FRONTIER.

FRONTIER RESEARCH / DECISION-GRADE ANALYSIS

The Token Crash Is Coming. Your AI Bill Is Going Up.

Through 2027, fixed-capability inference will keep collapsing in price while agents, premium reasoning, and outcome-priced breakthrough models push enterprise AI spend higher. The winning CIO strategy is not token thrift. It is intelligence arbitrage.

FR-027 / Dion Hinchcliffe / October 2, 2026

REPORT PROFILE

Built to change a decision.

Through 2027, fixed-capability inference will keep collapsing in price while agents, premium reasoning, and outcome-priced breakthrough models push enterprise AI spend higher. The winning CIO strategy is not token thrift. It is intelligence arbitrage.

REPORT	**FR-027**
AUTHOR	**Dion Hinchcliffe**
PUBLISHED	**October 2, 2026**
PROFILE	**frontier-ro-0.1**

This PDF is a publication rendering of the canonical Frontier Research Object. The web report remains source of truth; citations, figures, methodology, and rights metadata travel with this edition.

Executive brief

The price of tokens is falling. The cost of enterprise intelligence is not.

That is the central economic paradox CIOs will confront through 2027.

At a fixed level of capability, AI inference has been getting cheaper at a rate without precedent modern technology. Epoch AI estimates the cost of achieving a given level of AI performance h fallen about 47% per quarter since 2023 — roughly 13× per year.[1] Stanford's AI Index previo documented a more than 280-fold decline in the cost of GPT-3.5-class inference between late and late 2024.[2] The direction is not in dispute: commodity intelligence is deflating extremely f

But enterprises do not buy a fixed amount of fixed-capability intelligence.

They respond to lower prices by using more of it, moving up the capability curve, adding longer context, adding reasoning, adding tools, adding agents, adding retries, and automating workflows that were previously too expensive to automate. Stanford researchers found agentic coding tasks can consume roughly 1,000× more tokens than code chat or single-pass code reasoning, with repeated runs on the same task varying by as much as 30× in token consumption.[3] Gartner now expects inference cost per agentic workflow to increase more than fivefold through 2028 even as foundational model economics improve.[4]

This is why the wrong 2027 question is:

How cheap will tokens become?

The right question is:

How cheaply can the enterprise convert intelligence into completed work?

Frontier expects the market to split into three economic layers by the end of 2027:

1. Commodity intelligence — cheap, routable, often open-weight, heavily cached, increasingly local or privately hosted.
2. Premium reasoning — expensive frontier models used selectively for difficult decisions, coding, research, planning, and high-stakes agentic work.
3. Outcome intelligence — the beginning of a new pricing layer in which buyers pay for a successful resolution, qualified lead, completed transaction, verified code repair, scientific result, or other attributable business outcome rather than raw tokens.

   The CIO mandate is therefore not to minimize token prices. It is to build an intelligence market inside the enterprise: a control plane that can continuously arbitrage models, providers, deployment locations, latency, quality, risk, and price.

   This extends the operating model Frontier described in The CIO in 2030: people, agents, software, compute, data, and trust increasingly become resources dynamically allocated to outcomes.[5] It also sharpens the economic thesis in The Autonomous Enterprise: cost per completed business outcome matters more than cost per token.[6]

   Frontier's 2027 forecast is deliberately provocative:

> **Tokens will get dramatically cheaper. Enterprises will consume so many more of them that total AI inference spend will still rise sharply. The economic winners will be the organizations that can route intelligence as dynamically as hyperscalers route compute.**

The first mistake: confusing token deflation with AI deflation

There are now at least three different "prices of intelligence," and treating them as one number leads to bad decisions.

Price 1: cost at fixed capability. This is collapsing. Epoch AI's 2026 analysis finds approximately 47% quarterly price declines for a fixed level of model performance since 2023.[1]

Price 2: list price at the current frontier. This is not collapsing nearly as fast, because the frontier keeps moving. A current market tracker shows workhorse and small/fast model medians falling much faster than the frontier tier, while new top-end models can reset the ceiling upward.[7] OpenAI's current enterprise rate card illustrates the spread: GPT-6 Luna is priced at $0.10 per million input tokens and $0.50 per million output tokens, GPT-6 Sol at $2 and $10, while GPT-6 Astra is $10 and $50.[8] That is a 100× input-price gap between the cheap and premium tiers inside one vendor family.

Anthropic shows the same segmentation. Claude Sonnet 5 is $2 per million input tokens and $10 per million output tokens, while Opus 5.5 is $4 and $20, with a faster Opus mode priced higher still.[9][10]

Price 3: cost per completed outcome. This is the metric that will decide whether enterprise AI is economically successful. It includes retries, tool calls, retrieval, context accumulation, agent coordination, human review, failed trajectories, latency, infrastructure, and the cost of using a model that is more capable than the task requires.

The market is moving toward simultaneously cheaper tokens and more expensive workflows.

That is not a contradiction. It is Jevons' paradox with reasoning.

The 2027 Token Paradox

Frontier modeled a representative large-enterprise AI portfolio from Q4 2026 through Q4 2027. The model is not a market forecast of vendor revenue. It is a decision model intended to make the economic direction visible.

We assume four forces operate at once:

- fixed-capability inference continues to decline quickly;

- routing, caching, batching, prompt reduction, and open-weight substitution push the enterprise's effective blended token price down;
- agentic workflows multiply token demand through repeated context, tools, retries, and long-running execution;
- a small but economically important share of workloads moves upward into premium reasoning and breakthrough-class models.

Our likely case produces the following index, with Q4 2026 = 100:

Metric	Q4 2026	Q1 2027	Q2 2027	Q3 2027	Q4 2027
Effective enterprise token price	100	82	70	60	52
Enterprise token volume	100	160	240	350	500
Total inference spend	100	131	168	210	260
Cost per completed AI-enabled outcome	100	88	78	70	65

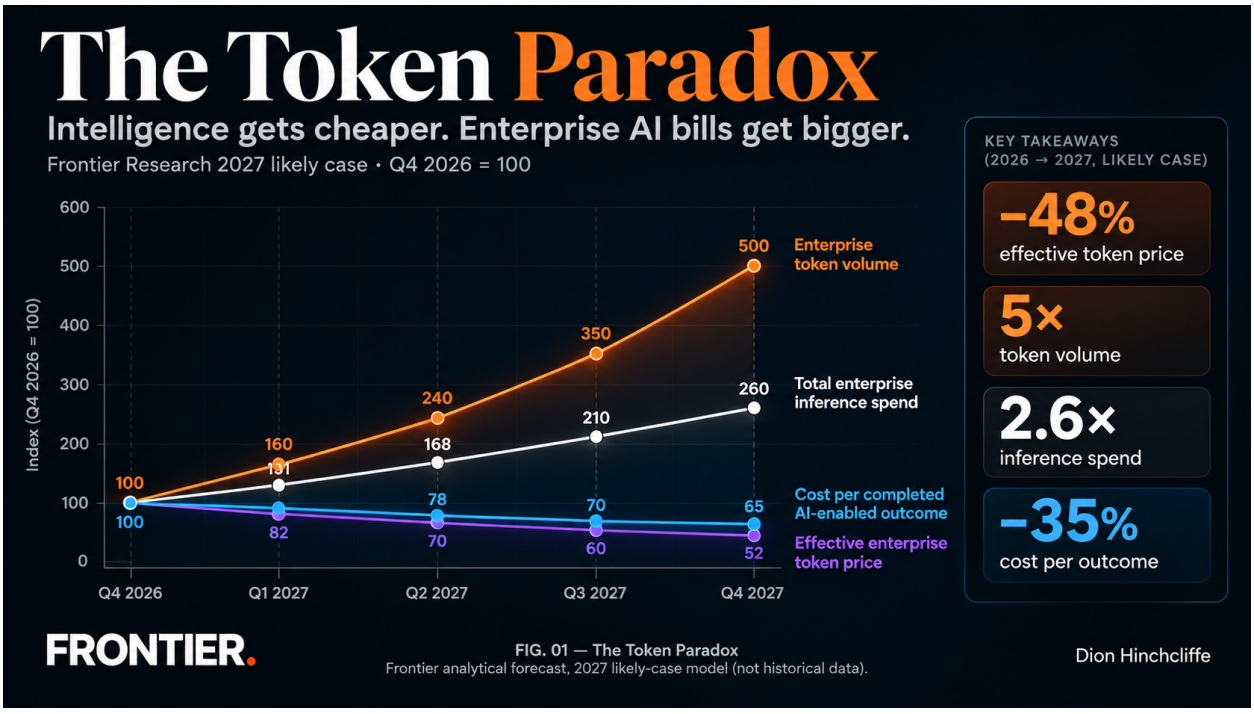


FIG. 01 The Token Paradox: effective token prices fall while enterprise token volume and total inference spend rise. Frontier Research likely-case model; Q4 2026 = 100.

The striking result is not that AI becomes more expensive. It is that AI becomes cheaper enough to be used vastly more often.

In the likely case, the enterprise pays about half as much per effective token by the end of 2027, consumes about five times as many tokens, spends roughly 2.6× as much on inference, and still lowers the cost of a completed AI-enabled outcome by about one-third.

This reconciles the apparently conflicting evidence. Gartner expects model and platform spending to grow rapidly even while inference economics improve.[11] McKinsey reports that enterprise AI spend rises nearly fourfold as organizations move from isolated use cases to enterprise-scale deployment, with 93% of surveyed organizations exceeding AI budgets and a majority expecting at least another 25% increase over the following year.[12]

The unit cost is deflationary. The demand curve is explosive.

All-you-can-eat AI is already retreating

The enterprise market spent 2023 through 2025 pretending AI could be packaged like SaaS: buy a seat, encourage usage, and let the vendor absorb the marginal compute.

That model becomes unstable when a user is no longer a human typing 30 prompts a day but an agent running continuously, opening tools, rereading context, delegating subtasks, checking work, retrying failures, and spawning other agents.

The retreat has already started.

OpenAI now offers token-based ChatGPT Enterprise agreements in which usage across Chat, Work, and Codex is metered separately from seat fees.[13] Anthropic's current Enterprise plan is even more explicit: the seat fee covers access, while all Claude, Claude Code, and Cowork usage is billed separately at standard API rates, with organization and individual spend limits available to control consumption.[14]

Frontier expects true unlimited frontier-agent usage to become economically rare in the enterprise by the end of 2027.

"Unlimited" will survive, but increasingly as one of four things:

- a marketing wrapper around fair-use and abuse controls;
- a bundle of cheap-model usage with metered premium-model overflow;
- a seat entitlement plus a separately metered agentic or reasoning pool;
- a committed-spend contract that merely disguises consumption pricing.

This is not a failure of the labs. It is the inevitable result of agentic demand.

When a single autonomous workflow can consume orders of magnitude more tokens than a chat interaction, no rational provider can expose uncapped frontier inference without either raising the seat price dramatically, throttling the experience, shifting the workload to cheaper models, or metering usage.

CIOs should stop negotiating "unlimited seats" as if that guarantees unlimited economics.

It does not.

Routing becomes procurement

In the cloud era, enterprise architects separated the application from infrastructure so workloads could move between compute pools.

In the intelligence economy, the equivalent abstraction is the enterprise model router.

The router sits between users, applications, agents, and model providers. It observes the task, identity, data sensitivity, latency target, budget, quality requirement, and provider health — then chooses where the request should go.

The strategic shift is that routing is no longer merely a reliability feature. It is becoming a real-time procurement engine.

AWS says Bedrock Intelligent Prompt Routing can reduce costs by up to 30% by dynamically choosing between models in the same family.[15] Cloudflare's AI Gateway now combines spend limits, custom negotiated costs, caching, dynamic routing, and an Auto Router intended to choose the most cost-effective model that can satisfy the task.[16] Microsoft Foundry offers a model-router layer across a large model catalog, reflecting the same architectural direction.[17]

Neutral gateways go further because they preserve bargaining leverage across providers. OpenRouter exposes hundreds of models and scores of providers through one API, with enterprise controls, spend limits, policy-based routing, regional routing, and BYOK support.[18] LiteLLM provides a self-hostable gateway with budgets, virtual keys, routing, provider failover, lowest-cost routing, caching, and prompt compression.[19] Portkey — now offered as Prisma AIRS AI Gateway — emphasizes enterprise routing, observability, key management, caching, hybrid deployment, and broad model access.[20] Kong extends the same control-plane idea into the API infrastructure many enterprises already operate, including cost attribution and budget enforcement based on model pricing dimensions.[21]

Frontier's view is that the enterprise router will become as strategically important as the API gateway.

The key is neutrality.

A router that can only choose among models sold through one cloud can optimize execution. A router that can choose among clouds, direct model APIs, open-weight inference providers, and private endpoints can optimize economics.

That difference will be worth real money.

The new arbitrage patterns

By 2027, mature enterprises will run several forms of intelligence arbitrage simultaneously.

Capability arbitrage

Route the task to the cheapest model that reliably clears the required quality threshold.

This sounds obvious, but most enterprises still overbuy intelligence. Frontier models are frequently used for classification, extraction, summarization, formatting, and routine transformation that smaller models can execute adequately.

The spread is already enormous. OpenAI's current enterprise rate card spans from GPT-6 Luna at $0.10/$0.50 per million input/output tokens to GPT-6 Astra at $10/$50.[8] Together AI lists many open models well below $1 per million input tokens, including models priced in the low tens of cents.[22]

Provider arbitrage

The same or similar model can be available through direct APIs, hyperscalers, routing exchanges, and specialized inference providers under different commercial terms.

The enterprise gateway should preserve the ability to send traffic to whichever endpoint provides the best combination of price, latency, region, availability, and negotiated terms.

Time arbitrage

Not every inference request needs to happen immediately.

Batch and flex processing can materially lower costs for asynchronous workloads. Google's Gemini pricing already separates standard, batch, flex, and priority tiers, illustrating how latency itself is becoming a pricing dimension.[23] Enterprises should treat "when must this answer exist?" as an economic policy.

Cache arbitrage

Caching converts repeated intelligence into stored intelligence.

OpenAI prices cached input dramatically below normal input on many models.[8] McKinsey estimates prompt caching can reduce repeated input-token costs by as much as about 90% for workloads with large stable prefixes, such as RAG and agents.[12]

Architecture arbitrage

The biggest savings often come from not asking a model at all.

Deterministic logic, SQL, rules, search indexes, vector retrieval, traditional classifiers, workflow engines, and cached computations remain cheaper and more predictable than repeated generative inference when the task is stable.

By 2027, efficient AI systems will use LLMs for ambiguity — not for every operation.

Deployment arbitrage

The same open-weight model can be consumed through a serverless provider, a provisioned throughput contract, a dedicated endpoint, a private cloud cluster, or an on-premises system.

Together AI now offers provisioned throughput with reserved inference capacity and says some profiles can run substantially below closed-model list prices.[24] The enterprise can increasingly choose not just which model, but which economic form factor should serve it.

Private AI will rise — and disappoint companies that treat GPUs as free tokens

More enterprises will move some inference into private infrastructure in 2027.

Cost will be one reason. Sovereignty, privacy, latency, customization, and negotiation leverage will be equally important.

The attraction is easy to understand: once a workload reaches sufficiently high and predictable volume, renting every token from a frontier API can look expensive relative to owning or reserving capacity. NVIDIA explicitly positions enterprise-controlled deployment as a way to reduce dependency on metered APIs and improve cost predictability for high-volume inference.[25]

But "private AI" is not a magic discount.

An API converts infrastructure risk into a variable price. Private deployment brings that risk back onto the enterprise balance sheet.

The true cost includes accelerators and depreciation, idle capacity, power and cooling, networking, storage, inference serving, quantization, model optimization, observability, security, MLOps, model refresh, staffing, failover, and rapid hardware obsolescence.

The economic variable that matters most is utilization.

A private cluster running at high sustained utilization on a stable workload can produce excellent unit economics. The same cluster sized for a spiky enterprise peak can become one of the most expensive ways to save money.

Frontier therefore expects a three-stage private-AI adoption pattern:

Stage 1: provider-hosted open models. Enterprises get most of the model-price hedge without operating infrastructure.

Stage 2: dedicated or provisioned inference. Stable high-volume workloads move to reserved capacity, improving price predictability and performance.

Stage 3: genuinely private AI. Only workloads with sufficient scale, sovereignty requirements, stable demand, or proprietary model value justify owning the full serving environment.

McKinsey makes the same distinction: enterprise-hosted open-weight models can improve control and unit economics at scale, but require materially stronger engineering, MLOps, security, and infrastructure capabilities.[12]

The default CIO move should therefore be private where economically or regulatorily justified, not private by ideology.

Open-weight models become the enterprise price hedge

Open-weight models have moved from ideological preference to procurement leverage.

That is a major change.

McKinsey's survey work found respondents frequently associate open-source AI with lower implementation and maintenance costs, while also flagging cybersecurity, compliance, and IP concerns.[26] Recent reporting shows large enterprises increasingly exploring open-weight models specifically to control AI costs, improve sovereignty, and reduce dependence on the frontier labs.[27]

Frontier expects open-weight models to play four distinct economic roles through 2027:

1. The price floor. Cheap capable open models prevent closed-model vendors from maintaining high prices on routine work.

2. The negotiation hedge. A production-ready open alternative changes the enterprise's posture in commercial negotiations even if most traffic remains on proprietary APIs.

3. The private-AI bridge. Open weights make dedicated and self-hosted inference possible when sovereignty or economics require it.

4. The specialization layer. Fine-tuned domain models can beat a general frontier model on narrow enterprise tasks at a fraction of the inference cost.

This is why the right architecture is not "open versus closed."

It is open plus closed behind a common control plane.

Frontier argued in The Autonomous Enterprise that autonomy scales process by process, not company by company.[6] The same will be true of open models. They will win workload by workload, not by replacing proprietary AI wholesale.

What enterprises are actually doing to control token spend

There is no single optimization lever. McKinsey says mature organizations are applying dozens of levers across model selection, prompts, workflows, infrastructure, sourcing, and governance, with active optimization already producing meaningful savings for some companies.[12]

Frontier assesses the practical enterprise playbook this way:

Lever	2027 economic potential	Best use	Main trap

Model right-sizing	Very high	Routine tasks with measurable quality thresholds	Teams default to the smartest model because evaluation is weak
Intelligent routing	Very high	Mixed portfolios with varied task complexity	Router optimizes list price rather than actual outcome quality
Prompt/context reduction	High	RAG, agents, long conversations	Cutting context can damage quality if retrieval is poor
Prompt caching	High	Stable system prompts, tool schemas, repeated context	Treating cache savings as true computational efficiency
Batch/flex processing	High	Back-office, analytics, enrichment, overnight jobs	Product teams label everything "real time"
Output constraints	Medium-high	Extraction, structured workflows, agent tools	Over-constraining tasks that need exploration
Agent loop budgets	Very high	Long-running agents and coding workflows	Optimizing tokens without measuring task success
Memory compaction	High	Persistent agents, support, coding, research	Summaries can silently discard critical state
Open-weight substitution	High	High-volume, bounded, domain-stable tasks	Security, quality, maintenance, and model-drift burden
Provisioned throughput	High at scale	Stable, predictable workloads	Paying for reserved capacity that sits idle
Private/self-hosted inference	Potentially very high	Sovereignty or very high steady utilization	Capex, utilization, staffing, rapid obsolescence
Edge/on-device models	Medium-high	Privacy-sensitive, repetitive local work	Capability ceiling and fleet management
Showback/chargeback	Indirect but essential	Large decentralized organizations	Treating cost allocation as cost optimization
Business-logic substitution	Very high	Repeatable deterministic substeps	Overusing LLMs where software already solves the problem
Commercial arbitrage	High	Multi-provider estates	Commit discounts that recreate lock-in

The most underestimated lever is agent architecture.

Stanford's work shows token consumption can vary up to 30× for the same agentic task and that higher spend does not reliably produce higher accuracy.[3] Separate 2026 research finds prompt and harness design can multiply reasoning costs without improving correctness.[28]

The implication is uncomfortable: a meaningful share of future token spend will be software inefficiency wearing an AI badge.

CIOs should require agent teams to measure cost per successful completion, not tokens per request.

Which enterprise-grade routers matter

The router market is still forming, but the strategic categories are now clear.

Router/control plane	Why it matters	Lock-in posture	Frontier assessment
OpenRouter Enterprise	Very broad model/provider market, auto-routing, BYOK, budgets, policy and regional routing	Strong provider neutrality; dependence shifts to router	Best fit when commercial optionality and rapid model access matter most
Cloudflare AI Gateway	Global gateway, caching, spend controls, dynamic/ auto routing, custom providers, identity/security adjacency	Strong multi-provider hedge; deeper Cloudflare platform gravity	Especially strong for distributed applications and organizations already using Cloudflare
Portkey / Prisma AIRS AI Gateway	Routing, observability, prompt controls, guardrails, hybrid/air-gapped patterns	Strong neutral-control-plane posture	Strong enterprise governance choice where security and observability dominate
LiteLLM Enterprise	Self-hosted, OpenAI-compatible, broad provider support, budgets, fallbacks, caching, lowest-cost routing	Excellent architectural portability	Strongest fit for enterprises willing to operate the gateway themselves
Kong AI Gateway	AI policy integrated into mature API management, detailed cost governance	Good neutrality within existing Kong architecture	Attractive where API governance is already standardized on Kong
AWS Bedrock Intelligent Prompt Routing	Native optimization and procurement inside AWS	Lower portability across commercial domains	Excellent local optimization; weaker as a true multi-cloud arbitrage layer
Microsoft Foundry Model Router	Broad catalog inside Microsoft enterprise control plane	Strong Azure/Microsoft gravity	Logical for Microsoft-centric estates; should be paired with an external escape route for strategic workloads

The strategic decision is not whether to use a router.

It is where the routing policy lives.

If model-selection logic is embedded separately inside every application and agent, the organization will not have meaningful procurement leverage. If it lives in a common gateway with telemetry, policies, evals, budgets, and multiple providers, the CIO can change the economics of intelligence without rewriting every system.

Frontier recently argued that Cloudflare is moving toward becoming a control plane for the machine Internet. Its new AI routing and cost controls make the token-economics dimension of that position particularly important.[29]

Outcome pricing is coming — but not everywhere

Token pricing has a conceptual flaw: it charges the buyer for the provider's computational effort rather than the buyer's business result.

The more inefficient the model or agent, the more tokens it may consume.

That is a strange economic relationship.

The market is already experimenting with alternatives. Intercom prices its Fin agent by successful outcomes in a conversation, including resolutions and selected workflow outcomes.[30] Zendesk bills AI-agent usage through automated resolutions, charging when a customer request is successfully resolved without human escalation.[31] Salesforce offers action-based Flex Credits and also documents business-metrics-based pricing for certain AI usage.[32][33]

These models are early, but they point to where enterprise AI economics is heading.

Frontier expects outcome-based pricing to expand first where four conditions are true:

- the outcome is observable;
- attribution is reasonably clean;
- completion happens quickly;
- the economic value can be bounded.

  Customer-service resolution is ideal. Lead qualification, collections, document processing, claims triage, code repair, fraud investigation, procurement negotiation, and selected autonomous operations are plausible next markets.

Outcome pricing will be much harder for strategy, research, product design, scientific discovery, or executive decision support because the value may appear months later and attribution is contested.

The likely 2027 model is therefore hybrid economics:

base access + metered intelligence + outcome premium.

The token will not disappear. It will become the cost basis underneath a higher-level commercial unit.

The superintelligence premium

The most underappreciated 2027 possibility is that commodity intelligence and breakthrough intelligence move in opposite pricing directions.

We already see the beginnings of a premium curve. OpenAI's GPT-6 Astra costs roughly five times GPT-6 Sol at current standard input and output rates, and about 100 times GPT-6 Luna on input tokens.[8][34] Anthropic similarly maintains an Opus premium above Sonnet.[10][9]

If 2027 models begin producing economically meaningful breakthroughs — a novel drug candidate, materially superior chip design, a patentable algorithm, a major contract win, a verified software migration, a successful autonomous negotiation — providers will have little incentive to price those capabilities as mere commodity tokens.

The market could develop a breakthrough lane.

Under Frontier's star-shot case, the enterprise might spend pennies on millions of routine tokens while paying thousands, tens of thousands, or more for a bounded attempt at a high-value outcome.

The pricing form could look like:

reservation fee + compute floor + success premium + audit evidence.

That is much closer to professional services, transaction fees, or performance contracting than to SaaS.

The constraint will be attribution. Providers will not want to accept unlimited risk, and buyers will not want to pay a percentage of value for a result they cannot prove the model caused.

But where outcomes are machine-verifiable, this pricing model is economically rational.

It also aligns incentives better than token billing. The provider gets paid for being effective, not verbose.

2027: likely, outlier, and star-shot cases

Likely case — The Token Paradox wins

Fixed-capability inference continues to fall sharply. Enterprise effective token prices fall roughly 40–50% through aggressive routing, caching, and open-model substitution. Token volume rises about fivefold as agents spread across coding, operations, analytics, service, security, and knowledge work. Total inference spend rises to roughly 2.5–3× its late-2026 level.

All-you-can-eat frontier usage largely disappears from serious enterprise agent deployments. Routers become a standard architecture layer. Open-weight models take a meaningful share of bounded workloads. Private AI grows, but mostly through dedicated/provisioned inference rather than enterprises buying giant GPU fleets.

Outcome pricing expands in customer service and transaction-like agent work.

Probability judgment: this is Frontier's central case.

Outlier case — The commodity collapse outruns demand

Inference efficiency, open models, custom silicon, batching, cache reuse, and model competition move faster than agent demand. Effective enterprise token price falls 70–80%. Open-weight models close enough of the capability gap that frontier models are used only for a minority of requests. Token volume may rise 8–10×, but total spend rises much more slowly because the cost curve collapses underneath it.

This is the best case for enterprise buyers — and the hardest case for frontier-lab margins.

The trigger would be rapid convergence of model quality plus aggressive open-model hosting economics.

Star-shot case — Commodity tokens go near-zero while breakthrough intelligence becomes expensive

Routine intelligence becomes almost ambient: heavily cached, local, open-weight, and cheap enough that token accounting is irrelevant for ordinary work.

At the same time, a small number of frontier systems become materially better at long-horizon research, invention, coding, negotiation, and autonomous execution. Vendors stop selling their highest-value capability primarily by token and begin charging for bounded high-value attempts or verified outcomes.

Enterprise AI spending rises fastest in this world — not because tokens are expensive, but because the addressable value of intelligence explodes.

The CIO's job changes from controlling compute consumption to allocating an intelligence investment portfolio.

This is the scenario most organizations are least prepared to budget for.

The CIO pricing playbook

The goal is not the lowest nominal token rate.

The goal is to preserve enough optionality that the enterprise can continuously force intelligence providers to compete for the next workload.

Frontier recommends ten commercial and architectural moves.

1. Put every production model behind an enterprise-controlled gateway. No business unit should hard-code a strategic application directly to a frontier provider without a deliberate exception.

2. Maintain at least three economic lanes. A cheap routine lane, a workhorse reasoning lane, and a premium frontier lane. Add a private/open lane where scale or sovereignty justifies it.

3. Negotiate on effective blended cost, not list price. Include cache rates, long-context multipliers, batch/flex discounts, regional premiums, tool fees, search fees, fast-mode surcharges, and committed-spend economics.

4. Demand price-decline protection. Contracts should not leave the enterprise paying yesterday's rate after public prices fall. Seek benchmark reopeners, periodic repricing, or the ability to redirect committed spend to newer models.

5. Make commitments portable. If the vendor launches a cheaper or better model, committed dollars should be transferable across the model family. If possible, commitments should also cover API, agent, and enterprise-workspace surfaces rather than trapping spend in one product.

6. Keep BYOK and direct-provider paths available. A router is most valuable when it can use negotiated enterprise keys rather than forcing the buyer onto the router's own markup or provider contract.

7. Separate gateway policy from model provider. If the same vendor owns the model, router, telemetry, evaluation, and spend controls, the organization has observability — but not necessarily leverage.

8. Establish cost-per-outcome benchmarks before negotiating outcome pricing. Otherwise the provider will know the unit economics better than the buyer.

9. Use open weights as a credible BATNA. The enterprise does not need to run every workload on open models. It needs enough production capability to make "we can move this workload" a believable negotiating position.

10. Build AI FinOps as a permanent operating capability. McKinsey finds only a minority of organizations have mature AI FinOps today, while better forecasting and active optimization correlate with measurable savings.[12] This function should own forecasting, routing economics, chargeback/showback, contract benchmarks, model evaluation, and cost per completed outcome.

The enterprise architecture should make switching routine.

The procurement process should make complacency expensive.

Frontier outlook

The 2027 token market will look less like SaaS and more like energy.

There will be spot prices, reserved capacity, premium grades, routing, arbitrage, regional constraints, transmission layers, private generation, hedging, and increasingly sophisticated demand management.

The difference is that intelligence demand is much more elastic than electricity demand.

Make reasoning 10× cheaper and the enterprise will not buy the same amount of reasoning for one-tenth the price.

It will redesign work around the assumption that reasoning is abundant.

That is the deepest economic insight in this report.

The token is not the final unit of value. It is the raw material.

By 2027, the most sophisticated CIOs will stop managing AI primarily as software licensing and start managing it as an intelligence portfolio — continuously buying the cheapest form of cognition that can produce the required outcome, while reserving expensive frontier intelligence for the few moments where it can change the trajectory of the business.

The price of thought will keep falling.

The demand for useful thought will rise faster.

And that is why the token crash is coming — while the AI bill keeps going up.

SOURCES / CITATIONS

Trace the argument.

Every source below is linked. Source numbering follows first appearance in the report.

01 **Epoch AI — The plunging price of thought**
https://epoch.ai/publications/the-plunging-price-of-thought

02 **Stanford HAI — AI Index 2025**
https://hai.stanford.edu/news/ai-index-2025-state-of-ai-in-10-charts

03 **Stanford Digital Economy Lab — How Do AI Agents Spend Your Money?**
https://digitaleconomy.stanford.edu/publication/how-do-ai-agents-spend-your-money-analyzing-and-predicting-token-consumption-in-agentic-coding-tasks/

04 **Gartner — AI inference costs per agentic workflow**
https://www.gartner.com/en/newsroom/press-releases/2026-08-17-gartner-predicts-ai-inference-costs-per-agentic-workflow-will-increase-more-than-fivefold-through-2028

05 **Frontier Research — The CIO in 2030**
https://www.thisisfrontier.ai/research/cio-2030

06 **Frontier Research — The Autonomous Enterprise**
https://www.thisisfrontier.ai/research/the-autonomous-enterprise

07 **AI Token Price — Token Price Index**
https://aitokenprice.com/token-price-index

08 **OpenAI — ChatGPT Enterprise token-based rate card**
https://help.openai.com/en/articles/20001415-chatgpt-rate-card-enterprise-token-based-pricing

09 **Anthropic — Claude Sonnet 5**
https://www.anthropic.com/research/claude-sonnet-5

10 **Anthropic — Claude Opus**
https://www.anthropic.com/claude/opus

11 **Gartner — AI platforms and models market forecast**
https://www.gartner.com/en/newsroom/press-releases/2026-07-20-gartner-forecasts-worldwide-ai-platforms-and-models-market-to-grow-63-percent-in-2026

12 **McKinsey — The cost of intelligence**
https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-cost-of-intelligence-how-cios-can-manage-ai-demand-at-scale

13 **OpenAI — Token-based billing for ChatGPT Enterprise**
https://help.openai.com/en/articles/20001520-token-based-billing-for-chatgpt-enterprise

14 **Anthropic — Claude Enterprise plan**
https://support.claude.com/en/articles/9797531-what-is-the-enterprise-plan

15 **AWS — Bedrock Intelligent Prompt Routing**
https://aws.amazon.com/bedrock/intelligent-prompt-routing/

16 **Cloudflare — AI Gateway features**
https://developers.cloudflare.com/ai-gateway/features/

17 **Microsoft — Foundry Model Router**
https://azure.microsoft.com/en-us/pricing/details/ai-foundry-models/model-router/

18 **OpenRouter — Pricing and enterprise controls**
https://openrouter.ai/pricing/

19 **LiteLLM — Enterprise AI Gateway**
https://www.litellm.ai/enterprise

20 **Portkey / Prisma AIRS — Enterprise AI Gateway**
https://portkey.ai/features/ai-gateway

21 **Kong — AI Gateway model cost management**
https://developer.konghq.com/ai-gateway/model-cost-management/

22 **Together AI — Inference pricing**
https://www.together.ai/pricing

23 **Google — Gemini Developer API pricing**
https://ai.google.dev/gemini-api/docs/pricing

24 **Together AI — Provisioned Throughput**
https://www.together.ai/blog/provisioned-throughput

25 **NVIDIA — Self-hosted confidential AI reference architecture**
https://docs.nvidia.com/enterprise-reference-architectures/deploying-proprietary-models-confidential-compute-self-hosted-kubernetes/latest/introduction.html

26 **McKinsey — Open source technology in the age of AI**
https://www.mckinsey.com/capabilities/quantumblack/our-insights/open-source-technology-in-the-age-of-ai

27 **Financial Times — Corporate America embraces cheaper open AI models**
https://www.ft.com/content/d9de4776-1fc9-4f2b-aaaf-9961c35d8acd

28 **Prompt-Induced Waste in Large Reasoning Models**
https://arxiv.org/abs/2608.01347

29 **Frontier Research — Cloudflare Ate the Edge. Now It Makes a Move for the Whole Cloud.**
https://www.thisisfrontier.ai/news/cloudflare-ate-the-edge-now-it-wants-the-whole-cloud

30 **Intercom — Fin AI Agent outcome pricing**
https://www.intercom.com/help/en/articles/8205718-fin-ai-agent-outcomes

31 **Zendesk — Automated resolution tiers**
https://support.zendesk.com/hc/en-us/articles/9570369117338-About-automated-resolution-tiers

32 **Salesforce — Agentforce pricing**
https://www.salesforce.com/agentforce/pricing/

33 **Salesforce — Agentforce and Generative AI Usage and Billing**
https://help.salesforce.com/s/articleView?id=generative_ai_usage.htm&language=en_US&type=5

34 **OpenAI — GPT-6 Astra**
https://developers.openai.com/api/docs/models/gpt-6-astra

FRONTIER.

TAKE THE NEXT STEP

Push this further.

Research is useful when it changes a decision. Bring us the question, the constraint, or the argument you want tested.

RESEARCH / BRIEFINGS

inquiries@thisisfrontier.ai

Briefings, advisory work, executive discussions, and research inquiries.

REPRINTS / LICENSING

reprints@thisisfrontier.ai

Reprints, enterprise distribution, excerpts, and licensing.

FRONTIER

www.thisisfrontier.ai

Research, Data Insights, News, and the live Frontier corpus.

All rights reserved. Contact Frontier for redistribution or reprint rights.

SEE FURTHER. MOVE SOONER.