SECOND ORDER

AI, Capital & Work · A Second Order research briefing

ISSUE 02 · JUNE 12, 2026

The Wrong Denominator

Since Issue 01 we have held that the binding constraint on AI is cost and diffusion. This week we push that argument to the unit itself. A token is priced by the seller, and it is a poor stand-in for the value it delivers to the buyer.

IN THIS BRIEFING

Tokens measure lab revenue, and buyers pay for value

THE THROUGH-LINE

Since Issue 01 we have read AI through cost and diffusion. Token consumption repriced the semiconductor complex and underpins a near-trillion-dollar IPO runway, and priced per completed task the cost of intelligence is collapsing. Both are true, and the gap between them drives the next 18 months.

WHY NOW

Flat-rate pricing ended June 1 (GitHub) and Anthropic filed its S-1 the same morning. The token spend index rolled over after doubling, and on June 10 Apollo's co-president called token talk “a lot of BS.”

WHAT YOU'LL LEARN

Why the bills rise as the unit cheapens, what enterprises cut to keep AI, what the buildout does to the GDP arithmetic (we ran it), and where the dollars pool when good enough will do.

Govern cost per completed task ahead of adoption. Token budgets have begun to behave like headcount budgets, capped and forecast and fought over.

THE CLAIM

Both sides of the standoff are right, in different units

“I think tokenmaxxing and token talk is - it’s a lot of BS, honestly. … Prices are collapsing per unit of IQ, if you did it that way.”

John Zito, co-president, Apollo · Morgan Stanley US Financials Conference, June 10, 2026

Goldman's desk named the metric five days earlier, useful task completion per watt and per dollar.

$37.50 to $0.175

per M tokens for the cheapest GPT-4-level model, in 23 months (Epoch AI; Exhibit 2)

about 13x

growth in average business token spend since Jan 2025 (Ramp); a typical agent job burns about 96,000 tokens (SemiAnalysis)

85%

of Q1 2026 GDP growth came from the IT-investment categories (Second Order calculation; Exhibit 5)

Token expenditure says little about value created, but it measures the labs' revenue directly, and that gap sets the next 18 months.

Seventy days repriced the cost of AI

EXHIBIT 1

Seventy days from tokenmaxxing to tokenpanic

Key repricing events, April 2 to June 11, 2026

Apr 2	OpenAI Codex moves to usage-based billing	
Apr 27	Developer backlash to Copilot repricing: “pay the same, get less”	
May 19	Google Gemini moves to usage-based billing	
May 26	Uber burns its full-year 2026 AI budget in four months	
May 28	Axios “AI sticker shock”; $500M single-month Claude bill reported*	
May 30	Amazon removes its internal token leaderboard (KiroRank)	
Jun 1	GitHub meters every Copilot plan · Anthropic files confidential S-1 · Walmart caps its agent	THE PANIC WEEKS
Jun 2	Uber caps agentic coding tools at $1,500 per engineer per month, per tool	
Jun 8	Citrini Research: “tokenmaxxing → tokenpanic”	
Jun 10	Apollo's Zito: price per unit of IQ, “token talk is BS” · Citadel's Tokenomics note	
Jun 11	Altman: token costs went from non-issue to “huge issue”; Microsoft pitches token-lean MAI models	

*Note: *Single-sourced (Axios, anonymous consultant); unconfirmed by Anthropic. Reported dates; leaderboard removal reported late May.*

Source: Company announcements; GitHub changelog; Bloomberg; Axios; Business Insider; Citrini Research; Citadel Securities.

SO WHAT

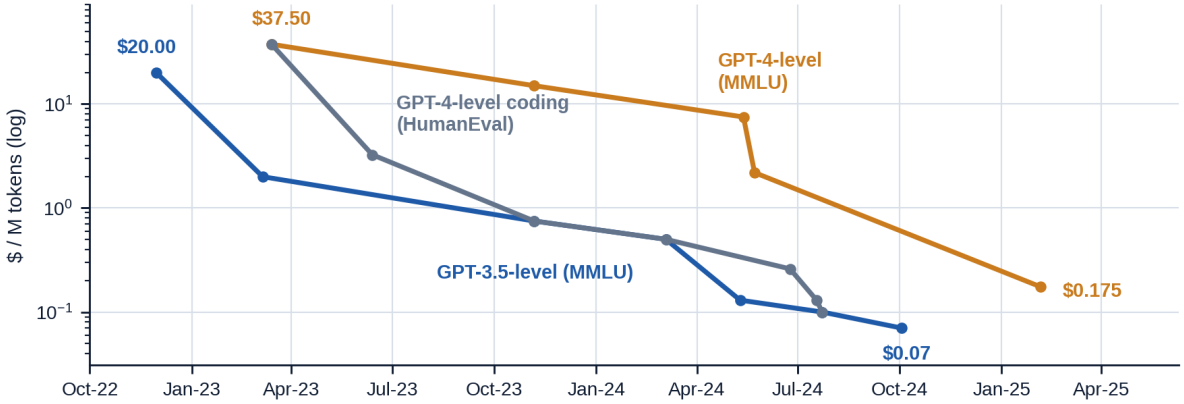
- All four major providers moved to usage billing within 70 days, and flat-rate pricing ended June 1.
- Uber capped agentic tools at $1,500 per engineer per month after burning its 2026 budget in four months.
- Buyers are metering and holding, with Amazon killing the token leaderboard, Walmart capping its agent, and Microsoft cutting Claude Code seats.

A fixed capability level deflates 100x or more

EXHIBIT 2

A fixed level of capability deflates 100x or more per year

$ per million tokens (log) for the cheapest model at or above each capability threshold



Note: Observed model prices only, 3:1 input/output blend; reasoning models excluded by Epoch's method. Series ends Feb 2025 (dataset vintage); no extrapolation drawn. Corroborating anchor (Stanford AI Index): GPT-3.5-class output fell $20.00 to $0.07 per M tokens in 18 months, 280x.

Source: Epoch AI / Artificial Analysis (CC-BY); data and vintage notes filed in data/i02_cost_curve/.

SO WHAT

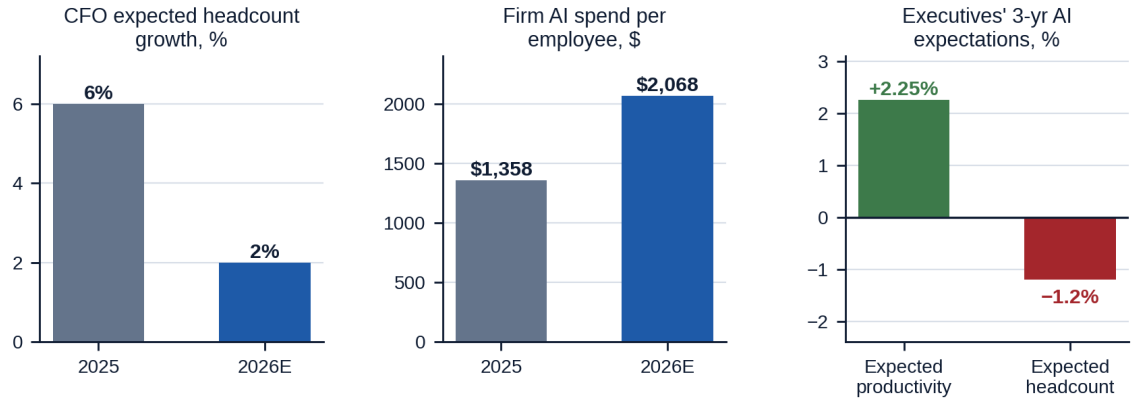
- Cheapest GPT-4-level model fell from $37.50 to $0.175 per M tokens in 23 months (Epoch AI).
- GPT-3.5-level fell from $20.00 to $0.07, the AI Index's 280x in 18 months.
- The steelman holds, and Gartner still sees bills rising, since agentic work burns 5-30x the tokens per task and consumption outruns the price decline.

Enterprises ration without retreating, and the capital swap runs through hiring

EXHIBIT 3

Compute budgets are eating headcount plans

CFO and executive survey readings, 2025 to 2026



Note: Panels mix different surveys, samples, and horizons; read as direction, not level. The $280B aggregate AI-investment figure in the text is the Atlanta Fed's flagged ballpark estimate.

Source: Gartner CFO surveys (Oct 2025, Feb 2026; n=303); Atlanta Fed Survey of Business Uncertainty macroblog (May 6, 2026); data/i02_surveys/.

SO WHAT

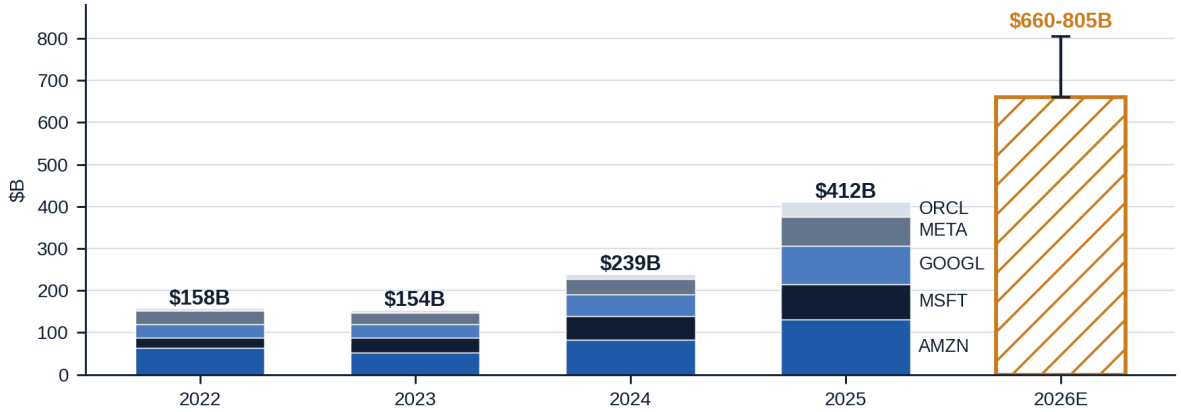
- About 60% of UBS checks call token costs a real issue, and none are slamming the brakes (via press coverage; see report caveat).
- They cut around AI, and CFO headcount growth expectations collapsed from 6% to 2% while 75% raise tech budgets (Gartner).
- Executives expect +2.25% productivity against a 1.2% headcount reduction, and AI spend per employee rose from $1,358 to $2,068 (Atlanta Fed).

The buildout nearly tripled in two years

EXHIBIT 4

The denominator debate is funded by a buildout that nearly tripled in two years

Big-5 hyperscaler cash capex (purchases of PP&E), $B by calendar year, with 2026 guidance range



Note: Cash PP&E only; excludes new finance leases, so levels sit below lease-inclusive compilations (Morgan Stanley's $805B basis). Oracle fiscal quarters mapped to calendar. 2026E bar = sum of company guidance (~$660B, estimate); whisker spans to Morgan Stanley's $805B.

Source: Second Order calculations from SEC EDGAR XBRL filings; 2026 guidance per company earnings calls; data and method in data/i02_capex/.

SO WHAT

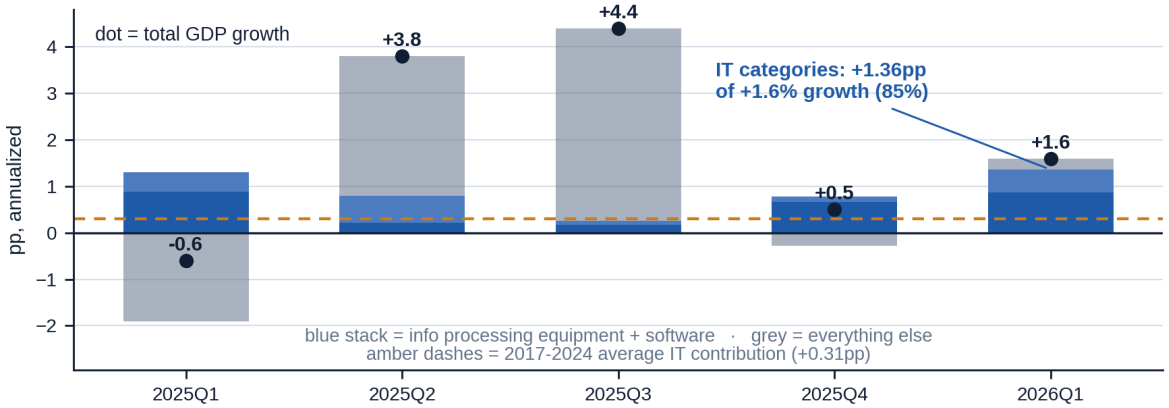
- Big-5 cash capex ran $158B (2022) to $412B (2025), computed from SEC filings and not a sell-side deck.
- 2026 company guidance sums to about $660B, and Morgan Stanley's lease-inclusive estimate is $805B.
- Slok finds about 60% of operating cash flow now goes to capex, and the top-10 S&P names trade richer than at the 1999 peak.

We tested the 75% claim ourselves

EXHIBIT 5

IT investment carried 85 percent of Q1 growth

Contributions to annualized real GDP growth, percentage points (BEA second estimate, May 28, 2026)



Note: Contributions are gross of imports (Q1 2025: equipment +0.89pp while GDP fell 0.6%, the tariff quarter). Not all IT investment is AI; data-center construction and power sit elsewhere and are excluded. Descriptive, not causal.

Source: Second Order calculations from BEA via FRED (A191RL1Q225SBEA, Y034RY2Q224SBEA, B985RY2Q224SBEA); data/i02_gdp_contrib/.

SO WHAT

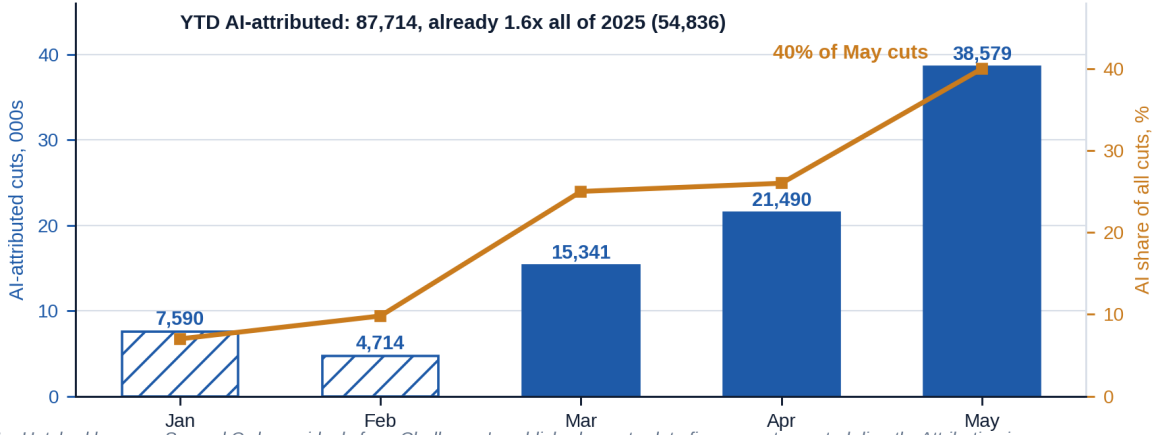
- Info-processing equipment plus software contributed 1.36pp of Q1 2026's 1.6% growth, 85% of it (BEA via FRED).
- The surge over the 2017-2024 trend is about 1.05pp, two thirds of all growth, and stripping it out leaves the economy at about 0.5%.
- Caveats cut both ways, since the measure is gross of imports, misses data centers and power, and includes non-AI IT. Our calc bounds the claim.

The narrative is firing ahead of the capability

EXHIBIT 6

AI is the No. 1 stated reason for US job cuts, up from 7 percent in January

AI-attributed announced job cuts (bars, thousands) and AI share of all cuts (line, %), 2026



Note: Hatched bars are Second Order residuals from Challenger's published year-to-date figures, not reported directly. Attribution is employer self-reporting; Challenger itself flags that AI may cover ordinary cost cuts.

Source: Challenger, Gray & Christmas monthly reports (Mar-May reported); Jan-Feb derived from YTD totals; data/i02_challenger/.

SO WHAT

- AI is the No. 1 stated layoff reason for a third month, at 38,579 May cuts, 40% of the total.
- Year-to-date AI-attributed cuts (87,714) already run 1.6x the whole of 2025.
- Self-attribution is the caveat, since firms cut ahead of demonstrated capability because investors reward the story (HBR), and some may cite AI to pay for AI.

WHAT WE'D DO

The operator's playbook for the next two quarters

1 · RE-DENOMINATE

Instrument cost per completed task for your top five AI workflows, and retire the “AI usage” KPIs. Amazon's leaderboard bought token burn and busywork, and no output.

2 · WRITE A ROUTING POLICY

Default to the cheapest model that clears your quality bar, and escalate to frontier only on documented ROI, a local, mid, frontier ladder.

3 · GOVERN TOKENS LIKE HEADCOUNT

Pair caps, alerts, and dashboards (Uber's $1,500 template) with workflow redesign, so caps do not freeze adoption at its least productive shape.

4 · RE-UNDERWRITE CONTRACTS

Stress-test anything priced on flat rates at 2-3x usage growth, and track renewal dates, since annual Copilot plans roll onto metered billing as they renew.

5 · SET THE LABOR-TO-COMPUTE RATIO

Decide what share of next year's capacity growth is hired against computed, and own that trade-off deliberately before you inherit it.

The capital swap is happening either way, and the only choice is whether you set the ratio or inherit it under bill shock.

WATCHLIST

What would change our mind

- **Silicon Data token index** · already 12% off the peak, and a sustained break lower separates deflation in the spend line from demand destruction
- **Challenger June report (early July)** · does AI attribution hold above 40% of cuts, or does the un-attributing begin
- **Q2 earnings (late July)** · hyperscaler capex guidance against softening token expenditure, the first quarter the two can diverge
- **BEA Q2 advance estimate (July 30)** · our Exhibit 5 bound updates mechanically, and the one-trade economy thesis hardens or cracks
- **Anthropic roadshow** · committed spend against meterable usage customers are learning to reduce
- **Falsifier** · a sustained break lower in the token index plus a hyperscaler guidance cut would flip our read from deflation to demand destruction

---

NEXT ISSUE · THE ENTRY-LEVEL PUZZLE Stanford says AI displaces 22-to-25-year-olds (16% relative decline). The NY Fed blames remote work (64% of the rise). Yale sees nothing. Someone is measuring wrong, and next week we show you who.