

WHITE PAPER · COST

Pay for outcomes, not tokens

Why consumption pricing rewards the labs, what it is costing UK businesses, and how an independent control layer changes the economics.

Steven Myers, GenWA

SUMMARY

The meter runs for the vendor

Every AI vendor bills by consumption. Consumption rewards the vendor whether or not anything got done. That is not a scandal; it is a pricing model. But it means nobody on the vendor's side of the table is paid to make your usage smaller, and almost nobody on your side can see it.

This paper sets out what consumption pricing is doing to UK businesses, why the choice of model has quietly become the largest controllable line in the AI budget, what an independent control layer can take out of the bill, and how outcome pricing realigns the incentive: you pay for work delivered and verified, and the token cost becomes someone else's problem to minimise.

IN SHORT

Token usage across one large customer base rose 1,001% in fifteen months while spend rose 497% (Ramp, April 2026, vendor data). Premium models now take 56% of AI cost, up from 6% ten months earlier, for tasks on which cheaper models perform "well enough" (a16z, 2025). Routing each task to the cheapest adequate model retains about 95% of frontier quality at up to 85% lower cost (LMSYS, 2024). Only 26% of companies fully track what AI costs them (KPMG, 2026). For a 300-person UK business we estimate the controllable saving at £16,000 a month, £19,900 with open-source models in the mix.

The answer is not to use less AI. It is to stop paying frontier prices for routine work, to see the meter before the invoice, and, where the work is an agent's, to pay per verified outcome with the token cost carried by the party whose incentive is efficiency.

+1,001%

token usage in fifteen months. Spend +497%.

Ramp, April 2026 · vendor data

56%

of AI cost goes to premium models, up from 6%.

Ramp, April 2026 · vendor data

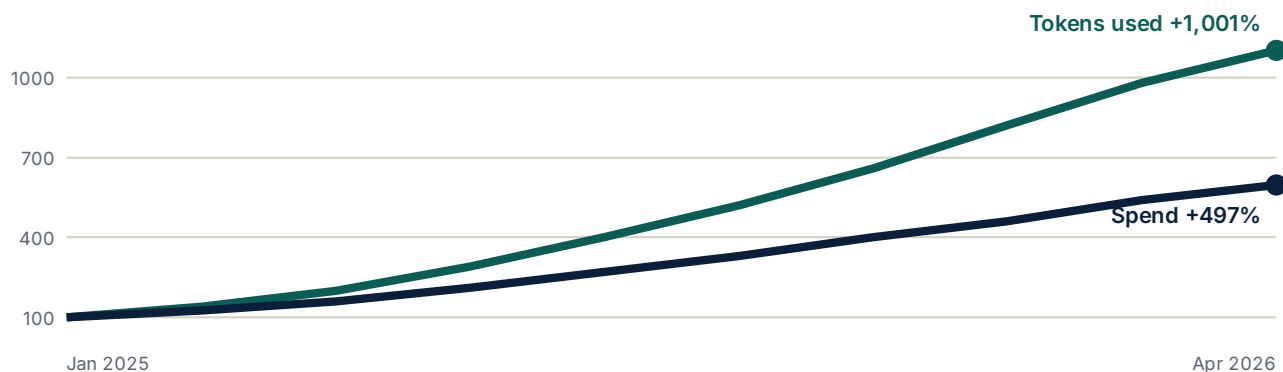
26%

of companies fully track what AI costs them.

KPMG, 2026

01 · THE METER

Software used to be a seat. AI is a meter.



Indexed, January 2025 = 100. Token usage and spend across Ramp's customer base. Ramp AI Index, April 2026 (vendor data).

For twenty years business software was bought by the seat: a known number of users at a known monthly price, renewed annually, negotiated once. AI broke the model in two ways at once. Most AI is now billed by consumption, per token in and per token out, and consumption is compounding, because every new capability creates new use. Unit prices are falling by 9 to 900 times a year for a fixed level of capability (Epoch AI, 2025) and bills are rising anyway, because use grows faster than prices fall and because each new generation of model consumes more tokens per task than the last.

The result is a cost line that behaves like cloud infrastructure in 2012: elastic, invisible to the people who authorise it, and discovered on the invoice. 62% of organisations have had unexpected AI costs large enough to change a business decision, and a third have imposed emergency spending freezes (Mavvrik, 2026, vendor survey). Only 11% can forecast AI spend to within 10%. Uber, having told its engineers to use AI as much as possible, spent its entire 2026 AI tooling budget in four months and capped every employee at \$1,500 a month per tool (Bloomberg, June 2026).

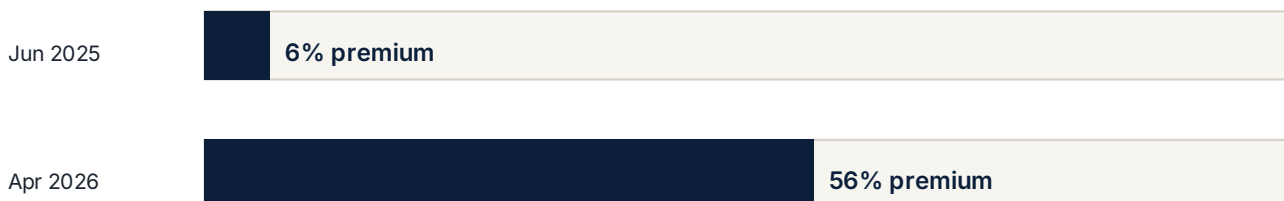
The structure of the problem is the same in every business we look at. The people who see usage, engineers and team leads, have no budget authority. The people who pay, finance, see the bill weeks later and cannot attribute it. Business units, not IT, control 81% of software spend (Zylo, 2026, vendor data). ChatGPT is now the single most-expensed application in business; expense-report software spend is up 267% year on year. In the UK, 42% of leaders have only partial visibility of AI spend and 33% say they do not understand AI cost structures, tokens included (KPMG UK, July 2026, n=116).

A meter nobody can read is not a cost control. It is a subscription with no ceiling.

None of this is the vendors' fault, and none of them will fix it. Their revenue is the meter. Anthropic's own documentation notes that unexpectedly high spend "usually traces back to long sessions that were never cleared or to Opus left as the default model" (Anthropic, 2026). Cursor's move to consumption pricing in June 2025 emptied users' allowances within days and ended in a public apology (TechCrunch, July 2025). The vendor tells you how the meter works. It does not turn it down.

02 · THE MODEL

Which model does the work is the largest decision nobody makes



Premium models' share of total AI cost. They account for 46% of tokens but 56% of cost. Ramp, April 2026 (vendor data).

Inside most businesses, which AI model handles a task is decided by whatever the tool defaulted to. The price spread between the cheapest and most expensive mainstream models is about 20 times, for tasks on which, in the words of a survey of 100 enterprise chief information officers, "all models perform well enough now" (a16z, 2025). When OpenAI switched on automatic routing in 2025, daily use of its expensive reasoning models among paying users jumped from 7% to 24% overnight (OpenAI, 2025). That is how little of model choice is a considered decision.

The economics of getting it right are well established. Routing each request to the cheapest model that will do the job retains about 95% of frontier quality at up to 85% lower cost on standard benchmarks (LMSYS RouteLLM, 2024). Most requests in a business are routine: summarising, drafting, classifying, extracting, answering from a document. A model that costs a twentieth of the frontier price handles them. Sprawl compounds the waste: companies using 26 or more models spend \$442 per employee per month on tokens against \$28 for those using four to ten (Ramp, 2026, vendor data).

Open-source and open-weight models change the arithmetic again. Run on your own cloud or a controlled one, they cost a fraction of a lab's API price for the same class of routine task, keep the data inside your boundary, and cannot be retired or repriced by a vendor. They are the right answer for a large share of routine work and the wrong answer for some. The decision has to be made per request, automatically, against a policy, which no employee at a keyboard can do.

85%

lower cost at 95% of frontier quality with routing.

LMSYS RouteLLM, 2024

20×

price spread between mainstream models.

a16z, 2025

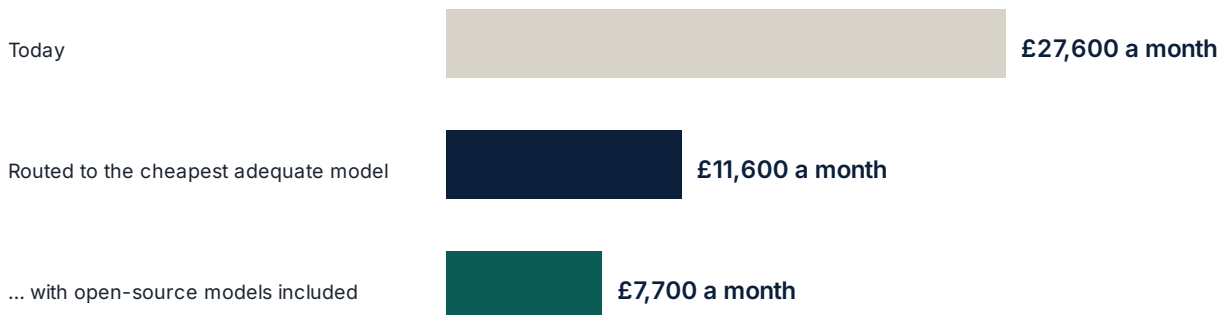
\$442 vs \$28

monthly token cost per employee, 26+ models against 4-10.

Ramp, 2026 · vendor data

03 · THE MONEY

What a 300-person business is leaving on the table



Monthly AI spend at £92 per employee (Ramp, 2026: AI at 15% of software spend; Zylo, 2026: £7,400 software spend per employee). Standard routing: 56% of spend on routable tasks × 75% routed × 85% saving, plus 36% unused seats on 40% of spend, plus 8% duplicates. Open-source models: 85% routable at 92% cheaper plus 45% of the standard tier at 80% cheaper. Capped at 72%. Run your own numbers at genwa.co.uk/#/businesscase.

FOUR SOURCES OF SAVING, IN ORDER OF SIZE

Routing. The largest line, and the one that requires no change in how anyone works. The request goes to the cheapest adequate model; the person asking sees the same interface and, on routine work, the same answer.

Open-source models. A second tier of routing for the work that a well-run open model does as well as a lab's standard model, at a fraction of the price and inside your boundary.

Unused seats. 36% of software licences go unused (Zylo, 2026, vendor data). AI seats are no different, and AI features inside software you already own are often paid for twice.

Duplicates. Two teams paying for the same capability is the norm, not the exception; generative AI entered the list of most redundant software functions for the first time in 2026 (Zylo, 2026).

£16,000

a month back for a 300-person firm at standard routing. £192,000 a year.

GenWA calculator · assumptions above

£19,900

a month with open-source models in the mix. £238,000 a year, at the 72% cap.

GenWA calculator · assumptions above

WHY THE NUMBER IS NOT CONSERVATIVE

The assumptions come from vendor telemetry across thousands of businesses rather than from our own accounts, and the audit replaces them with yours. In practice the first month's register usually finds spend that nobody had counted at all, which is not in this figure.

04 · THE CONTROL LAYER

Someone has to be paid to make the bill smaller



The vendor will not do it. The employee cannot. A tool that sits between the two, and is paid by the business rather than by the volume, can. The gateway is one door between everything that uses AI, people, agents and systems, and every lab that provides it. On every request it identifies who is asking, checks the data against the rules, chooses the model that satisfies the task and the policy at the lowest cost, records the request and the outcome, and applies the budget.

Because the rules are written about tasks, data and risk rather than about a vendor, a price change on one model is a routing change, not a bill; a retirement is a test and a change record, not a rewrite. That matters more each quarter: OpenAI retired four models from its consumer product in February 2026 with about two weeks’ notice; Anthropic retired six between February and August 2026; the leaderboard has changed hands nineteen times since 2023.

WHAT THE CONTROL LAYER DOES TO THE BILL

LEVER	EFFECT
Routing by task class	Routine work to the standard tier or an open model; premium by policy, not default
Caps per person, team and agent	Work pauses at the limit rather than billing over it; exceptions by approval
Caching and de-duplication	The same question, asked twice, costs once
Seat and duplicate retirement	The register shows what is unused and what is paid for twice
Model Watch and Change Control	Price, limit and terms changes tracked before they land; a vendor change is a governed event
A meter the board can read	One number a month, by team and tool, against budget, before the invoice

Pay for outcomes, not tokens

Where the work is an agent’s, the business should pay for the work, verified, and the tokens should be the problem of whoever is paid to minimise them.

The principle. Token consumption rewards the labs whether or not anything got done. Tokens therefore need policing by a party whose interests are aligned with the customer’s outcomes rather than the vendor’s volume. GenWA is that party: its job is to drive token cost down and to be paid for what was delivered.

The model. A fixed platform fee per organisation for the control layer that outcome pricing depends on: the gateway, the ledger and the console, because you cannot verify outcomes without the record. Then a published price per verified outcome for each agent, by agent and by sector, with a monthly cap the customer sets. An outcome is verified when the ledger records it and the customer’s named owner signs it off. Only signed-off outcomes are billed.

The incentive. Token cost sits with GenWA inside the outcome price. Model routing, caching and sandbox testing are therefore our margin levers, which is the point: our incentive is efficiency, the labs’ incentive is volume. The value register shows outcome count, unit price, token cost and GenWA’s margin to the customer, transparently. That transparency is the proposition.

The definition. What counts as an outcome is written into each agent’s specification with the owner before go-live. A customer communication sent that passed the compliance check. A KYC case prepared for decision with all checks complete. An invoice matched to a purchase order or a documented exception. A complaint acknowledged inside the regulatory deadline. A contact-centre call summarised and wrapped.

GUARD-RAILS

No charge for held, blocked or rejected outputs.

No charge for outcomes the owner reverses within 30 days.

A cap breach pauses the agent; it never bills over the cap.

Outcome count, unit price, token cost and margin visible to the customer at all times.

INDICATIVE RATE CARD [CONFIRM]

OUTCOME	PRICE
Customer communication sent, compliance-checked	£0.35–£0.60
KYC case prepared for decision	£4–£9
Invoice processed to PO or exception	£0.40–£0.80
Complaint acknowledged and triaged	£2.50–£4
Contact-centre call summarised and wrapped	£0.20–£0.35

Ranges to be confirmed against measured token cost. Floors: platform fee plus a minimum monthly commitment.

Seven questions for your AI vendors, and one for your finance director

1

What is our default model, and who set it?

If the answer is “the vendor did”, the premium share of your bill is not a decision.

2

What did we spend last month, by team, and how does it compare with the seats we pay for?

If the vendor cannot tell you, you have a meter you cannot read.

3

Can we set a hard cap per user and per agent that pauses rather than bills?

Most vendors offer alerts. Alerts are not caps.

4

How much notice do you give on a price, limit or model change?

Two weeks is the recent norm. Ask what happens to work in flight.

5

Can the same task run on a cheaper or open model, and who decides?

If the answer is “no”, the vendor is selling you the meter, not the outcome.

6

What do we get for the tokens?

Ask for outcomes: documents produced, cases prepared, calls wrapped. Tokens are an input.

7

Would you price this per outcome?

Watch how quickly the conversation turns to how hard outcomes are to define. They are not.

AND FOR THE FINANCE DIRECTOR

“Which of our AI spend lines are defended by a measured return, and which by fear of falling behind?”

64% of chief executives admit that the risk of falling behind drives investment before value is understood (IBM, 2025). 56% report no significant financial benefit from AI yet (PwC, January 2026). 92% of finance chiefs feel pressure to demonstrate a return (Avalara, 2026, vendor survey).

£0

for the AI health check: ten minutes online, 48 hours with your usage data, or half a day on site.

genwa.co.uk/#/healthcheck ·
genwa.co.uk/#/businesscase

Sources

Official statistics, regulators and courts first; vendor-commissioned surveys and vendor telemetry are marked as such.

1. Ramp, AI Index, April 2026 (vendor data).
2. Epoch AI, LLM inference price trends, 2025.
3. Mavvrik, AI cost survey, 2026 (vendor survey).
4. Bloomberg, "Uber Caps Employee Spending on AI Tools Like Claude Code to Manage Costs", 2 June 2026; TechCrunch, 2 June 2026.
5. Zylo, SaaS Management Index, 2026 (vendor data).
6. KPMG UK, AI cost visibility survey, July 2026 (n=116); KPMG, Global AI Pulse, 2026.
7. Anthropic, Claude Code documentation on cost, 2026.
8. TechCrunch, "Cursor apologizes for unclear pricing changes that upset users", 7 July 2025.
9. a16z, "How 100 enterprise CIOs are building and buying gen AI", 2025.
1. OpenAI, GPT-5 routing usage note, August 2025.
2. LMSYS, RouteLLM, 2024.
3. OpenAI, "Retiring GPT-4o and older models", 29 January 2026; Anthropic, model deprecations documentation, 2026.
4. BenchLM leaderboard tracker, September 2026 (secondary).
5. IBM Institute for Business Value, CEO Study, 2025.
6. PwC, Global CEO Survey, January 2026.
7. Avalara, CFO survey, 2026 (vendor survey).
8. GenWA, Pricing Strategy and ROI Model, \$7 Outcome pricing, September 2026.

© GenWA Ltd 2026. This white paper is general guidance and not legal advice. Figures are attributed to their sources and were correct to the best of our knowledge at publication. Outcome prices are indicative and subject to confirmation.



All the upside of AI. None of the surprises.

[GENWA.CO.UK](https://genwa.co.uk)