

MARKET VIEWPOINT

Beyond the Model

How falling token costs and rising usage are shifting AI value

Christian Grunt

Director, Private Equity

Miranda Silver

Analyst, Private Equity

The views and opinions expressed in this article are those of BCI and do not necessarily represent the views of its clients or investee companies. Any copying, modification or circulation is prohibited without prior written consent from BCI. The information in this document is provided as of the date hereof. Contents copyright © BCI® 2026.

Contents

1 Executive summary

2 How we got here: what changed, and why now

2.1 Overview of tokens and open- & closed-weight models

2.2 Open weight model vendors: closed for business?

2.3 As model choice grows, routing becomes more valuable

2.4 Rising inference demand creates winners beyond the model layer

2.5 AI infrastructure spending is high, and its funding is increasingly circular: more value AND more risk?

3 Where we are now

3.1 Cost per outcome matters more than cost per token

3.2 Token growth is becoming a physical infrastructure constraint

3.3 AI is changing which workers are needed, not simply reducing jobs

4 Future outlook

4.1 Enterprises will mix models, but compute and power demand will persist

4.2 Capital is moving beyond models into the infrastructure around them

4.3 Which industry is next?

References

1 Executive summary

Token prices are falling, but AI spend is rising: the winners will be the companies that convert growing usage into durable returns, not the model layer itself. For investors, identifying companies that can create value from AI and adapt as the focus moves from the model to workflow, routing and eventually industrial applications is key.

The Challenge

Token price is a poor proxy for AI cost or value. As agentic systems consume far more tokens per task, falling unit prices hide rising total spend, a measurement gap that matters for how enterprises budget and how investors allocate capital.

Approach

In this paper, we synthesized company filings, pricing data, and third-party research (Goldman Sachs, RBC, Stanford/Together AI), tested against real cases, including Klarna, Anthropic's own code metrics, and BCI's Project Ledger, and tracked recent financing and M&A activity to see where capital is actually moving.

Key Findings

Tokens have become cheaper but companies are spending more. Most forecasts show AI usage rising by as much as 24x by 2030 even as costs fall 60–70% annually. The open-closed model competition has open-weight models closing the gap. However, this may be short-lived as tightening licences are making access to the leading ones more conditional.

Routing has taken on a heightened role and is now a value layer in its own right as enterprises mix models by cost, quality, and task. In addition, inference hosting, deployment platforms, and AI-spend management tools are all becoming sources of value.

While not a new story, power, not compute, is the real bottleneck. The US needs ~139 GW of new capacity by 2030. AI is reshaping labour, but not erasing it. Labour demand is shifting toward skilled trades and higher-judgment roles. Overall, cost-per-outcome is the right metric for evaluating AI investments.

Why it matters

Durable value sits with companies that own the workflow, control routing, or supply the power and compute beneath rising usage, reinforcing our conviction in Vertical SaaS and extending it toward infrastructure and, longer-term, robotics.

Recommendations

- Evaluate AI exposure on cost-per-completed-task, not price-per-token.
- Track licensing shifts among open-weight labs for hosting and routing risk.
- Make power availability a first-order underwriting criterion for infrastructure deals.
- Extend the vertical-software playbook to early robotics opportunities.
- Monitor circular financing among hyperscalers and chipmakers to size true demand.

2 How we got here: what changed, and why now

2.1 Overview of tokens and open- & closed-weight models

A token is a unit a model uses to process and generate information. When a user submits a prompt, the model breaks it into input tokens and generates a response with output tokens. In English, 100 tokens are roughly 75 words. Token costs depend on how they are accessed: consumers pay a seat-licence for access, while developers and enterprises are charged by usage through the Application Programming Interface ("API"). Anthropic's pricing illustrates the difference between subscription-based consumer access and usage-based API access. API-based Claude Code usage averages \$150 - \$250 per developer per month, compared with \$25 per month for a Claude Team seat. The implied API cost is ~8x the monthly seat price.

Until November 2025, enterprise AI spending was concentrated in seat-based subscriptions such as Copilot, ChatGPT, and Claude. Since then, it has shifted to usage-based API pricing as enterprises embed AI into applications and deploy automated workflows at scale.

Models are broadly divided into closed and open-weight systems. Closed models, including ChatGPT, Claude, and Gemini, remain controlled by their providers and are accessed through an application or API. Open-weight models, including Llama, Mistral, and DeepSeek, can be downloaded, modified, and run on infrastructure chosen by the user (including computers and phones). This can provide greater control and customization. However, enterprises must still pay for hosting, engineering, security, and maintenance. Figure 1 summarizes the trade-offs and the companies competing across both approaches. Providers such as Fireworks AI, Together AI, and Baseten, as well as large cloud providers, reduce this burden by hosting open-weight models, centralizing API access and support. At scale, however, these hosts are among the businesses targeted by the new flagship-model terms, and the largest may need to negotiate directly with the labs.

FIGURE 1: AI MODEL LANDSCAPE, FUNDING AND SPONSORSHIP

SEGMENT	ILLUSTRATIVE PROVIDERS & MODELS	OWNERSHIP, FUNDING OR SPONSORSHIP	TYPICAL ACCESS & DEPLOYMENT	ENTERPRISE IMPLICATIONS
China open-weight	DeepSeek; Moonshot AI (Kimi); Alibaba (Qwen); Zhipu AI (GLM)	DeepSeek: Financed by High-Flyer, the quantitative hedge fund founded by DeepSeek founder Liang Wenfeng. Moonshot AI: Backed by Alibaba, Tencent, and IDG Capital. Alibaba: Qwen developed and funded internally. Zhipu AI: Backed by Alibaba, Tencent, Ant Group, and several state-linked investment funds.	Model weights can be downloaded and self-hosted or accessed through a managed API.	Materially lower token costs and greater control over hosting and customization; savings depend on infrastructure and operating costs. Licence terms scale with size: free at small and mid scale, a negotiated agreement with the lab past a revenue threshold, and internal use typically exempt. Some provenance, data-governance, and geopolitical diligence may apply.
Western open-weight	Reflection AI; Poolside; Arcee AI (Trinity); NVIDIA (Nemotron)	Reflection AI: NVIDIA-backed; reported March 2026 to be seeking \$2.5B at a \$25B valuation. Poolside: NVIDIA agreed to invest \$1B at a \$12B pre-money valuation and entered a separate \$6B licensing agreement. Arcee AI: ~\$50M raised; third-party valuation estimate of \$240M as of March 2025. NVIDIA: Nemotron developed and funded internally.	Model weights can be self-hosted or accessed through cloud providers, specialized hosting platforms, and APIs.	Greater customization, data control, and vendor diversification. However, the enterprise or hosting provider must manage infrastructure, security, monitoring, and updates. Total costs may therefore be comparable with closed models.
Closed frontier	OpenAI (GPT); Anthropic (Claude); Google (Gemini)	OpenAI: Microsoft has invested ~\$13.75B and holds a 27% stake valued at ~\$135B. Anthropic: Amazon has invested \$13B and may invest up to an additional \$20B; Google has committed \$10B and may invest up to an additional \$30B; Anthropic also raised a \$65B Series H in May 2026. Google: Gemini developed and funded internally by Alphabet.	Proprietary models are accessed through the provider's application or API, using subscription-based or usage-based pricing.	Easier deployment, integrated products, vendor support, and access to leading performance. Trade-offs include greater vendor dependency, less control over infrastructure and customization, and exposure to usage-based costs.

Source: company filings and announcements; BCI analysis.

In June 2026, a US export-control directive forced Anthropic to pull public access to Claude Fable 5 and Mythos 5 for 18 days, while OpenAI's GPT-5.6 Sol launched gated to government-vetted customers only. For enterprises now dependent on these tools for core operations, both instances made the vendor-dependency trade-off in Figure 1 concrete, resulting in accelerated interest in open-weight models that can be hosted and governed independently.

2.2 Open weight model vendors: closed for business?

Open-weight models are narrowing the gap with closed frontier systems, led by Chinese labs. The catch-up time has halved each generation, while GLM-5.2's token prices ran 65% to 82% below leading OpenAI and Anthropic models as of mid-2026; GLM-5.3 has since become the flagship and, unlike GLM-5.2, is not released under a permissive open licence. Chinese models now account for close to half of the tokens consumed through OpenRouter but only a single-digit share of model revenue. Lower prices expand use while compressing model economics. Most enterprises will therefore mix models: lower-cost options for routine, verifiable work and frontier models where performance and reliability justify the premium. This has pressured frontier labs' pricing and margins but should not reduce overall AI consumption, as seen in OpenAI's July 2026 price cuts of 80% for GPT-5.6 Luna and 20% for GPT-5.6 Terra. Instead, lower costs make more applications economical, shifting value toward software companies, routing and deployment platforms, and infrastructure providers as token demand grows.

Although the labs have not stated their motivation, the licence structure suggests a commercial objective: preserve the adoption benefits of open weights while gaining leverage to negotiate with the largest third-party hosts that resell access. Moonshot and Alibaba require these providers to negotiate separate agreements once group revenue exceeds \$20 million and \$50 million, respectively. Z.ai requires a security review only for Model-as-a-Service providers above \$10 billion. MiniMax goes further, requiring written permission for any commercial use of M2.7. Baseten and Fireworks could therefore be affected if they host these models and cross the relevant thresholds, while internal enterprise use remains exempt under the Moonshot, Alibaba, and Z.ai licences. The commercial terms have not been disclosed. The weights remain downloadable, but large third parties may no longer be free to resell access to them.

Distillation is one mechanism behind this catch-up. OpenAI made it a standard API feature in 2024, and DeepSeek's open-sourced distilled R1 variants (1.5B to 70B parameters) showed that smaller models can retain most of a frontier model's capability at a fraction of the compute. This lets providers push near-frontier performance onto cheaper-to-serve models, reinforcing the price declines above. This catch-up may get harder from here: matching the newest frontier gains increasingly requires distilling from expensive reasoning-and-reinforcement-learning training runs, one reason Chinese labs are now raising larger sums themselves.

Cost unpredictability has compounded the shift, as average enterprise API spend reached roughly \$385,000 per year by April 2026, with most IT leaders reporting unexpected AI-related charges. Combined with open-weight models running 60% to 90% cheaper than flagship closed alternatives, this pushed US enterprises to route over 30% of their OpenRouter traffic to open-weight models by February 2026, up from just 4.5% a year earlier.

COMPANY SPOTLIGHT: META'S AI RESET

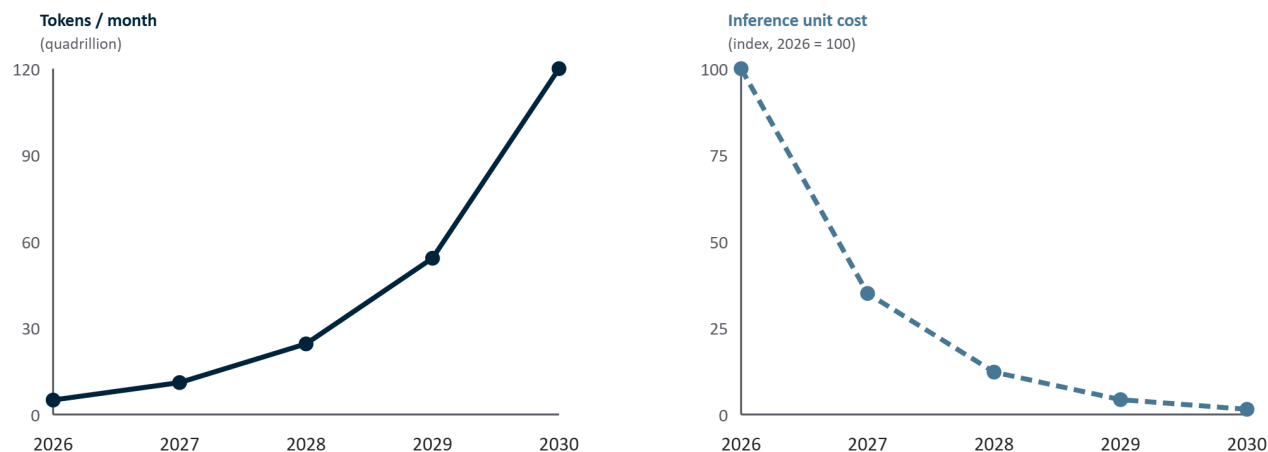
Meta illustrates both the potential and the risks of an open-weight strategy. Llama surpassed 1B downloads, giving Meta enormous developer reach, but Llama 4's mixed reception and benchmark controversy showed that distribution did not guarantee performance. Meta responded by investing \$14.3 billion for a 49% stake in Scale AI, recruiting its CEO Alexandr Wang and researchers from rival labs, and later cutting roughly 600 roles across its existing AI teams as it rebuilt around Meta Superintelligence Labs.

2.3 As model choice grows, routing becomes more valuable

Enterprise AI is becoming a hybrid market in which applications select models based on the task. GitHub Copilot routes simpler work to faster, lower-cost models and reserves more expensive reasoning models for harder problems, while Microsoft Foundry supports similar selection across proprietary and open-weight models. This makes routing based on cost, quality, speed, security, and licence terms increasingly valuable. Goldman Sachs expects monthly token consumption to increase 24x by 2030 even as inference costs fall 60% to 70% annually, while RBC found that Kimi K3 used 5x as many tokens to produce a comparable result. Lower prices therefore do not guarantee lower spending, particularly when agentic workflows trigger multiple model calls. Simply put, all tokens are not equal.

OpenRouter illustrates the value of this routing layer. Its platform connects developers to more than 400 models from over 80 providers, helping customers direct workloads and monitor spending. Routing works best on well-scoped, verifiable tasks. Shifting context to a cheaper model forfeits the cached tokens built up on the pricier one, a cost that's harder to justify for open-ended knowledge work.

FIGURE 2: TOKEN DEMAND RISES AS INFERENCE UNIT COST FALLS



Source: Goldman Sachs Global Investment Research; RBC Capital Markets; BCI analysis.

2.4 Rising inference demand creates winners beyond the model layer

Selecting the model is step one, and then every routed request must be executed on compute infrastructure. Training is how a model learns; inference is what happens when someone uses it. Training is compute-intensive upfront. Inference demand grows with usage, increasing demand for chips, servers, networking, and power.

Enterprises must optimize where the model runs. They can use a hyperscaler (AWS, Microsoft Azure, Google Cloud), rent capacity from an AI-focused provider such as CoreWeave, or run their own infrastructure. The best approach depends on workload volume, cost, latency, and data control. Deployment platforms help companies run and monitor models across these environments, while specialized systems compete on speed and efficiency.

This opens more ways to monetize AI usage beyond the model provider. Some earn directly by hosting and processing model calls; others provide the data, application, deployment, or financial-control layers that support that usage. The companies below illustrate this across the inference ecosystem.

FIGURE 3: MONETIZING AI USAGE BEYOND THE MODEL LAYER

COMPANY	DESCRIPTION	HOW TOKEN USAGE DRIVES REVENUE	LATEST FUNDING	DATE
Baseten	Inference infrastructure for deploying and running models in production.	Direct exposure. Model APIs are priced per million tokens, while dedicated deployments are priced according to compute usage. Revenue rises as customers generate more tokens and consume more GPU capacity. Baseten is a tracked per-token provider of newer open-weight models such as Kimi K3, so the weights it serves are no longer free to host now that the latest models carry licence terms above a revenue threshold.	Series F, \$1.5B	June 22, 2026
Databricks	Enterprise data and AI control layer with model routing and governance.	Direct and indirect exposure. Databricks charges for data and compute consumption through Databricks Units, while hosted foundation-model APIs can be billed per token. Higher AI activity increases consumption across the platform. Databricks is also a tracked per-token provider of newer open-weight models such as Kimi K3, so the weights it serves are no longer free to host now that those models carry licence terms above a revenue threshold.	Strategic financing, \$5.0B	August 13, 2026
Vercel	Application and deployment layer with a multi-model AI gateway.	Primarily indirect exposure. Vercel passes model tokens through at provider prices without a markup. It earns revenue from the hosting, compute, requests, enterprise features and developer infrastructure surrounding AI applications.	Series F, \$300M	September 30, 2025
Ramp	Financial operations layer that routes and controls AI spend.	Currently indirect. Router is free through 2026 and its long-term pricing has not been disclosed. Ramp can monetize growing AI activity through finance subscriptions, token-spend management and payment volume generated by agents.	Series F, \$750M	June 4, 2026

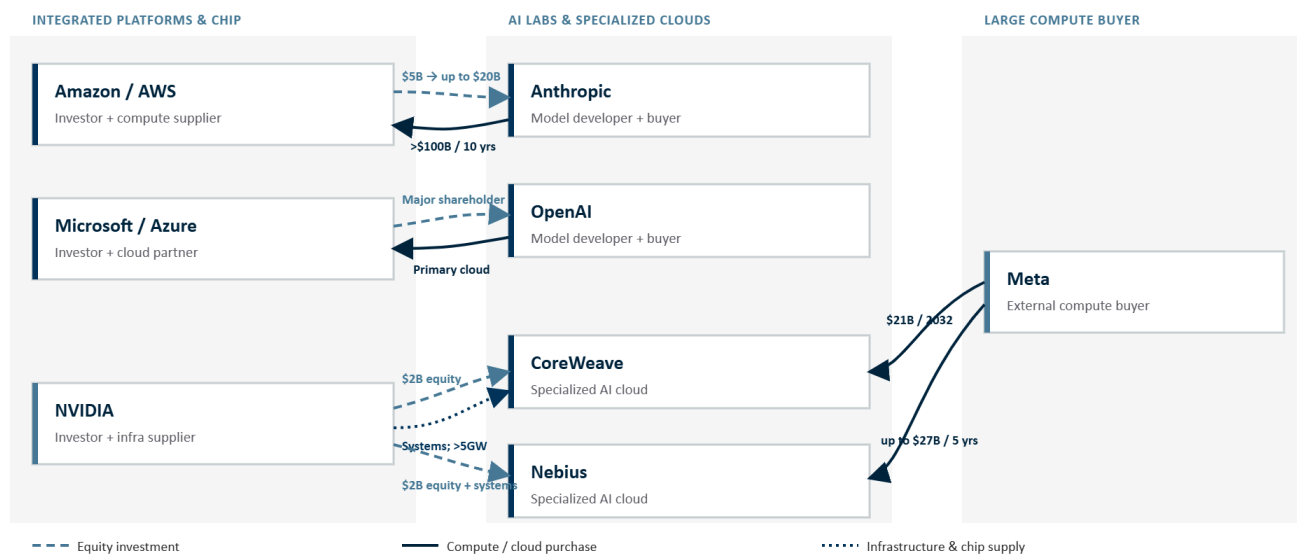
Source: company announcements; BCI analysis.

2.5 AI infrastructure spending is high, and its funding is increasingly circular: more value AND more risk?

The companies above sit on top of a much larger infrastructure buildout. Goldman Sachs raised its estimate for growth in global data-center power demand through 2030 from 117% to 170%, while RBC increased its forecast for aggregate hyperscaler capital-spending growth in 2026 from ~70% to 80%. The scale is visible in company guidance: Amazon, Alphabet, Microsoft, and Meta together forecast \$730B in 2026. This figure still omits capacity added through leases, third-party projects, and long-term cloud-provider agreements.

This also makes the AI ecosystem increasingly circular, as the same companies act as investors, suppliers, and customers. Hyperscalers invest in model developers that buy their cloud, chip firms invest in infrastructure providers that buy their chips, and hyperscalers contract with neoclouds whose financing depends on those deals. These ties can accelerate the buildout and provide committed revenue, but also blur independent demand from spending by connected parties. The key question is not whether the relationships are circular, but whether each transaction creates distinct economic value and whether reliance on a few counterparties raises risk.

FIGURE 4: THE AI BUILDOUT AND INTERDEPENDENCE



Selected announced relationships only. Commitment values may be conditional, span different periods and are not directly comparable or additive. Source: company filings and announcements through August 29, 2026.

COMPANY SPOTLIGHT: NVIDIA, THE INFRASTRUCTURE LAYER

NVIDIA's economy reflects the token economy. Closed and open-weight models both require compute, memory, networking and software, and NVIDIA supplies an integrated platform across these layers. It controls an estimated 80% or more of the AI-chip market, reinforced by its CUDA software ecosystem. The company reported \$89 billion of Data Center revenue last quarter, up 117% year over year. NVIDIA is also expanding into open models, robotics, autonomous vehicles and infrastructure financing. Its reach connects the main themes and the risks: rising power and capital needs, custom-chip competition and if demand will justify the buildout.

3 Where we are now

3.1 Cost per outcome matters more than cost per token

Models tokenize text and use compute differently; price per token is not a like-for-like measure of the cost to complete a task. The same task can cost materially more depending on model selection, harness sophistication and the user's know-how. Enterprises should therefore compare the full cost of completing a task successfully, at the required quality and speed, with the time, cost, and risk of the process it replaces, including integration, human review, error correction, and maintenance. A model with cheaper tokens can still produce a more expensive outcome if it requires more calls or rework. BCI's Project Ledger, highlighted below, provides an early example of this outcome-based approach.

3.2 Token growth is becoming a physical infrastructure constraint

The power required for the next wave of AI data centers is not available today. Current estimates suggest the US will need ~139 GW of additional generation capacity to serve data-center demand through 2030. Globally, data centers could consume 1,100 TWh more electricity in 2030 than in 2023, an increase larger than Japan's total electricity consumption in 2023. Much of the required generation, transmission, and infrastructure still needs to be built.

Adding this capacity will require workers, equipment, grid connections, and regulatory approvals. McKinsey estimates that the United States will need roughly 520,000 additional electricians, construction labourers, and supervisors by 2030. The International Energy Agency notes that new transmission lines can take four to eight years to complete, while wait times for critical grid equipment have doubled in the past three years.

Even where power is available, data centers must be redesigned to handle much greater density. Leading NVIDIA racks likely require 120 to 150 kW today, compared with 40 kW in 2022, and could exceed 1 MW by 2027. This shift requires more advanced power distribution and cooling, with liquid cooling and higher-voltage systems expected to become increasingly common. The constraint is not simply generating enough electricity but delivering and managing it inside the data center.

An emerging industry lens applies this trade-off in physical terms. As power becomes the binding constraint on AI capacity, not every dollar of compute is equally valuable. What matters is how much revenue a company can generate per gigawatt of power it controls. A gigawatt of AI data center capacity costs roughly \$8.5 billion a year to run, generating about \$25 billion in gross token revenue at blended industry pricing. But that average masks a widening gap. Frontier labs that can charge premium prices per token extract far more revenue from the same gigawatt than commodity providers can. As power and chips stay scarce, capacity will increasingly flow to whoever can generate the most revenue per gigawatt, putting weaker-pricing-power models at a structural disadvantage regardless of how much capital they raise.

3.3 AI is changing which workers are needed, not simply reducing jobs

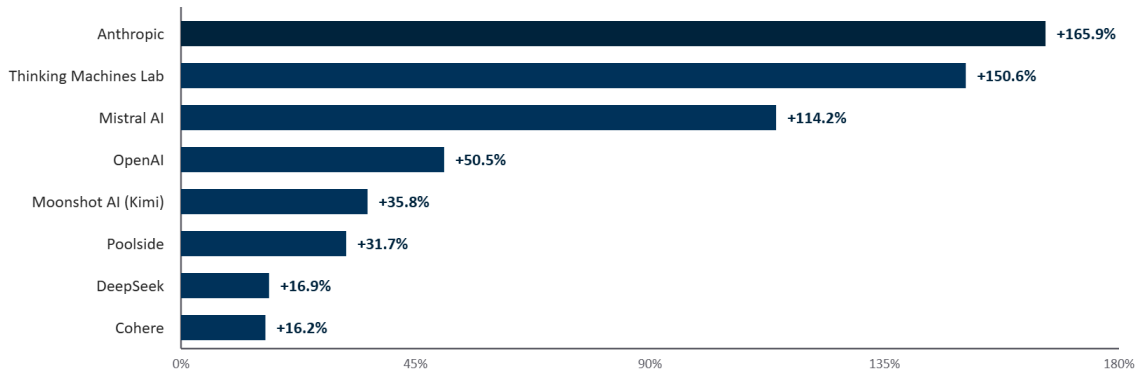
The buildout is one side of AI's labour-market impact: even as AI reduces the labour demand for select professionals, it increases demand for the skilled trades and technical workers who build and operate data centers, power systems, and networks. Across industries, analysts find no statistically significant relationship between AI adoption and overall employment, which suggests any effect is concentrated in particular roles rather than visible in aggregate hiring. The point estimates that circulate are contested and should be read as indicative rather than settled: some analysts attribute a reduction in US monthly payroll growth of roughly 25,000 jobs to AI-related substitution, partly offset by about 9,000 jobs added through AI-supported work.

Klarna shows how this plays out inside a company. Its AI assistant initially handled two-thirds of customer-service conversations, work equal to 700 full-time agents and an estimated \$40 million in annual profit gain. Klarna later added a small team of specialists for complex, sensitive cases. AI can reduce routine work without eliminating the need for human judgment. Shifting spending from payroll to tokens creates value only if service quality holds.

Software development shows a similar shift. Anthropic reported that Claude wrote more than 80% of its merged code, while engineers merged 8x as much per day as in 2024. Yet AI-native firms keep hiring, suggesting engineers are shifting toward directing, reviewing, and validating AI-generated work. Despite the narrative on AI job

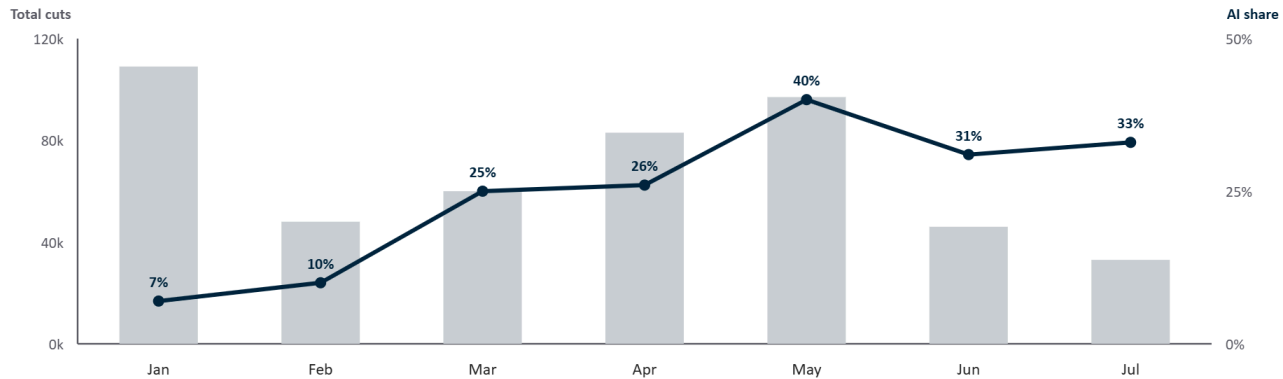
displacement in software engineering, Citadel Securities found US software engineer job postings up 11% year-over-year, suggesting that cheaper code generation expands demand for engineers rather than contracting it.

FIGURE 5: 12-MONTH HEADCOUNT CHANGE PER COMPANY



Source: company disclosures and reporting; BCI analysis.

FIGURE 6: AI SHARE RISES AS TOTAL CUTS FALL (2026)



Source: Challenger, Gray & Christmas, January to July 2026.

4 Future outlook

4.1 Enterprises will mix models, but compute and power demand will persist

Enterprise AI is becoming a hybrid market rather than a single-model one, with model selection increasingly embedded within applications and infrastructure. Model efficiency may cut the electricity per task but is unlikely to close the energy gap if total AI use keeps rising. Closing that gap will require both new power supply and more flexible use of existing grids.

In the near term, developers are adding generation directly at data-center sites so projects can begin operating before grid connections are ready. These on-site solutions, led by natural gas, could support roughly 30% of US data-center growth through 2030. Over time, hyperscalers are also supporting additional nuclear generation and

making some AI workloads flexible enough to reduce or shift electricity use when grids are strained. Microsoft's 20-year power agreement is supporting the planned restart of an 835 MW nuclear reactor in Pennsylvania, while Google has agreements with two utilities to reduce machine-learning loads during periods of grid stress. SpaceX has filed for an orbital data-center system and K2 Space is testing high-power satellite technology for computing in space, but neither has demonstrated a commercially viable data center at scale. Power availability and demand management are therefore likely to play a larger role in where and how new data centers are built.

4.2 Capital is moving beyond models into the infrastructure around them

Growth and strategic investors are financing inference and deployment platforms, including Together AI's \$800M round. After NVIDIA's acquihire of Poolside's team, Jason Warner is pursuing a data center development in Texas. Infrastructure managers and private-credit firms are funding data centers and GPU fleets, and Apollo, BlackRock, Blackstone, Brookfield, Goldman Sachs, and KKR have announced financing platforms with NVIDIA intended to mobilize more than \$500B of third-party capital. Investor demand appears close to unlimited.

The financing mechanics follow a consistent pattern. Private lenders underwrite AI data centers and GPU fleets against customer contracts and equipment. Once a project is operating, the borrower refinances those loans with a broader group of institutional investors. CoreWeave did this through a \$3.1B GPU-backed syndicated facility, while DataBank used a \$1.1B asset-backed securitization to repay construction loans. This lets the original lenders recover capital and fund new projects, accelerating the buildout. It also transfers risk rather than eliminating it. If contracts weaken, capacity sits idle, GPUs depreciate faster than expected, or lenders pull back, borrowing costs rise and construction slows. That risk is already visible: CoreWeave reportedly offered nearly 9.1% yield and stronger lender protections on a later \$2.6B loan after weak initial demand.

Recent M&A points to the same shift (see Figure 7). Three of the six transactions shown target power delivery and management, Stripe's agreed acquisition of OpenRouter targets token routing across models, and NVIDIA's moves on Poolside and Hugging Face reach into the model and open-source distribution layers. Research from Stanford and Together AI helps explain why that routing layer commands a premium: across 1M real-world queries, a router directing traffic between small local models and a frontier model delivered roughly 60% reductions in inference energy, compute and cost at realistic routing accuracy, with the savings driven primarily by right-sizing the model to the query rather than by any hardware advantage. Strategic buyers appear to expect value to accrue to the infrastructure that powers, routes, and operates AI workloads, not only to the labs that develop the models.

FIGURE 7: STRATEGIC BUYERS ARE TARGETING AI'S POWER AND ROUTING BOTTLENECKS

DATE	ACQUIRER	TARGET	DEAL VALUE	STRATEGIC RATIONALE
Aug 26, 2026	NVIDIA	Hugging Face	\$12.9B (reported)	Reported agreement to acquire the largest open-source model hub, extending NVIDIA beyond hardware into model distribution and hosting and back into cloud inference. Owning the main venue where open models reach developers deepens the usage that drives demand for NVIDIA compute. The timing is notable: NVIDIA is moving to control the main distribution point for open models in the same month that licence terms on those models began to tighten, just as access to them becomes more conditional. Reported, not yet formally confirmed.
Aug 20, 2026	NVIDIA	Poolside	\$6.0B licensing; separate \$1.0B investment	Licenses model technology and recruits more than 100 engineers to develop open-weight models, expanding NVIDIA from the hardware used to produce tokens into the model layer itself. Increased usage of the model layer drives demand for more compute, favouring NVIDIA's core business.
Aug 19, 2026	Stripe	OpenRouter	\$7.5B	Adds a platform that routes workloads across many models and manages token cost and performance, placing Stripe closer to the control point for AI usage and payments. The new licence terms fall on hosting rather than routing: Moonshot's licence exempts the mere relaying of requests to models that others host, so a pure router such as OpenRouter plausibly sits outside the revenue thresholds now binding the hosting layer directly beneath it.
Jul 21, 2026	AIP, MGX and BlackRock GIP	Aligned Data Centers	~\$40.0B	Adds 51 data-center campuses and more than 6.4 GW of operating and planned capacity, expanding the physical infrastructure available to generate tokens.
May 19, 2026	Analog Devices	Empower Semiconductor	\$1.5B	Adds integrated voltage regulators and silicon capacitors that deliver power closer to AI processors, supporting greater compute density and lower power loss.
Mar 10, 2026	Alphabet	Intersect	\$4.75B, plus assumed debt	Adds energy and data-center development capabilities, helping new power generation come online alongside rising demand for AI compute.

Source: company announcements; BCI analysis.

4.3 Which industry is next?

The re-architecture of Vertical SaaS is here. In our view, this will be a multi-year process as companies embed models into agentic workflows, partially augmenting human review. EliseAI automates leasing and resident communications; GovDash supports government contractors from proposal development to contract management; Cyera classifies and prioritizes sensitive-data risks; and Huntress pairs automated threat detection and remediation with a 24/7 human augmented security operations center. These businesses share the same economic structure: costly, high-volume work can be automated in stages while people retain control over ambiguous or high-risk decisions. As models become cheaper and more interchangeable, vendors that own the workflow, integrations, and customer relationship benefit from lower costs. Point-solution applications where value is attributable to a front-end are at risk of being commoditized. This distinction underpins our conviction in Vertical SaaS.

Beyond software, AI's expansion into physical work is starting. Stanford reports that robots complete only 12% of real household tasks, compared with 89.4% in simulation. Autonomous vehicles are further along: Waymo reached ~450,000 weekly trips in 2025 and reported more than 220 million autonomous miles through March 2026, with 94% fewer serious or fatal injury crashes than comparable human drivers. This demonstrates scale and safety, but Waymo continues to burn cash. Although not all physical-AI processing is billed as tokens, every robot or vehicle continuously processes data and selects actions, extending token economics from cost per digital task to compute per safe mile or physical task. Robotics adopting tokens are driving our early thesis work in industrial applications.

As the token economy cascades beyond software into the physical world, usage will continue growing while open-weight models keep frontier labs honest. Over the next five years, we expect value to keep shifting away from the model itself toward the applications, routing platforms, and infrastructure that turn falling token costs into useful work and lasting demand. The winners will be the companies that convert rising usage into durable returns, by owning the workflow, controlling how work is routed, and supplying the compute and power beneath it. That discipline guides where we invest today in vertical software. We expect it to hold as the same economics begin to reach robotics and other physical work, where our early thesis work is now underway.

References

- "Anthropic's Fable Shutdown Is a Big Moment for Open-Source AI." CNBC, June 16, 2026. "OpenAI Cuts GPT-5.6 Sol Prices 20% in AI Model Price War." Enterprise DNA, August 21, 2026.
- Amazon, Alphabet, Microsoft, and Meta Platforms. Second Quarter 2026 Earnings Calls and Results. July 2026.
- AMD. "AMD Completes Acquisition of ZT Systems." March 31, 2025.
- Anthropic. "How Am I Billed for My Enterprise Plan?" Accessed August 26, 2026.
- Anthropic. "Manage Costs Effectively." Claude Code Documentation. Accessed August 26, 2026.
- Anthropic. "Plans and Pricing." Accessed August 26, 2026.
- Anthropic. "Pricing." Claude Platform Documentation. Accessed August 26, 2026.
- Apollo. "Apollo Leads \$35 Billion Capital Solution for Broadcom AI Infrastructure." June 9, 2026.
- Artificial Analysis. "Kimi K3 (max): API Provider Performance Benchmarking & Price Analysis." Accessed August 31, 2026.
- Baseten. "Announcing Our Series F." June 22, 2026.
- Baseten. "Overview." Accessed August 26, 2026.
- Blackstone. "NVIDIA Partners with Apollo, BlackRock, Blackstone, Brookfield, Goldman Sachs and KKR to Establish AI Compute Infrastructure Financing Platforms." August 10, 2026.
- "Chinese AI Models Gain Ground with U.S. Companies as Costs Surge." CNBC, July 7, 2026.
- Citadel Securities. "The 2026 Global Intelligence Crisis." Citadel Securities Market Insights, February 24, 2026.
- Claburn, Thomas. "Anthropic Squeezes Enterprises by Ejecting Bundled Tokens from Seat Deal." The Register, April 16, 2026.
- Cloutier, Evan, Max Kan, Jordan Nanos, and Dylan Patel. "Are Open Models Catching Up?" SemiAnalysis, August 21, 2026.
- Constellation Energy. "Crane Clean Energy Center." September 20, 2024.
- CoreWeave. "CoreWeave Closes \$3.1 Billion Loan Facility." May 18, 2026.
- CoreWeave. "CoreWeave Closes Landmark \$8.5 Billion Financing Facility." March 31, 2026.
- CoreWeave. "CoreWeave Completes Acquisition of Weights & Biases." May 5, 2025.
- DataBank. "DataBank Raises \$1.1 Billion in Hyperscale Asset Securitization." September 24, 2025.
- DeepSeek. "DeepSeek License FAQ." Accessed August 31, 2026.
- Erickson, Brad, et al. "Kimi K3 Rocks the Frontiers' World, but Does It?" RBC Capital Markets, July 19, 2026.
- Fernholz, Tim. "K2 Space to Launch Its First High-Powered Satellite." TechCrunch, March 19, 2026.
- Fielding, Mark, et al. "RBC Imagine: Building AI: Datacenter Equipment Growth Drivers Remain Strong." RBC Capital Markets, March 20, 2026.
- Financial Times. "CoreWeave Bows to Investor Pushback on Debt Linked to Anthropic Contracts." July 30, 2026.
- Fireworks AI. "Inference." Accessed August 26, 2026.
- Flight, Frank. "The 2026 Global Intelligence Crisis." Citadel Securities, February 2026.
- Google. "Understanding and Counting Tokens." Gemini API Documentation. Accessed August 23, 2026.
- Hoover, Amanda. "The Real AI Talent War Is for Plumbers and Electricians." Wired, January 15, 2026.
- International Energy Agency. "Energy and AI: Executive Summary." 2025.
- Jaluria, Rishi, et al. "The Software Investor's Handbook to AI." RBC Capital Markets, May 9, 2025.
- Keung, Ronald, et al. "Navigating China Internet/AI Models: Key AI and Mega-Cap Debates." Goldman Sachs Global Investment Research, June 9, 2026.
- Meta Platforms. "Meta Announces New Strategic Venture with BlackRock to Develop Data Center in El Paso." July 2026.
- MiniMax. "MiniMax-M2.7 License." GitHub, MiniMax-AI/MiniMax-M2.7. Accessed August 31, 2026.
- "Model Distillation for LLMs: Cut Costs and Boost Speed." Redis, 2026. "Why a Cheaper Model Won't Lower Your AI Bill." CIO, 2026. "Scaling Token Factory Revenue and AI Efficiency by Maximizing Performance per Watt." NVIDIA Technical Blog, 2026.

Moonshot AI. "Kimi K3 License." Hugging Face, moonshotai/Kimi-K3. Accessed August 31, 2026.

Nathan, Allison, ed. "An AI Job Apocalypse?" Top of Mind, no. 149. Goldman Sachs Global Investment Research, June 25, 2026.

Nebius Group. "Nebius Raises \$775 Million in First Secured Debt Financing." July 17, 2026.

The New Stack. "Z.ai's GLM-5.3 Goes Open Weight, but Its New License Aims at Hyperscalers." August 2026.

NVIDIA. "NVIDIA Announces Financial Results for Second Quarter Fiscal 2027." August 26, 2026.

NVIDIA. "NVIDIA Expands Open Model Families." March 16, 2026.

NVIDIA. "NVIDIA Partners With Leading Investors to Establish AI Compute Infrastructure Financing Platforms." August 10, 2026.

OpenAI. "Advancing the Price-Performance Frontier with GPT-5.6." July 30, 2026.

OpenAI. "Tokens." OpenAI API Documentation. Accessed August 23, 2026.

OpenRouter. "Kimi K3: API Pricing & Benchmarks." Accessed August 31, 2026.

Qwen (Alibaba). "Qwen3.8-Max License." Hugging Face, Qwen/Qwen3.8-2.4T-A95B. Accessed August 31, 2026.

RBC Capital Markets. "2Q26 Hyperscale Capex Update: Spending Slowdown Concerns Overdone." July 30, 2026.

Romei, Robert. "Project Ledger: GP Transaction Notice Agent." BCI internal presentation, June 2026.

Saad-Falcon, Jon, et al. "Intelligence per Watt: Measuring Intelligence Efficiency of Local AI." Stanford University and Together AI, November 2025. arXiv:2511.07885.

Schneider, James, Eric Sheridan, Noah Naparst, and Emma Huang. "Decoding the Agentic Economy." Goldman Sachs Global Investment Research, May 5, 2026.

Singer, Marina, et al. "GS SUSTAIN: AI/Data Centers: Power Demand, Cyclical Progression and Sustainability Implications." Goldman Sachs Global Investment Research, August 6, 2026.

Stanford Institute for Human-Centered Artificial Intelligence. "Technical Performance." The 2026 AI Index Report, 2026.

Stripe. "Stripe Agrees to Acquire OpenRouter to Help Businesses Optimize Token Routing and Usage." August 19, 2026.

Terrell, Michael. "Making Data Centers More Flexible to Support the Grid." Google, August 4, 2025.

Together AI. "Announcing Our \$800M Series C." July 1, 2026.

Together AI. "Serverless Inference." Accessed August 26, 2026.

Vercel. "Kimi K3 and Kimi K3 Fast with ZDR and US-Based Providers Now on AI Gateway." July 27, 2026.

The Wall Street Journal. "Nvidia Draws Antitrust Scrutiny as Enforcers Signal Early Interest in AI."

Waymo. "From the Road: June 2026 Safety Data." June 24, 2026.

Willison, Simon. "moonshotai/Kimi-K3." July 27, 2026.

Yakefu, Adina, Apolinário Passos, and Irene Solaiman. "State of Open Models: Summer 2026 Observations." Hugging Face, August 14, 2026.

Z.AI. "GLM-5.3 License." Hugging Face, zai-org/GLM-5.3. Accessed August 31, 2026.



HEAD OFFICE
VICTORIA

750 Pandora Ave
Victoria, BC
V8W 0E4

VANCOUVER

666 Burrard St
Vancouver, BC
V6C 2X8

NEW YORK

16B-767 Fifth Ave
New York, NY
10153

LONDON

20 Carlton House Terrace,
London, SW1Y 5AN

Visit us at

BCI.ca

Follow us on

[LinkedIn](#)