit is expected to be most productive".

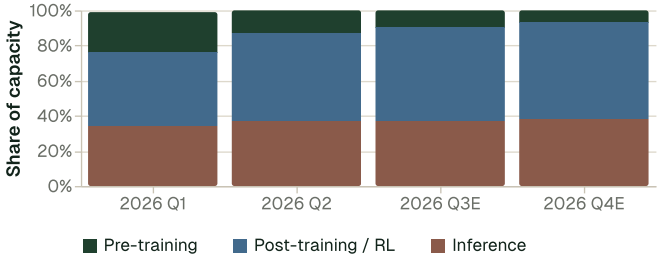
A gigawatt therefore has two competing uses, because serving earns revenue now while research buys a better model that earns revenue later. Anthropic says customer demand has consistently exceeded its available compute, so power given to one use is taken from the other. The narrative around recursive self-improvement and the race between the labs further complicates this choice.

At Anthropic, serving rose to 38% of its compute spending in 2025 and fell to 26% in 2026, while revenue grew 13-fold. The chart shows SemiAnalysis's estimate for both labs.

That fall can be read in two ways. Dylan Patel sees a rising share of each new megawatt going to research, and puts a frontier lab today at 40% inference, 50% research and 10% development. Anthropic itself points to timing, because capacity arrives in large steps before the revenue it supports, and in the third quarter about half of its compute bill was short-term capacity bought at a premium.

A third of lab compute runs inference

OpenAI + Anthropic; SemiAnalysis capacity shares



13 September estimates; E = forecast. Q1 rounds to 99%. Lab compute-capacity shares, not whole-fleet IT-energy shares. The source ends in 2026.

Today's use against next year's supply

Outside China, AI services process about 0.6 quadrillion tokens a day for their users. By our method that takes 8%–11% of today's fleet.

Next year the installed fleet rises from 29 GW to 66 GW, and each gigawatt serving customers processes 0.44 quadrillion tokens a day instead of 0.21, thanks to newer chips and better software. Together these raise capacity 4.8 times, so the same share of the fleet could process 2.7 quadrillion tokens a day. Capacity grows only 3.4 times without the software gain we project, and 3.7 times if models keep growing as measured earlier.

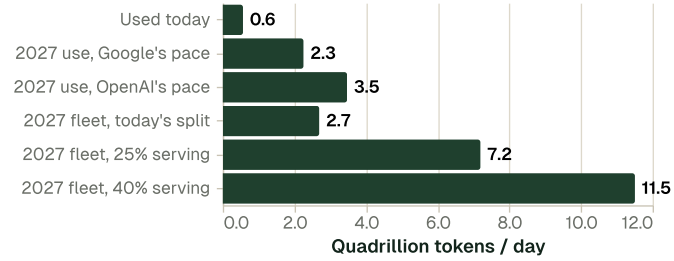
Total token use is hard to measure, so we take the growth that Google and OpenAI report for their own APIs as a guide. Google reports that its API volume more than quadrupled over the past year, and OpenAI's rose 2.5 times in the six months to April, a rate of about 6.2 times a year. At Google's rate total use rises from 0.6 to 2.3 quadrillion tokens a day, and at OpenAI's rate it reaches 3.5. The 2.7 that next year's fleet could process at today's split lies between the two, so supply and use stay roughly in balance. Token counts overstate the growth in compute, because cached prompts cost far less to process and rose from about 60% to just under 90% of agent tokens on OpenRouter this year.

Far more could be served if the split itself changed, since a quarter of the end-2027 fleet serving text only could process 7.2 quadrillion tokens a day, about 13 times today's use.

That power would have to come from research or from other owners. On our estimates, OpenAI and Anthropic hold 27%–31% of the non-China fleet at the end of 2026 (Patel's share of new compute at the low end, the labs' targets at the high end), and 34%–42% a year later if they take 40–50% of new compute, the upper end of what SemiAnalysis expects.

Today's use against next year's supply

Quadrillion tokens a day served to users · outside China



Supply: installed end-2027 fleet × serving share × our tokens per GW. Use: today grown one year at API disclosure paces. Today's prompt-to-output ratio.

How demand and the mix move the split

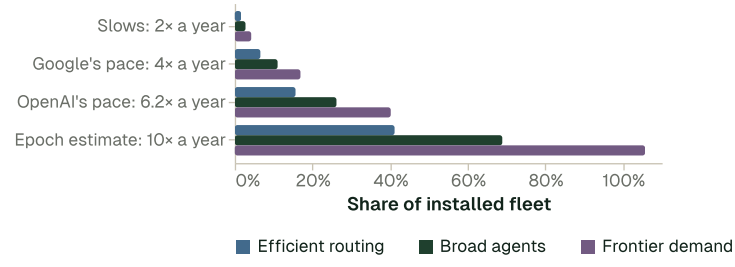
How much of the 2028 fleet text serving needs depends first on how fast demand grows over the next two years, and we consider four paces. Use could slow to a doubling each year, hold the more than fourfold a year that Google reported for its API, hold OpenAI's 6.2 times, or reach the roughly tenfold a year that Epoch AI estimates for demand at current prices. With the broad-agent mix, these paces need 2%–3%, 10%–12%, 23%–30% and 60%–78% of the 127 GW fleet.

The model mix moves each figure, because the same tokens need 2.6 times as much power under frontier demand as under efficient routing. At the fastest pace, frontier demand would need 92%–119% of the fleet and leave nothing for other work.

Demand and research therefore compete for the same power, so the faster demand grows and the larger the models users choose, the less compute is left for research and other workloads, while slower demand frees power for them. We consider faster growth more likely than slower, because Google and Anthropic both reached their recent growth while reporting that demand exceeded their capacity. The paces rest on a few company disclosures, and the slow case is our own assumption.

Faster demand leaves less of the fleet for research

Share of the 2028 fleet needed for text · by pace of demand and model mix



Today's use grown two years at each pace, at our tokens per gigawatt; midpoints. Slow pace is our assumption; 10× is Epoch AI's estimate of demand at current prices.

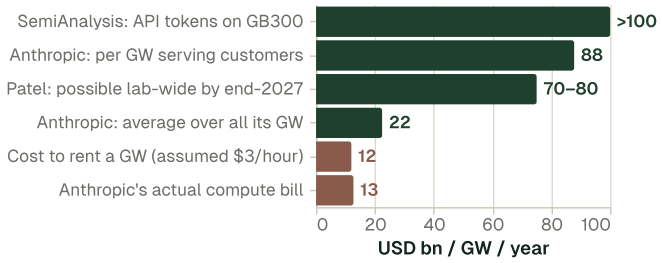
What a gigawatt earns

Two layers earn money from a gigawatt: a compute provider rents out the infrastructure, and a model company sells the work its models do on it. SemiAnalysis's August scenario for a GB300 cluster puts API revenue above $100B per gigawatt a year against about $12B of rent at $3 per GPU-hour, and Patel considers $70–80B per gigawatt across all of a lab's compute possible by the end of 2027 if it keeps the best model.

Anthropic's filing lets us check these scenarios against a real company, whose revenue in the first nine months of 2026 averaged about $44B a year. We estimate that it had about 2.0 GW on average over those months, between a reported 1.4 GW at end-2025 and a target of 3.5 GW for end-2026, and its filed serving share puts about 0.5 GW of that on serving customers. That gives roughly $88B per gigawatt serving customers, below SemiAnalysis's scenario. Spread over all its compute, the same revenue comes to only $22B per gigawatt, well below Patel's figure, because most of its compute does not serve paying customers. Its compute bill came to about $13B per gigawatt, which is close to the rent SemiAnalysis assumes.

Revenue and compute cost per gigawatt

USD billion per gigawatt a year · revenue in green, compute cost in brown



Anthropic: nine months to September 2026, annualised, over our estimates of its serving and total GW. Scenarios: SemiAnalysis and Patel, August 2026. Not profit.

What a gigawatt has to earn

The owner of a gigawatt has to recover what it cost to build, which Epoch AI puts at about $38B for Nvidia GB200 systems and BCG at about $42B. Annualised over the life of the equipment at an industry cost of capital, and including power and upkeep, Epoch arrives at about $8.5B a year. BCG's model, which also covers debt and a return for equity investors, puts the earnings needed at $15B.

The owner receives the rent and not the lab's revenue, and the $12–13B that labs pay covers Epoch's figure but not BCG's. On Epoch's costs, with chips and network written off over five years and the building over fourteen, that rent leaves the owner $5.1–5.8B a year after depreciation and running costs if all capacity is rented. That is an operating margin of 42%–45% and a return of 13%–15% a year, before tax and interest, on the $38B invested.

Anthropic's $22B per gigawatt is above both figures although only a quarter of its compute spending serves customers, and it would equal the cost of ownership if 10%–17% of a gigawatt served paying customers at 2026 prices.

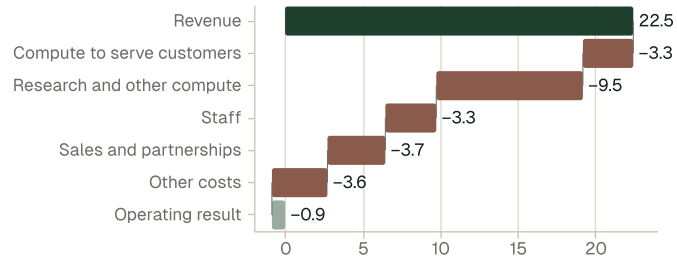
Where a lab's revenue goes

The revenue is not profit for the lab that rents the gigawatt, because staff, sales and financing have to be paid as well as compute. Compute takes $12.7B of the $22.5B that each gigawatt earns, of which $9.5B does not serve paying customers. Staff, sales and other costs take a further $10.6B, and the company ends slightly below break-even.

These figures hold prices at 2026 levels, and since tokens per gigawatt roughly double next year in our estimate, revenue per gigawatt rises only if the price per token falls by less than half.

Where the revenue of an Anthropic gigawatt goes

USD billion per gigawatt of all its compute · 2026, annualised



S-1/A operating expenses, annualised, over our estimate of Anthropic's total gigawatts. Serving compute is approximated by cost of revenue. Operating result before interest.

Conclusion

The buildout earns a return on the compute that serves customers, and the overall return depends more on demand and on how compute is split than on the supply of chips. Serving today's users takes about a tenth of the fleet. If serving keeps that share, capacity for text grows 4.8 times next year while disclosed use grows 4.0–6.2 times, so supply and use stay roughly in balance only as long as demand keeps its pace and serving software keeps improving.

A gigawatt that serves Anthropic's customers earns several times what it costs to own, but most of the labs' compute does not serve customers, and Anthropic as a whole runs slightly below break-even. The rest has to pay through research that yields better models, through workloads we do not count, or through faster growth in demand.

The model mix is among the larger effects, since a shift to the largest models needs 2.6 times the power for the same tokens and takes it from research.

These results rest on figures we cannot observe or hold fixed: fleet utilisation and future prices are not disclosed, model sizes and request shape stay at today's levels, and our large-model rate is fitted to one filing.

Sources

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 for discussion purposes and not investment advice. Nothing here is a recommendation, an offer or a solicitation to buy or sell any security, and it takes no account of any reader's circumstances. Sources are believed reliable but are not guaranteed, and views are as at the date shown and may change without notice. Scenario estimates are not forecasts unless explicitly identified as such.