

01AI LTD

WHITEPAPER

The Token Paradox

Why AI Gets Cheaper per Token and Dearer per Month, and How to Run It at a Cost You Can Defend

Five cost leaks in enterprise AI, and the control plane that closes them.

October 2026
01AI LTD · Dublin, Ireland
www.01ltd.com

Abstract

The price of a token has collapsed; the AI bill has not. Agents, reasoning models and long contexts multiply the tokens behind every task faster than prices fall, and the spend is scattered across API keys, seat licences and features inside other software. Five cost leaks explain most of the gap, and one control plane closes them.

The unit price of AI has fallen remarkably fast. Stanford's AI Index found that querying a model as capable as GPT-3.5 cost \$20 per million tokens in November 2022 and \$0.07 by October 2024, a more than 280-fold drop in about eighteen months. Yet in McKinsey's 2026 survey about one respondent in five reports that AI-related operating costs, including token costs, have constrained their organisation's AI use. And in the FinOps Foundation's State of FinOps 2026, 98% of practitioners now manage AI spend, against 31% two years earlier.

The two facts are compatible. A per-token price measures one unit; a bill measures units times price, and the units have grown faster than the price has fallen. McKinsey's Michael Chui puts it in one sentence: "Even as per-token costs have declined, the number of tokens consumed and generated has increased even faster."

This paper names five places where AI spend leaks: the loop tax of agents, frontier models used by default, real-time pricing for work nobody waits on, seats nobody uses, and spend nobody can see. It then looks at the two decisions that set the long-run cost, switching providers and owning hardware, and sets out a control plane that closes the leaks: one gateway, one unit metric, one review.

The argument is not that AI costs too much. Most organisations in McKinsey's survey plan to invest more in AI, not less. It is that an unmanaged AI bill grows with activity rather than with value, and that the fix is architecture and governance, not a cheaper model.

Cheaper Tokens, Bigger Bills

Budgets for AI were set in the vocabulary of the supplier: a price per million tokens, a price per seat. Both numbers have behaved well. What has not behaved is the number of tokens behind each piece of work, and four multipliers explain most of the growth.

1. **Loops.** An agent calls the model again at every step, and each call carries the instructions, the tool definitions and the history so far. The input grows with every step of the task.
2. **Reasoning.** Reasoning models generate tokens before they answer, and those tokens are billed as output. Output is the expensive side of the price list: five times the input price on Anthropic's current models.
3. **Context.** Retrieval, long documents and conversation history put more input into each request, and a larger context window invites teams to fill it.
4. **Adoption.** More users, more use cases, and agents that run without a person typing each request. Volume grows with success, which is the point, as long as the value grows with it.

The effect lands where the work is heaviest. McKinsey found that among AI high performers, operating costs have most commonly constrained the use of software coding agents, the tools that run the longest loops over the largest contexts.

It also lands in three different shapes of spend, each needing its own control. Per-token API spend is variable and grows with use. Seat licences for assistants are fixed and renew whether or not anyone uses them. And AI built into existing software arrives as a price rise on a contract nobody reopens. A cost programme that looks at only one of the three moves the waste into the other two.

1. Cost Leak One: The Loop Tax

An agent completes a task by calling the model again and again: plan, call a tool, read the result, decide what to do next. In most implementations every one of those calls resends the instructions, the tool definitions and the history so far. A twenty-step task pays for its opening context twenty times, and pays again for every failed step, retry and dead end.

None of this is visible if spend is measured per request. Each call looks cheap. The cost of the task only appears when the calls are added up, and an agent that loops on an error can run for a long time before anyone adds them up.

The corrective

Measure cost per completed task, not per call. Cache the stable part of every prompt: providers bill cache reads at a fraction of the input price, a tenth on Anthropic's current list. Trim tool output before it goes back into the context, and give every agent a step and token ceiling that ends in an alert, not in an open-ended run.

2. Cost Leak Two: Frontier by Default

Teams build the prototype on the most capable model available, because that is how you find out whether the idea works. Then they ship it unchanged. Every request, including the classification, extraction and routing steps that a small model handles well, now pays the frontier price.

The gap is not marginal. On Anthropic's current price list the most capable model costs ten times the smallest per input token, \$10 against \$1 per million. Reasoning effort is a second dial that most teams leave at its default, and on reasoning models it changes the output bill directly.

The corrective

Route by task, and decide routes with evidence. Keep an evaluation set for each use case, test smaller models and lower reasoning effort against it, and move a route only when quality holds. Judge the result by cost per completed task: a cheaper model that needs more retries is not cheaper. Re-run the test when models change.

3. Cost Leak Three: Real Time for Everything

A large share of AI work has no one waiting for it: overnight document enrichment, report generation, classification of a backlog, evaluation runs. It is usually sent through the same interactive endpoint as a chat message, at the same price.

Providers price waiting differently. Anthropic's batch API charges half the standard price for input and output in exchange for asynchronous processing. The same logic runs the other way for premiums: faster tiers cost more, and some providers charge extra to pin processing to one region, 1.1 times the standard price for US-only inference on Anthropic's list.

The corrective

Classify every workload by how long its user can wait. Send what can wait to batch, schedule it, and keep interactive pricing for interactive paths. Pay for premium speed and for regional pinning where the use case or the data requires them, not as a default inherited from the first prototype.

4. Cost Leak Four: Seats Nobody Uses

A pilot goes well and an assistant is licensed for a whole department. Use concentrates in the people whose work suits it, the rest open it occasionally, and the licence renews on schedule. McKinsey's commentary notes that spend on horizontal AI tools such as chatbots is increasingly managed as a necessary cost of doing business, like office productivity tools. Spend managed that way is rarely revisited seat by seat.

The same pattern hides in AI features added to software the organisation already pays for. The capability arrives as a price rise at renewal, and the question of whether anyone uses it is never asked.

The corrective

Read active use per seat from the admin consoles every quarter and reclaim idle seats before renewal. License by role on measured use rather than by department on enthusiasm. When a supplier adds an AI uplift, ask for usage data on your own tenant before accepting it.

5. Cost Leak Five: Spend Nobody Can See

AI spend rarely sits on one invoice. API keys are created by individual teams, some paid on cards and expensed, some in the cloud bill under a generic line. Personal accounts on consumer plans carry company work. None of it is tagged by use case, so none of it can be set against the value it produces.

The FinOps Foundation's 2026 survey names the consequence. The top challenges practitioners cite for AI spend are visibility into AI costs, allocating them to business units and determining their

value. One practitioner quoted in the report sums it up: “Is your AI providing value? No one can answer that question yet.”

The corrective

Give model calls one entry point and tag every call by team, use case and environment. Find the spend outside it with a shadow AI inventory of keys, expense lines and consumer accounts. Show each team its AI cost every month, next to the measure its use case was approved to move.

The Switching Cost

The cheapest model next year is worth nothing if you cannot move to it. Providers retire model versions, change prices and revise terms, and an application tuned to one model degrades or breaks when that model goes. The bill then arrives as engineering work under a deadline set by someone else.

Switching cost is set long before the switch. Applications that call one provider's API directly, with prompts tuned by trial and error and no record of what good output looks like, cannot be moved quickly. Applications that call models through one internal interface, with an evaluation set per use case and a tested fallback, can treat a retirement notice as a test to run rather than a project to staff. The fallback can be a second provider or an open-weight model on infrastructure you control.

When Owning the Hardware Pays

Rising bills invite the conclusion that the answer is to run models on your own GPUs. On cost alone it usually is not. Our guide to self-hosting an LLM in production (01ltd.com/learn/self-hosting-llm-production) works the arithmetic for a 70-billion-parameter model on owned hardware and finds that, in 2026, it does not beat per-token APIs for the same model below a break-even volume: annual self-hosting cost divided by the API cost per request. Against the highest list price in that comparison the owned node breaks even at about 40 million requests a year; against the cheapest listing it never does.

Self-hosting wins for other reasons: latency budgets a round trip cannot meet, data that must not leave your environment, and very high, steady utilisation. Below the break-even volume, idle GPUs are the most expensive tokens an organisation can buy. Most organisations end up with both, routed by workload, which is one more reason the routing layer matters more than the choice of side.

The Control Plane: One Gateway, One Unit Metric, One Review

The five leaks have the same root: spend is decided at the edges, by whoever wrote the integration or signed the licence, and nobody sees the total until the invoice. The control plane moves the decisions to one place without slowing the teams down.

1. **One gateway for every model call.** Authentication, routing, caching, budgets and logging in one place, with every call tagged by team, use case and environment. The gateway reads every prompt, so run it yourself or contract it as a processor, as our whitepaper Follow the Prompt sets out.
2. **One unit metric.** Cost per completed task, set against the business measure each use case was approved to move. Cost per token and cost per seat describe the supplier's price list, not your economics.
3. **Budgets that stop runaway work.** A spend budget per use case with alerts, and step and token ceilings on agent loops, so a stuck agent ends in an alert rather than an invoice.
4. **A monthly review with owners.** The largest routes by spend, candidates for a smaller model or for batch, idle seats before renewal, models due for retirement. Each line has an owner and a decision.
5. **Purchasing rules.** Seats bought on measured use, AI price uplifts in existing software accepted only with usage data, and new model providers reached only through the gateway.

The unit metric is the piece most often skipped and the one everything else depends on. Without a cost per completed task there is no way to tell a cheaper route from a worse one, and without the value measure beside it there is no way to tell spend that should be cut from spend that should grow. That second half is the subject of our whitepaper The AI ROI Mirage.

Conclusion: Budget the Task, Not the Token

Token prices will keep falling, and for organisations that only watch the price per token the bill will keep rising. McKinsey describes the organisations moving fastest as treating operating costs as a design constraint, not an afterthought. That is the whole discipline: decide at design time what a task is allowed to cost, measure what it does cost, and route, cache, batch or retire until the two agree.

Cheap tokens are a gift. Unmetered loops, frontier defaults, real-time pricing for batch work, idle seats and invisible keys are how organisations give it back.

Five questions for next month's review

1. Can you state last month's AI spend by team and by use case?
2. Do you know the cost per completed task of your three largest AI workloads?
3. Which workloads run at real-time prices although nobody waits for the answer?
4. How many paid AI seats were idle last month?
5. If your main model were retired next quarter, how long would the switch take?

Related use cases at 01ltd.com/use-cases: AI Spend Control, AI ROI Measurement, Model Retirement & Provider Exit Plan, and Shadow AI Inventory.

References

1. Stanford Institute for Human-Centered AI (2025). "AI Index 2025: State of AI in 10 Charts": the cost of querying a model scoring at GPT-3.5 level on MMLU fell from \$20 to \$0.07 per million tokens between November 2022 and October 2024.
 2. McKinsey & Company (2026). "The state of AI in 2026: On the road to ROI." August 2026. McKinsey Global Survey, 1,719 participants, 4 May to 8 June 2026.
 3. FinOps Foundation (2026). "State of FinOps 2026." 1,192 respondents representing more than \$83 billion in annual cloud spend.
 4. Anthropic (2026). "Pricing", Claude Platform documentation, accessed 11 October 2026: Batch API at a 50% discount on input and output tokens; prompt-caching multipliers of 1.25x (5-minute write), 2x (1-hour write) and 0.1x (cache read) on the base input price; a 1.1x multiplier for US-only inference.
 5. 01AI LTD (2026). "Self-hosting an LLM in production: hardware, cost, latency." 01ltd.com/learn.
 6. 01AI LTD (2026). "The AI ROI Mirage" and "Follow the Prompt." 01ltd.com/blog.
-

This whitepaper is part of 01's research series on enterprise AI posture. 01 helps enterprises run AI at a cost they can defend, from spend visibility and model routing to the gateway, metrics and review that keep the bill tied to value.

www.01ltd.com/contact