emma

INSIGHT BRIEF · ENTERPRISE AI — VALUE, COST & CONTROLE

Your AI Spend Is Up. Can You Prove the Value, See the Cost, or Show the Control?

A financial-services leader's playbook for closing the gap between what AI costs and what it returns — why spending more, piloting more, and locking down all widen it, and the one discipline that closes it.

CTOS, CIOS, CHIEF DATA

AI OFFICERS, AND HEADS OF AI

PLATFORM AT BANKS, INSURERS, AND ASSET MANAGERS.

Your AI Spend Is Up. Can You Prove the Value, See the Cost, or Show the Control?

Your AI budget went up again this year. Almost everything else, the value you can prove, the cost you can attribute, the control you can demonstrate to a regulator, did not keep pace.

That gap is the problem. And the three reflexes you'll reach for to close it — spend more, pilot more, or centralise and lock down — each tend to widen it.

The spending is not in doubt. Enterprise generative-AI spend roughly tripled in a year, from $11.5 billion in 2024 to $37 billion in 2025 (Menlo Ventures), and Gartner puts total worldwide AI spending on the order of $1.5 trillion in 2025, rising toward $2.5 trillion in 2026.

The return is another matter. You've seen the headline that 95% of enterprise GenAI pilots fail. It's worth knowing where it comes from: a 2025 MIT study built on 52 interviews and 153 survey responses, counting any pilot short of rapid P&L impact as a "failure" — striking, but thinner than its fame suggests, and already contested by analysts who study the market.

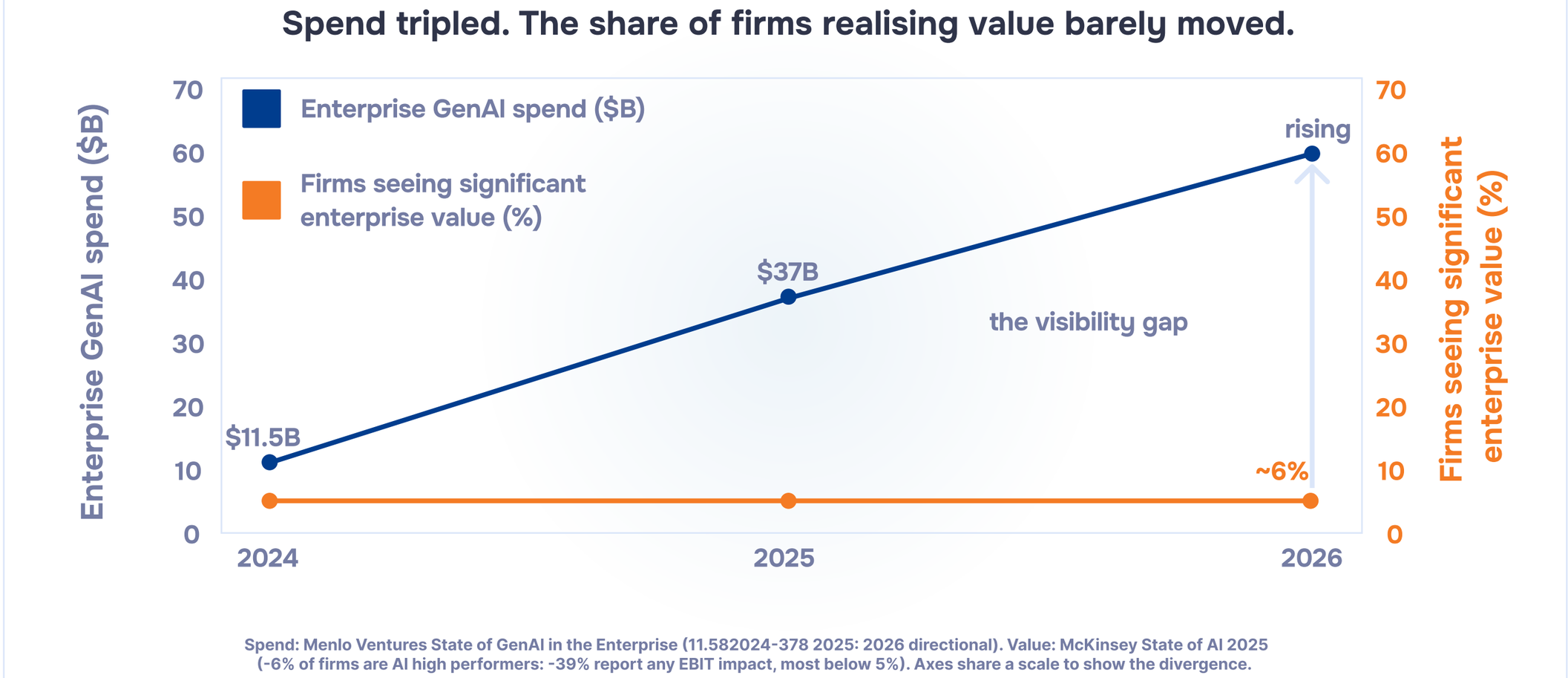
The more carefully-built numbers land in the same place. McKinsey's 2025 State of AI (nearly 2,000 respondents) finds only about 6% of organisations are AI "high performers," and roughly 39% report any enterprise EBIT impact at all — most of them below 5%. BCG finds only 26% of companies have moved past proof-of-concept to tangible value, and just 4% are generating significant value across functions.

Part of that gap is a value problem, and it's worth saying so plainly. Picking the right use cases, redesigning the workflows around the model instead of bolting the model onto them, winning adoption — that is the work McKinsey's high performers are defined by, and no platform, vendor or framework does it for you. But it isn't the whole story, and it isn't the neglected part. Every institution is already working that lever.

From the field

Across 400+ recent cloud and AI buying conversations we analyzed, security and compliance came up in more than four out of five (83%), and cost in roughly three out of four (73%).

The lever almost nobody works is the other one: what each AI outcome actually costs to produce, where it runs, and whether you can show it's controlled. The first lever decides whether your AI creates value. The second decides whether you can see it, afford it at scale, and defend it to a supervisor — and it's the one that is fully within your operating model's control.



Spend: Menlo Ventures State of GenAI in the Enterprise (11.582024-378 2025: 2026 directional). Value: McKinsey State of AI 2025 (-6% of firms are AI high performers: -39% report any EBIT impact, most below 5%). Axes share a scale to show the divergence.

Three reflexes — and why each one widens the gap

Look at your title questions again: prove the value, see the cost, show the control. Under pressure, institutions reach for an obvious answer to the first and the third. "Spend more" and "pilot more" are attempts to buy and to prove value. "Centralise and lock down" is an attempt to show control.

Each is intuitive. Each, done the obvious way, widens the gap it was meant to close — and notice what none of them even touches: the middle question. Nobody's reflex is "see the cost."

Reflex 1 — "Spend more to stay competitive."

The evidence is blunt: the strongest single correlate of AI value in McKinsey's data isn't model quality or budget — it's fundamental workflow redesign. So a bigger budget fails twice over. It buys more models bolted onto unchanged processes, which is the spend that shows up in the bill and not in the EBIT. And because none of it is attributed, you can't even see which of the new spend is failing. More budget without measurement doesn't close the value gap. It funds it — and hides where.

Reflex 2 — "Run more pilots." Usually proliferates proofs-of-concept that never reach production.

Faced with the control gap, the reflex is to route every AI initiative through one central committee. It feels like governance. It behaves like a queue.

Grant Thornton's 2026 survey found that centralised review bodies "become overwhelmed, creating bottlenecks that slow execution without reducing risk," and that 78% of leaders lack confidence they could pass an independent AI governance audit within 90 days. Meanwhile, Gartner expects 40% of agentic-AI projects to be scaled back or scrapped by 2027, governance gaps among the named causes.

Treating governance as a single on/off gate doesn't deliver control. It delivers a slower version of the same exposure.

> **The one-sentence version:**
>
> your AI bill tells you what you spent — not whether it worked, what each use cost, or whether you could explain it to a supervisor. Those three blanks are your three questions — and the reflexes above are what happens when you try to answer the first and third without ever answering the second.

The reframe: value, cost and control are one ledger, not three programmes

Every reflex above broke on a missing record.

Spending more failed because nothing tied the new spend to an outcome — you bought more of what you couldn't measure. Piloting more failed because nothing carried a pilot's cost, owner and success metric into production — the proof died at the cliff. Locking down failed because a committee is not a record — it can queue decisions, but it can't show a supervisor what's actually running.

Three failures, one missing artefact: a complete, current account of every AI system you run — what it costs, who owns it, whether it works, and how it's controlled. Proving value, attributing cost, and demonstrating control look like three programmes owned by the business, by finance, and by risk. They are three views of that one ledger.

01

Did it work?

Measured outcomes per use case. This is what graduates a pilot, kills a zombie, and tells you which of the value lever's bets are paying. The ledger doesn't create value — that's workflow and use-case work, and it stays yours — but it is the only thing that shows you where value is being created and at what price.

02

What did it cost?

every AI call mapped to a feature, a team, a customer, and the infrastructure it ran on. This is the neglected middle question, and it's the precondition for the other two: an outcome you can't cost can't be judged, and a system you can't cost can't be governed.

03

Can you show it's controlled?

owner, risk rating, documentation, and monitoring for every entry. This is the same record the EU AI Act, model-risk supervision and DORA third-party obligations will ask for — not three parallel filings.

Build it once and finance, model risk, and the regulator read from the same page. Build it three times, in three silos, and you'll reconcile nothing and scale nothing.

For a regulated institution, that changes the metric. Not spend, and not pilot count. The number on the wall is **cost per successful outcome** — what it costs to produce one correct, accepted result — trended against the value that outcome creates, with every system in the inventory tied to an owner and a control.

Why "per outcome" and not "per token" or "per call": agentic and reasoning workloads are non-deterministic and expensive in ways per-token pricing hides. A Stanford and Microsoft Research study found agentic tasks can consume **orders of magnitude more tokens** more tokens than a simple query. Multi-step reasoning, tool calls, retries, and long context accumulation can easily push usage 100×–1000× higher than a single-shot prompt. Token usage for the same task can also vary dramatically run-to-run. Two runs of the same task can differ by up to 30× in tokens, and models systematically underestimate their own cost. Two features with identical token cost can have wildly different cost-per-successful-outcome once you count retries and failures. Per-token budgets will lie to you exactly as agents scale.

Where are you on the ladder?

AI maturity in a regulated institution is a climb from "we're spending and hoping" to "we know what each AI outcome costs, that it works, and that we can prove it's controlled."

Most teams are stuck low — spending is near-universal, but the inventory, the attribution and the outcome metric are not. Spend you can't attribute to an outcome is just faster burn with a better story.

STAGE	WHAT IT LOOKS LIKE	THE QUESTION YOU CAN FINALLY ANSWER
Spending	AI budget rising; activity everywhere; value, cost and control all anecdotal.	"How much are we spending on AI?"
Inventory	Every model, prompt and agent is catalogued, owned, and tied to a use case — including the shadow ones.	"What AI do we actually run, and who owns each piece?"
Attribute	Every AI call maps to a feature, team and customer; cost per successful outcome is visible.	"What does each AI use cost — and per result?"
Prove	Outcomes are evaluated; pilots graduate only with a value metric and an owner; impact is measured, not asserted.	"Did it work, and was it worth it?"
Govern to scale	Proportional, federated controls and economic guardrails run in the execution path; one inventory serves finance, model risk and the regulator.	"Can we scale this and prove value, cost and control on demand?"

Most institutions plateau at **Spending**. The leap that changes the economics is **Spending → Inventory → Attribute** — the point where AI stops being an act of faith and becomes a managed portfolio.

A 12-question diagnostic for your next AI and risk review

Read these with your AI/platform, finance and risk leads in the room together — the same room, because the answers are interlocked. If you can't confidently answer eight, the gap is visibility, not ambition or budget.

Value (did it work?)

01 **For our top five AI use cases, what measurable business outcome has each delivered — in numbers, not anecdotes?**

02 **What share of our AI use cases has a disclosed impact metric at all?**

03 **How many of last year's pilots reached production with a named owner — and how many quietly lapsed**

Cost (can we attribute it?)

04 **What did each AI feature cost last month, mapped to team and customer?**

05 **What is our cost per successful outcome on our largest AI use case — and which way is it trending?**

06 **What did our agentic workloads cost, and how much did the most expensive single run cost (the p99)?**

07 **What share of our AI spend can we not attribute to any owner at all?**

Control (can we prove it?)

08 **Do we have a complete, current inventory of every model, prompt and agent — including ones bought inside SaaS tools?**

09 **Could we pass an independent AI governance audit within 90 days?**

10 **Which of our AI systems make or materially inform high-risk decisions (credit, pricing, underwriting), and are they documented to that standard?**

11 **How much of our AI capability depends on a single foundation-model or cloud provider — and what's our exit?**

12 **How much AI is running through unsanctioned, personal accounts right now — and do we know what data it sees?**

From insight to action: six moves that close the gap

Skip the table stakes ("measure your ROI," "stand up a governance committee," "spend to keep up"). These are the moves that separate institutions that scale AI from institutions that keep funding it — and each one pays off in value, cost and control at the same time.

Move 1. Build the AI inventory first — it's the artefact everything else reads from

Before you optimise or govern anything, catalogue it. A live inventory — sometimes called an AI bill of materials — lists every model, prompt, agent, dataset and dependency, with its purpose, owner, risk rating, third-party status, and the decisions it informs.

Do this: stand up the inventory and tie each entry to a cost tag and an owner. This single artefact is what your finance team needs for attribution, your model-risk team needs for validation, and your EU AI Act and DORA obligations need for documentation. When a foundation model is later found vulnerable or non-compliant, the inventory is what lets you find every place you use it in minutes, not weeks.

Move 2. Measure cost per successful outcome, not per token or per call

Put one number in front of the business: what it costs to produce one correct, accepted result. Not tokens, not API calls — outcomes that passed a quality bar.

Do this: instrument a quality score (an eval) on each AI use case, and divide spend by results that pass it. Track the median and the p99, because a runaway agent loop hides in the tail. This is the metric that ties FinOps to value — and it routinely surprises: in instrumented estates it's common to find a small number of features or customers driving the majority of AI cost while contributing a fraction of the value.

And the same eval does double duty as control evidence: a documented, versioned, regularly-run quality test on a model's outputs is exactly the kind of ongoing-monitoring artefact a model-risk function needs to validate an AI system. One instrument, two obligations — the value metric and the validation record are the same record.

Move 3. Make pilots earn production from day one

The cliff between a working pilot and a production system is where most AI budget dies. Close it by refusing to start a pilot that isn't designed to graduate.

Do this: require three things before a pilot is approved — a named production owner, a feature-level cost forecast (model the cost at full rollout, where one prompt change can multiply cost across every customer and turning on agentic mode can multiply it again), and a single measurable success metric. A pilot that can't define those isn't ready to spend on. Kill or redesign the ones already running that can't.

Move 4. Route to the cheapest model that does the job — and instrument quality while you do

Model choice is the first cost lever; placement is the second, and most estates only pull the first. Not every workload needs frontier-API pricing, and not every workload needs on-demand capacity: batch inference and evaluation runs tolerate interruption and can run on discounted capacity; steady high-volume workloads on open-weight models often cross the threshold where self-hosted GPU capacity beats per-token API pricing; latency-critical paths earn premium placement, the rest don't. The same non-determinism that breaks per-token budgets breaks capacity planning — plan for the p99 run, not the median.

Do this: classify each AI workload the way you'd classify any other — latency tolerance, interruption tolerance, volume, data residency — and place it accordingly, across providers where the economics or the residency requirement says so. A workload classification record does for AI cost what the placement record does for sovereignty: it replaces one blanket decision with a per-workload one. And it feeds the same exit question your supervisor is already asking — concentration on one model provider or one GPU source is lock-in with a newer name.

Move 5. Treat the EU AI Act deadline as the date you owe the work — not the date to start it

The EU AI Act's high-risk obligations land squarely on financial services: credit scoring and life/health insurance pricing are explicitly named high-risk uses. The original date for those obligations is August 2026; a deferral to December 2027 was politically agreed in 2026 but is not yet law.

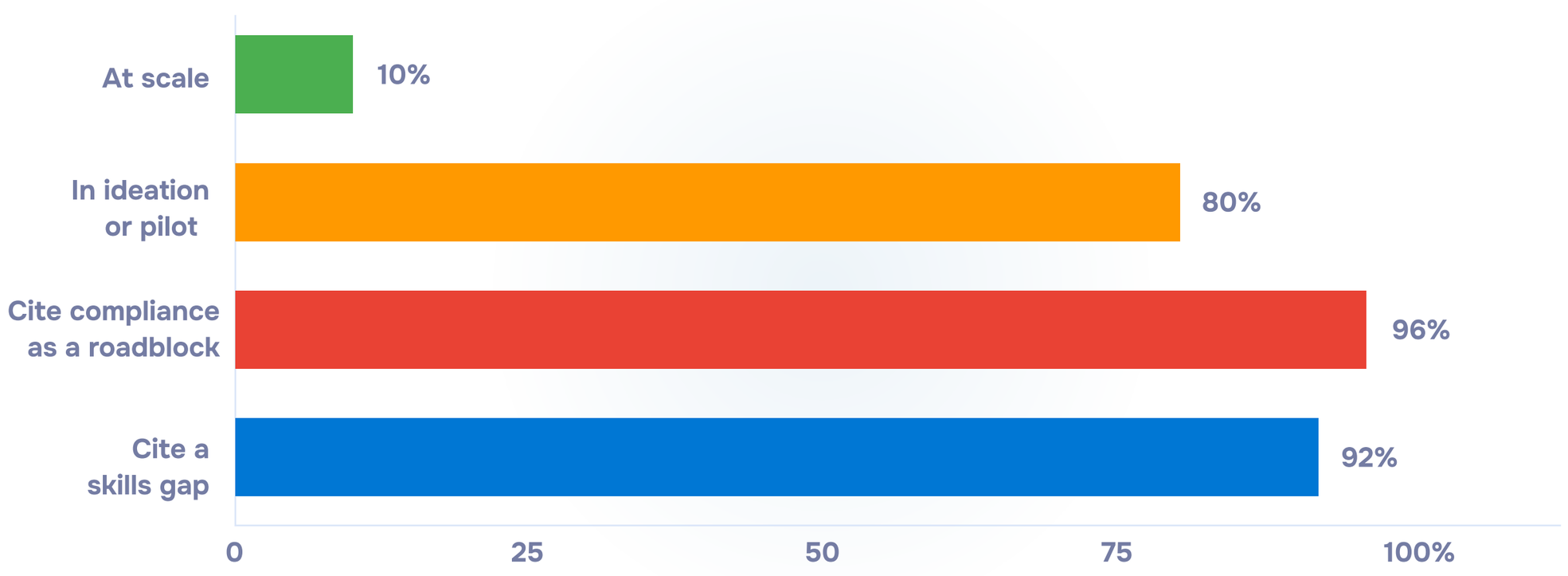
Do this — and here's the point that matters more than the date: plan to the earlier date, because the work doesn't change if the deadline moves. The obligations — a risk-management system, data governance, technical documentation, human oversight, a registered system — are very nearly the same artefacts your existing model-risk governance and DORA third-party rules already require. Build the inventory and the documentation once and you satisfy all three regimes. The institutions that wait for the date confirmation will do the same work later, under more time pressure, for no benefit.

Move 3. Govern proportionally and put the guardrails in the execution path

Central committees don't scale, and humans can't review agent decisions at machine speed — even large enterprises run FinOps with a handful of people. So control has to be designed, not staffed.

Do this: set policy and risk thresholds centrally, then delegate proportional review to the divisions — match the depth of review to the risk of the use case, rather than gating everything equally. And move the real-time controls into the execution path: per-agent and per-team budgets with hard ceilings, anomaly detection, and the ability to automatically pause, cap or reroute behaviour before it compounds. Governance that enables scale looks like guardrails, not a queue — and the evidence is that firms with mature governance are markedly more likely to report AI-driven growth, not less.

Financial services: lots of AI ambition, very little at scale



Capgemini World Cloud Report - Financial Services 2026 (1,100 FS executives, 14 markets): 10% run AI agents at scale; 80% in ideation or pilot; 96% cite regulatory/compliance burden; 92% cite a skills gap. Ambition is near-universal; production is rare.

Seven anti-patterns to avoid

Budgeting AI per token on infrastructure you never chose

runs vary by up to 30×. So, defaulting every workload to frontier APIs on premium capacity pays the maximum price for the maximum variance. Track cost per successful outcome and place workloads deliberately.

Counting pilots as progress

a pilot with no production owner or cost model is a demo, not a step toward value.

Budgeting AI per token in an agentic world

runs vary by up to 30×; track cost per successful outcome, including the p99.

One central gate for all AI

it slows the safe and the risky equally and reduces neither's exposure.

Treating governance as a compliance tax

done as enablement it correlates with growth; done as a blocker it correlates with shadow AI.

Ignoring shadow AI

ungoverned tools on personal accounts are invisible spend and a live data-leak path; you can't inventory what you pretend isn't there.

Single-provider dependence with no exit

concentration on one foundation-model or cloud provider is the AI version of the lock-in your supervisor is already asking about.

Start before your next board or risk report

Don't launch a programme. Build one artefact and take it to the table you already have scheduled.

Pick your five largest AI use cases and put each into a one-page record: what it costs (mapped to an owner), what measurable outcome it has produced, and whether it could survive an independent governance review. Add the high-risk ones — anything touching credit, pricing or underwriting — with their documentation status.

You will almost certainly surface at least one expensive use case with no measurable value, one material AI spend with no clear owner, and one high-risk system you couldn't yet document to standard. That is your agenda — concrete, defensible, and legible to the business, finance and your supervisor in the same breath.

The one number to put on the wall

Stop staring at the AI budget. Put **cost per successful outcome, trended against the value that outcome creates**, where the team can see it — with every system behind it sitting in an owned, auditable inventory.

It turns "is our AI spend worth it?" from a board-meeting argument into a question you can answer and improve. It exposes the pilots that never pay back, the agents whose cost runs away, and the shadow tools nobody owns — in one view.

And it's the number that aligns the three people who otherwise talk past each other: when the business, finance, and risk all argue from cost-per-successful-outcome over an inventory they trust, the conversation shifts from "spend more" to "scale what works, prove it, and control it." Which is the only version of enterprise AI that survives both the budget review and the supervisor.

This brief was put together by the team at emma. If a second set of eyes on your own estate would help — a quick read on what your AI actually costs by use case, where value is and isn't landing, what's running ungoverned, and how your inventory would stand up to a review — that's a conversation we're glad to have. No deck required.

Sources

- Menlo Ventures, State of Generative AI in the Enterprise 2025 (enterprise GenAI spend $11.5B in 2024 to $37B in 2025). Gartner, worldwide AI spending forecasts 2025–2026 (~$1.5T in 2025 rising toward ~$2.5T in 2026; AI in the "Trough of Disillusionment"; ROI predictability a precondition for scaling).
- McKinsey, The State of AI 2025 (~1,990 respondents: ~6% AI high performers; ~39% report any enterprise EBIT impact, most below 5%; workflow redesign the strongest correlate of value). BCG, Where's the Value in AI? 2024 (1,000 executives, 59 countries: 26% beyond proof-of-concept to tangible value; 4% generating significant value across functions). Deloitte, State of Generative AI in the Enterprise, Wave 4 (more than two-thirds expect 30% or fewer experiments to scale within 3–6 months; compliance the rising barrier). MIT NANDA, State of AI in Business 2025 (the widely-cited "~95% of pilots show no measurable P&L impact" figure — presented here as contested and directional, given a small qualitative base and a narrow definition of failure).
- Recurring theme across 400+ analyzed sales conversations at emma. Security and compliance came up in more than four out of five (83%), and cost in roughly three out of four (73%).
- Capgemini, World Cloud Report – Financial Services 2026 (1,100 FS executives across 14 markets: 10% run AI agents at scale; 80% in ideation or pilot; 96% cite regulatory/compliance burden; 92% cite a skills gap). Evident AI Index 2025 (independent bank-AI benchmark: minority of disclosed use cases report any impact metric).
- FinOps Foundation, State of FinOps 2026 (98% of teams now manage AI spend, up from 31% in 2024; AI cost management the top desired skill; token/inference pricing breaks traditional attribution). Stanford Digital Economy Lab & Microsoft Research, "How Do AI Agents Spend Your Money?" 2026 (agentic tasks consume up to ~1,000× more tokens than simple queries; same-task runs vary up to ~30×; models underestimate their own cost).
- IBM / Ponemon Institute, Cost of a Data Breach 2025 (shadow AI added ~$670,000 to average breach cost; 20% of breached organisations compromised via shadow AI; 63% lacked or were still building AI governance). IDC and Netskope 2025–2026 (majority of employees use unsanctioned AI tools; a large share of GenAI access via personal, unmanaged accounts).
- EU Artificial Intelligence Act (Regulation (EU) 2024/1689): high-risk obligations for credit scoring and insurance pricing originally from August 2026; a deferral toward December 2027 was politically agreed in 2026 but not yet enacted at the time of writing. DORA (Regulation (EU) 2022/2554, in application since January 2025). Revised US interagency model-risk guidance (2026, succeeding SR 11-7). Bank of England / FCA, AI in UK Financial Services 2024 (75% of firms use AI; only 34% report complete understanding of the AI they use; third-party dependency the largest expected systemic-risk increase). ECB supervisory commentary 2026 (generative AI sourced from a small number of major providers; concentration, lock-in and exit-strategy risk). Gartner 2025–2026 (prediction that ~40% of agentic-AI projects will be scaled back or cancelled by 2027; AI governance platforms correlate with markedly higher governance effectiveness). Grant Thornton, 2026 AI Impact Survey (centralised review bodies create bottlenecks; 78% not confident of passing an independent AI governance audit within 90 days; fully AI-integrated firms far likelier to report revenue growth).
- emma proprietary analysis of 400+ recorded cloud and AI buying conversations (security/compliance raised in 83%, cost in 73%; recurring customer language on proving value and demonstrating control).