Candoris

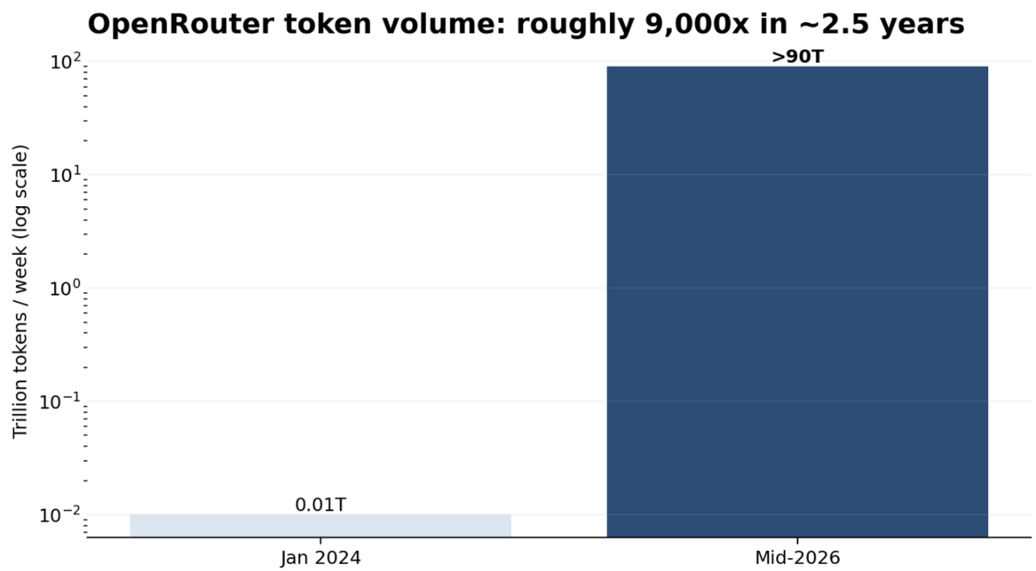
Access to the world's best investment strategies.

AI demand, productivity and the next macro regime

September 2026

Why this presentation starts with AI

- AI can reshape productivity, costs, margins, inflation and economic growth.
- Token demand shows adoption moving rapidly from experiments to always-on workloads.

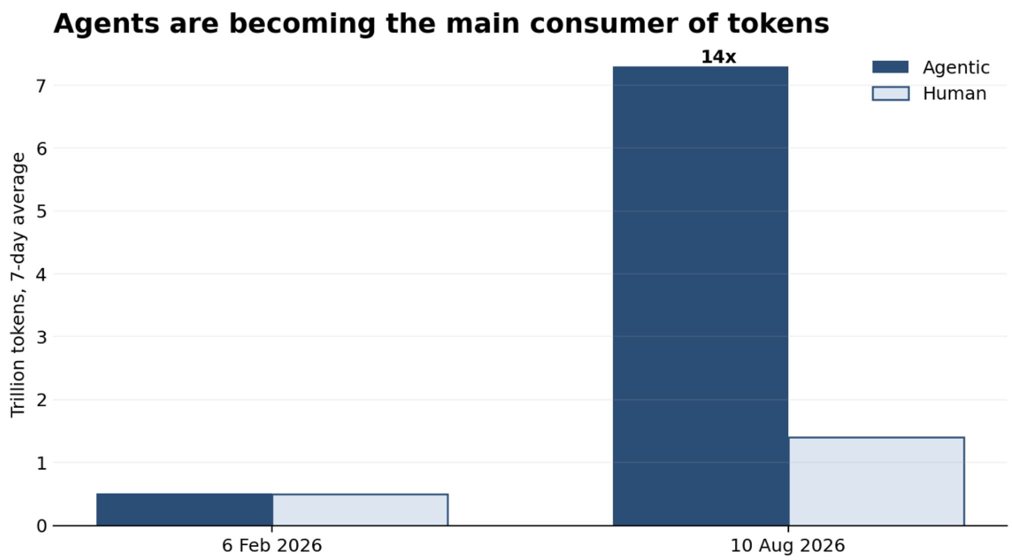


- ~10 billion tokens/week at the start of 2024.
- >90 trillion/week by mid-2026 in reported OpenRouter data.
- OpenRouter itself says it has seen at least 10x inference-volume growth every year.
- This is the core Jevons setup: lower cost expands the addressable set of tasks.

Source: OpenRouter, 19 Aug 2026; CryptoBriefing summary, 24 Aug 2026. OpenRouter is a partial market sample, not total global inference.

Agents are the main demand multiplier

- Agentic token usage increased around 14x between February and August.
- By August, agents used about five times as many tokens as direct human usage.



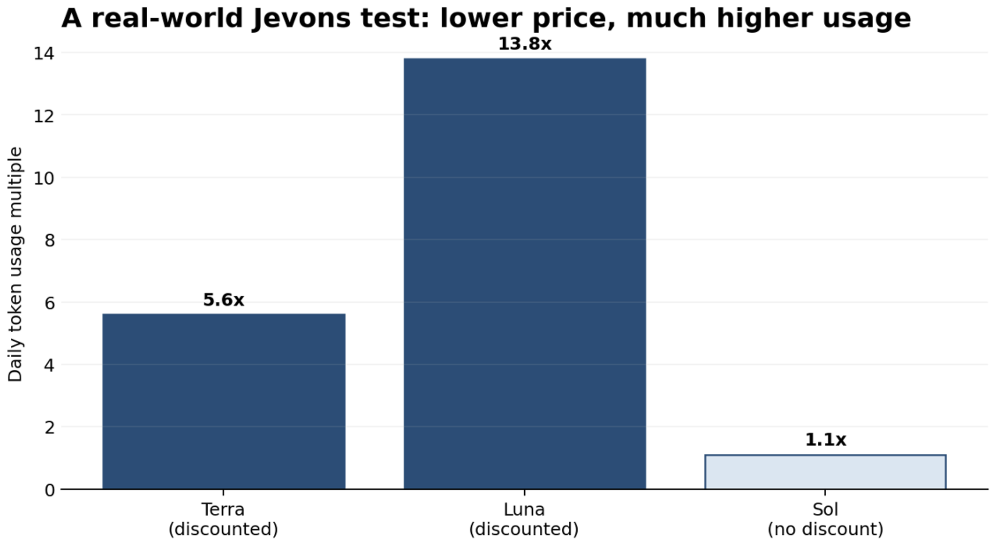
- Agentic usage: ~0.51T -> 7.3T tokens in six months, roughly 14x.
- Human usage: ~0.5T -> 1.4T, roughly 2.8x.
- Agents now consume around 5x the tokens of direct human use in this dataset.
- More than 85% of agentic token burn is cached prompt traffic: cheaper per token, but extremely memory intensive.

Source: a16z Charts of the Week, Aug 2026, using OpenRouter data; OpenRouter/ExoBrain, 28 Aug 2026.

Lower prices created much more usage

- Discounted OpenAI models saw daily token usage increase by 5.6x and 13.8x.
- A comparable model without a discount increased only 1.1x.

When price fell, usage exploded.



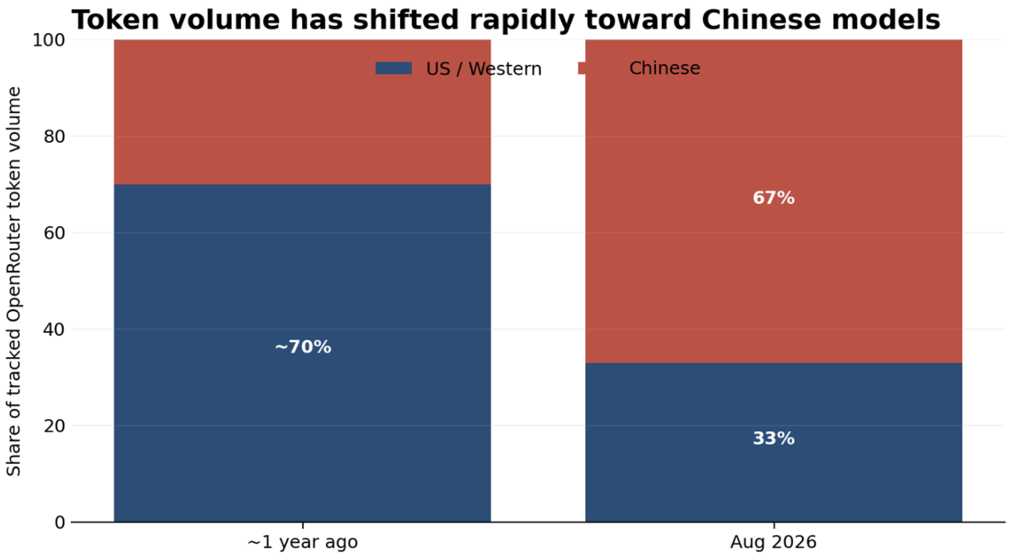
- OpenAI discounted Terra and Luna from 27 Jul to 14 Aug.
- Terra daily token usage rose 5.6x; Luna rose 13.8x.
- Sol stayed at list price and moved only 1.1x over the same window.
- Cheaper intelligence is not simply a margin story. It is a demand-creation mechanism.

Source: OpenRouter, “GPT 5.6 Discounts & Jevons Paradox”, 25 Aug 2026.

Token volume is shifting toward China

- Chinese-origin models accounted for approximately 67% of routed token volume in the latest August snapshot.

The shift happened much faster than expected.

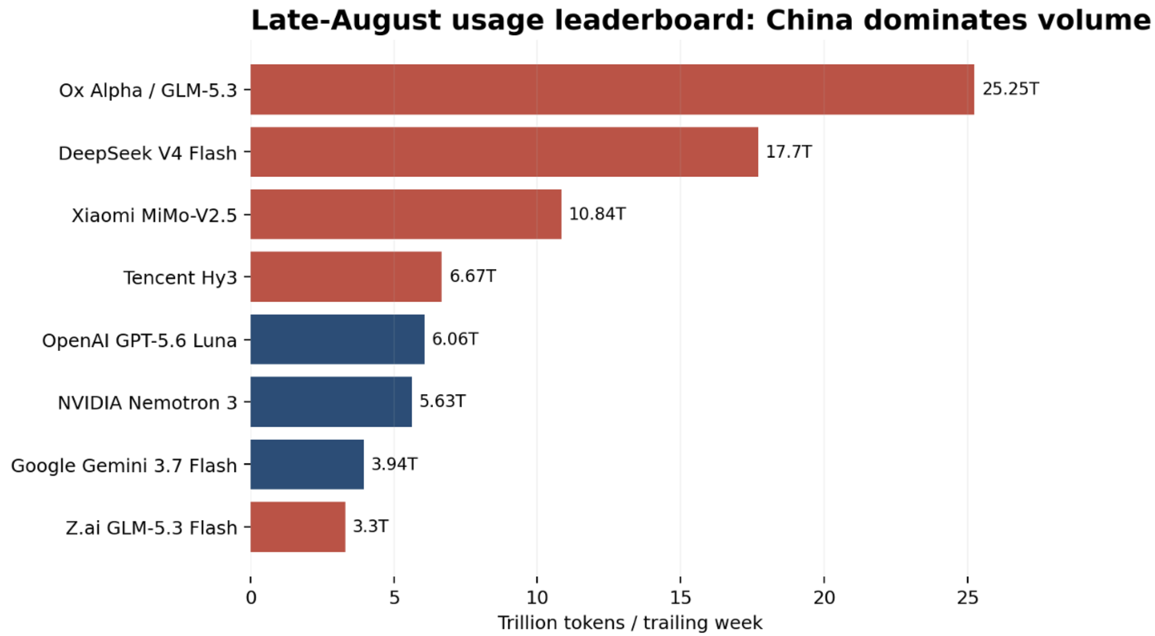


- Roughly a year ago, US models carried about 70% of OpenRouter traffic.
- Current independent tracking: Chinese-origin models ~67%, Western models ~33%.
- Chinese models crossed Western models in usage during 2026 and have continued to gain share.
- The key driver is not nationalism. It is price/performance plus open weights.

Source: Fortune, 21 Aug 2026; CodeSOTA, updated 31 Aug 2026, based on OpenRouter frontend statistics.

China's open-model ecosystem is scaling rapidly

Late August is dominated by Chinese-origin models by token volume.



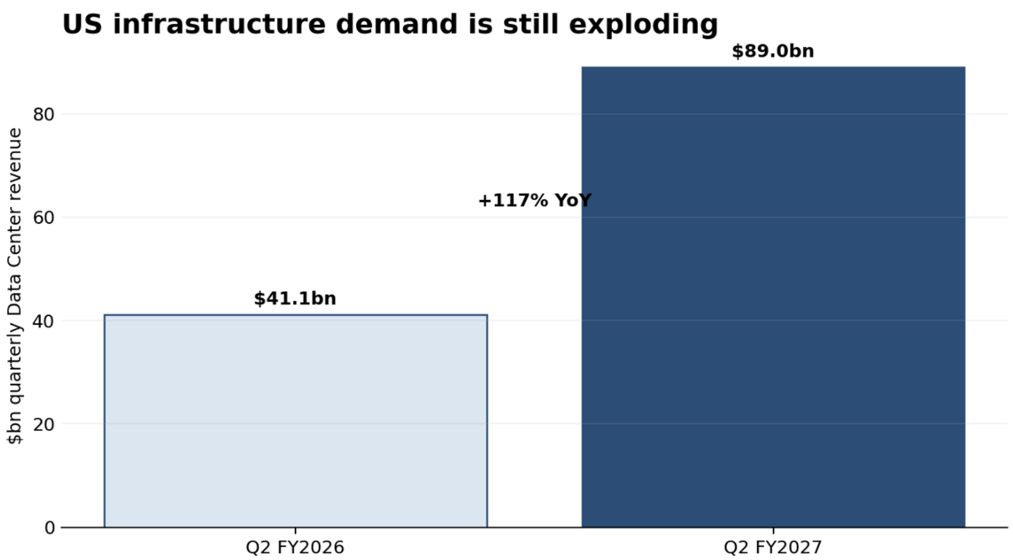
- Five of the eight largest models in this late-August snapshot are Chinese-origin.
- Ox Alpha was revealed on 27 Aug as Z.ai GLM-5.3-Flash and had become the most-used model of the week.
- DeepSeek, Xiaomi and Tencent also sit near the top.
- US leaders remain important, but token share is fragmenting quickly.

Source: What's Trending AI / OpenRouter snapshot, 29 Aug 2026; Business Insider, 27 Aug 2026.

China’s gains do not weaken the US compute-demand outlook

- China gaining model share does not mean US compute demand is weak.

China gaining model share does not mean US compute demand is weak.



- NVIDIA Data Center revenue reached $89.0bn in the July quarter, +117% YoY.
- Supply and capacity commitments increased from $119bn last quarter to $279bn.
- But China Data Center shipments were less than 1% of current Data Center revenue.
- NVIDIA explicitly warns that China exclusion helps local competitors build larger developer and customer ecosystems.

Source: NVIDIA Q2 FY2027 10-Q filed Aug 2026, SEC.

Conclusions

- AI demand remains underestimated. Lower costs are driving faster adoption and significantly more usage.
- AI agents are a major multiplier: AI increasingly uses other AI, making human user numbers a poor measure of demand.
- Infrastructure remains a key beneficiary, particularly chips, memory, networking, data centres and power.
- US model developers remain important, but open models and automated routing will put pressure on pricing.
- China is building a competitive AI ecosystem across low-cost models, domestic chips, cloud infrastructure and distribution.
- Chinese hyperscalers are starting to monetise AI, although high investment costs will weigh on near-term earnings.

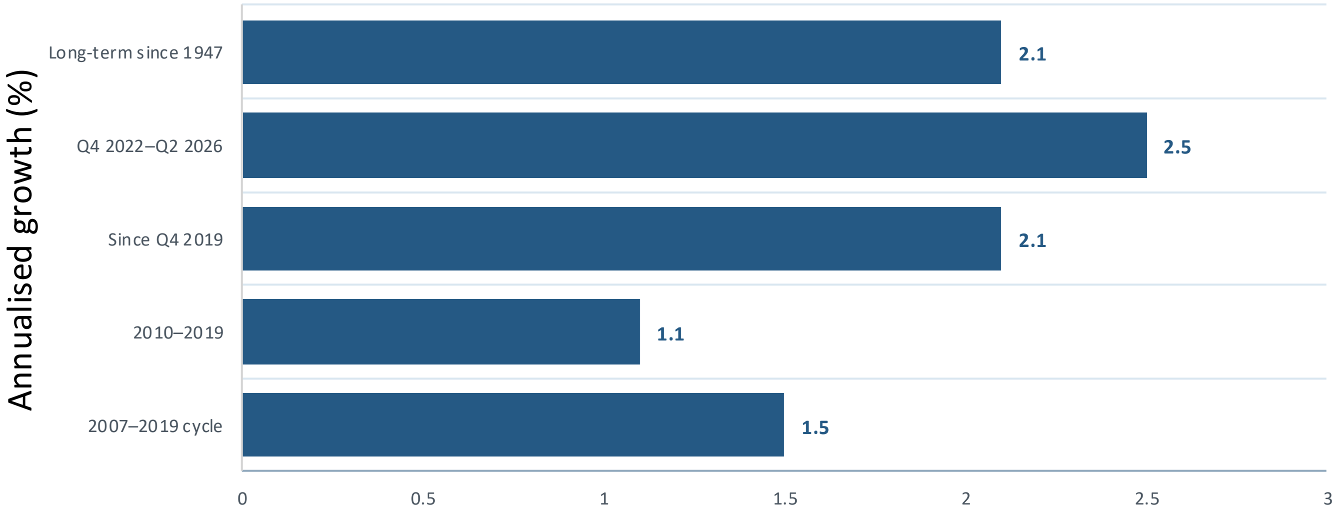
Macro view

- Higher labour productivity.
- Lower unit labour costs.
- Stronger corporate margins.
- Faster potential growth.
- Structurally lower inflation.
- **Overall: stronger growth with lower inflation.**

US productivity has moved above its pre-pandemic trend

- Since Q4 2022, output per hour has risen by 9.1%, equivalent to approximately 2.5% per year.
- Since late 2019, productivity has grown 2.1% annually versus 1.5% in the previous cycle and 1.1% during 2010–2019.

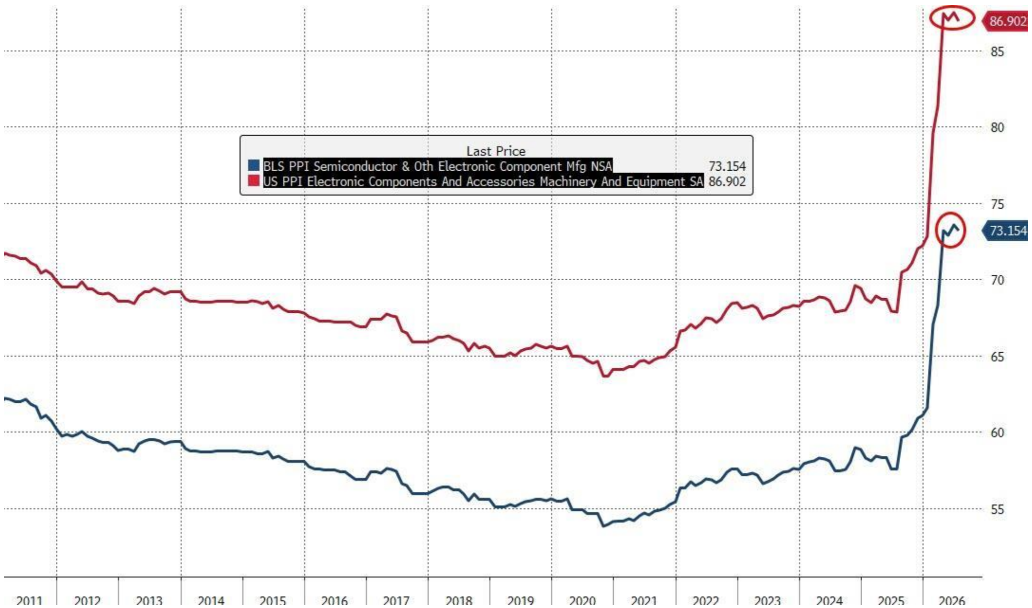
The improvement is meaningful, but revisions and other drivers mean it cannot yet be attributed primarily to AI.



Source: BLS and FRED. Post-ChatGPT rate calculated from the BLS productivity index.

Inflation

- Conflict with Iran or further escalation in Ukraine could temporarily lift inflation.
- Higher memory-chip and semiconductor prices add pressure.
- Commodity and weather risks, including a strong El Niño, remain.
- But the longer-term trend should remain AI-driven disinflation.

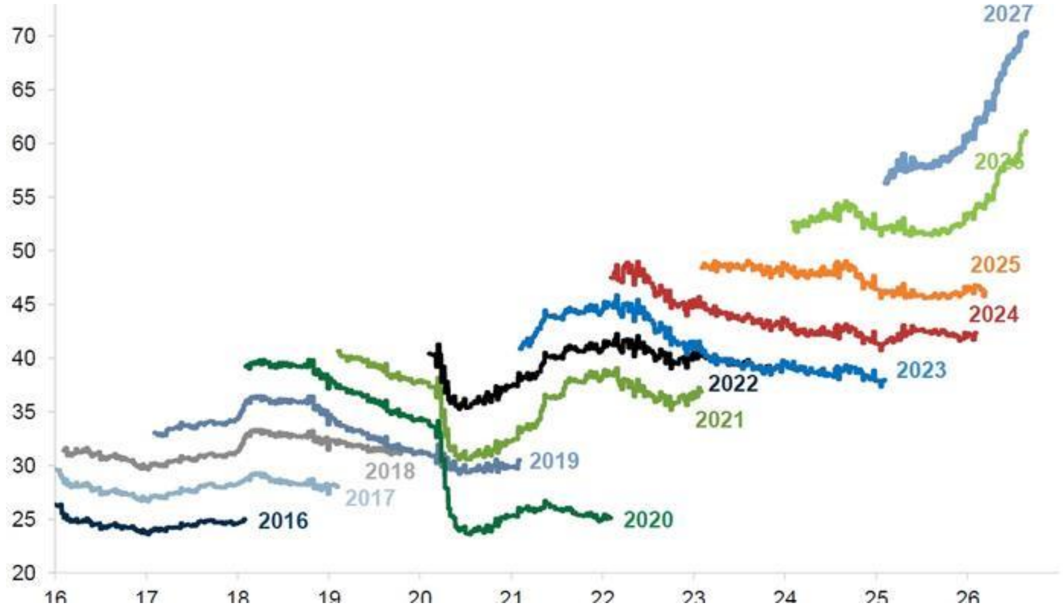


Source: Bloomberg and Zero Hedge, 13 August 2026.

Corporate earnings

- AI supports earnings growth and margins.
- The US leads AI investment.
- Benefits should gradually spread across sectors and regions.
- Europe is likely to lag the US and Asia.

Unlike in most years, 2026 earnings expectations have been revised higher and the 2027 trajectory is strengthening. Profits remain the key valuation support.

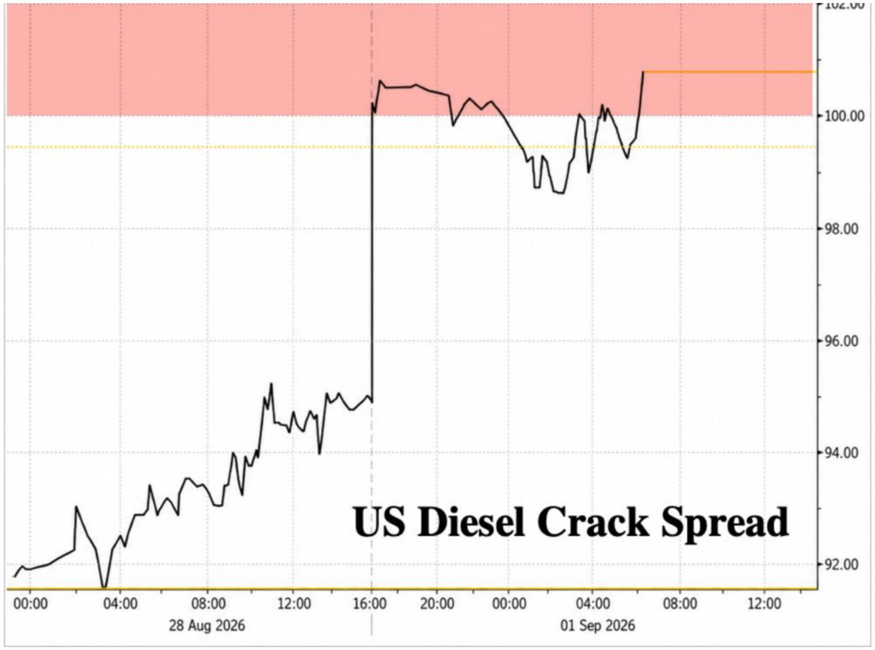


Source: Goldman Sachs and Zero Hedge, 28 August 2026.

Risks

- Geopolitical escalation and oil-price shocks.
- Commodity and weather shocks, including El Niño.
- AI valuation risk and market concentration.
- Regulatory intervention or tighter export restrictions.

Macro: geopolitics and energy can delay disinflation.



Source: Bloomberg, and Zero Hedge, Sept 1, 2026.

Bottom line

- AI remains the dominant structural investment theme.
- Token demand is growing faster than unit costs are falling.
- Productivity and earnings should strengthen and broaden.
- Medium-term inflation should decline, despite near-term supply shocks.
- China is taking model volume; the US still leads infrastructure and premium AI.
- Be selective: adoption can surge while model-layer margins fall.
- Watch token volume, pricing and workload complexity rather than benchmark leadership alone.