EXPONENTIAL VIEW

The State of the *AI* Economy

June 25, 2026

Azeem Azhar, William Gildea, Hannah Petrovic, PhD, Nathan Warren & Marija Gavrilov

Exponential View
www.exponentialview.co

Independently produced by Exponential View. Built from public disclosures and Exponential View's own models; all conclusions are our own.

Why we've done this

There is a visibility problem in the AI economy. Until now, it has been impossible to deconstruct real customer demand.

The supply side of the AI economy is well-documented. Most semiconductor companies and hyperscalers are public and disclose their activities in some detail. Sell-side analysts have done a great job decomposing their performance.

The demand side, **what customers are actually *paying* for and if the revenues are *real*,** has been obscure.
The largest labs are private, and even public companies bury AI revenue inside segment totals.

Without understanding genuine demand, it is impossible to judge the health of the AI economy that underpins $22.7 trillion of stock market valuation and has driven US GDP growth in the past six quarters.

We hope that this report serves as a reference source on the current state of play, free of hype and fear, while helping us all have a more informed conversation about the gravitational pull AI is exerting on the economy and the world at large.

Special thanks to those who kindly reviewed an early draft of this presentation and gave us feedback:
Alex Imas, Shanu Mathew, Patrick Rutherford, Jaime Sevilla and Amy Sutter.

– Azeem and the Exponential View team

A proprietary line-level revenue model: Sourced, scored, triangulated, deduplicated

1 Source

Bottom-up, 1,000+ firms

Every revenue line traced to primary filings, audited accounts, transcripts and credible reporting; plus cloud-attribution where a private firm's revenue surfaces in a public firm's accounts (OpenAI via Azure, Anthropic via Bedrock).

Additional **soft signals we use include unofficial sources** such as:

- Public comments by executives and related parties.
- Proxy and sample metrics.
- Commentary and unverified estimates and leaks in traditional, new and social media.

We flag, investigate and maintain our datasets using **analyst research, augmented** by a proprietary system that scans, crawls and synthesizes insights.

2 Confidence score

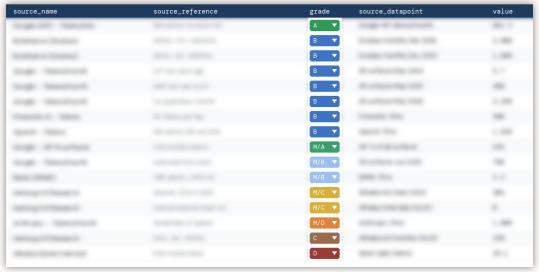
Confidence-scored before it counts

Each line carries a rigorous **confidence score** before it enters any model, so weak inputs can't inflate the number.

We **grade filed figures highest**, above other primary sources, corroborated 3rd-party estimates and single-sourced claims.

All derived numbers **inherit the lowest grading** from input sources.

Audit trail: Sample source table



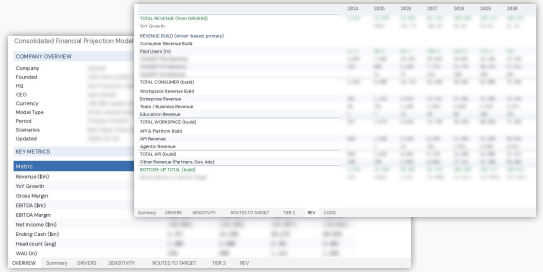
3 Model and triangulate

Company financial models checked against top down

We build **full per-company models** specifically for GenAI financials (split out from top-line reporting), covering key drivers of revenue, profitability and cost in the **P&L, cash flows and balance sheets**.

These models are reconciled against **independent proxies**: silicon (chip-maker revenue), build cost, segment mix, industry research, traffic and capacity.

Audit trail: Sample company revenue model

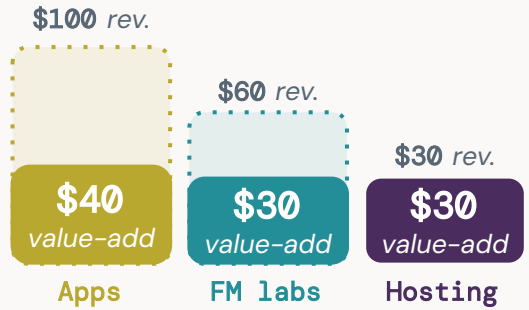


4 Deduplicate

Spend only counted once

Revenue is counted at every layer but never summed across them: attributed by value-add so the same dollar isn't double- or triple-counted.

e.g. **$100 app spend** that sends $60 to a model provider, which spends $30 on inference hosting, is counted as **$100, not $190** :



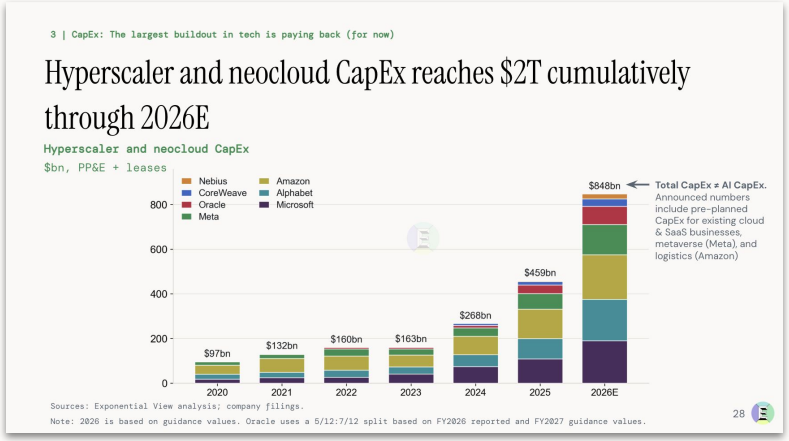
Scope: Global ex-China · App, model & infrastructure revenue counted · Excludes chips, AI ad-uplift, legacy-software features and financing.

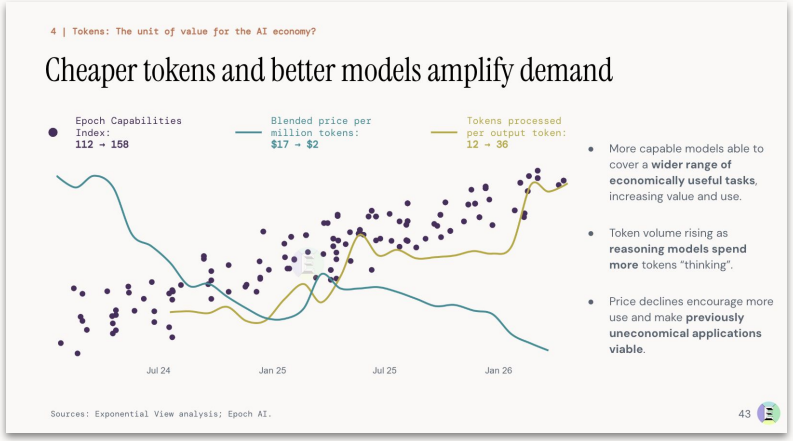
The top line

AI demand is more clearly validated by realized revenue than previous platform shifts. Generative AI ecosystem revenue has already surpassed $175 billion annualized (after removing double-counting from provider revenues).

CapEx intensity is growing well above historical large-cap technology norms to deliver the AI buildout. And third-party financing is increasingly entering the financing mix.

The open question is whether cheapening artificial intelligence can create enough volume and margin to service the buildout.





Contents

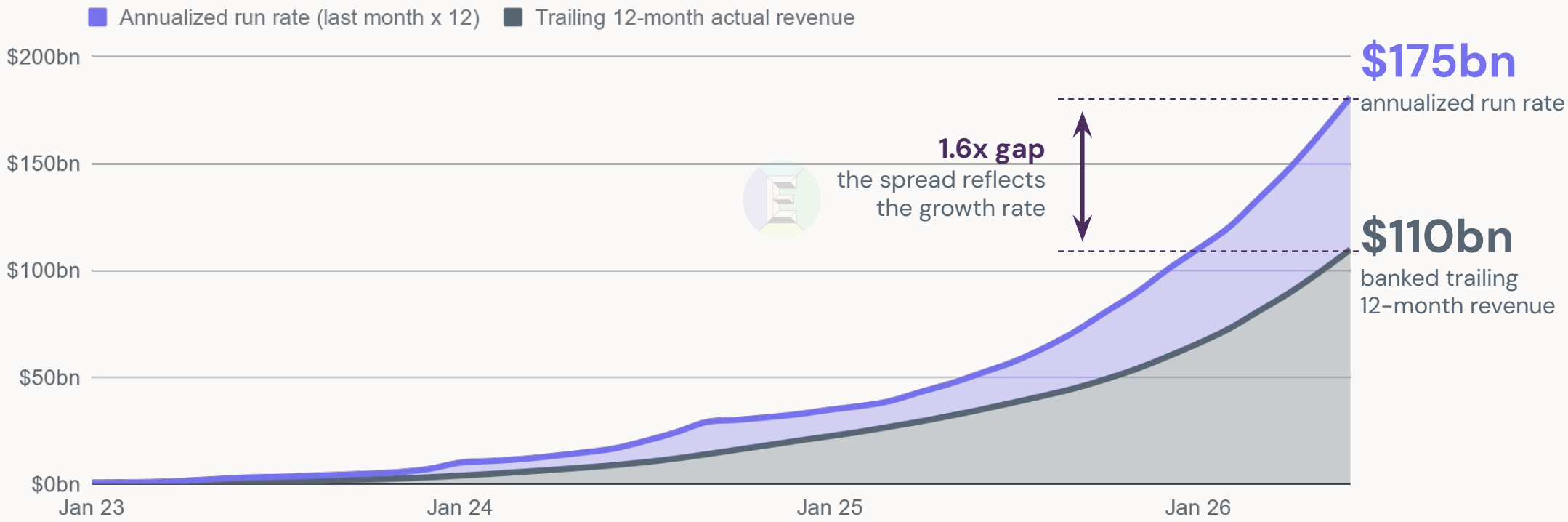
1 | Demand :
It’s real, big & fast

Revenues are driven by real external customers.
The sector is growing 3x faster than any IT wave before it.
This demand has created a compute supercycle: 10x more compute, new energy generation, larger data centers and mounting backlogs where supply cannot keep up.

$110bn trailing 12-month revenues – now at a $175bn pace

Generative AI economy revenue, deduplicated

$bn/year, Jan 2023 - Jun 2026



Source: Exponential View analysis.
Note: Global ex-China. Deduplicated app, foundation model, and infrastructure hosting revenues. Excludes chip manufacturing.

Real, external demand drives AI revenues

Illustration of how we model and deduplicate revenues between providers

Customer

pays app license/fees, or buys tokens directly

Revenues flow down the stack

App layer

Model lab's apps

e.g. Claude Code, Codex

Value add: Revenue minus token costs

Standalone GenAI apps

e.g. Cursor, OpenRouter, Harvey

Value add: Revenue minus token costs

Foundation model layer

Closed-weight models

e.g. Opus 4.8, GPT-5.5

Value add: Revenue minus inference costs

Open-weight models

e.g. Deepseek-v4, MiniMax M3

Value add: $0 (revenue counted in hosting)

Hosting layer

e.g. Azure, CoreWeave, Nebius

$: AI revenues (full)

Not counted:

Non-AI-native apps *(Counted via token spend)*

Chip sales *(CapEx from hosting layer)*

Ad uplift *(Google/Meta AI ad revenue)*

We source, triangulate, model & audit to verify & deduplicate:

Sourced from official filings, 1st-party disclosures, leaks, government stats, 3rd-party analysts, and proxy metrics; all sources quality-graded.

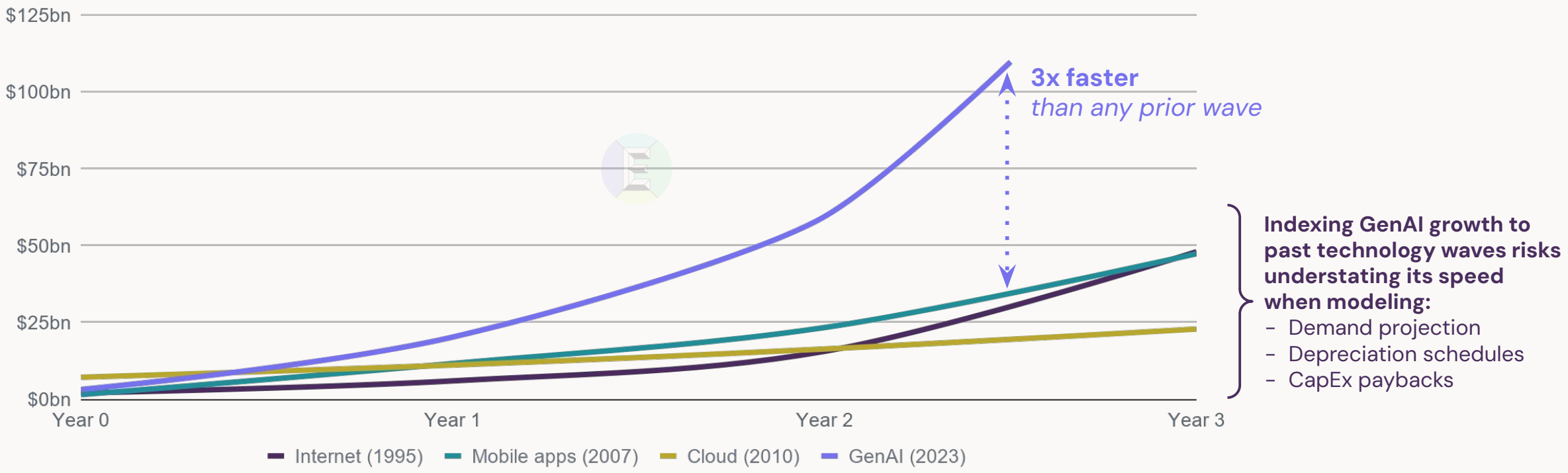
Rigorous company-by-company financial modeling to build a bottom-up deduplicated revenue model.

Continuous scanning and crawling across hundreds of sources to maintain and adjust the dataset.

AI is scaling three times faster than any IT wave

Realized revenue trajectory time-aligned to year zero

$bn/year, adjusted for inflation

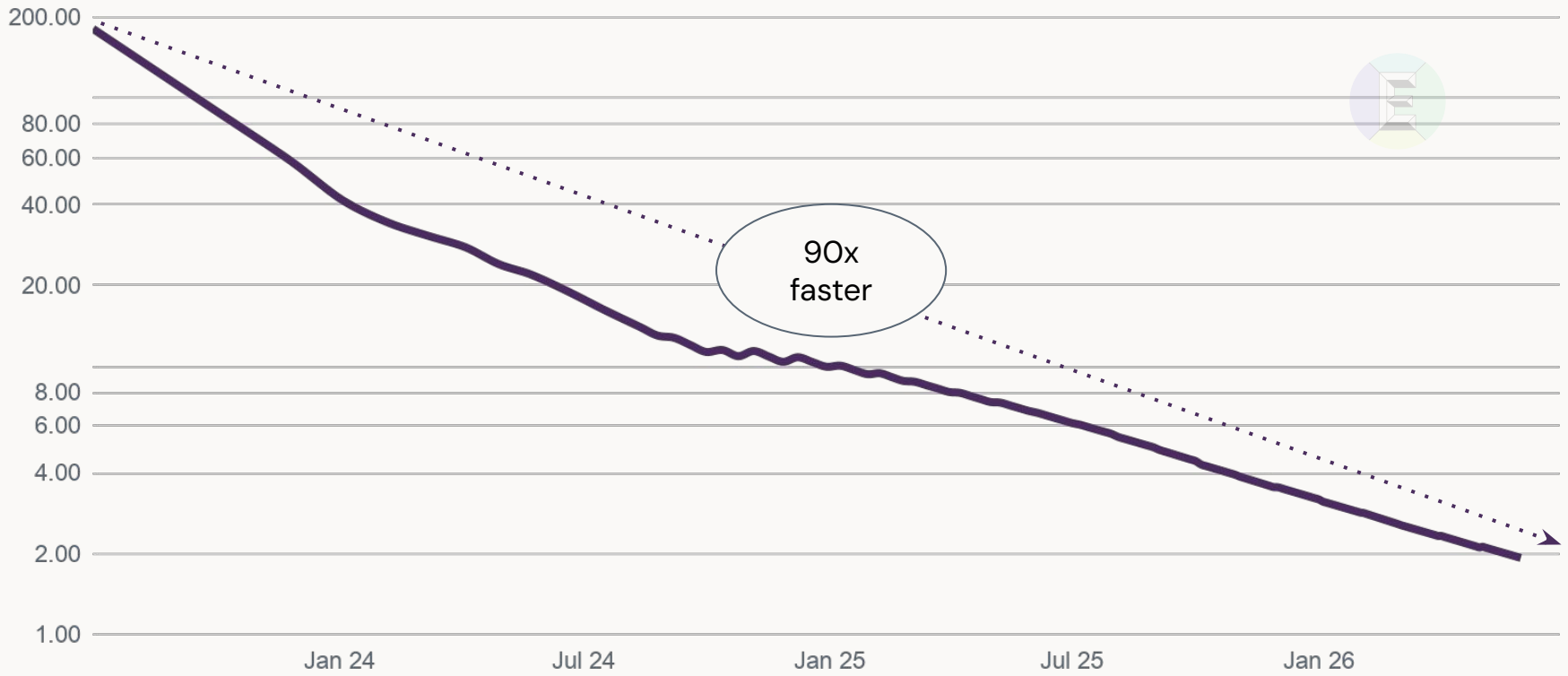


Sources: Exponential View analysis; US Commerce; company filings; UBS; US Bureau of Labor Statistics.
Note: We use the first full year of revenues, so have started GenAI measurement from January 2023.

Each new $1 billion of revenue arrives faster than the last

Time to add $1bn additional cumulative revenue

days, log scale



In 2023, the AI industry needed 180 days to add $1 billion in cumulative revenue.

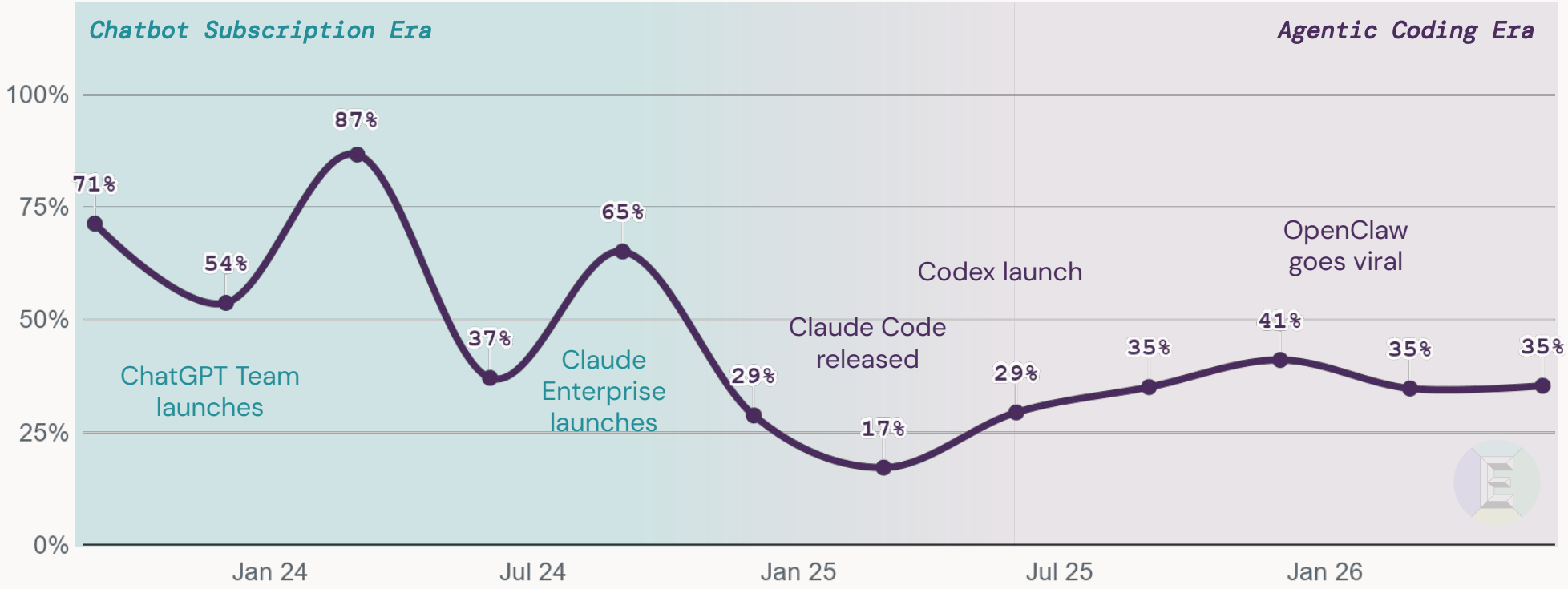
It now needs less than two days.

Source: Exponential View analysis.
Note: Global ex-China. Deduplicated app, foundation model, and infrastructure hosting revenues. Excludes chip manufacturing.

Growth has held across each adoption phase

Revenue growth quarter-on-quarter

% change since prior quarter

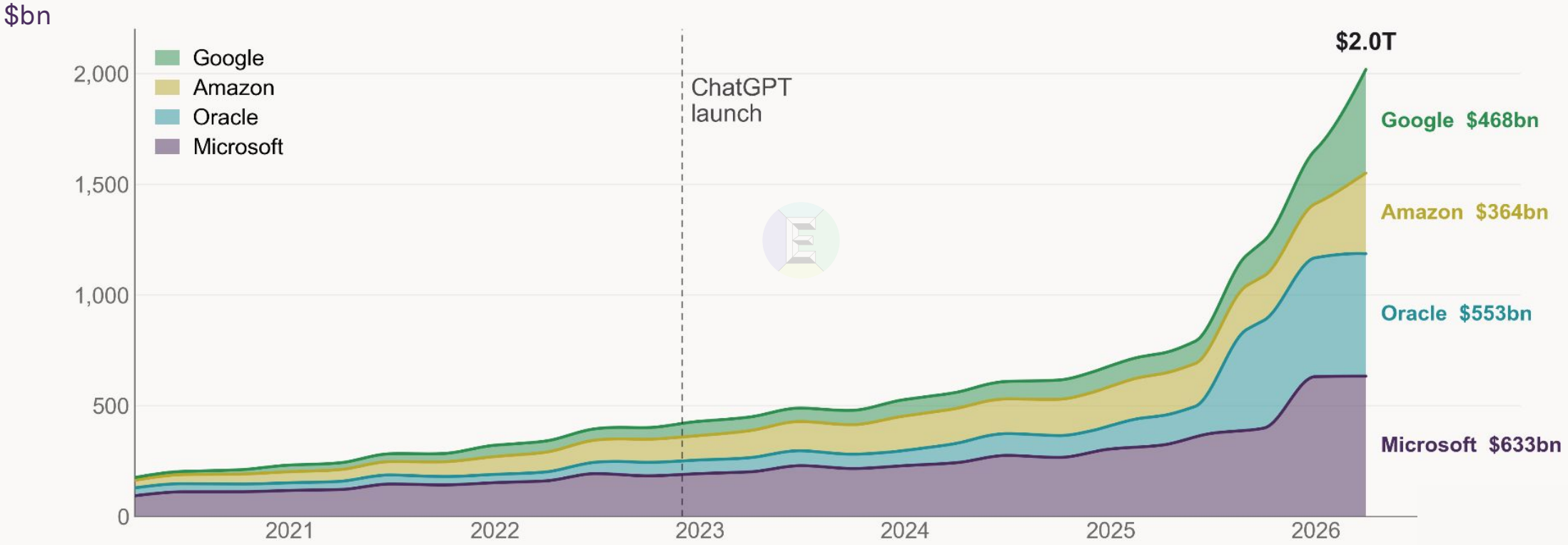


35% QoQ
equivalent to
3.2x annually

Source: Exponential View analysis.

This rapid demand growth is showing as a contract backlog for hyperscalers

Combined hyperscaler backlog (remaining performance obligations)



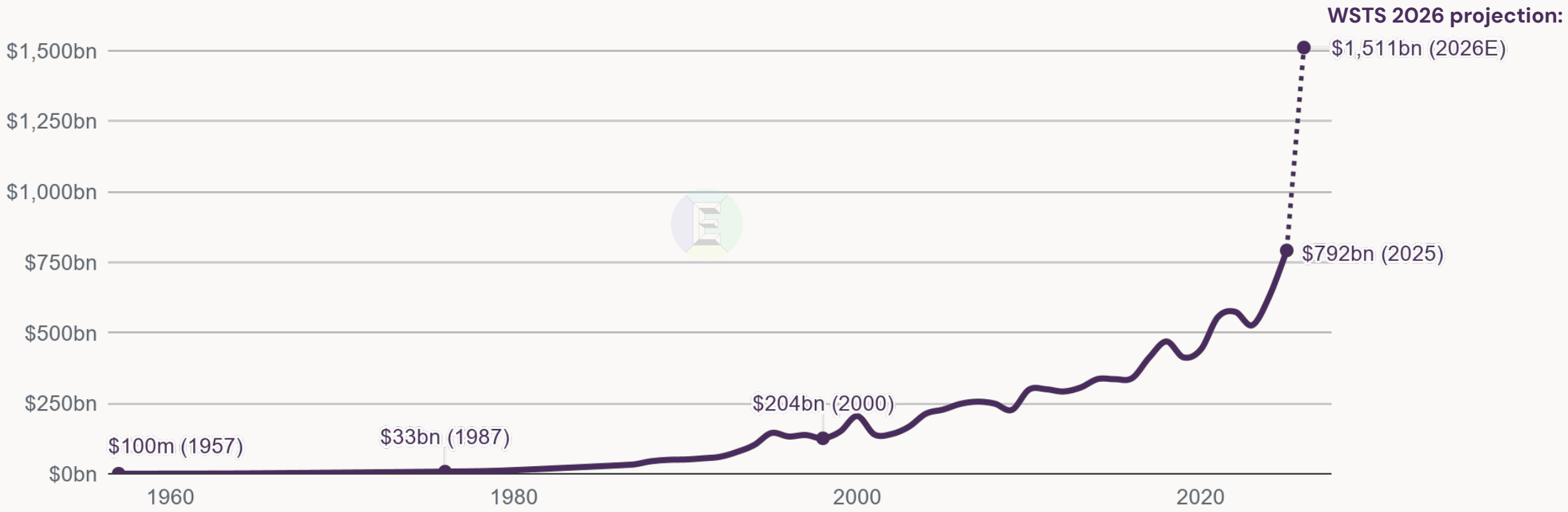
Sources: Exponential View analysis; company filings.

Note: Microsoft = total RPO (incl. M365/Dynamics); Amazon = total company RPO (mostly AWS); Google = revenue backlog (mostly Cloud). Oracle quarter ends one month earlier.

Demand has launched a compute supercycle

Global semiconductor market revenues

$bn/year

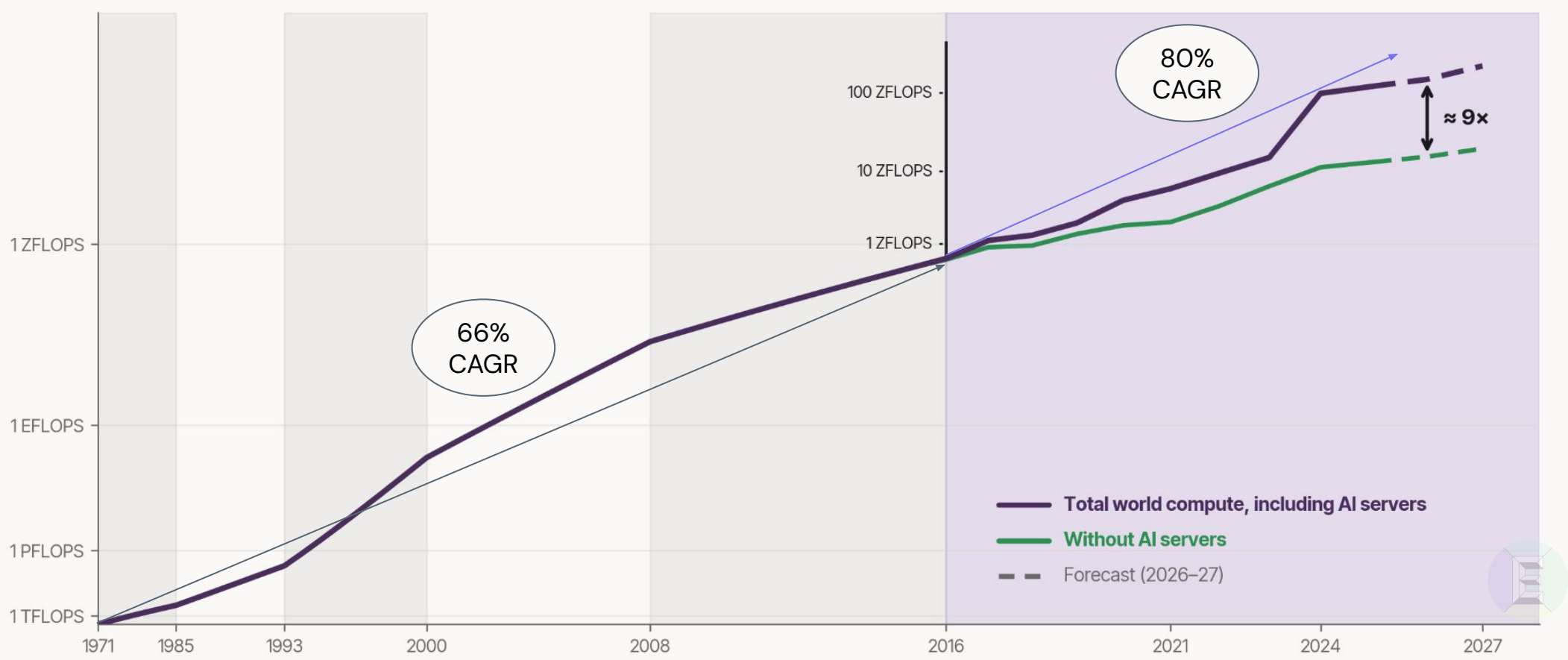


Sources: Exponential View analysis; World Semiconductor Trade Statistics; Japanese Semiconductor History Museum of Japan.

Note: Nominal $US

AI has spurred an uptick in a 50-year trend of compute growth

Global compute since 1971



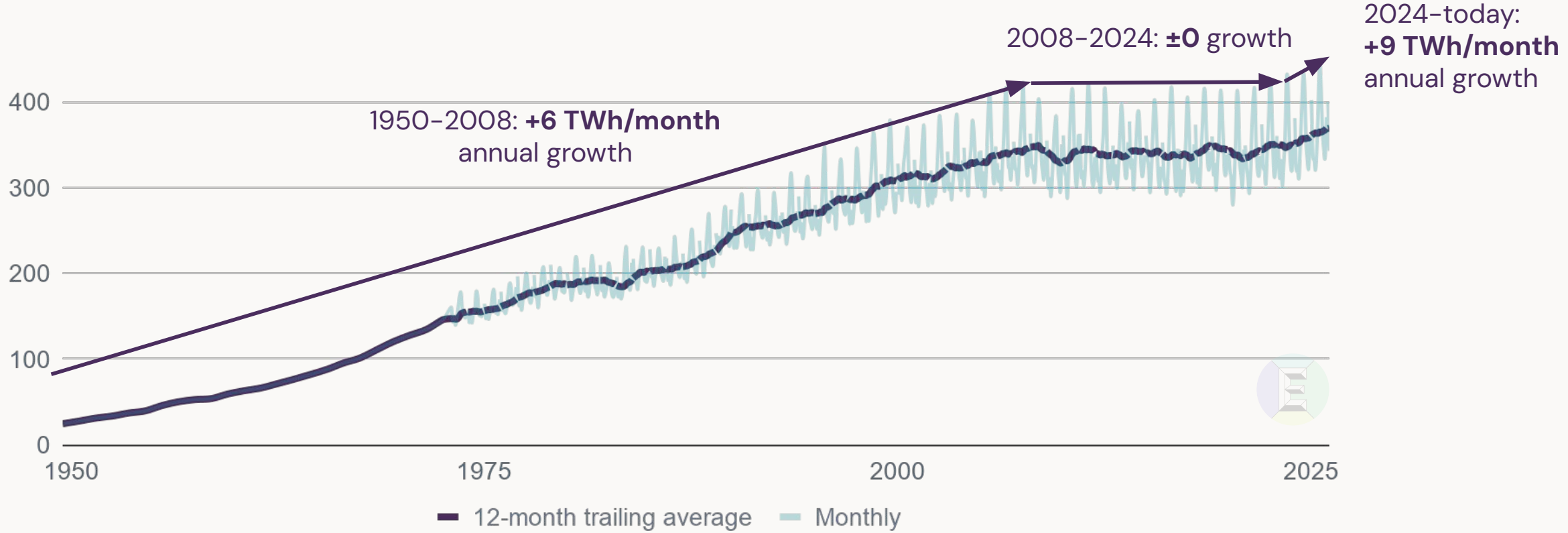
Source: Exponential View analysis.

Note: Includes mainframes, minicomputers, PCs, servers, smartphones, IoT and AI compute. AI-server FLOPS derived from the installed base of Nvidia GPUs by generation, using FP8 from Hopper onward.

AI demand is reigniting a moribund US power sector

US electricity net generation

TWh/month

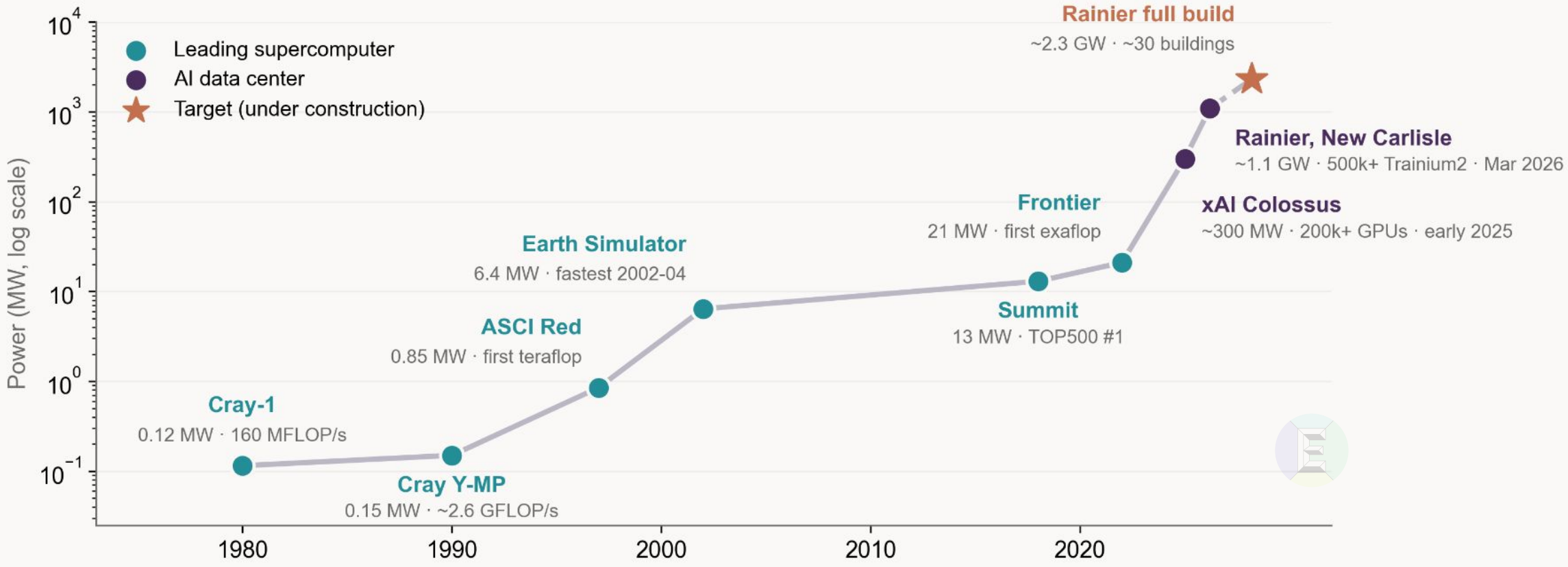


Sources: Exponential View analysis; US Energy Information Administration.

The size of the largest data centers has grown 50x in four years

Power of the most powerful computers over time

MW, log scale

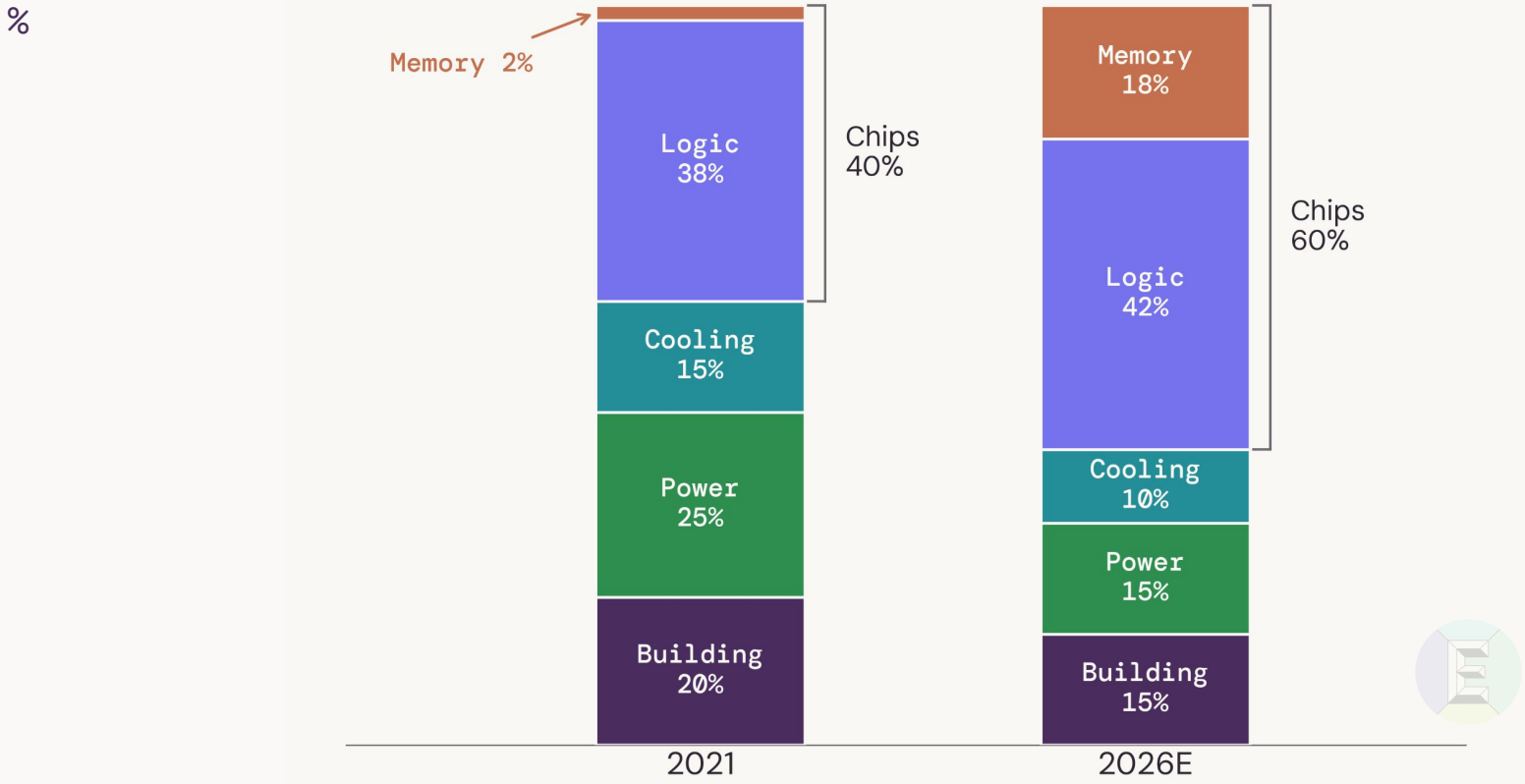


Sources: Exponential View analysis; TOP500 / Green500; Epoch AI; OpenAI; Oracle.

Note: Rainier full build is a target.

Memory and compute now take a majority of every dollar spent on the data center buildout

Share of total data center build cost by component, 2021 vs 2026E



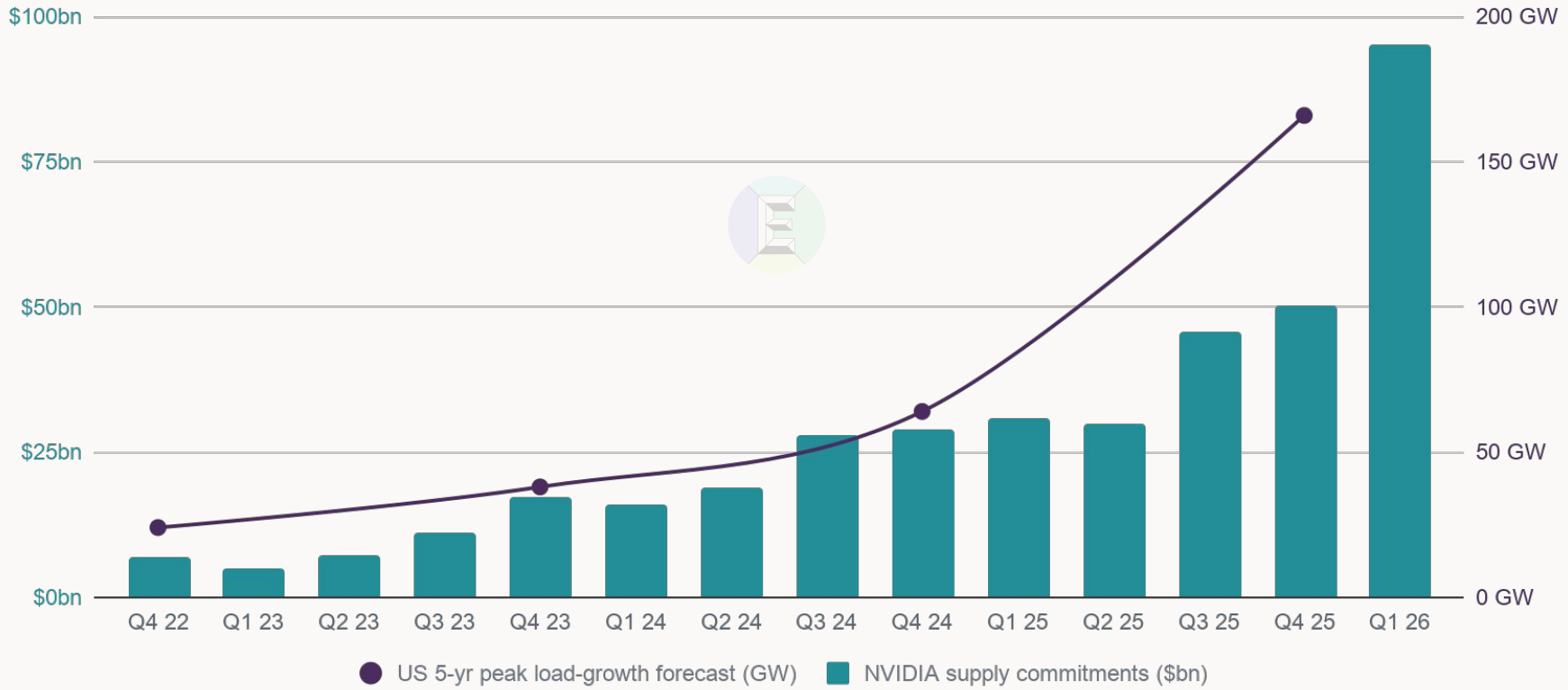
- Each new dollar buys **more silicon and less concrete**: chips' share of spend is up 50% (40%→60%).
- **Memory** is the single biggest mover, from a 2% rounding error to ~18%.

Sources: Exponential View analysis; Goldman Sachs; Epoch AI; Semi Analysis.
Note: Figures may not sum to 100% because of rounding.

Leading to growing commitments for compute and energy

Compute & power commitments

Nvidia supply commitments ($bn, left axis) & US load-growth (GW, right axis)



- Nvidia supply commitments have grown from **$31bn to $95bn** in the last year.
- The extra electricity the US grid is expected to need by 2030 has **grown ~7x since 2022** (24GW→166GW), with data centers accounting for ~55% of this growth.

Sources: Exponential View analysis; Grid Strategies; Nvidia filings.

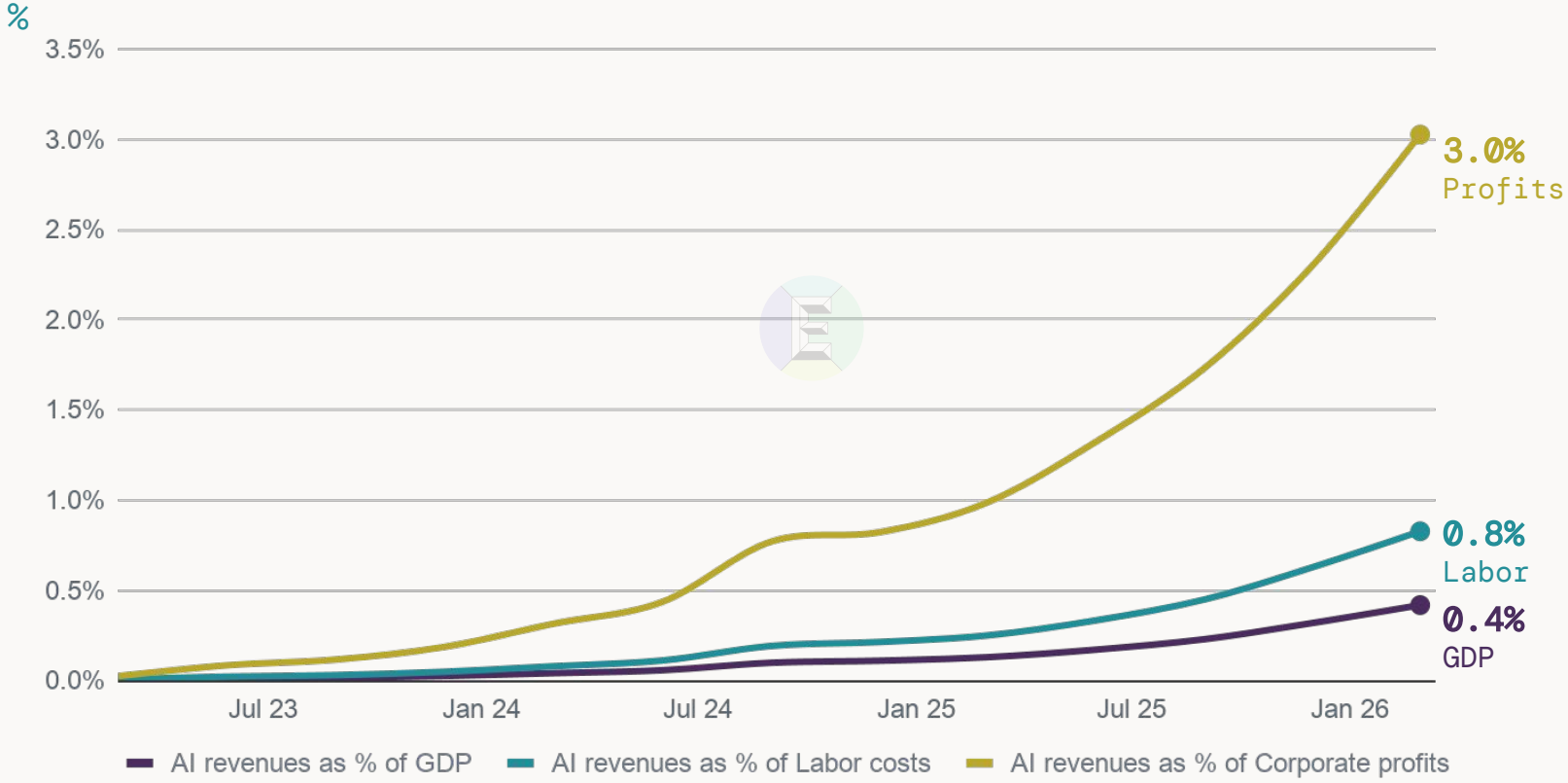
2 | Economy :

Big is still small, and early

Even for the highest corporate spenders, AI is a rounding error in the P&L. It still looks early. Initiatives have focused on efficiency & cost savings, although the mix is changing. And measured revenue may understate the social gains, as consumers report benefits that don’t yet show up in the data.

Against GDP, AI revenue is still a rounding error

Global AI revenues (ex. China), relative to US GDP, labor costs & corporate profits

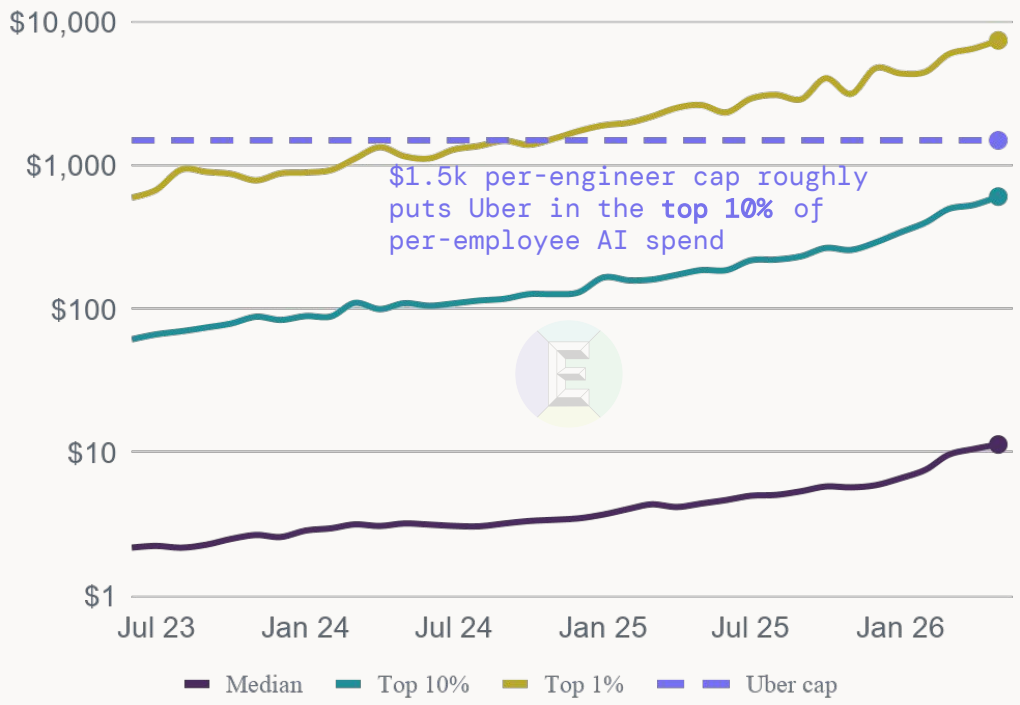


- **Still tiny:** AI revenue is equivalent to 0.42% of US GDP (vs IT sector's 9.4%).
- Even a generous yardstick (corporate profits) is **32x larger** than all GenAI revenues.
- **Still early:** AI revenue relative to GDP has risen **3x vs Q1 2025** (0.13%), and **10x vs Q1 2024** (0.04%).

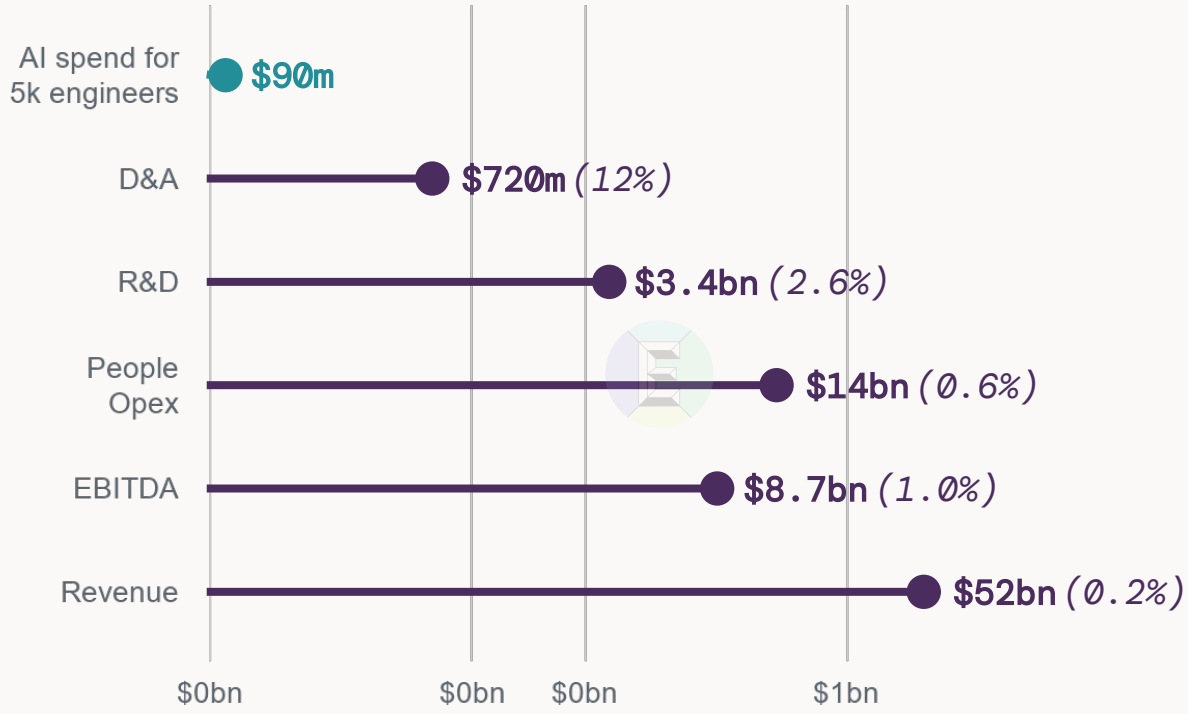
Sources: Exponential View analysis; St. Louis Fed.

At a company level, AI spending is still relatively small: e.g. Uber's $1.5k per engineer barely dents the P&L

AI spend per employee
$/month, log scale, Ramp customers vs Uber cap



Uber AI spend (maxed cap) vs P&L line items
$/year (AI spend %), log scale, vs FY2025



Sources: Exponential View analysis; Ramp Economics Lab (n = 70,000 US businesses), Uber filings.
Note: Top 1% / top 10% / median defined by level of AI spend compared across Ramp's customer base. Uber figure is a max per-engineer cap, benchmarked against Ramp per-employee AI spend.

Like previous general-purpose technologies, some gains may escape GDP measurement

AI impacts

Consumer surplus

Value that reaches people directly at a near-zero price. Little is sold, so GDP under-represents consumer benefit, e.g.:

- Free replacement of software & service purchases
- Learning, leisure & convenience

Producer surplus

Value embedded in sold goods and services. More is transacted and **recorded in GDP**, e.g:

- AI-enabled features that drive revenue
- Faster (valuable) releases
- Service firm margins

Historic cases

1880-1920: Electric lighting

≈ $0 direct GDP impact

Light became **~99.97% cheaper**: an hour's wage buys **~40,000x more**. Prices didn't record this gain.

Nordhaus (1996)

2000-2020: Free digital goods

≈ $0 direct GDP impact

Free search, encyclopaedias and maps displaced paid services. Search alone is worth **~$17.5k/yr/person**.

Brynjolfsson et al. (2019)

1850-1870: Steam

+0.4 pp/yr GDP

Mechanized factories and railways, so one worker could produce and move far more to market.

Crafts (2004)

1980-2000: Automation

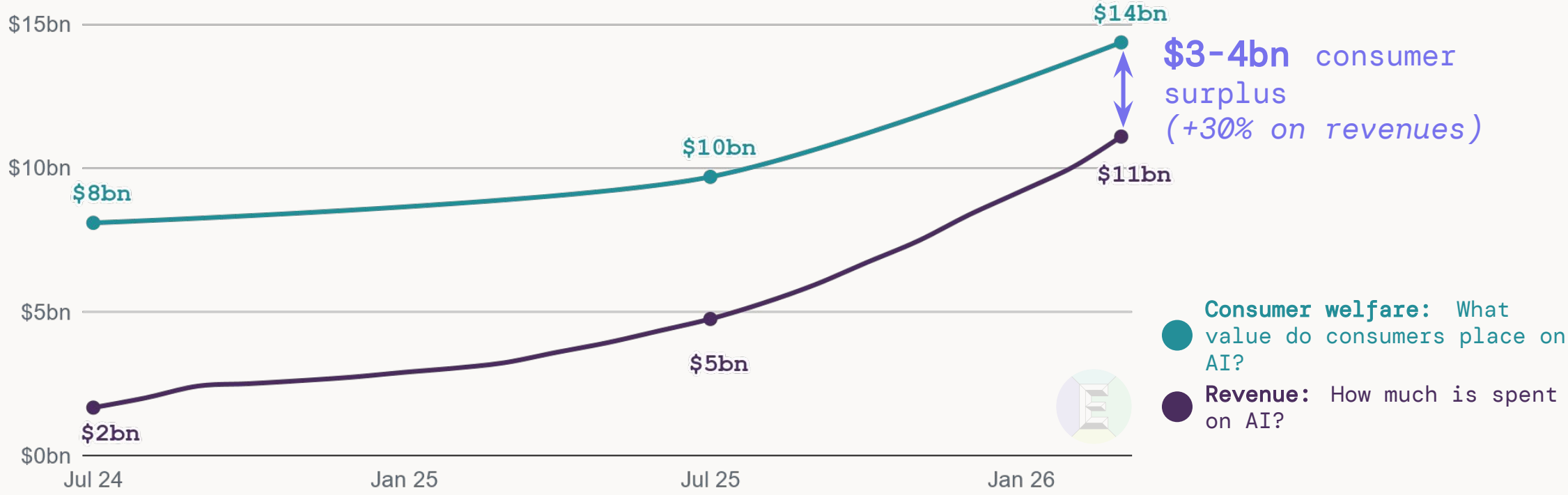
+0.37 pp/yr GDP

Programmable machine tools cut labor in every unit, raising output per worker.

Graetz & Michaels (2018)

GDP knows the price of everything but the value of nothing. AI's economic value exceeds measured revenue

Monthly GenAI revenues vs US consumer welfare
$bn/month, Jul 2024 - Mar 2026

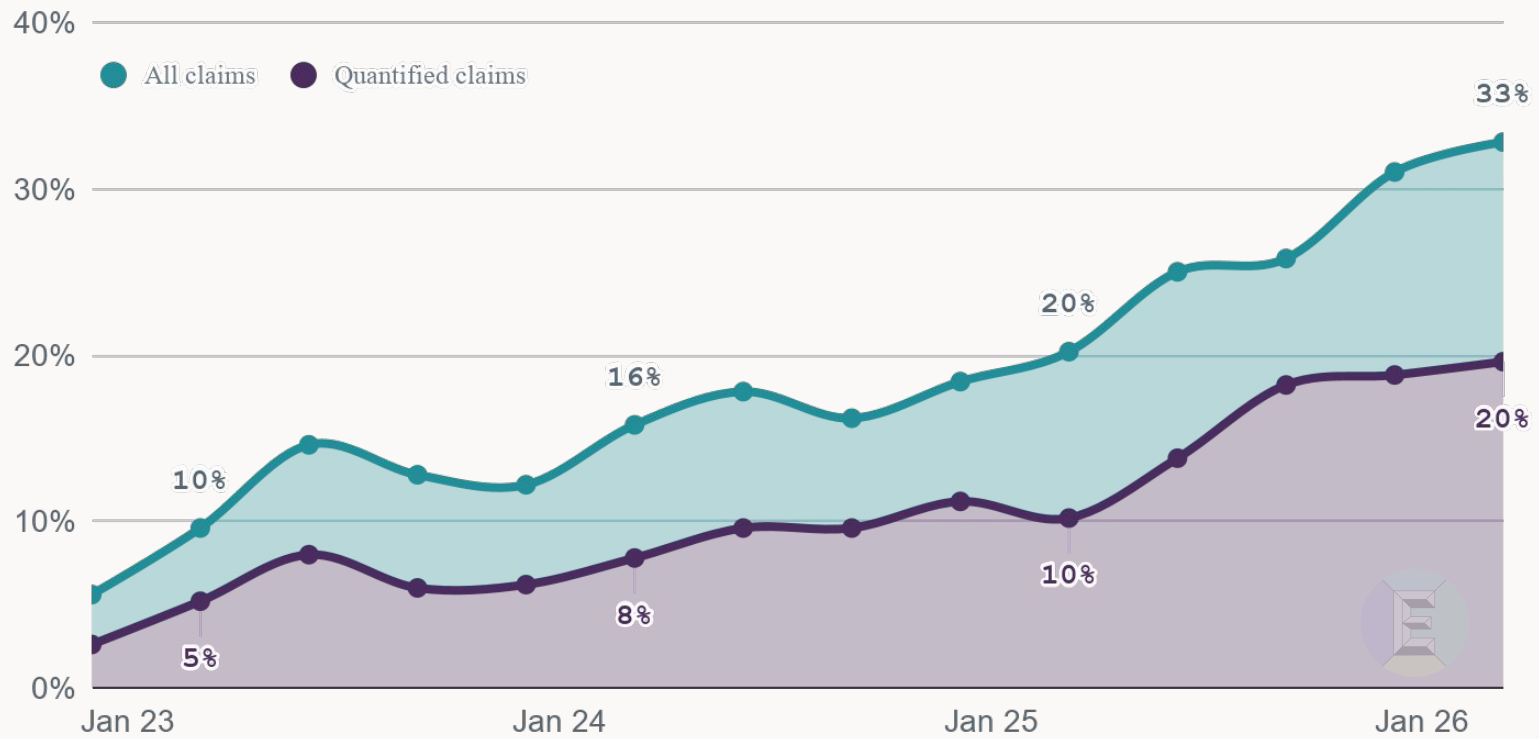


Sources: Exponential View analysis; Stanford Digital Economy Lab

Note: Welfare value determined from responses to "Would you give up access to all AI tools like ChatGPT, Gemini, Claude, or Copilot for one month starting tomorrow morning in exchange for [$US]?". Revenue includes global (ex-China) consumer and enterprise spend.

Public companies are reporting increased impact of GenAI

Companies making claims of AI impact on earnings calls
S&P 500, Q4 2022 - Q1 2026

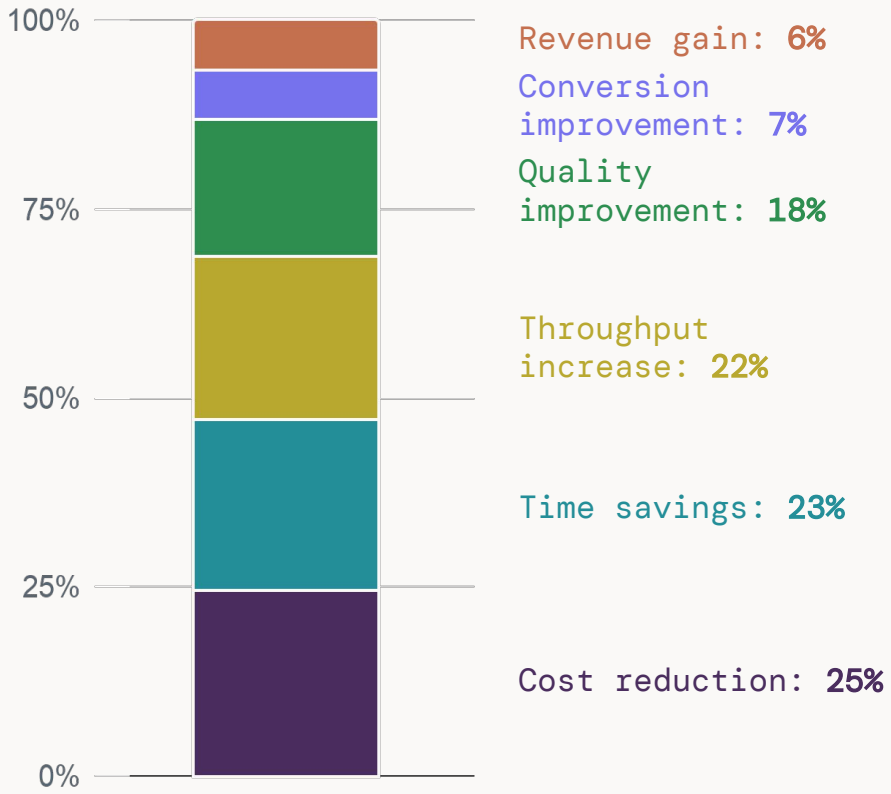


- **Growing attention:** Firms see AI as an opportunity to improve earnings. We tracked a 3-4x rise in mentions of AI's impact across the S&P 500 since 2023.
- **Put a number to it:** 50-60% of claims are now quantified, but TBD how large and meaningful these are for companies' bottom line.
- **Still a minority:** A majority of firms have not yet reported a quantified impact from AI use.

Sources: Exponential View analysis; earnings calls.

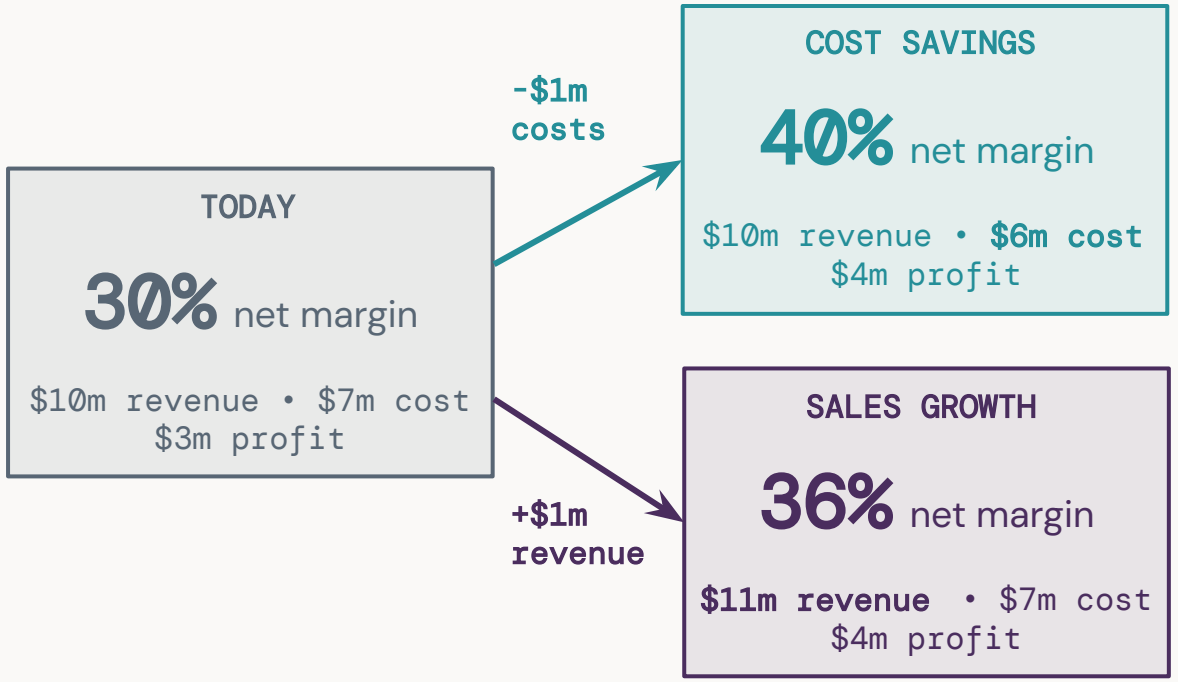
Seven in ten GenAI claims focus on cost savings or efficiency

Claimed AI outcomes
S&P 500, Q4 2022 - Q1 2026



Sources: Exponential View analysis; earnings calls.
Note: Figures may not sum to 100% because of rounding.

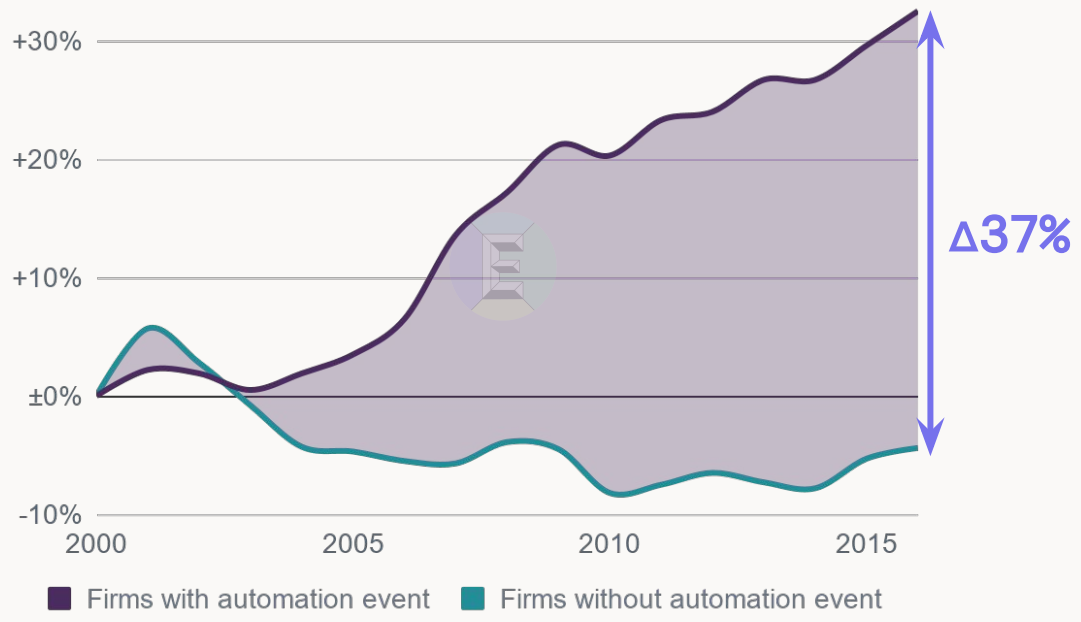
Why initial projects prioritize efficiency, an illustrative example:
The same pure $1m impact has ***4pp better profit margin from savings*** vs sales growth



Note: For illustrative purposes, revenue is added at $0 cost.
Adding costs-of-sales would dampen margin growth further.

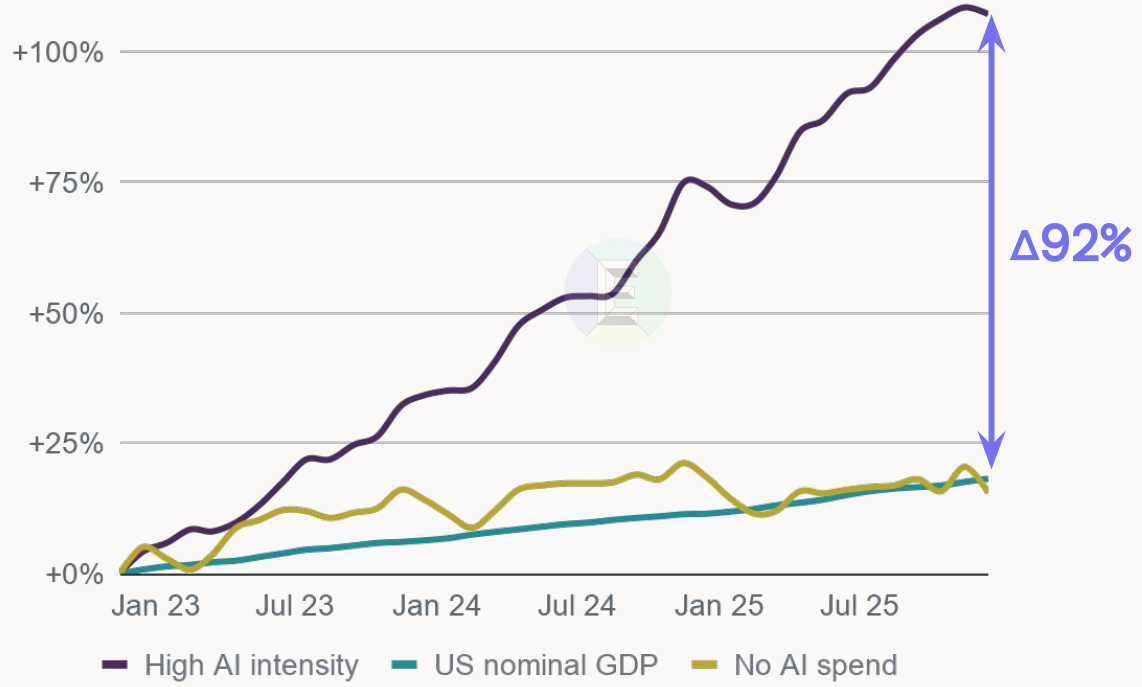
As in prior waves, early adopters are outgrowing their peers

The historic case study:
Firm-level employment with/without automation
% change 2000-2016



Sources: Exponential View analysis;
Bessen, Goos, Salomons & Van den Berge (2020):
"What Happens to Workers at Firms that Automate?".

The AI economy today:
Revenue growth with high vs no AI usage
% change since Nov 2022



Sources: Exponential View analysis; Ramp Economics Lab
(n = 70,000 US businesses).
Note: High intensity = Top 25% AI spenders by share of revenue.

3 | CapEx:

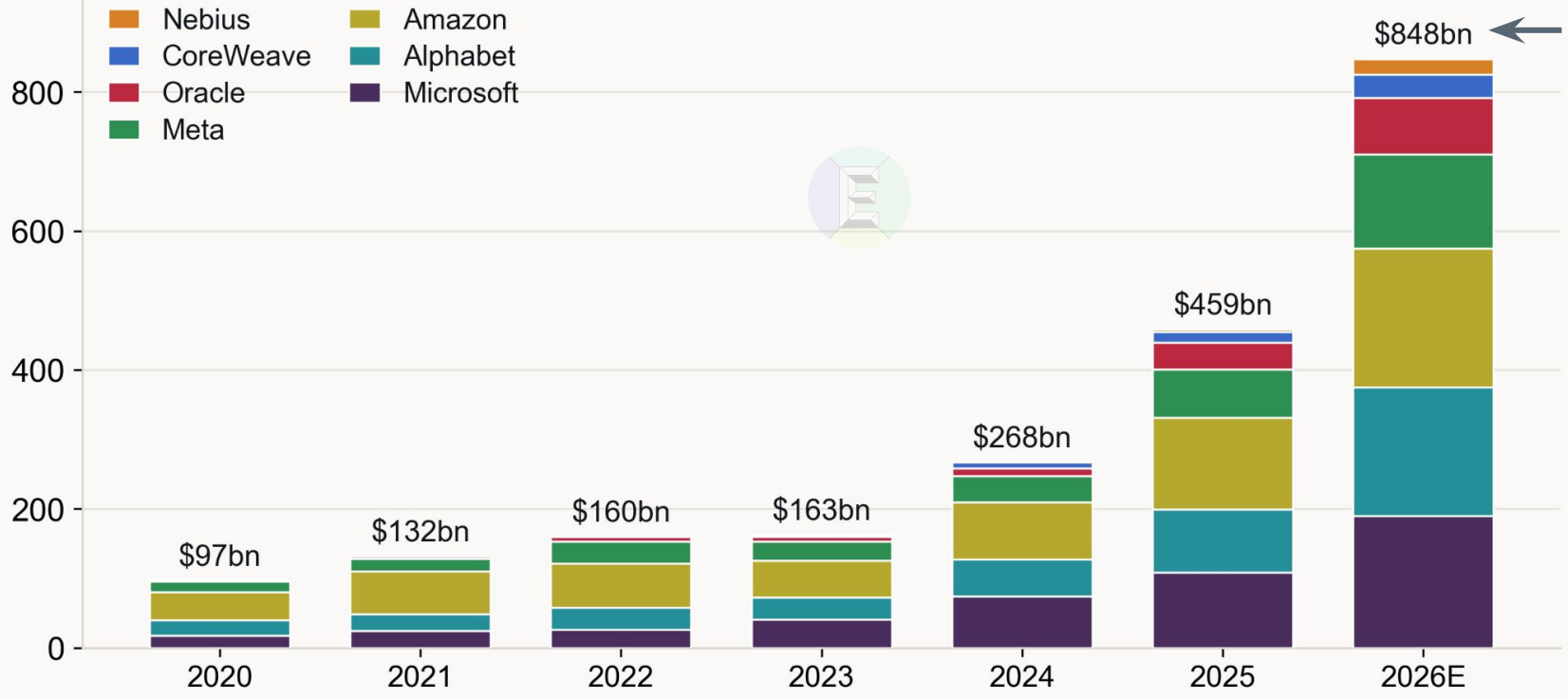
The largest buildout in tech history is paying back (for now)

Hyperscalers & neoclouds have committed to $2 trillion of cumulative CapEx to 2026, putting pressure on growing revenues to pay back, especially as more is funded by external capital. These economics set the tone for data center and token production finances.

Hyperscaler and neocloud CapEx reaches $2T cumulatively through 2026E

Hyperscaler and neocloud CapEx

$bn, PP&E + leases



Total CapEx ≠ AI CapEx. Announced numbers include pre-planned CapEx for existing cloud & SaaS businesses, metaverse (Meta), and logistics (Amazon)

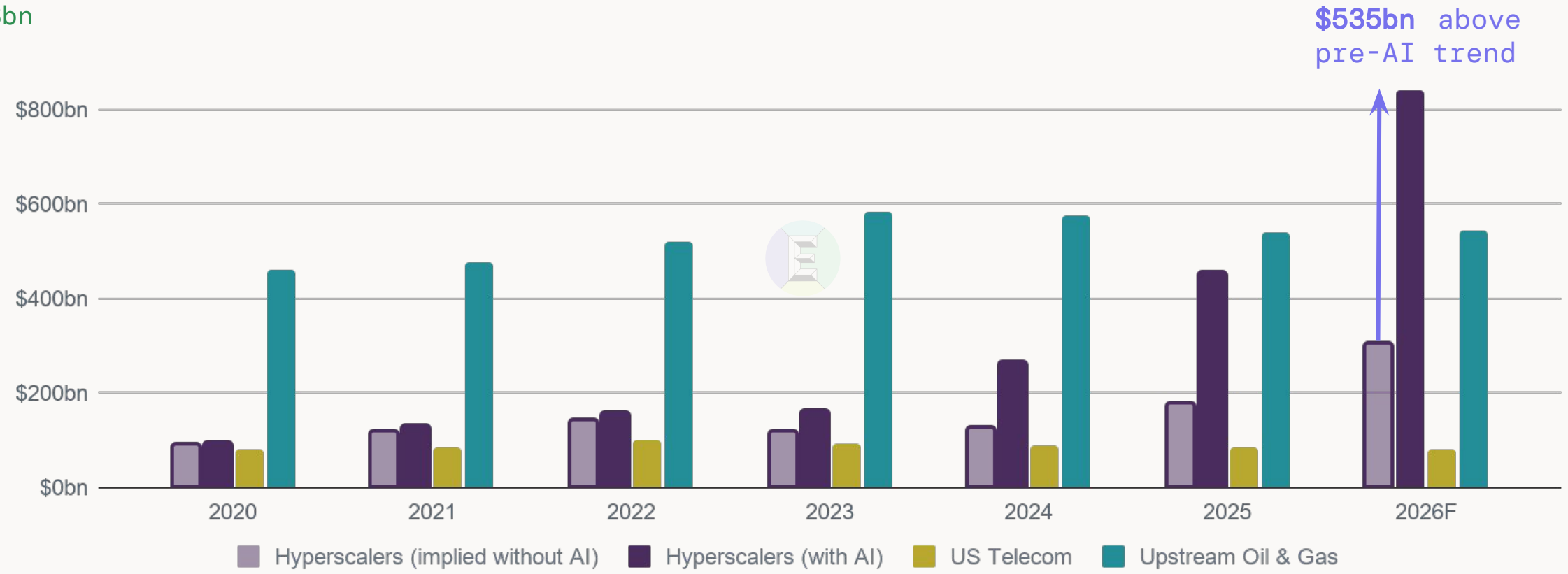
Sources: Exponential View analysis; company filings.

Note: 2026 is based on guidance values. Oracle uses a 5/12:7/12 split based on FY2026 reported and FY2027 guidance values.

AI-linked CapEx adds $535bn above the pre-AI trend by 2026E

Annual CapEx per industry

$bn

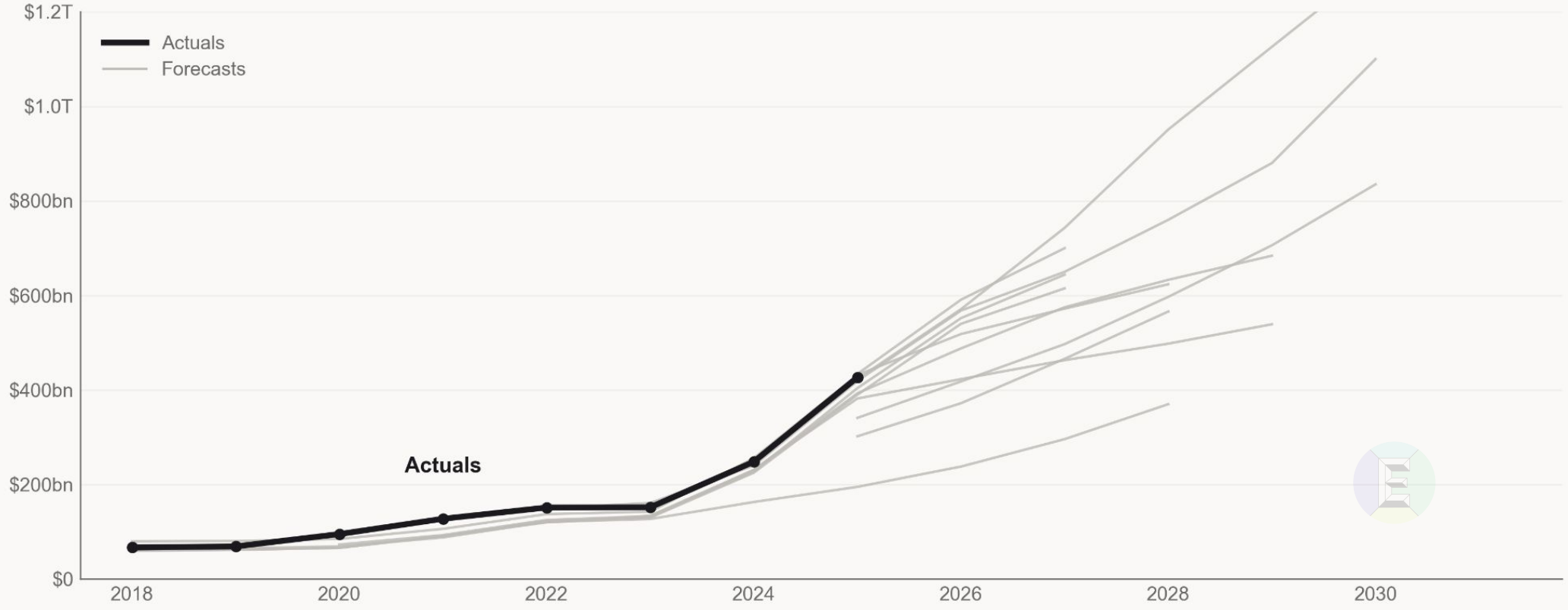


Sources: Exponential View analysis; Hyperscaler earnings & guidance; US telecom & carrier guidance; The International Energy Agency.

Forecasts have chased the CapEx curve higher

Hyperscaler / AI infra CapEx forecasts by analyst and forecast date

$/year



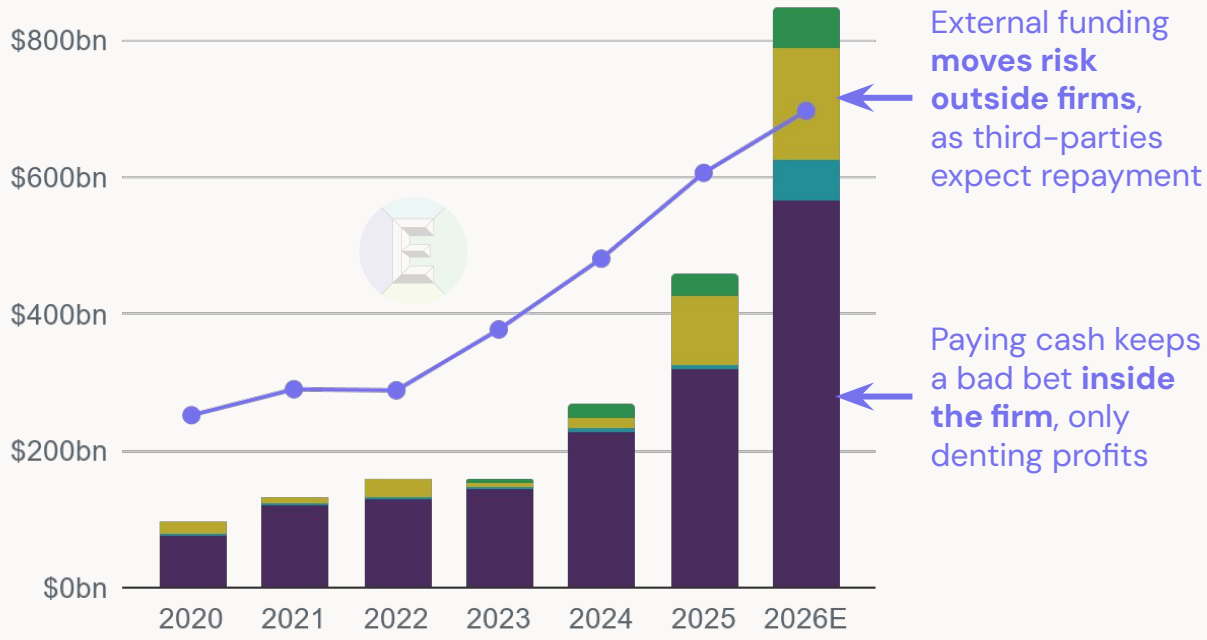
Sources: Exponential View analysis; Barclays, Citi, Goldman Sachs, JP Morgan, Morgan Stanley, New Street Research, SemiAnalysis, UBS; company filings.

Note: Forecasters use slightly different baskets (the Big Five hyperscalers vs broader AI infrastructure). Actuals here correspond to the cash CapEx (purchases of property and equipment) for Microsoft, Alphabet, Amazon, Meta, Oracle, CoreWeave, and Nebius.

The marginal AI-infra dollar is increasingly externally financed

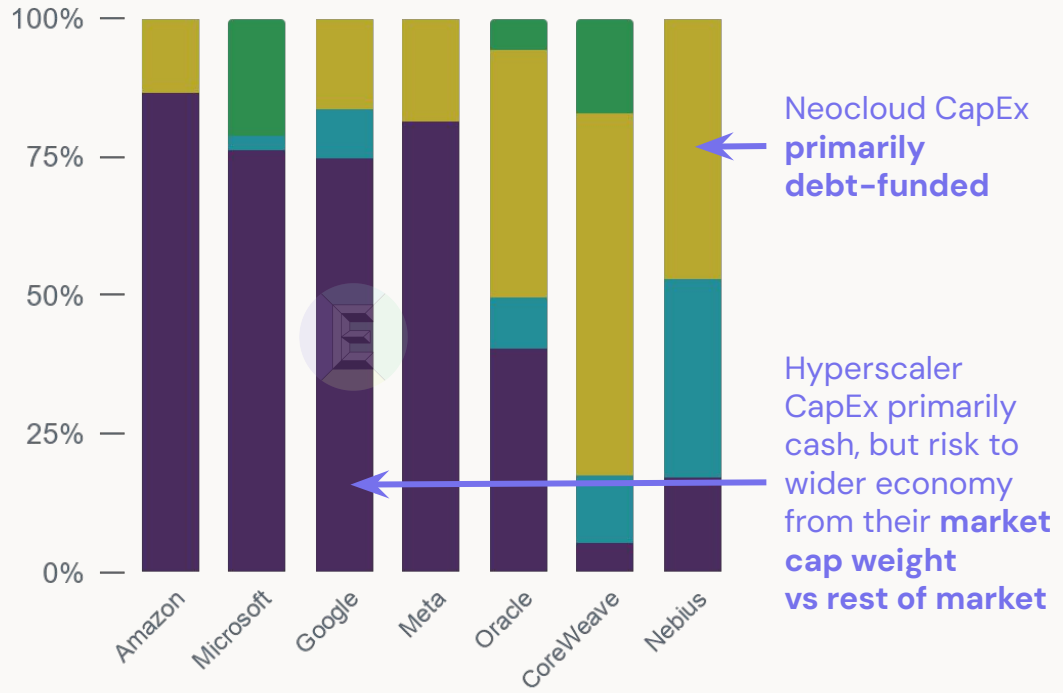
Hyperscaler and neocloud CapEx by funding source

$, 2020-2026E



CapEx by funding source

%, 2020-2026E total

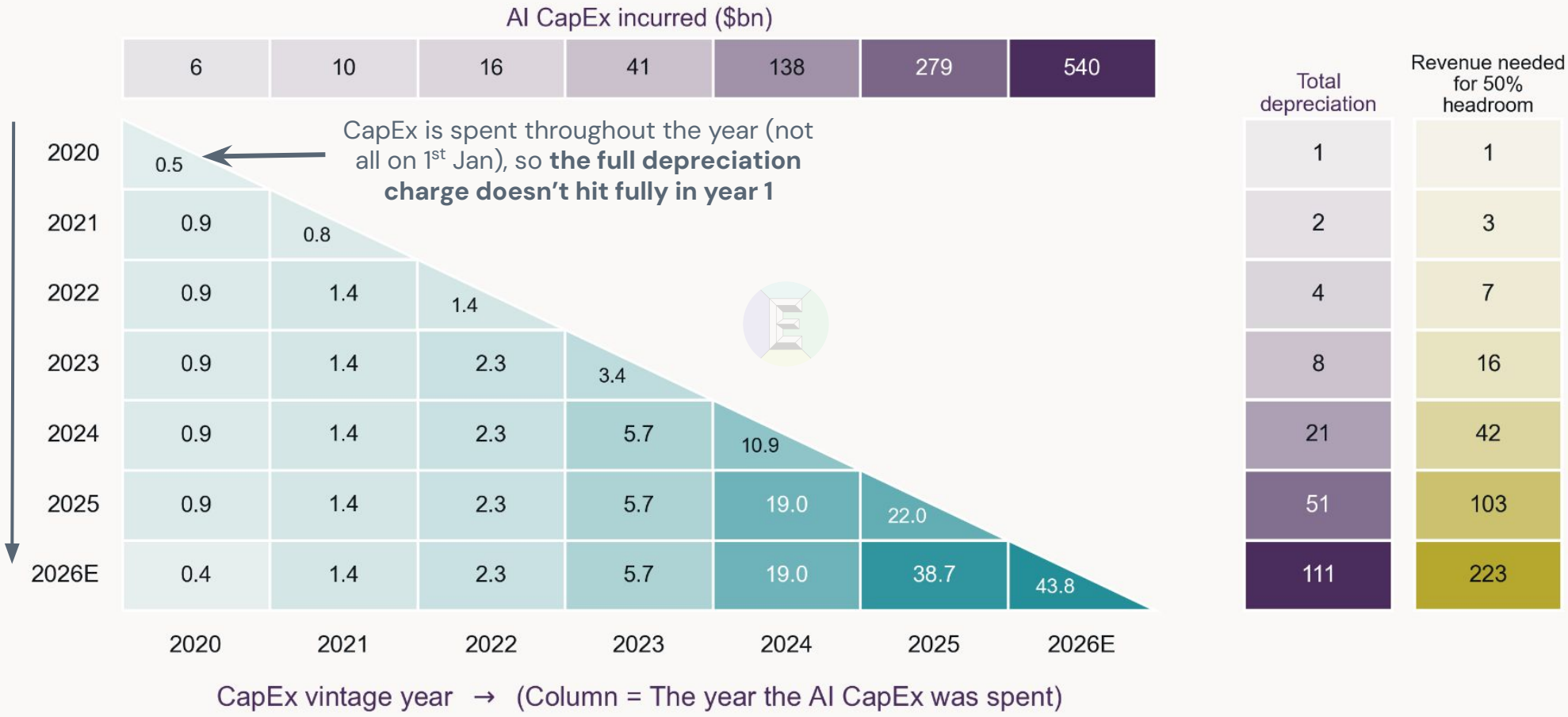


Operating cash flow | Leases | Debt (net new) | Equity | Cash

Sources: Exponential View analysis; company filings.
Note: Debt is net of repayments (not gross issuance) and includes all debt instruments (bonds, commercial paper, etc.).

The 2026E depreciation charge approaches $lll billion

CapEx is expensed through depreciation over the assets' useful life. So **the cost is spread** and doesn't need to be recognized instantly



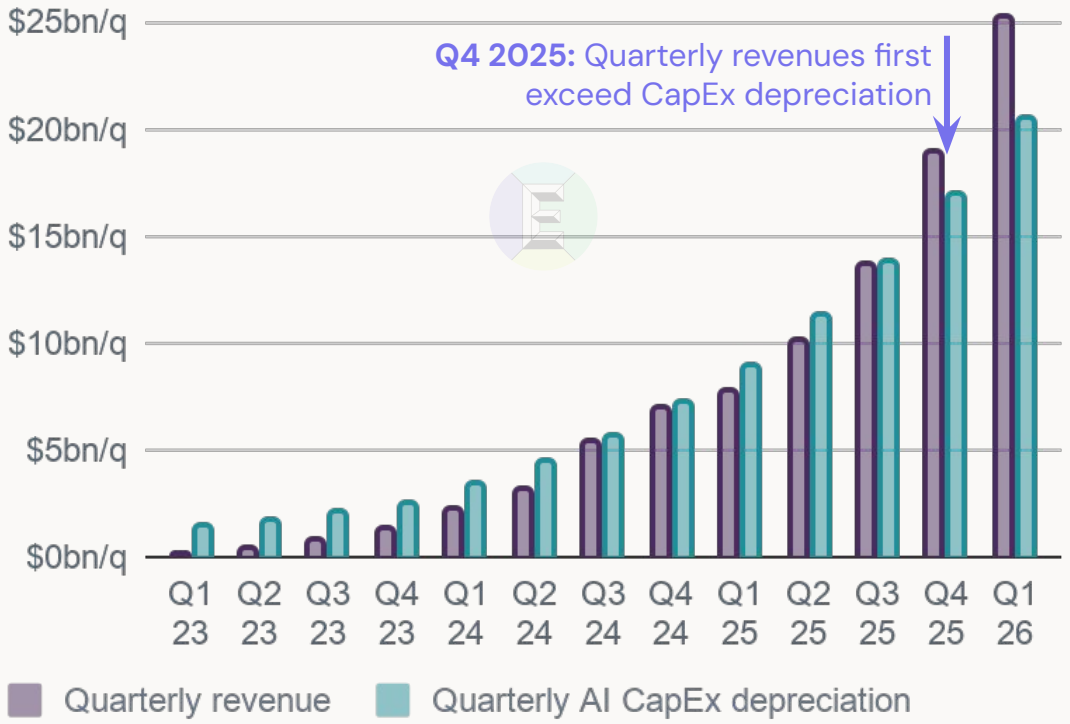
Total depreciation	Revenue needed for 50% headroom
1	1
2	3
4	7
8	16
21	42
51	103
111	223

Source: Exponential View analysis.
Note: IT equipment is depreciated over 6 years, and buildings over 14 years. Required revenue values exclude OpEx. Headroom is the portion of revenue beyond that required to meet the depreciation expense.

Revenues cover the ongoing expense, not yet the cumulative bill

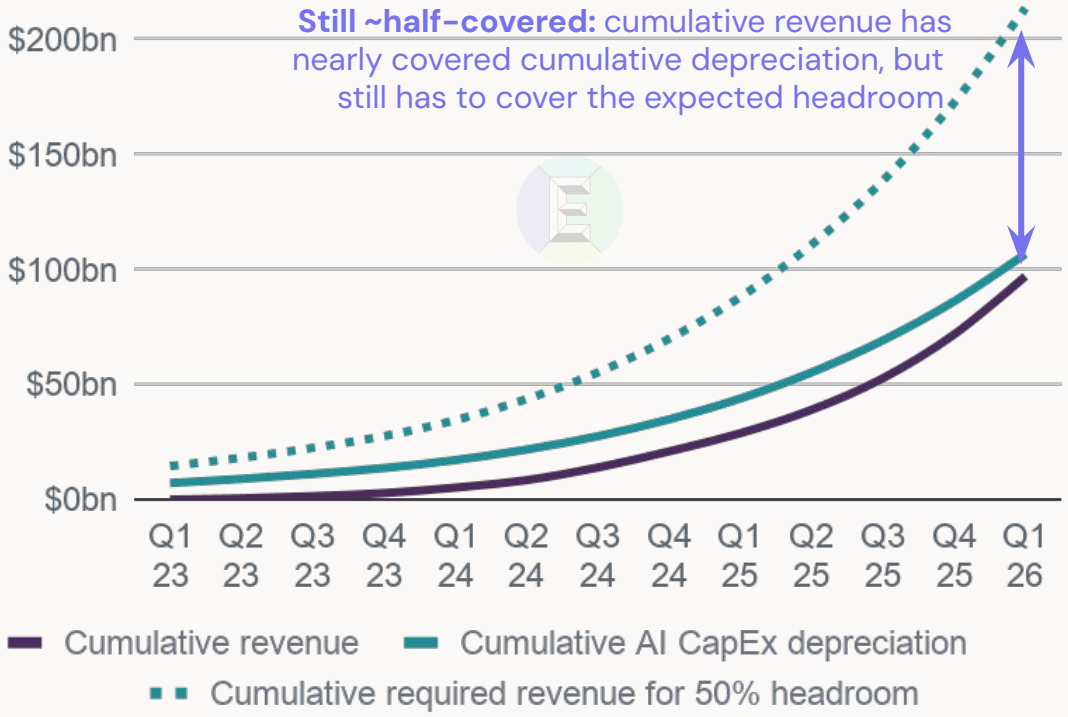
Quarterly AI revenues & CapEx depreciation

$bn/quarter, hyperscalers & neoclouds only



Cumulative AI revenues & CapEx depreciation

$bn, hyperscalers & neoclouds only



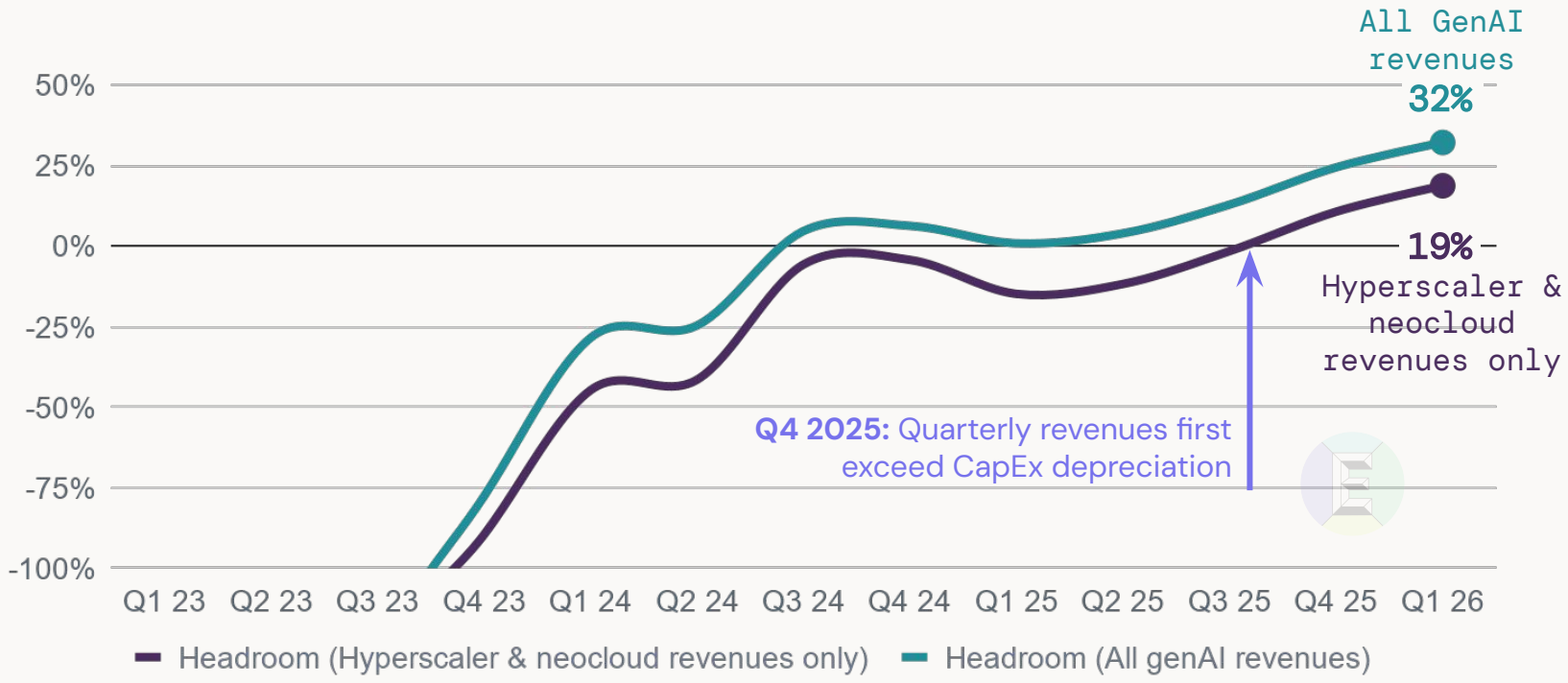
Sources: Exponential View analysis; company filings.

Note: Meta contributes to industry CapEx but initiatives are focused on ad uplift, so not recognized as pure GenAI revenue, or currently have minimal direct monetization (e.g. Meta AI assistant, Muse Spark).

AI infra revenue now just clears today's depreciation hurdle

Headroom after quarterly CapEx depreciation

% = (Revenue - Depreciation) ÷ Revenue



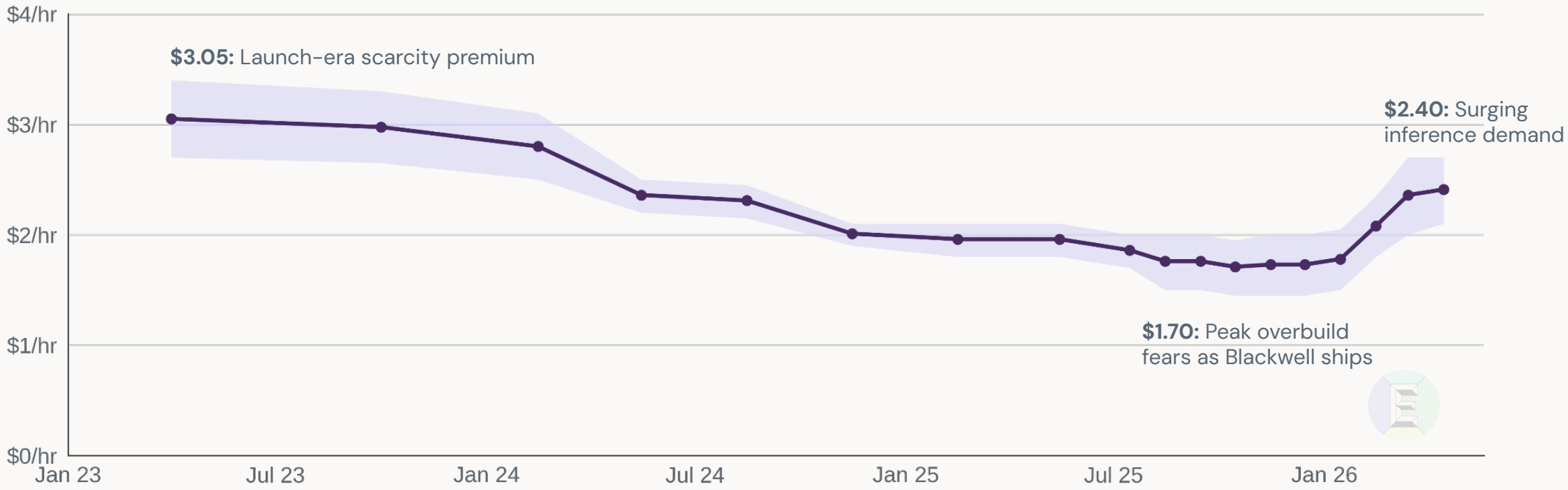
- **GenAI revenues now cover the quarterly depreciation of AI infrastructure.** Q1 26 headroom reached 19% for hyperscaler/neocloud revenues and 32% across all GenAI revenues.
- **Coverage remains thin.** Depreciation absorbs roughly 81% of hyperscaler/neocloud GenAI revenue and 68% of total GenAI revenue before additional costs.
- **The next test is incremental coverage.** As committed AI capex enters service, the depreciation base will rise. Revenue growth, utilization and pricing must continue to compound or headroom will compress again.

Source: Exponential View analysis.

Rental rates suggest demand is absorbing existing supply

H100 1-year rental contract price

$/hour/GPU. H100 is the most liquid, most-traded GPU in the merchant market: a useful signal.



Sources: Exponential View analysis; SemiAnalysis.

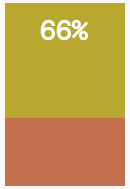
Data center economics set the hurdle for token pricing

$7.9bn annual cost to own and operate 1 GW of AI capacity

Capital costs $7.0bn/year · 89%

- **Servers $4.6bn**
  480,000 GPUs in 6,700 systems · 6-year life
- **Facility $1.4bn**
  Building shell, power & cooling · 14-year life
- **DC network $1.0bn**
  Switching & interconnect fabric · 6-year life
- **Land + utility $33m**
  Land at cost of capital · utility 14-year life

OpEx $900m/year · 11%



- **Energy $594m**
  Electricity to run the fleet · 66% of OpEx
- **Other OpEx $308m**
  Staff, maintenance & overhead

● Kimi K2.5 (1T) ● Kimi-class model under closed licensing (illustrative)

DIVIDE $7.9bn/yr BY TOKEN OUTPUT:

÷ TOKENS PER GW / YEAR

75–85 quadrillion

= COST FOR INFERENCE PROVIDER *per 1m tokens*

$0.10	$0.42
	$0.32 licensing fee (~25% of the blended selling price of $1.29)

REQUIRED PRICE FOR 50-75% GROSS MARGIN: *per 1m tokens*

$0.20 – $0.40	$0.84 – $1.68

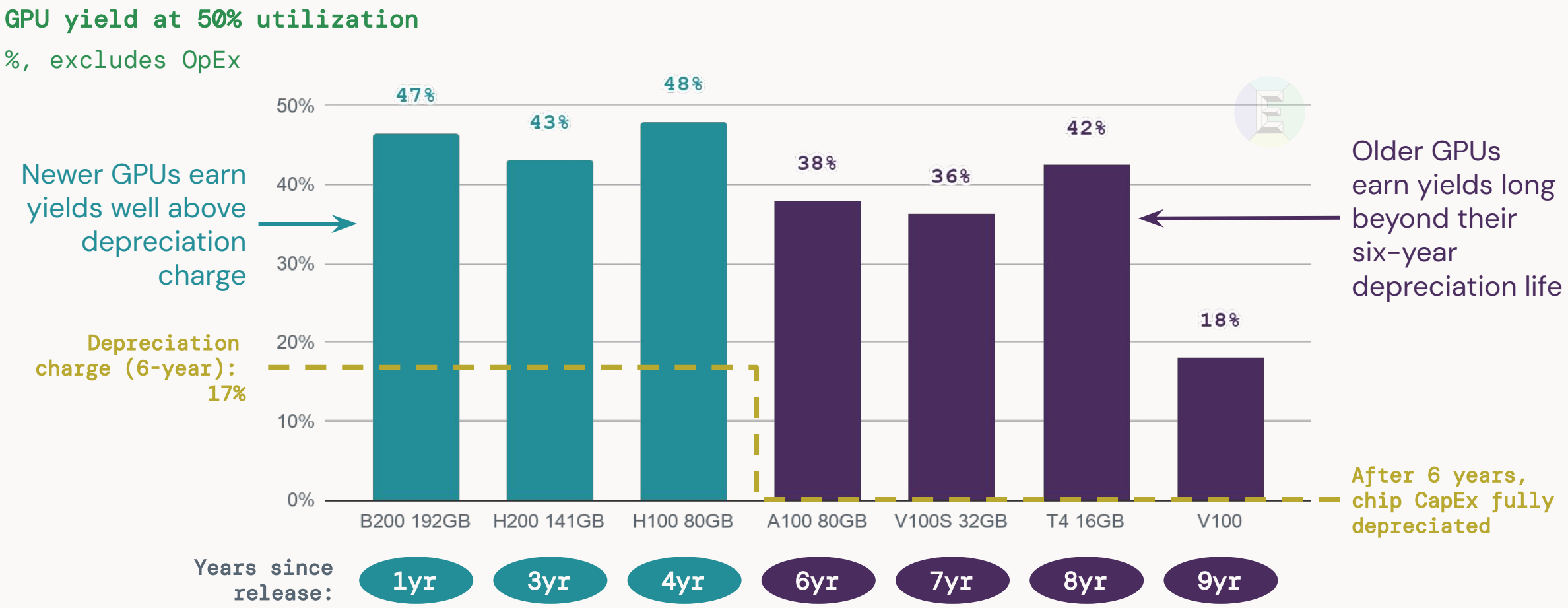
REQUIRED CUSTOMER VALUE FOR 25% ROI *per 1m tokens*

$0.25 – $0.50	$1.05 – $2.10

Sources: Exponential View analysis; Epoch AI; SemiAnalysis.

Note: Illustrative model. 1GW of IT capacity (6.7k GB200 NVL72 systems, 480k GPUs), cost of ownership per Epoch AI (May 2026), annualized incl. cost of capital, 6-year IT life. Token output from SemiAnalysis InferenceX: FP4, 8k-in/1k-out, 50 tokens/sec/user, 65% utilization (Apr 2026). Token output range reflects +10-25% throughput uplift from speculative decoding. Open-weight models incur no model-licensing fee. Closed-weight column adds a 25% licensing fee of Kimi's $1.29 blended price.

Gross rental yields suggest useful lives extend past six years



Sources: Exponential View analysis; Silicon Data.
Note: Yield = (On-demand rate x 50% utilization x 8760 hours) ÷ original list price.

Longer GPU useful life stretches headroom

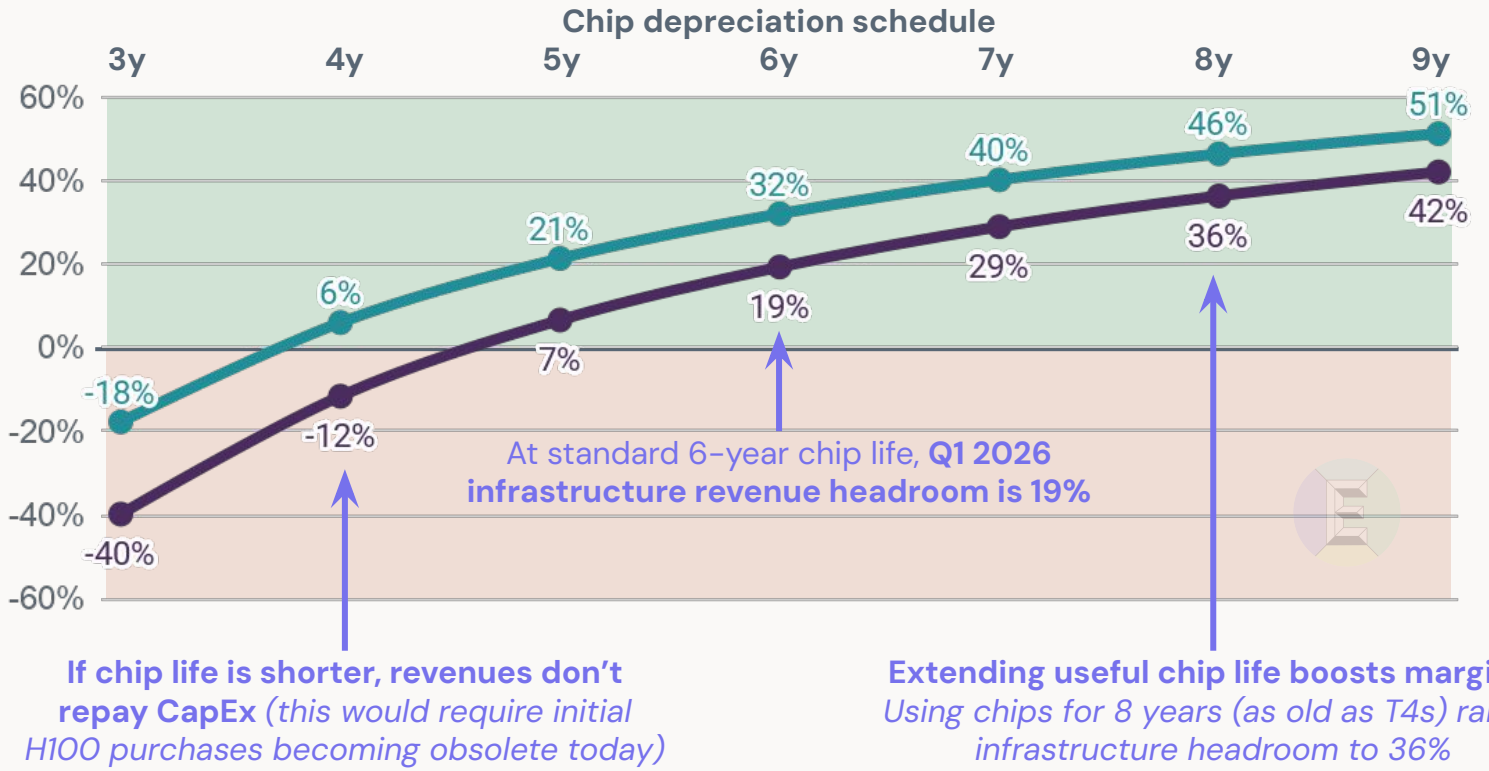
Mark Zuckerberg
Meta Q3 2025 earnings call

"... the kind of very worst case would be that we effectively have just prebuilt for a couple of years, in which case, of course, there would be some loss and depreciation, but we'd grow into that and use it over time."

Overbuild can be a bet on longer chip depreciation

Range of headroom after CapEx per chip depreciation schedule

Q1 2026, 3-9-year schedules, % = (Revenue - Depreciation) ÷ Revenue



Sources: Exponential View analysis; Meta.
Note: Buildings depreciated over 14 years as constant.

Headroom (infrastructure revenues only)
Headroom (whole market revenues)

4 | Tokens :

The unit of value for the AI economy?

Token volumes are growing 14x annually, propelled by agentic workloads and highly elastic demand. Token-based pricing has made this especially pertinent, but it also represents an opportunity for the industry to attribute and evaluate the output from token consumption.

> *"The input is electrons, the output is tokens. In the middle is Nvidia."*
>
> – Jensen Huang

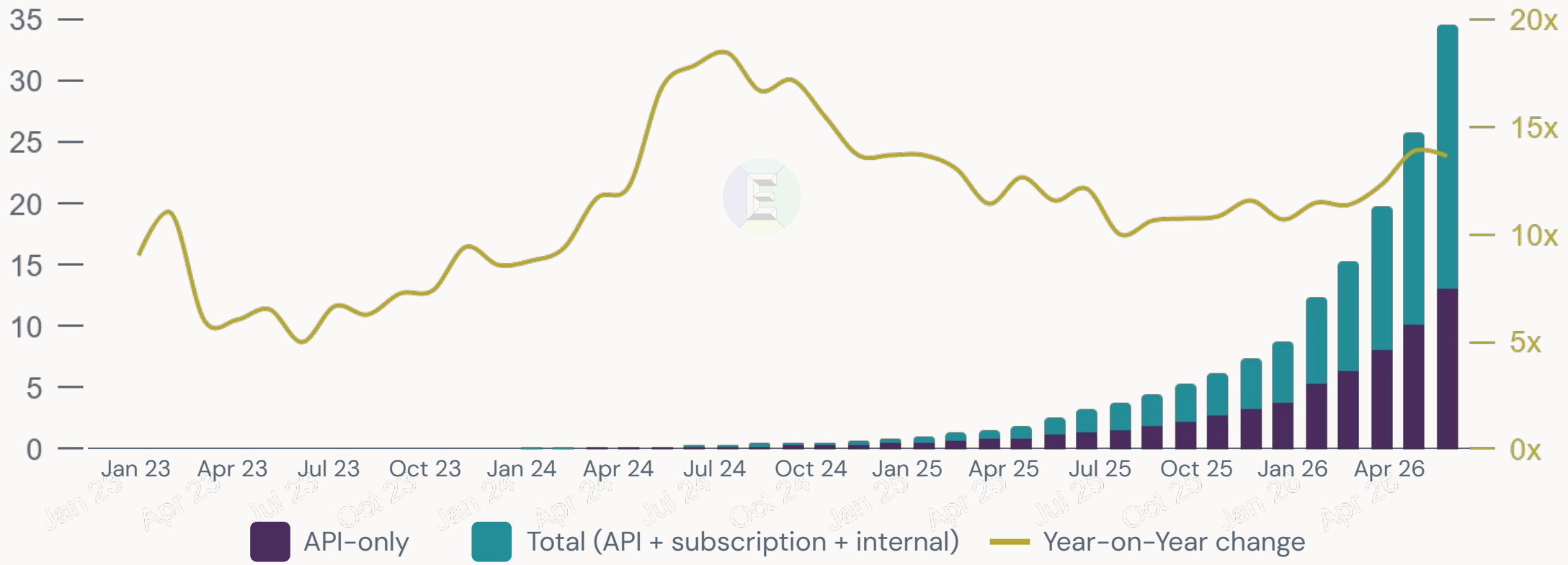
> *"Tokens, the fundamental units of data our models process..."*
>
> – Sundar Pichai

Is the GenAI economy a token economy? Sort of.

Global token volumes exceed 30Q/month, growing 14x YoY

Inference tokens processed

Quadrillion tokens per month (left axis), growth rate multiple (right axis)

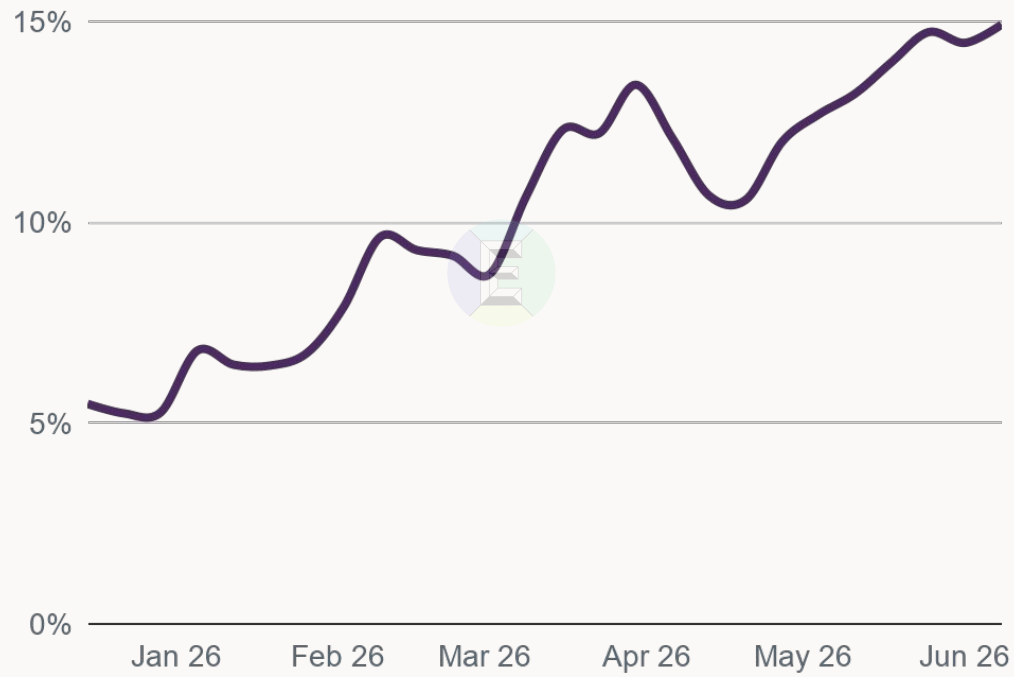


Sources: Exponential View analysis.
Note: Global, inc. China

The transition from chat to agents is multiplying token use

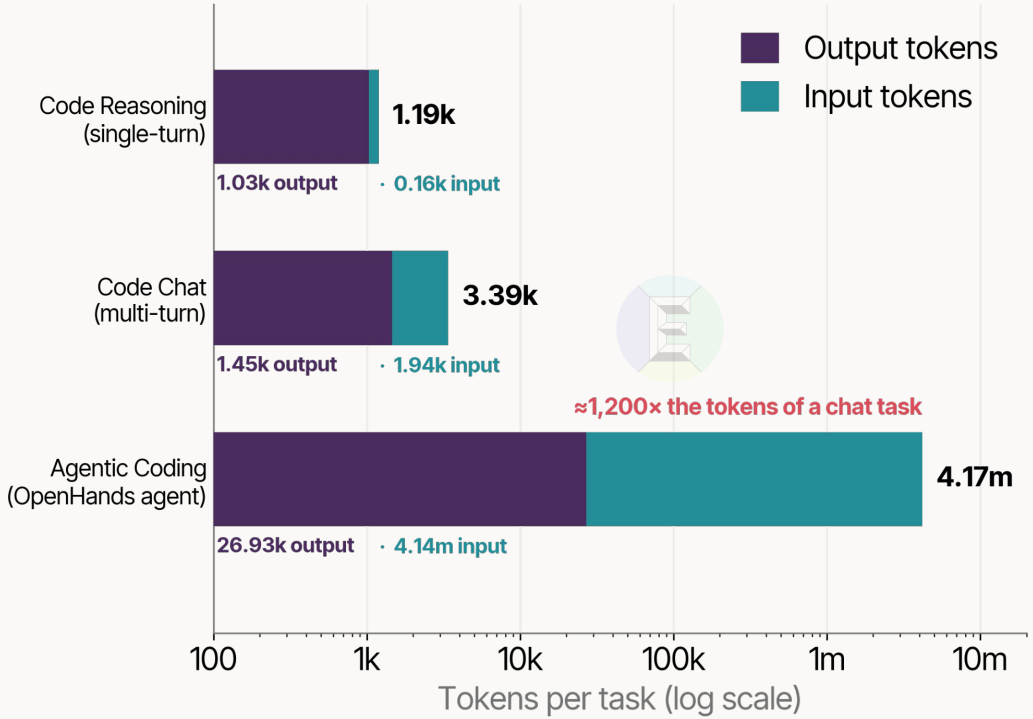
Agent coordination density

% tool use per prompt, OpenRouter



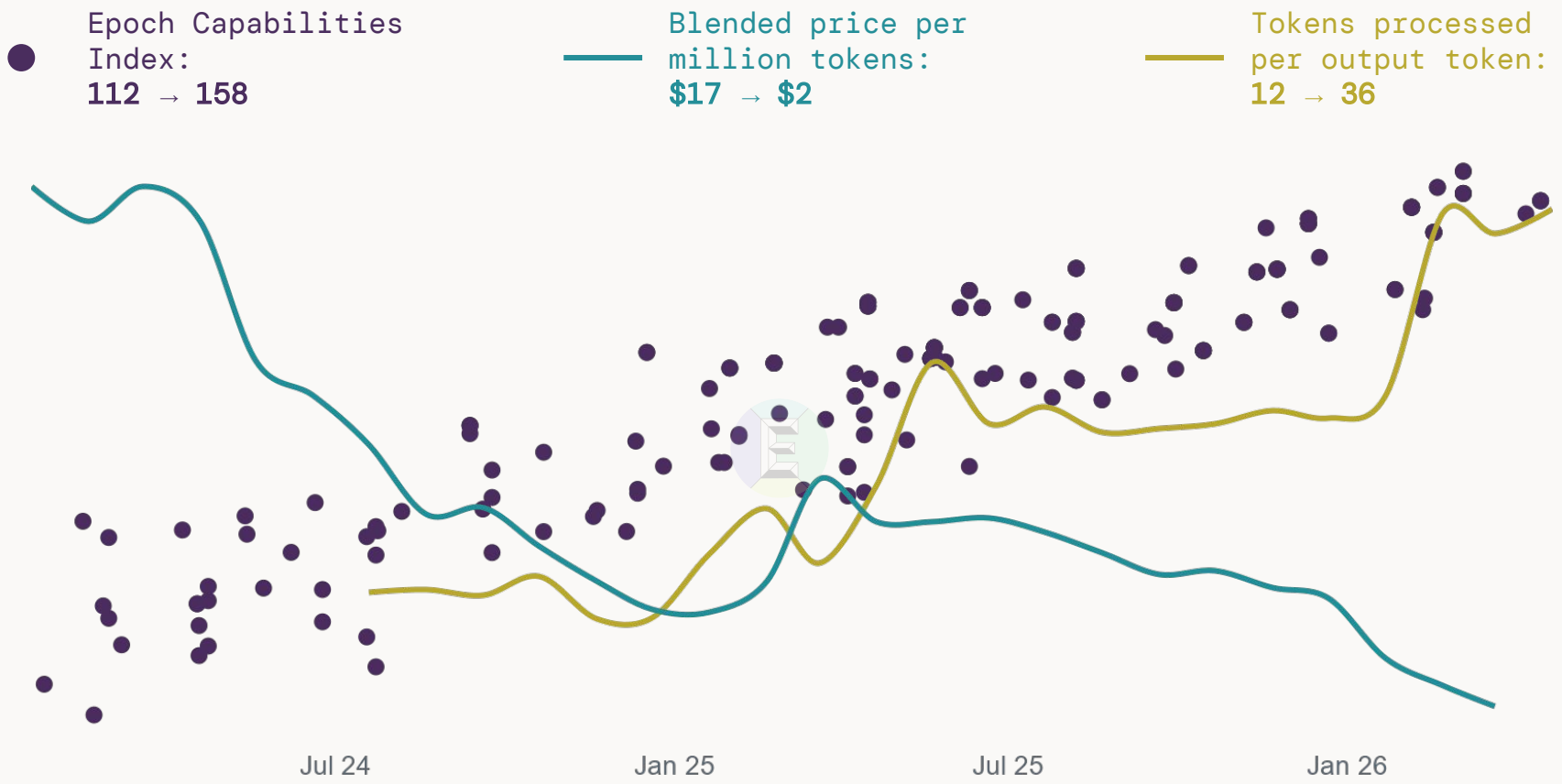
Token consumption per task

Average tokens consumed per task, log scale



Sources: Exponential View analysis; OpenRouter; Bai, Huang, Wang, Sun, Mihalcea, Brynjolfsson and Pentland 2026.

Cheaper tokens and better models amplify demand



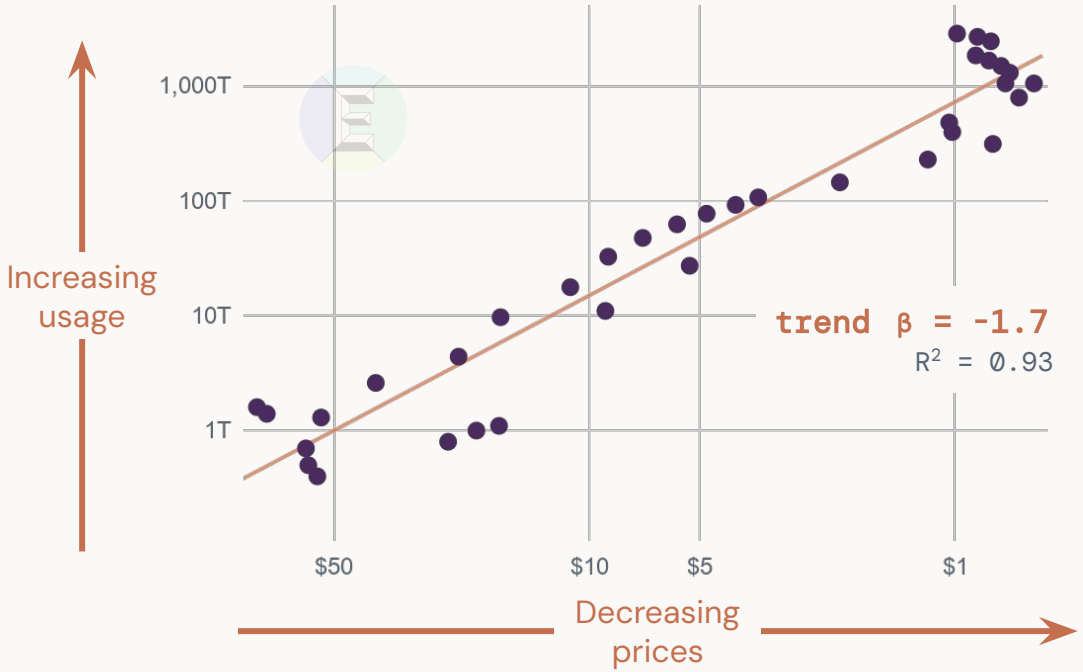
- More capable models able to cover a **wider range of economically useful tasks**, increasing value and use.
- Token volume rising as **reasoning models spend more** tokens “thinking”.
- Price declines encourage more use and make **previously uneconomical applications viable**.

Sources: Exponential View analysis; Epoch AI.

Token demand appears elastic: As prices fall, usage grows faster

Google price elasticity (avg. price vs volume)

$/million tokens vs trillion tokens/month, log-log scale, 2023-2026



Sundar Pichai Google: I/O 2025

"...we were processing 9.7 trillion tokens a month. Now, over 480 trillion — 50x more."

Price -97% | Volume 50x

Sam Altman OpenAI: "Three Observations", 2025

"The cost to use a given level of AI falls about 10x every 12 months, and lower prices lead to much more use"

Price -90% | Volume ↑

Tan Dai Volcengine / ByteDance, 2025

"Doubao's daily token usage exceeded 50 trillion this month, up from 4 trillion in Dec 2024."

Price -50% | Volume 12x

Across providers, magnitude of **elasticity ≈ 1.2-1.8** :
every 10% price cut → 12-18% more tokens → total token spend still rises

Sources: Exponential View analysis; Google; OpenAI; ByteDance.
Note: Time-series, not cross-sectional: price and usage both trend with time, so β may overstate pure price-elasticity.

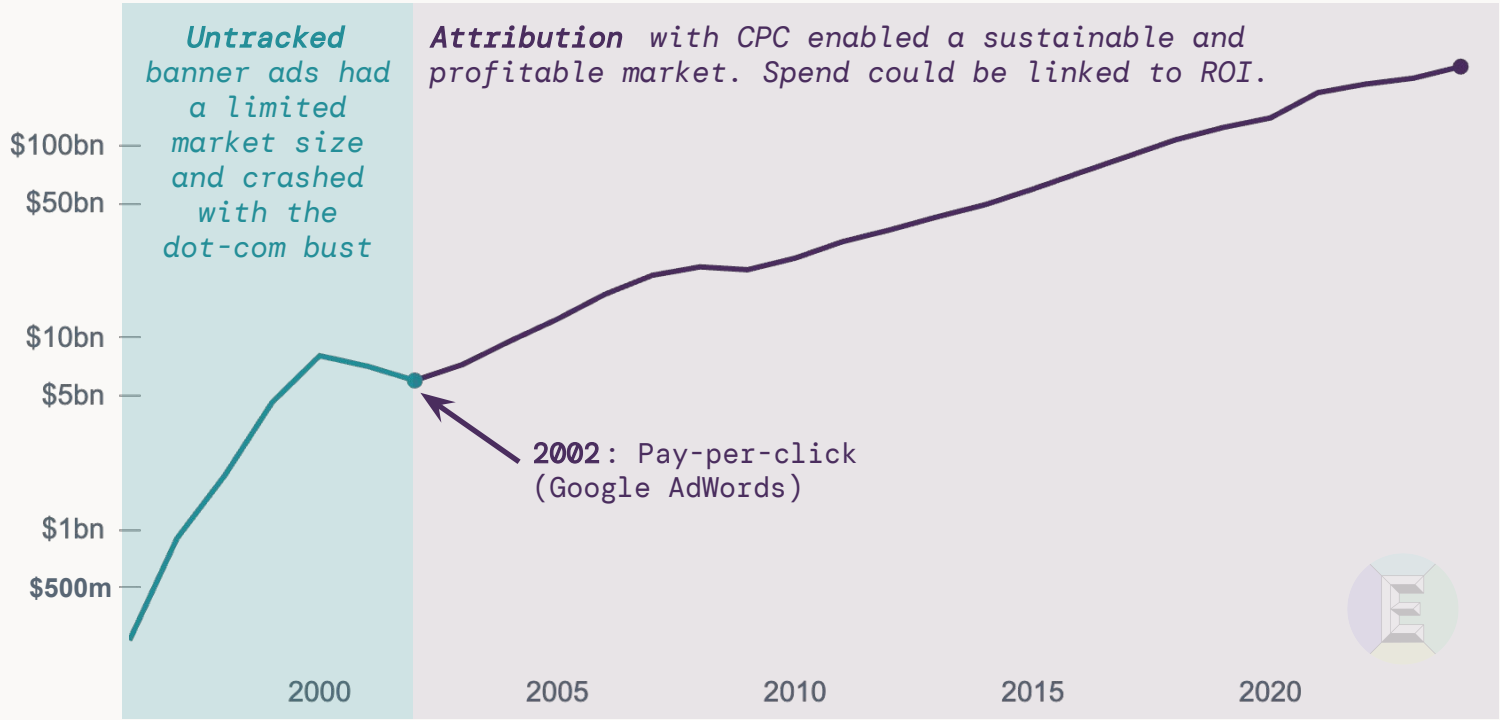
Token-based pricing is AI's 'pay-per-click' moment

Evolution of AI pricing models

1. **FREE**
   Unrecognized in budgets
2. **SUBSCRIPTION**
   Seat-based pricing without tracing to specific value
3. **TOKEN-BASED PRICING**
   Metered usage enables (and requires) attribution to projects

Annual digital ad revenue

$bn/year, log scale, 1996-2024

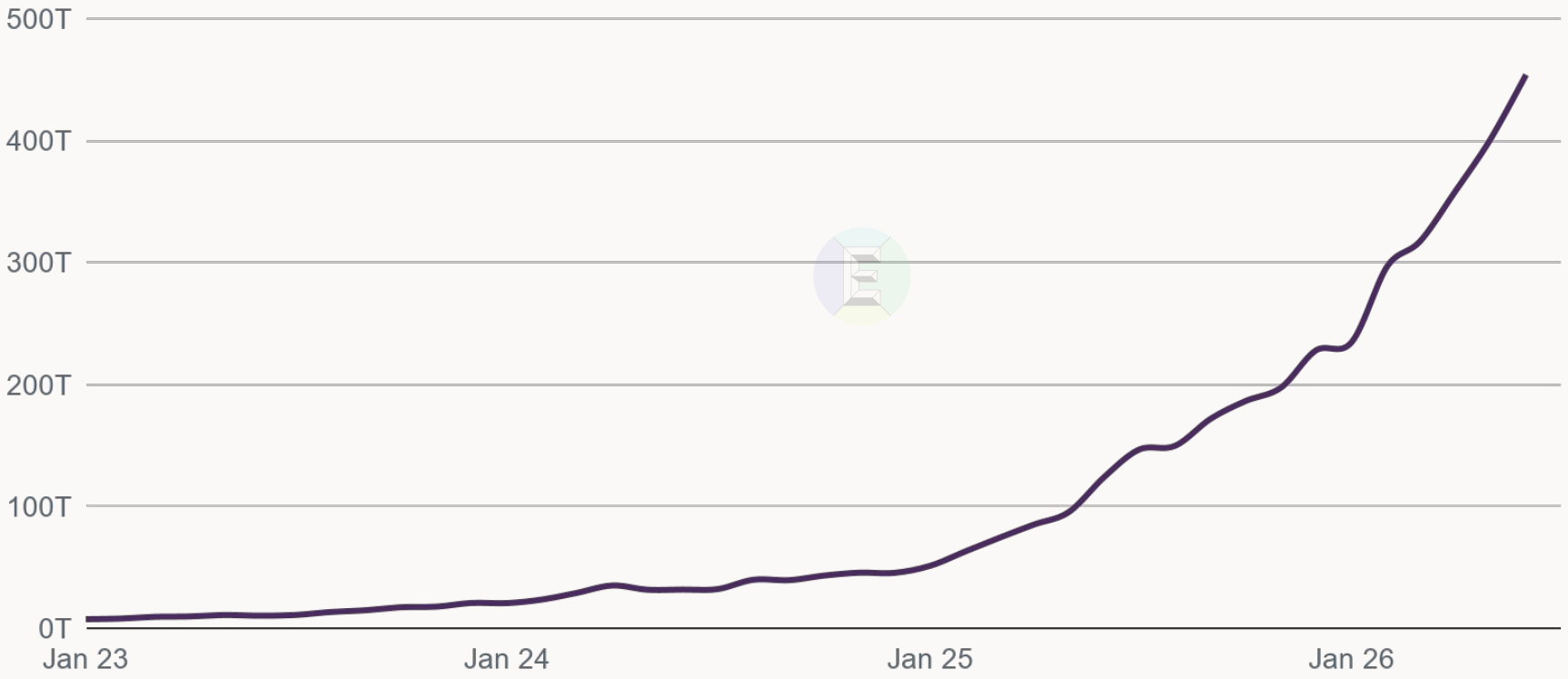


Sources: Exponential View analysis; IAB/PwC Internet Advertising Revenue Report.

Each generation lifts token output per gigawatt of capacity

Tokens produced per gigawatt of data center capacity

Trillion tokens per GW per month



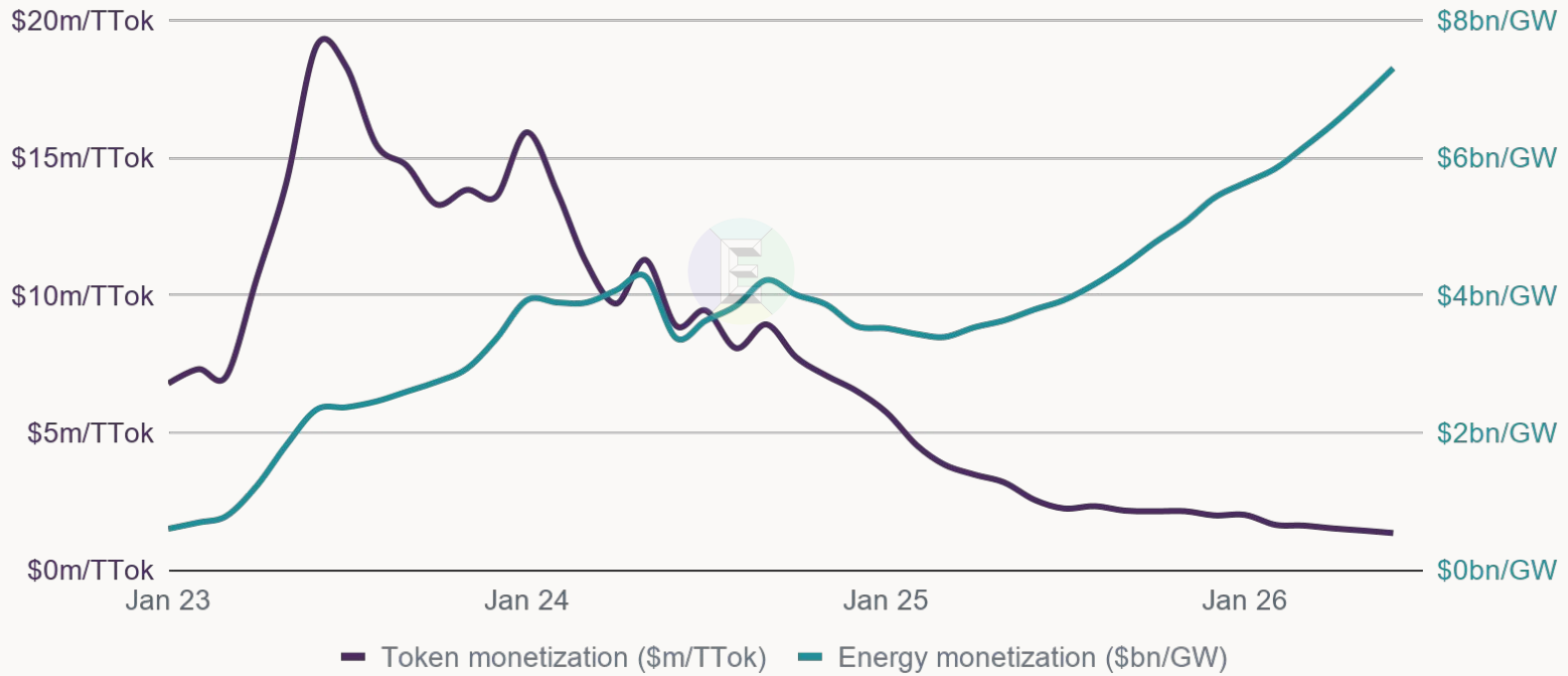
- Every gigawatt buys more token output each month.
- This enables a supercycle, even as physical constraints put pressure on the AI economy.
- Driven by:
  a) Labs achieving higher efficiency through smaller models and better serving.
  b) Hardware gains, although slower moving (e.g. Hopper → Blackwell → Rubin).
  c) Workload mix shifting to inference vs training.

Source: Exponential View analysis. Global ex-China.

This efficiency is increasing monetization per GW of capacity while revenues per token fall

Revenue generated from tokens & data center capacity

$m/TTok (left axis), $bn/GW (right axis)



- Revenue per trillion tokens has fallen since its 2023 peak, mirroring price declines.
- Efficiency gains drive lower token prices, which are more than offset by higher demand.
- Industry-wide revenue per GW of data center capacity passed $7bn/GW.

Source: Exponential View analysis. Global ex-China. Annualized revenue.

Tokens are AI's billing metric, but not yet a unit of value

ELECTRICITY

Lamps

Kilowatt-hours

Edison's first customers paid per light bulb installed. Metering came later.

INTERNET

Pageviews

Sessions & clicks

Advertisers pay on attention and conversion:
CPM → CPC → CPA

MOBILE INTERNET

Megabytes

Active users

Apps valued on engagement:
DAU / MAU, retention

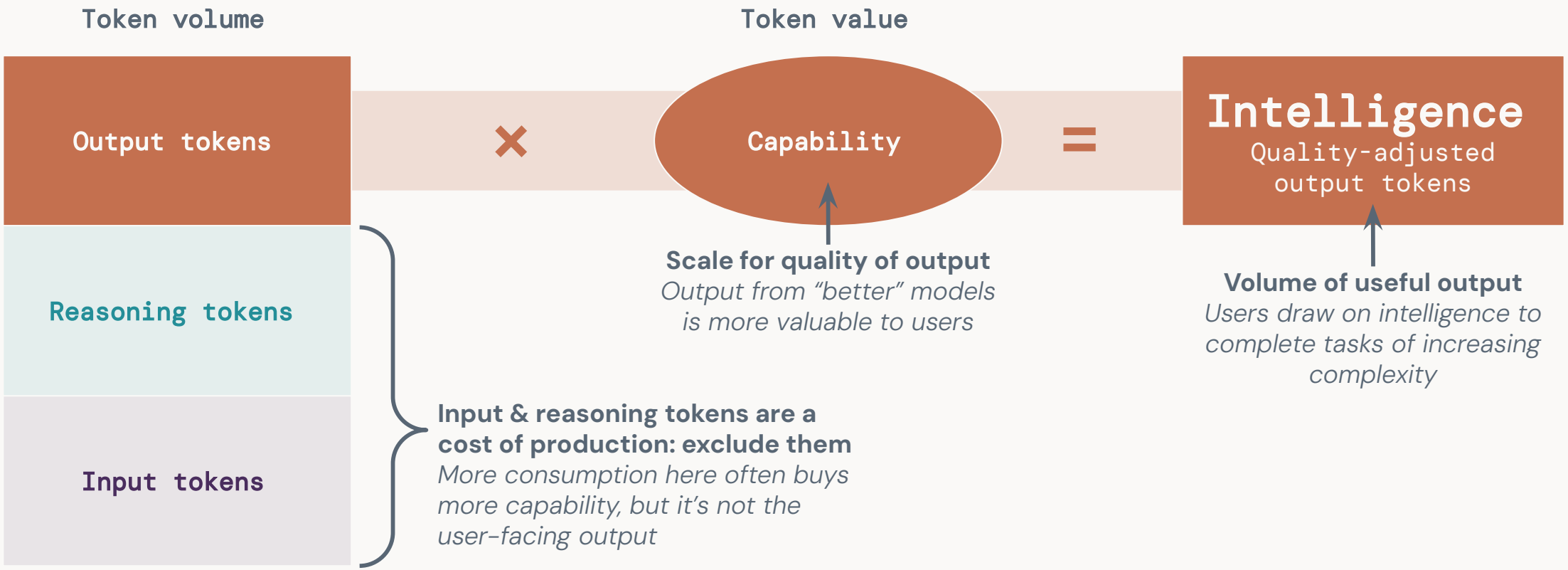
GEN AI

Tokens

"Intelligence"?

The value-producing unit is still undefined

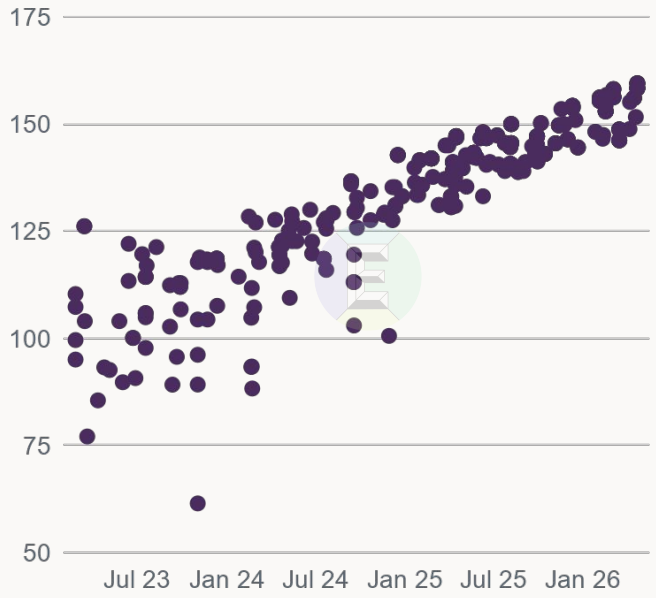
Quality-adjusted tokens come closest to a usable unit of value



Regardless of the measure you pick, the trend is on the up

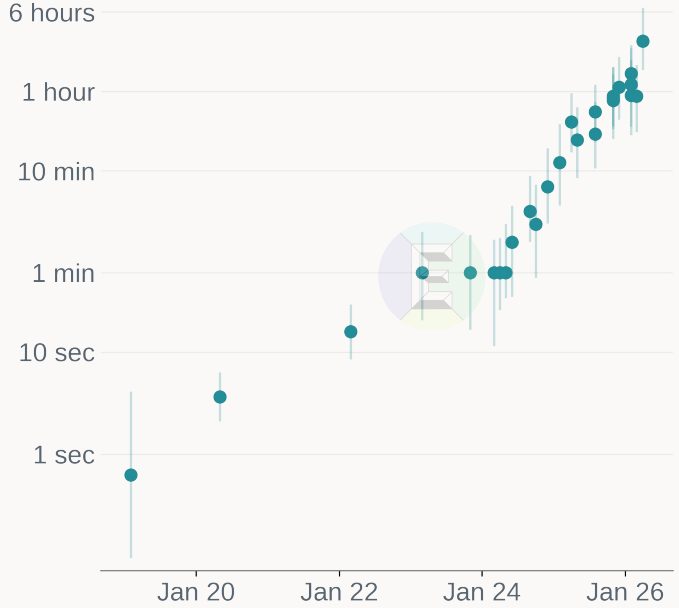
Epoch Capabilities Index

Score: GPT-5 = 150,
Claude 3.5 Sonnet = 130



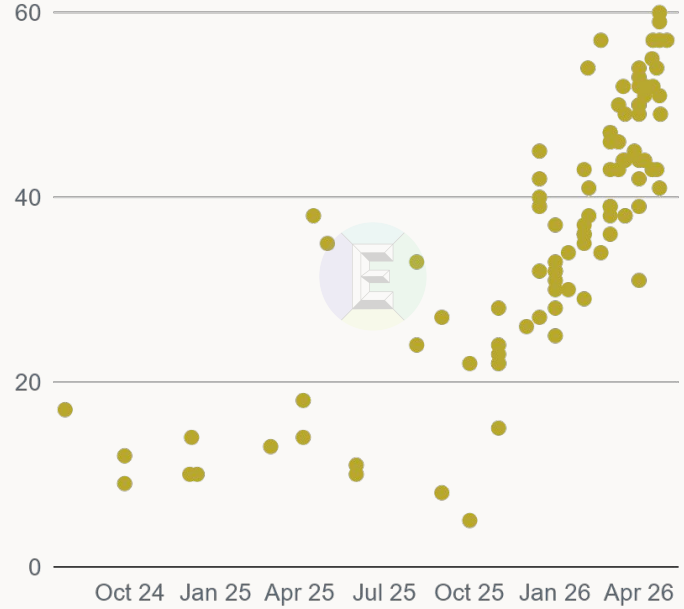
METR Task Horizon

Human task duration with 50% model success rate, log scale



Artificial Analysis Intelligence Index

Score: 0-100

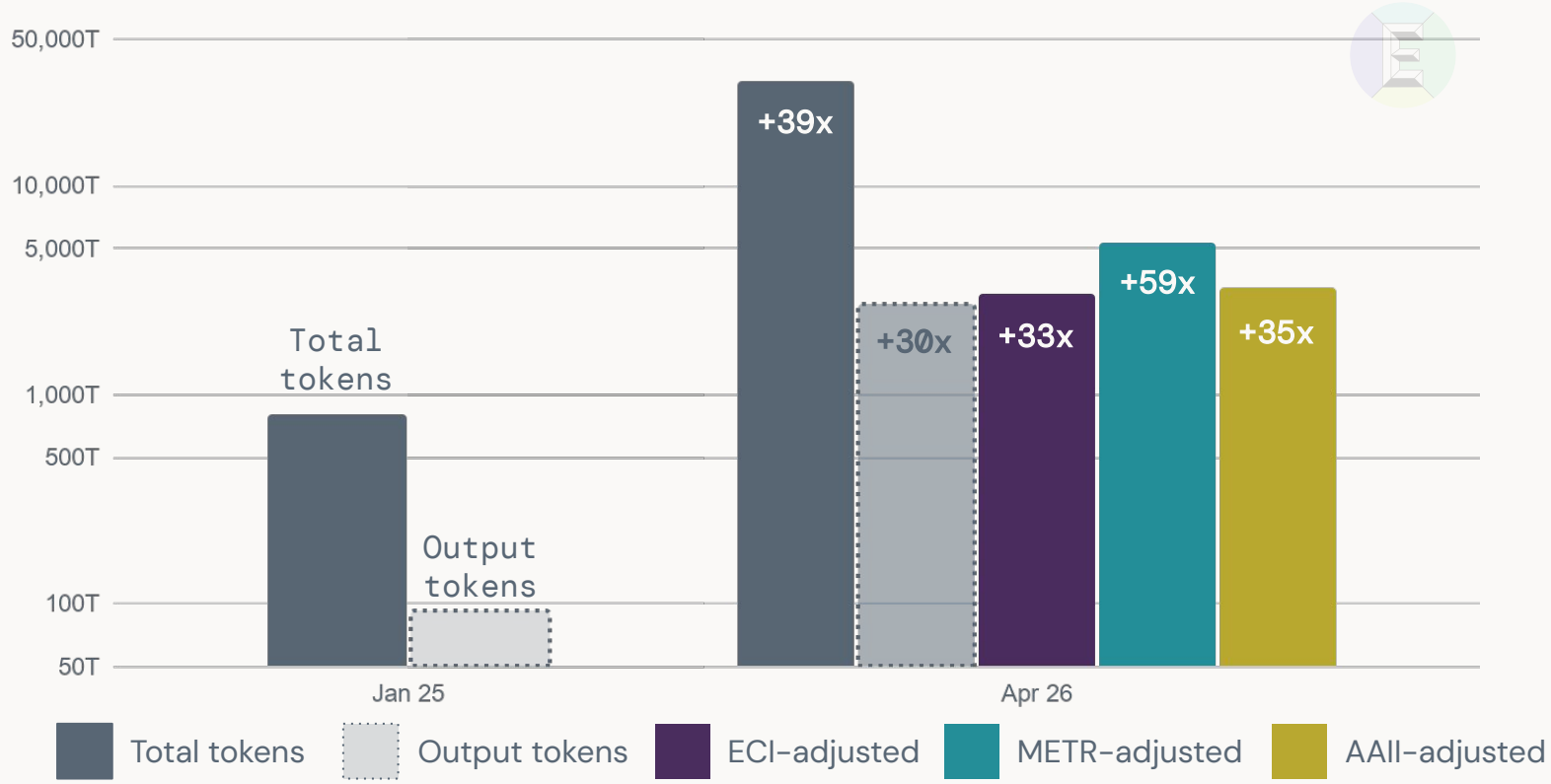


Sources: Exponential View analysis; Epoch AI; METR; Artificial Analysis.
Note: METR tasks primarily consist of software engineering, machine learning, and cybersecurity tasks.

Quality-adjusted output kept pace with raw volume growth

Tokens: Total, output & quality-adjusted

Trillion tokens per month, Jan 2025 vs Apr 2026



- **Raw output tokens grew more slowly than total tokens (30x vs 39x):** Increasing volumes spent on input and reasoning.
- **Wide spread in score improvement (33x–59x):** Different measures and different scoring scales mean we can only draw directional conclusions.
- **Quality-adjusted output tokens seem to be growing:** Output volumes and capabilities both higher than January 2025.

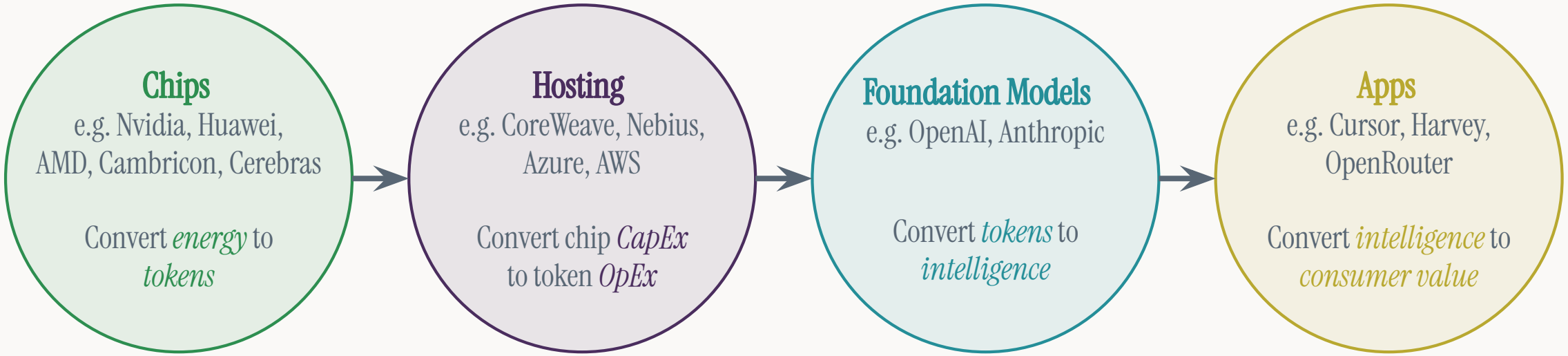
Sources: Exponential View analysis; OpenRouter; Epoch AI; METR; Artificial Analysis.
Note: ECI, METR & AAII indexed to average Jan 2025 scores for comparative purposes.

5 | Stack:

Where the value is captured

Revenue is concentrated, but apps and models are gaining share. Labs retain pricing only while they hold the frontier; the economic value of yesterday's frontier diminishes quickly into open weights. Where margin accrues across the stack will depend on technical advancements, which cut across competitive dynamics.

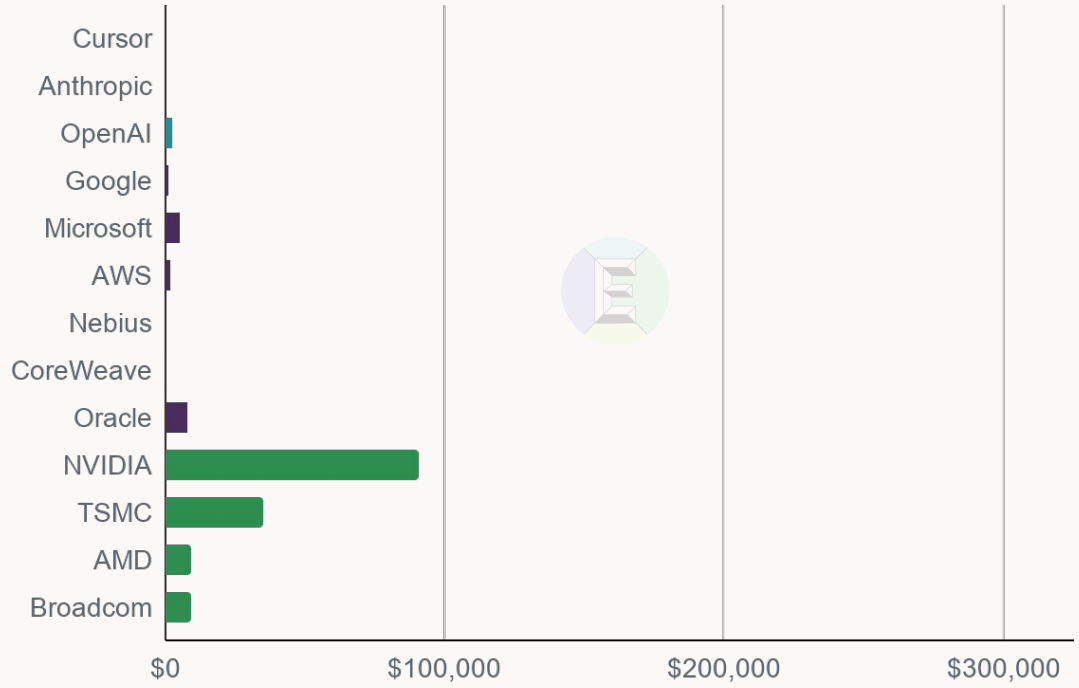
The stack turns capital and energy into cognitive work



Revenue is concentrated today, but the mix is shifting

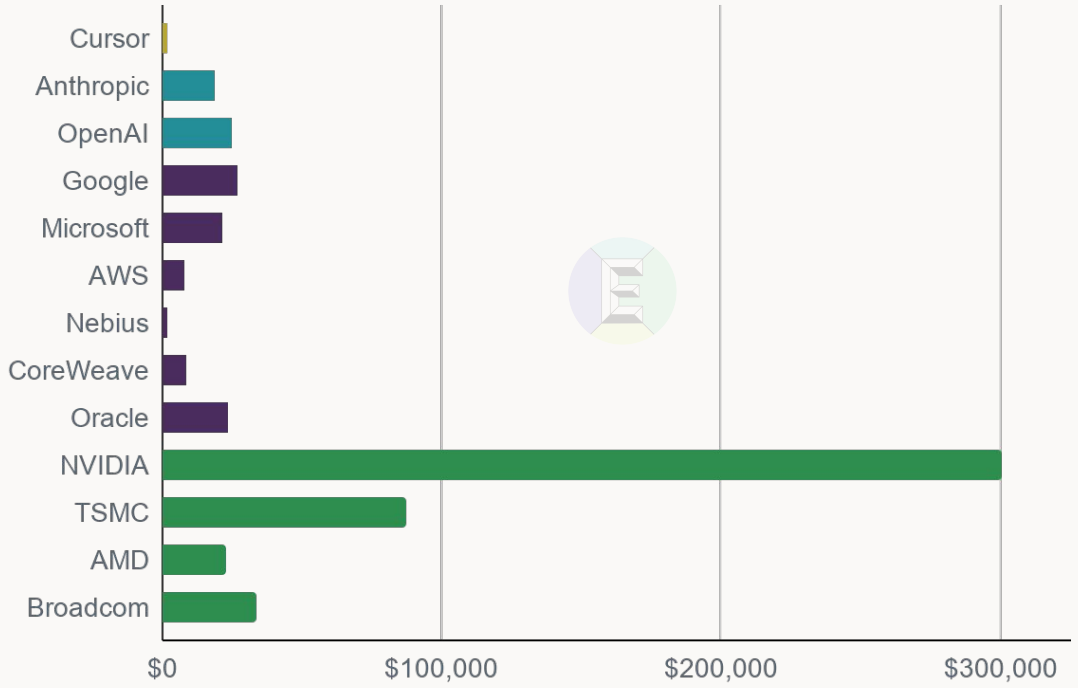
Annualized GenAI revenue

$million, Q1 2024



Annualized GenAI revenue

$million, Q1 2026



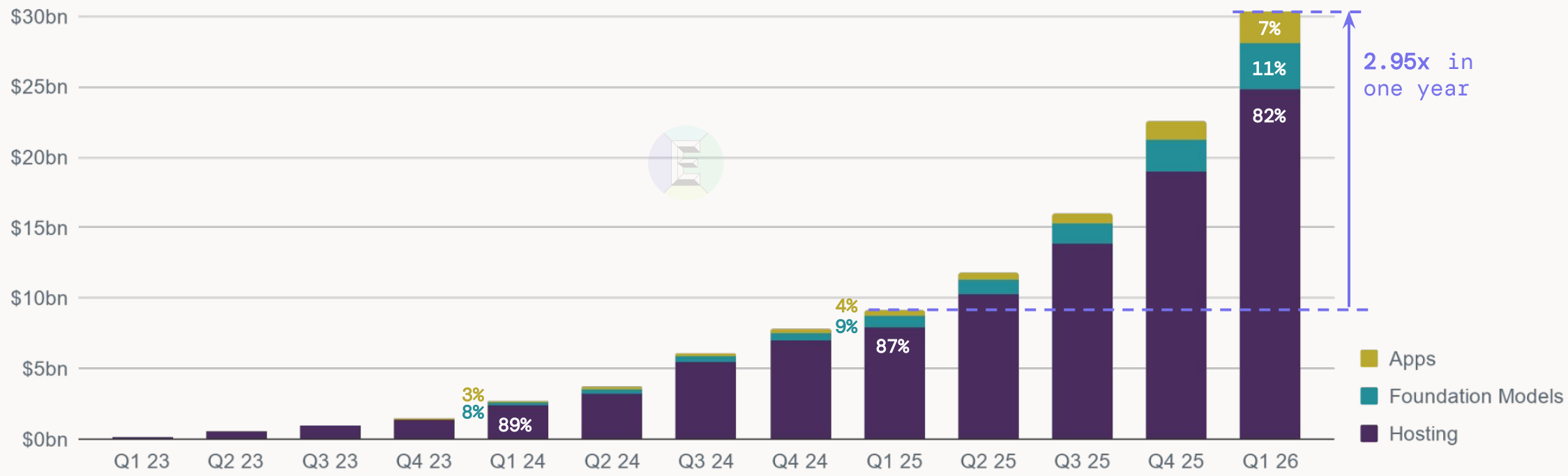
Apps · Foundation Models · Hosting · Chips

Sources: Exponential View analysis; company filings.
Note: Revenues not subject to deduplication adjustment.

Value is moving up the stack, towards apps and models

Quarterly GenAI revenues, deduplicated by layer to reduce double-counting

$bn per quarter

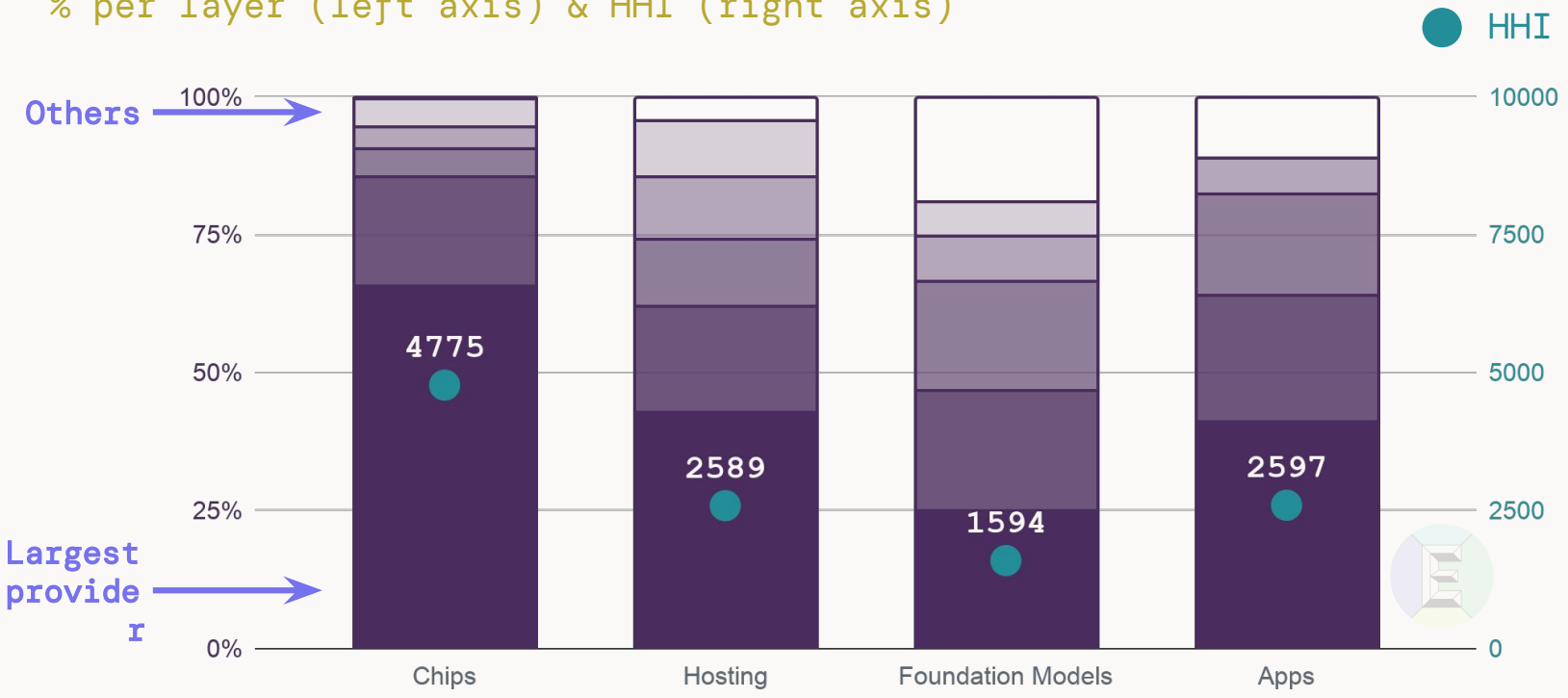


Source: Exponential View analysis.
Note: Global ex-China. Excludes chips (which are capitalized as CapEx by the hosting layer).
Figures may not sum to 100% because of rounding.

Pricing power follows competitive pressure, not stack position

Market share of leading companies & Herfindahl-Hirschman Index (HHI)

% per layer (left axis) & HHI (right axis)



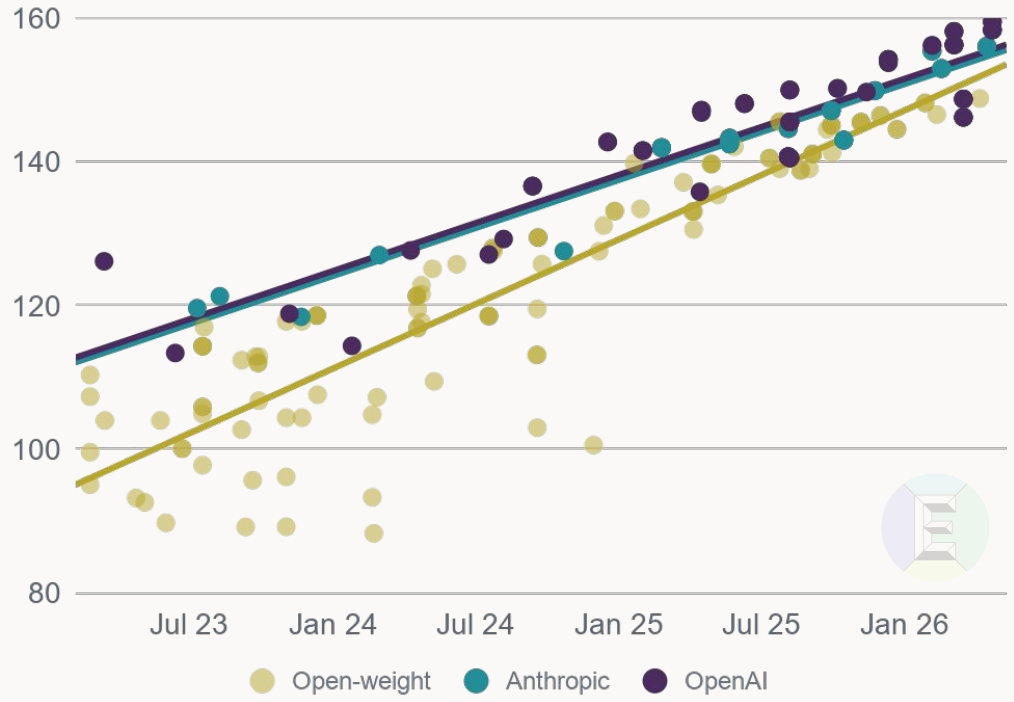
- Upstream suppliers price tokens to **capture all the margin available**, absent competition downstream.
- Nvidia is the largest chip provider, but **vertical integration by AWS & Google into custom silicon** may reduce its prominence.
- FM revenues are concentrated with OpenAI & Anthropic, but open-weight models offer **low-cost competition for quality tokens.**

Source: Exponential View analysis.
Note: Hosting & Apps are defined by share of revenues. Foundation Models are defined by share of tokens (due to open-weight competition). Chips are defined by share of compute (H100-eq) due to vertical integration of dedicated hyperscaler chips. Herfindahl-Hirschman index is a measure of market concentration.

Frontier labs can defend premium pricing, for now

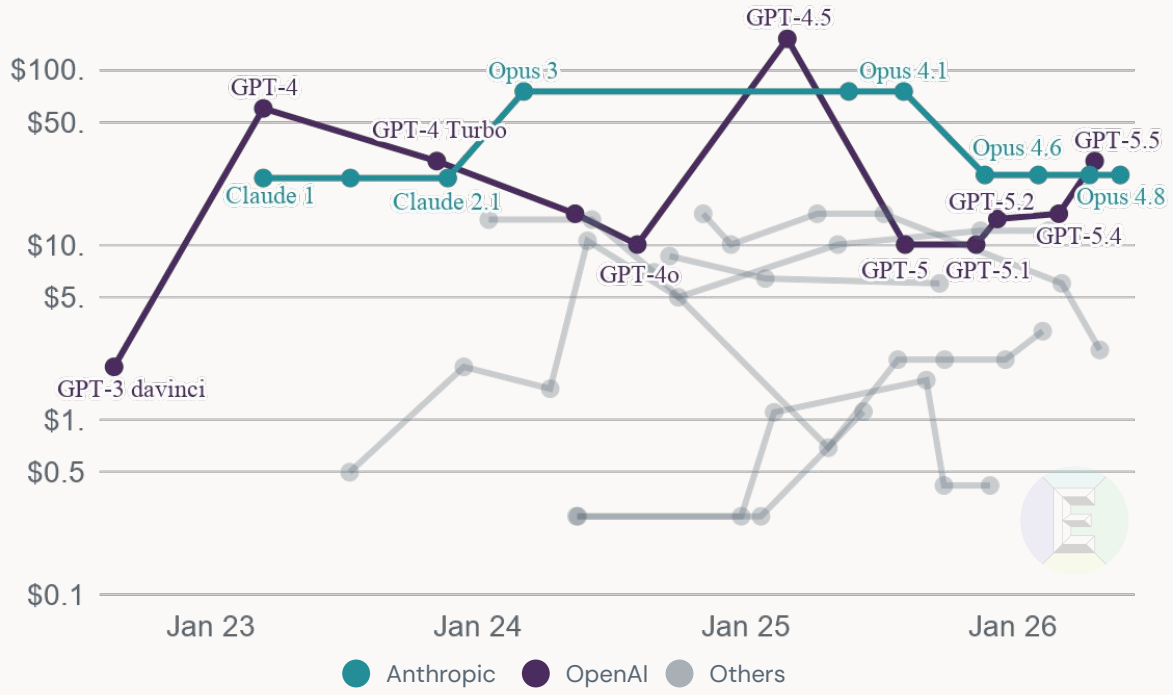
Epoch Capabilities Index

Score at release, OpenAI/Anthropic/open-weight



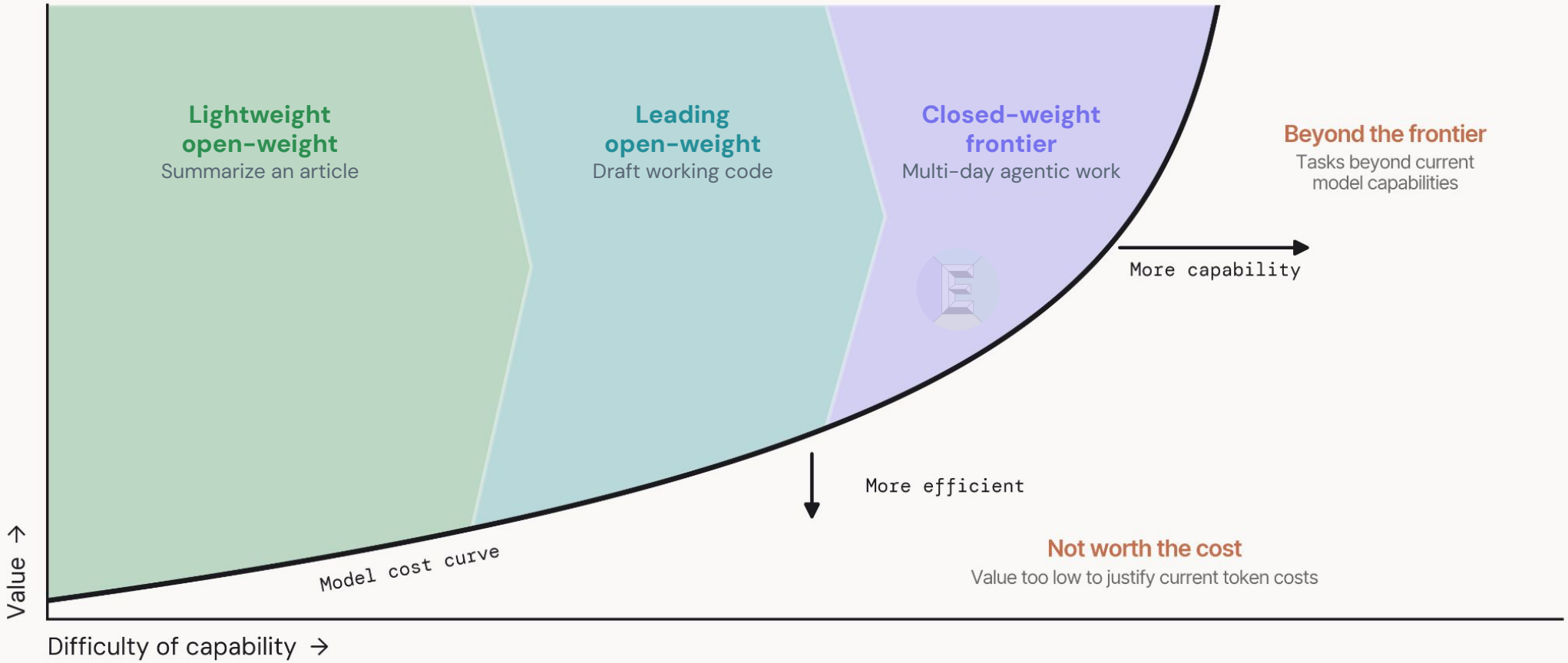
Output price

$/million output tokens



Sources: Exponential View analysis; Epoch AI.

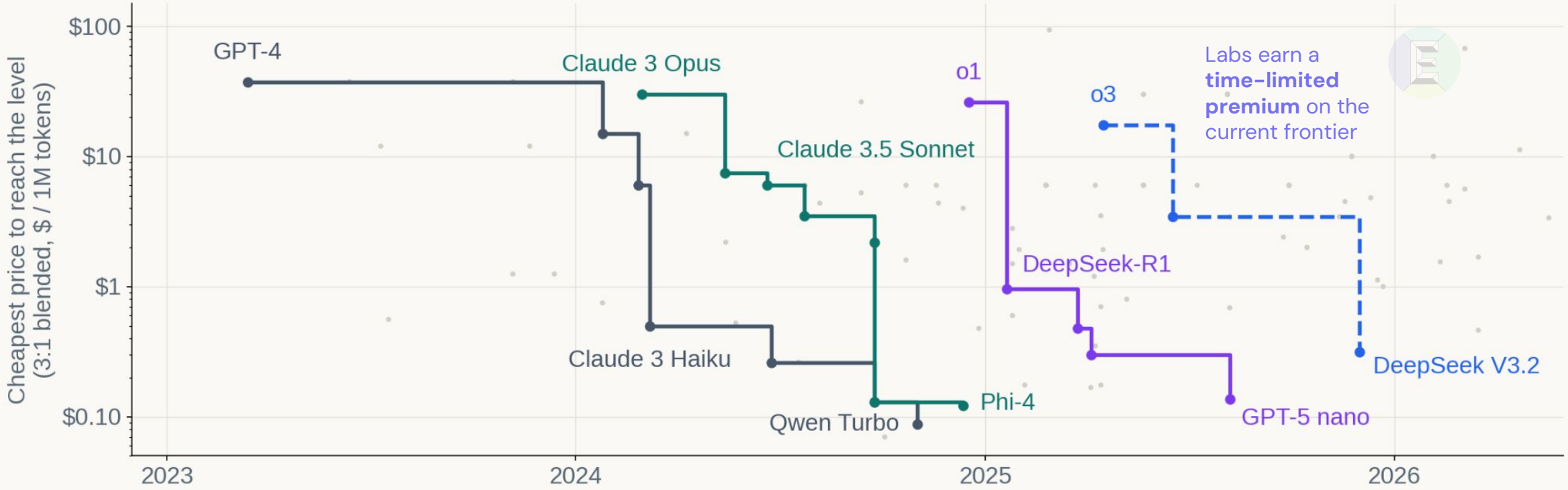
Labs must outrun open-weight commoditization to hold margin



Last year's frontier is commoditizing fast

Price per capability frontier

Blended price $, grouped by performance at GPQA Diamond (PhD-level science)

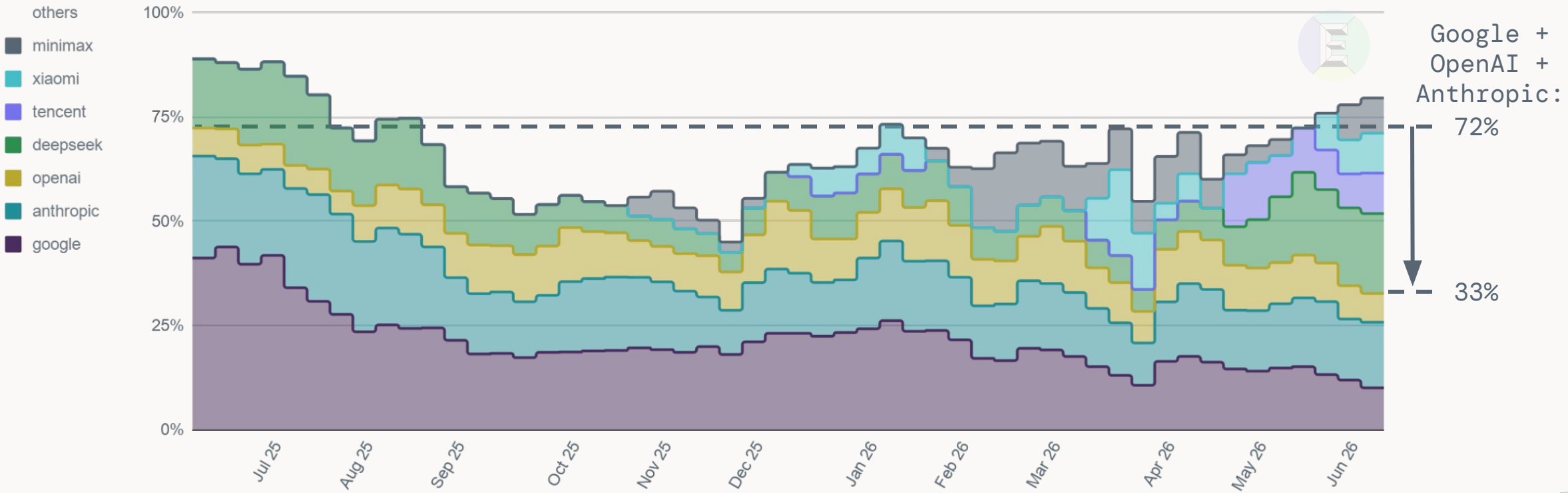


Sources: Exponential View analysis; Epoch AI.

Among self-selecting OpenRouter users, token share is moving to open-weight

Weekly OpenRouter token share

% per model author



Sources: Exponential View analysis; OpenRouter.
Note: While OpenRouter is not a cross-section of the market, its data shows the behavior of self-selecting “model-routing” users.

Under pricing pressure, labs push into apps and infrastructure

Building vertical apps: Law

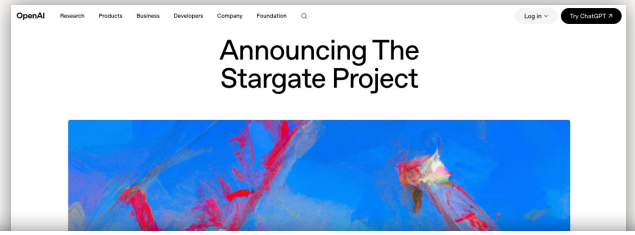
LEGORA

Clio

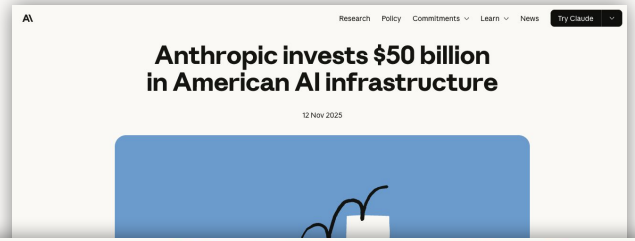
Codex for Legal

Claude for Legal

As Anthropic put it in its announcement: "An engagement might begin with the company's engineering team sitting down with clinicians and IT staff to build tools that fit into the workflows that staff already use... Engagements like this will run across mid-sized companies across industries, each shaped by the people closest to the work."



The Stargate Project is a new company which intends to invest $500 billion over the next four years building new AI infrastructure for OpenAI in the United States. We will begin deploying $100 billion immediately. This infrastructure will secure American leadership in AI, create hundreds of thousands of American jobs, and generate massive economic benefit for the entire world. This project will not only support the re-industrialization of the United States but also provide a strategic capability to protect the national security of America and its allies.



Today, we are announcing a $50 billion investment in American computing infrastructure, building data centers with Fluidstack in Texas and New York, with more sites to come. These facilities are custom built for Anthropic with a focus on maximizing efficiency for our workloads, enabling continued research and development at the frontier.

Sources: Exponential View analysis; TechCrunch; OpenAI; Anthropic.

Compress every layer, and consumers capture the surplus

Today

Scenario 1: **Open-weight models catch up**

Labs unable to charge license fee for frontier models

Tokens are cheaper for apps to consume & hosting to serve

Foundation model labs build application layer and their own infrastructure

Scenario 2: **General-purpose models**

Frontier models can replace AI application/integration layer

Frontier model labs absorb integration work; generic "wrapper" pricing power is squeezed

Apps defend with proprietary data, domain-specific workflows & own models

Scenario 3: **Reduced compute needs**

Models become smaller, and compute more efficient

Models and apps cost less to run, growing the applicable market

Some local hosting, with chips distributed to edge devices

Scenario 4: **All 3 combined**

The perfect storm of efficiency improvements

Cost reductions grow the applicable market

Significant consumer surplus as benefits delivered beyond price

Genuine demand and price elasticity of demand:
Cost reductions grow the market and result in increased consumer surplus.

AI demand is more revenue-validated than
any prior platform shift.
The investment case comes down to whether falling prices can
move enough token volume to earn a return on CapEx.

Methodology: How we count revenues, CapEx, tokens

What we count in, and what we exclude

- We count **revenue** at every layer of the stack: apps (subscriptions and AI-first software), foundation-model APIs, and AI cloud and compute sold as discrete services. Each layer represents real spend by a paying customer. A company that operates across layers (e.g. foundation model providers with customer-facing apps) has its revenue split across them. We count global ex-China revenues, and exclude chips and hardware (a cost to the compute layer, not a customer payment), AI features in legacy software, advertising uplift from AI (primarily Alphabet and Meta). We also exclude CapEx and financing from measures of revenue.
- **CapEx** is counted as the AI-attributable portion of the seven stack-builders' infrastructure spend (hyperscalers and neoclouds), including both cash PP&E and leases.
- **Tokens** are counted as every token processed, input and output, across all major providers and surfaces.

How we deduplicate, and why

- **Revenue** is counted at each layer but never summed across them: $100 of app spend that sends $60 to a model provider which in turn spends $30 on cloud hosting for inference is attributed (in line with added value) $40/$30/$30 to sum to the same $100 without double-/triple-counting, which would otherwise result in an erroneous $190 figure.
- **CapEx** is counted once, on the balance sheet of the entity that actually owns the asset. Compute that is jointly leased, or rented by a foundation model provider from a hyperscaler, sits with the owner-operator: not also with the renter.
- For **tokens**: when inference is served by a foundry running another company's foundation model, those tokens are attributed once, to the model actually run, so a model offered through foundries isn't double-counted.

Sources

All figures are built **bottom-up from primary and specialist sources**, and triangulated against top-level estimates and proxies. **Revenue and CapEx** are grounded on company filings (SEC 10-K, 10-Q and 8-K) and executive disclosures, cross-checked through cloud attribution (e.g. Azure and Bedrock) where a private company's revenue appears in a public firm's accounts. Our systems daily scan and crawl available sources to create, maintain and improve the breadth and depth of our data, with source **attribution and confidence grading**. This high-quality dataset is used to build up **full financial models** (including P&L) for major companies, and driver-based models for smaller companies.

CapEx's AI share is carved per company and reconciled against silicon (chip providers' revenue), build cost, segment composition, and sell-side research. **Token volumes** are reconciled from executive statements, third-party sampling and analysis, traffic volumes, and triangulated against revenues and available compute capacity.

Authors

Azeem Azhar
Founder

William Gildea
Product Manager

Hannah Petrovic, PhD
Senior Researcher

Nathan Warren
Senior Researcher

Marija Gavrilov
Managing Director

We welcome feedback and contributions at aieconomy@exponentialview.co

For advisory requests and institutional inquiries, please contact helen@exponentialview.co

Follow our analysis on Exponential View (www.exponentialview.co)

Subscribe to Exponential View to receive our research in your inbox each week