Artificial Intelligence: Physically Constrained, Financing as the Weak Point

Eight themes, three indicators and our positioning · As of 14 August 2026 · Research Ahead GmbH

Artificial intelligence is one of our six structural drivers* for the macroeconomic environment of the coming years. This summary organises our assessments of recent quarters by theme and gives, for each, the key points and our conclusion. Further themes will follow.

Core thesis

The investment boom is real, physically constrained and initially inflationary. The bottleneck sits not in demand but in manufacturing capacity and, increasingly, in financing. We regard the monetisation concerns of early summer as overdone: spending per token has fallen, while volumes processed and spending per employee have risen faster. Competitive pressure from China is limited in the Western enterprise market for now — constrained less by model quality than by regulation and data protection. We consider the correction that began in early June to be complete.

Positioning

- **Strategic.** Overweight AI-exposed equities, with an emphasis on the infrastructure and supplier layer. The base case stands as long as none of the three indicators below turns.
- **Tactical.** Currently positive. A renewed financing wave may weigh in the autumn: the Anthropic IPO in October and SpaceX lock-ups of roughly USD 500bn expiring by December. Q3 results are again likely to be dominated by the question of capex cuts.
- **Company-level risk focus.** Oracle and, at some distance, Meta are the most likely to adjust their investment plans. We expect no cuts at Alphabet, Amazon or Microsoft.
- **Macro.** The capex cycle and its financing needs go hand in hand with a higher neutral real policy rate and higher real bond yields. Monetary policy is too loose given macroeconomic conditions.

What would change our view

- Falling rental rates for GPU hours while token usage continues to rise strongly.
- Further extensions of depreciation schedules while rental rates fall.
- Falling prices for older DRAM memory chips such as DDR3 and DDR4.

The eight themes

- **1. The AI capex cycle.** Hyperscaler capital spending doubles to over USD 800bn in 2026. The boom reaches Taiwan and South Korea through the trade channel.
- **2. Is AI inflationary or disinflationary?** General-purpose technologies are inflationary in the investment phase and disinflationary only at a mature stage. We are at the beginning, and the impulse is increasing.
- **3. Chinese models and systems competition.** The widely cited 60% comes from a sample covering only around 1% of the market. Geopolitical competition results in a duplication of the technology stack.
- **4. AI usage and monetisation.** The price per unit falls, volume rises faster. Segmentation rather than a price war.
- **5. Capital requirements.** Supply pressure in bond and equity markets persists and lifts real yields.
- **6. Compute as an asset class.** Rental rates are rising across all three GPU chip generations at once. With a reference price, a residual value and financing vehicles, the preconditions for a distinct asset class are in place.
- **7. Strategic and tactical market view.** Tactical correction complete, strategic overweight confirmed.
- **8. The three indicators.** How we test the base case on a continuing basis.

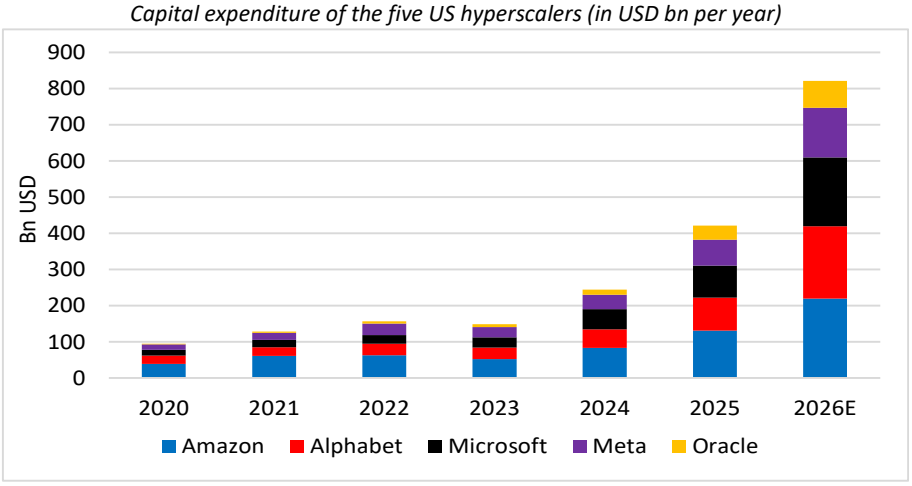
*: The other structural drivers are geoeconomics, China Shock 2.0, productivity, demographics, and fiscal policy and debt.

daniel.pfaendler@researchahead.com · www.researchahead.de · Wildenbruchstraße 50 · 60431 Frankfurt am Main · Germany
Research Ahead GmbH — independent macro and financial market strategy research for professional investors

1 The AI capex cycle

Investment in IT infrastructure and data centres has reached a scale that acts as a capex cycle in its own right, above all in the United States. The five large hyperscalers — Amazon, Alphabet, Meta, Microsoft and Oracle — invested around USD 420bn in 2025; more than USD 800bn is planned for 2026, close to a doubling within a year. Makers of compute and memory chips benefit most, alongside industrial suppliers to data-centre construction. Because the US imports a substantial share of its semiconductors, a parallel boom is emerging in Taiwan and South Korea. Demand for copper, silver and uranium rises with power requirements and electrification.

Capital expenditure of the five US hyperscalers (in USD bn per year)



Source: Company reports (quarterly filings), Research Ahead

Conclusion. *The build-out is carrying the cycle — through corporate investment in the US and through the export channel in North Asia. It is real, but it is concentrated: on few investors, few supplier countries and few stocks. If the high growth expectations disappoint, the sell-off would hit asset prices first and, via cuts to investment plans, the real economy afterwards.*

2 Is AI inflationary or disinflationary?

Artificial intelligence is a general-purpose technology. Historically such innovations are structurally disinflationary at a mature stage — they lower unit costs and expand supply. Before that, however, they pass through an investment and adoption phase in which demand for capital, power, hardware and specialist labour rises faster than productivity. That early phase is inflationary. Solow's paradox of 1987 resolved only in the late 1990s; electrification took some thirty years to deliver its productivity gains. We are clearly at the start of that curve, and the inflationary impulse is currently increasing.

Major technologies and the lag until disinflationary effects appear

Technology	Diffusion	Lag until productivity boost	Inflation regime in the mature phase
Steam engine	~1780–1850	40–60 years	Structurally disinflationary (industrial goods)
Electricity	~1890–1930	~30 years	Disinflationary (consumer goods, lighting)
Internal combustion engine	~1900–1950	20–30 years	Disinflationary (transport, logistics)
IT/internet	~1980–2010	~15–20 years	Strongly disinflationary (goods, ICT)
AI (today)	~2022–?	plausibly 5–15 years	open

Source: Research Ahead

US electricity consumption stagnated from 2007 to the pandemic and now runs about 8% above that period's average; consumer power prices rose less than 8% across the previous decade and more than 40% since — decoupled from gas prices. US producer prices for electronic components were deflationary from the mid-1990s to the pandemic; in July they stood almost 30% above a year earlier. The software component of PCE fell by roughly 5% a year from the late 1990s and now runs some 17% above the prior year.

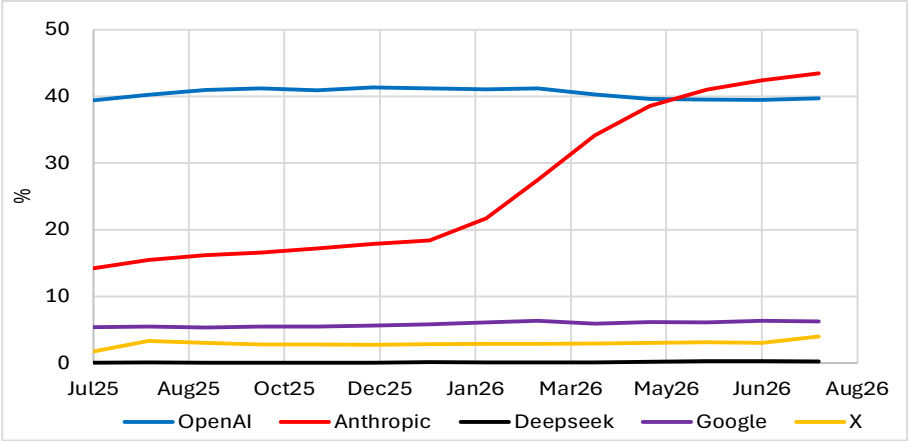
Conclusion. *AI is in its inflationary phase. Estimates of its contribution to US core inflation over the past twelve months range from 0.1 to 0.4 percentage points, depending on the channels included. The consensus expectation of a swift and durable return of inflation towards 2% underestimates this. We expect the balance to tip disinflationary only once productivity effects broaden, probably from 2028/2029.*

3 Chinese models and US–China systems competition

The concern that Chinese models could undermine the cash flows of US providers helped trigger the summer correction in AI equities. It rested on the rising token share of Chinese models on the aggregator platform OpenRouter — from under 10% in early 2025 to roughly 60%. But this captures only routed requests; anyone buying directly from OpenAI or Anthropic does not appear. Its share of total AI usage is around 1%, and the platform is popular chiefly with developers.

The enterprise market looks different. Card and invoice data from over 70,000 US companies show Anthropic at 43.5% and OpenAI at 39.7% adoption, with Chinese labs such as DeepSeek at 0.3%. Two markets are drifting apart: a price-sensitive developer market open to Chinese models, and an enterprise market tied to US providers for compliance and data-protection reasons.

Ramp AI adoption indices (share of US companies with AI transactions by provider)



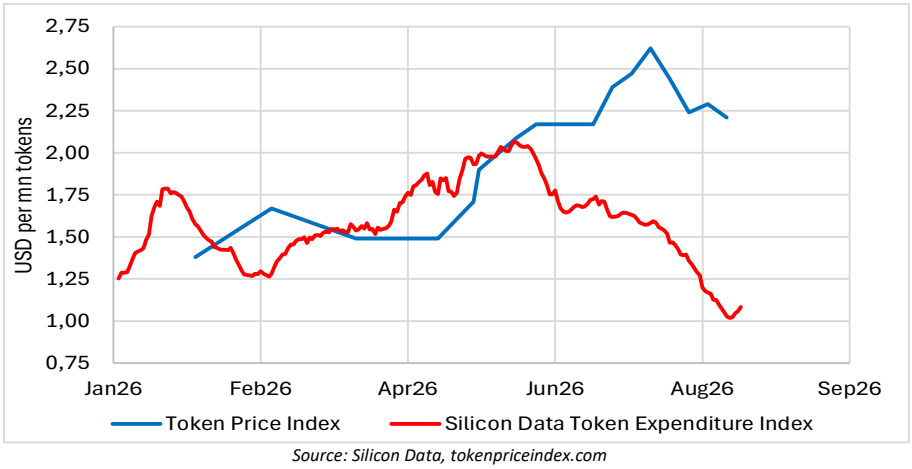
Source: Ramp

Conclusion. *Chinese competition is currently not a cash-flow risk for US providers, but it is a structural margin cap in the commodity segment below the frontier. The bifurcation of technology stacks supports growth and equipment suppliers in both blocs, yet lowers the capital efficiency of the system as a whole and raises the long-run risk of misallocation.*

4 AI usage and the monetisation question

To judge monetisation, price and volume must be separated. The widely watched token expenditure index has fallen since early June, but it is not a price index: it is the usage-weighted average price per million tokens, and it also falls when users move to cheaper models. The token price index, which measures average prices for frontier models, moved sideways. There is no price war between the labs, but a segmentation of the market — the top models hold their price while a commodity segment grows beneath them.

Silicon Data Token Expenditure Index and Token Price Index (in USD per mn tokens)



Source: Silicon Data, tokenpriceindex.com

Volumes keep growing. Alphabet processed around 9.7 trillion tokens per month in spring 2024, 480 trillion in May 2025 and over 3.2 quadrillion a year later. Revenues follow: the median of AI spending per employee has more than doubled this year. One caveat belongs here — none of these series is official, audited or consistently compiled. We use them as orders of magnitude, not measurements.

Conclusion. *The decline in unit costs is more than offset by volume growth. Cheaper tokens do not shrink budgets; they permit more applications. The open question is no longer whether monetisation happens, but where in the value chain the value accrues — over time it is likely to shift from infrastructure to vertically integrated adopters, for instance in biotechnology, robotics or defence.*

5 Capital requirements

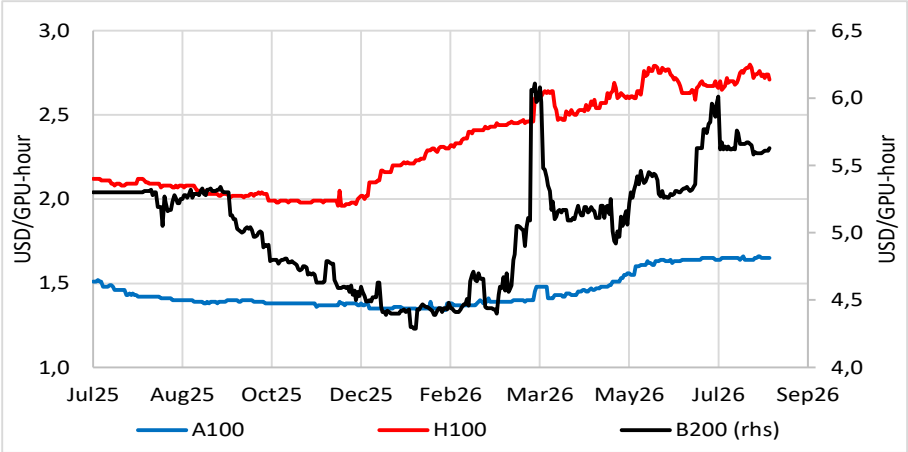
The flip side of high and rapidly rising investment is a high and rapidly rising financing need, and the source of that financing has shifted. Last year the hyperscalers funded capex largely through bonds; issuance is rising again this year and has already exceeded the 2025 full-year total, but it no longer suffices. Meta and Alphabet have scaled back share buybacks, and equity raising via US exchanges has risen sharply since June. Alphabet's placement of roughly USD 85bn was the largest equity transaction in US market history to that point. Since early June, AI-related equity transactions of around USD 220bn have been placed — against USD 270bn of total US exchange equity raising in all of 2025.

Conclusion. *Supply pressure in bond and equity markets will persist. The capex cycle and its financing needs imply both a higher neutral real policy rate and higher real bond yields. The most likely trigger for cuts to investment plans is constrained access to capital, not weaker AI demand. At the aggregate level we expect no cuts; at company level Oracle and, somewhat less, Meta are most exposed.*

6 Compute as an asset class

Rental rates for older graphics cards normally fall over time, particularly once the next generation becomes widely available. Since late last year they have risen across all three chip generations simultaneously — a pure volume shortage, because investment is rising rapidly and because compute chips serve two functions: new chips train frontier models, and as they age they migrate into inference. The hyperscalers have extended useful lives for servers and network equipment from three to 5.5–6 years since 2020.

Rental rates for NVIDIA compute chips (in USD per GPU hour)



Source: Silicon Data, Bloomberg, Research Ahead

The three preconditions for an asset class are now in place: a tradable reference price through compute futures based on daily rental indices; a demonstrable residual value, evidenced by rising rental rates even for five-year-old cards; and financing vehicles being assembled with major capital providers. The result is a structure familiar from real estate, aircraft or ships: recurring rental income, a tradable reference price, a residual value and debt secured against the asset.

Conclusion. *Compute is likely to establish itself as a distinct asset class. That brings external capital into the ecosystem rather than circulating existing capital, and it makes the residual value verifiable at a market price for the first time. The constraint: that residual value depends entirely on inference demand. Should it collapse, rental rates would come under pressure too. The indicators currently give no sign of this.*

7 Strategic and tactical market view

AI-focused equities entered a sharp correction in early June. Three factors combined: very heavy equity issuance via US exchanges; doubts about monetisation, triggered by improved Chinese models and falling spending per token; and leveraged positions built during the preceding bullish trend. The leverage has since left the market — leveraged ETF volumes fell from more than USD 50bn at the end of May to around USD 15bn two months later.

The market low coincided with Q2 results: Microsoft, Meta and Amazon announced no cuts to investment plans, and Alphabet had raised its plans slightly beforehand. Valuations support the view, with one qualification: the Nasdaq trades at a forward P/E of about 26 — not cheap, but not bubble territory. The question is therefore not the level of valuation but the durability of the earnings it rests on.

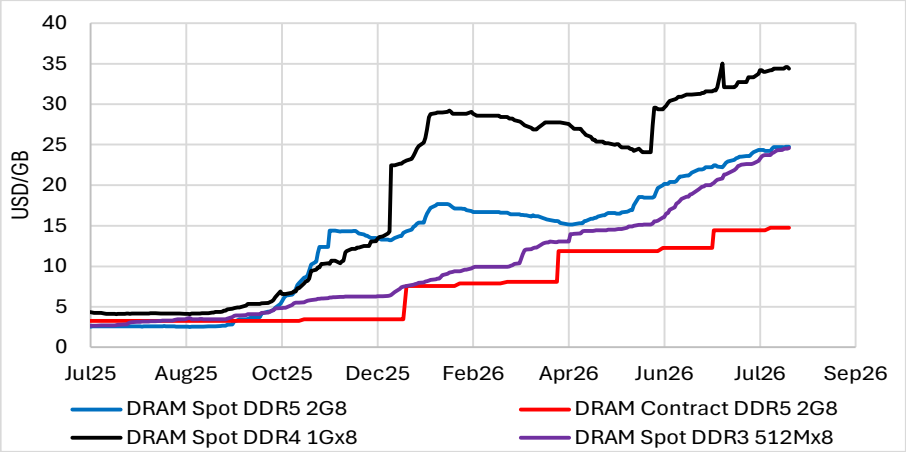
Conclusion. *We regard the correction that began in early June as complete. Strategically we maintain the overweight in AI-exposed equities with an emphasis on the infrastructure and supplier layer, and we keep this base case as long as the three indicators below do not turn. Tactically, a renewed financing wave could weigh in the autumn.*

8 The three indicators we test the base case against

AI is a new theme for markets and for the economy, and public, objectively compiled indicators — particularly official statistics — are largely absent. We therefore maintain our own set covering token prices, usage, provider shares, and memory and compute chips. The deeper reason is the reflexivity of the situation: rental rates rise because capacity is scarce; capacity is scarce because investment is high; investment is high because financing is available; financing is available because residual values are high; and residual values are high because rental rates rise. That loop sustains itself while it turns one way, and can equally run the other. At some point supply will catch up with demand. When that happens, we cannot say — which is precisely why we defined these three indicators.

- **Falling rental rates for GPU hours while token usage continues to rise strongly.** This would indicate that capacity has overtaken demand. The residual value of the installed base — and with it the financing of further investment — would be called into question.
- **Further extensions of depreciation schedules while rental rates fall.** This would be accounting cosmetics no longer covered by residual values, and reported hyperscaler earnings would be overstated to a degree that justifies a re-rating.
- **Falling prices for older DRAM chips such as DDR3 and DDR4.** This would be the first sign that crowding-out in memory fabrication is easing — an early signal for the entire memory cycle.

DRAM prices per generation, spot and contract (in USD/GB)



Source: inSpectrum, Bloomberg, Research Ahead

Conclusion. *The indicator suite does not replace a forecast. It tells us not when the cycle will turn, but where we would see it first. At present none of the three points in the critical direction: rental rates are rising across all generations, depreciation schedules have been unchanged since the spring, and prices for older memory chips continue to firm.*

Important information