BEYOND THE BLACK BOX

The Boardroom Briefing

Five short, board-ready reads on AI governance, spend accountability, and vendor risk — sized for one flight, not one quarter.

Curated by Prashant Bhardwaj · beyondtheblackbox.in

Contents

01 **The Four Questions a Board Should Ask About Any AI ROI Claim**

AI Strategy · 4 min read · August 2026

02 **The AI Spend Governance Gap: Why CIOs Are Accountable for Budgets They Can't See**

Enterprise AI · 7 min read · August 2026

03 **Prompt Injection & Enterprise AI Security: Why Your CVE Dashboard Is Lying to the Board**

AI Governance · 7 min read · August 2026

04 **The EU AI Act Just Got Teeth — And It's Proof CLEAR Was Right About Vendor Risk**

AI Governance · 7 min read · August 2026

05 **How to Choose an AI Vendor Without Marrying One**

AI Governance · 6 min read · September 2026

AI STRATEGY

The Four Questions a Board Should Ask About Any AI ROI Claim

90% of CEOs expect measurable AI ROI this year. Only 29% of executives can actually measure it. A non-technical board doesn't need to close that gap by learning the technology — it needs four questions that work regardless of AI fluency.

August 2026 · 4 min read · Prashant Bhardwaj

Two 2026 numbers sit uncomfortably next to each other. BCG's AI Radar survey found that roughly 90% of CEOs believe AI agents will deliver measurable ROI this year. IBM's own research found that only about 29% of executives can actually measure AI ROI with any confidence, even though 79% report seeing productivity gains. That's not a technology gap. It's a claim outrunning the ability to verify it, at exactly the moment boards are being asked to approve bigger checks on the strength of that same claim.

The instinct is to assume the fix is board members getting more technical. It isn't, and it can't be — a board's job is oversight, not model evaluation. The real gap is closer to what Writer's 2026 enterprise AI survey found: 58% of executives admit their fellow leaders lack the fundamental AI knowledge to make strategic decisions about it. That's not a call for a crash course in transformer architecture. It's a call for a small number of non-technical questions that work precisely because they don't require any.

Four questions do most of the work. First: which kind of value is this actually claiming — lower cost, more revenue, reduced risk, or a strategic option this opens up later? Most business cases quietly argue only the first and let the room assume the other three are implied. Naming which one narrows what should even be measured. Second: measured against what baseline, and who agreed that baseline was fair before the project started? A number with no pre-agreed starting point is not a result, it's a claim. Third: who owns this number six months after go-live, not who owned the pitch deck at approval? A ROI figure with no owner after launch is a number nobody has to defend. Fourth: what does "this isn't working" look like, concretely, and on what date do we find out? If the answer is vague, the project doesn't have a failure mode — it has a slow fade nobody will be first to name.

None of these four questions requires knowing what a model is, how it was trained, or what architecture sits underneath it. They require only the discipline the CLEAR framework already argues for at the system level — a named owner, a documented baseline, continuous evaluation, not a one-time claim — translated for the reader in the room who will never open a model card and shouldn't have to. The board's job was never to evaluate the AI. It's to make sure someone, specifically named, is accountable for proving the claim was true — and to keep asking until that person can.

Key takeaways

- 90% of CEOs expect measurable AI ROI this year (BCG); only ~29% of executives can actually measure it today (IBM) — a claim-versus-verification gap, not a technology gap.
- 58% of executives say their fellow leaders lack the fundamental AI knowledge for strategic decisions (Writer, 2026) — the fix isn't more technical fluency, it's better non-technical questions.
- Most AI business cases argue cost reduction and leave revenue, risk reduction, and strategic optionality implied rather than stated — naming which one is being claimed narrows what should be measured.
- Four board-ready questions: which value channel is claimed, what's the pre-agreed baseline, who owns the number after go-live, and what does failure look like and by when.
- These questions require no technical fluency — they require the same accountability discipline CLEAR argues for at the system level, translated for a non-technical reader.

Read the full piece at beyondtheblackbox.in/blog/non-technical-framework-enterprise-ai-roi

ENTERPRISE AI

The AI Spend Governance Gap: Why CIOs Are Accountable for Budgets They Can't See

Enterprises can approve an AI budget line. Very few can see, in real time, what an agent actually spends against it — and that blind spot is a governance failure wearing a finance costume.

August 2026 · 7 min read · Prashant Bhardwaj

Ask a CIO in mid-2026 whether their organisation approved an AI budget this year and almost all of them will say yes. Ask the same CIO whether they can see, right now, what their AI agents are actually spending against that budget, and the confidence drops sharply. A recent IBM Institute for Business Value study of 2,000 technology executives put a number on that gap: two-thirds of surveyed CIOs and CTOs report being held accountable for AI systems they do not fully control, and 85 percent still lack full visibility into real-time AI spend. That is not a rounding error in a finance process. It is the same accountability-without-control problem this series has already spent four pieces describing for credit decisions and the systems that produce them, just showing up in the ledger instead of the decision log.

The instinct is to file this under budgeting discipline — tighten the forecast, add a dashboard, ask finance to review invoices more often. That instinct undersells the problem. A budget miss on traditional software is a forecasting error: seats are counted, contracts are annual, the number was wrong but it was knowable in advance. AI agent spend does not fail that way. Zylo's 2026 SaaS Management Index found enterprise AI spending up 108 percent year-over-year, hitting an average of $1.2 million per organisation, with 78 percent of IT leaders reporting charges they never budgeted for at all. Databricks rolled out enterprise spend caps in June 2026, according to Axios, specifically because some customers found themselves with AI bills reaching tens of millions of dollars a month before anyone noticed. These are not organisations that budgeted badly. They are organisations that could not see the meter while it was running.

The reason traditional budgeting breaks is structural, not procedural. Software licensing meters seats — a known, slow-changing number a finance team can forecast a year out. Agentic AI meters consumption: every prompt, every tool call, every reasoning loop, every retrieved document. A single agent working a single case can trigger dozens of model calls, and the cost of that case is not fixed the way a per-seat licence is — it scales with how hard the problem turns out to be, which is precisely the thing no one knows in advance. Multiply that across an estate of agents running continuously across a business, and the spend curve stops looking like a subscription and starts looking like an unmetered utility bill arriving after the fact.

The consequence shows up in project outcomes, not just finance reports. Gartner projects that more than 40 percent of agentic AI projects will be cancelled by the end of 2027, and names escalating costs, unclear business value, and inadequate risk controls as the primary drivers — not model quality. Only 21 percent of organisations, by Gartner's count, have a mature governance model in place. Read against the IBM and Zylo numbers, a consistent picture emerges: enterprises are not failing at agentic AI because the agents don't work. They are failing because no one built the control layer that would let the organisation trust an agent with a budget the way it already trusts a person with one.

This is worth naming precisely, because the framing determines the fix. The CLEAR framework was built for a different question — whether an AI system producing a credit decision can be trusted continuously, not just on the day it was validated — but two of its five checks map onto spend almost without modification. Lineage asks whether, for any output, you can trace exactly which model version, prompt version, and data fed into it. Applied to cost, the same question is: for any dollar an agent spent, can you trace exactly which task, which model call, and which business justification produced it? Most enterprises can answer this for a single flagged invoice after a manual investigation. Almost none can answer it in real time, at the level of an individual agent action, which is the only level at which a runaway loop can actually be stopped before it becomes a headline.

Resilience is the second check that transfers directly. In CLEAR's original context, resilience means a documented fallback when a foundation model updates without warning: a rollback path, an automatic human-review trigger, a kill switch that works in minutes. Applied to spend, resilience means the same three things exist for cost, not just for output quality — a per-agent budget ceiling that halts execution automatically when crossed, an alert that fires before the ceiling is reached rather than after the invoice arrives, and a kill switch a human can pull in minutes when a monitoring signal says something is wrong. The TRUST framework's accountability principle adds the piece both of those checks assume but don't state outright: a named owner for every agent's spend, the same way a named owner exists for every agent's decisions, so that when a cost anomaly appears there is a specific person accountable for investigating it rather than a shared inbox no one checks until finance escalates.

None of this requires waiting for a vendor to solve it. The practical starting point is the same discipline this series has argued for at the decision level, applied one layer down, to cost. Set a budget ceiling per agent, not per department — departments don't run loops, agents do. Report cost per handled case, the same metric this series recommended for justifying an agent's value in the first place, so a spend spike is visible against the work it produced rather than buried in an aggregate monthly total. Alert on rate of spend, not just cumulative spend, because a runaway loop shows up as a slope change long before it shows up as a large number. And route any agent that exceeds its ceiling to a hard stop, not a warning email — the enterprises in the IBM study that built control into their AI systems deploy sixteen times more agents than those relying on manual governance, which is the opposite of what most finance teams assume tighter controls cost them.

The organisations that get ahead of this are not the ones with the most sophisticated cost dashboard. They are the ones that stopped treating AI spend as a procurement line item — approved once, reviewed annually — and started treating it as a governed system with the same lineage, resilience, and named accountability this series has already argued belongs on the decision side. Gartner's 40 percent cancellation figure is not a prediction about the technology failing. It is a prediction about how many organisations will still be governing agent spend the way they governed a software licence, a year after that stopped being enough.

Spend is most exposed when the people who would notice it are away. The holiday freeze checklist for AI agents applies these controls to the year-end break.

Key takeaways

- Two-thirds of CIOs and CTOs are accountable for AI systems they don't fully control, and 85% lack real-time visibility into AI spend — an IBM finding that mirrors the accountability-without-control gap this series has already traced through credit decisioning.
- AI agent spend breaks traditional budgeting because it's consumption-metered, not seat-metered — cost scales with how hard a task turns out to be, which is unknowable in advance and compounds across an entire agent estate.
- Gartner projects 40%+ of agentic AI projects cancelled by 2027, driven by cost, unclear value, and weak governance — not model quality. Only 21% of organisations have a mature governance model.
- CLEAR's lineage and resilience checks transfer directly to spend: trace every dollar to the task and model call that produced it, and build a per-agent budget ceiling, alert, and kill switch — before the invoice arrives, not after.
- Enterprises with built-in AI controls deploy 16x more agents than those relying on manual governance — spend discipline is what enables scale, not what limits it.

Read the full piece at beyondtheblackbox.in/blog/ai-agent-spend-governance-gap

AI GOVERNANCE

Prompt Injection & Enterprise AI Security: Why Your CVE Dashboard Is Lying to the Board

OWASP's own data shows prompt injection ranks #1 by expert vote and #12 in real-world incidents — a gap that's about visibility, not danger, and it should change how you fund AI security controls next year.

August 2026 · 7 min read · Prashant Bhardwaj

Your CVE Dashboard Is Lying to Your Board

A CISO who checks the vulnerability scanner, sees a low count for prompt injection, and moves the budget line elsewhere is reading the wrong instrument. Prompt injection has held the #1 spot on the OWASP Top 10 for LLM Applications for three years running. When two of the people who run that list checked the expert ranking against a database of real-world incidents, prompt injection came back at #12.

That's not a typo, and it's not the incident data proving the experts wrong. It's the two signals looking at different things.

Two witnesses, one contradiction

The comparison comes from Kyriakos "Rock" Lambros and Steve Wilson — both leaders of the OWASP GenAI Security Project — published on arXiv on August 18 as an exploratory analysis, not the official OWASP release. They built a real evidence base: 7,714 documented LLM security incidents pulled from CVE records, GitHub Security Advisories, OSV, and the AIAAIC AI-harm database, with 6,639 of them labeled against a 20-entry risk taxonomy.

The statistical agreement between the 29-person expert vote and that incident record: a Cohen's kappa of 0.20, with a confidence interval that crosses zero. In plain terms — you cannot rule out that the two rankings agree by chance alone. As Lambros put it, you have two witnesses to the same event, and they're contradicting each other. Neither one is automatically the truth.

Why the scoreboard misses the actual danger

The reason isn't that prompt injection got safer. It's that the attack is structurally invisible to the tools built to catch it.

Prompt injection hides instructions inside content a model reads — a support ticket, a log entry, a document pulled back through retrieval. The agent then executes the action the attacker wanted, using credentials it already legitimately holds. Nothing in that chain is a product defect. There's no patch to ship, no CVE to file. A scanner has nothing to find, because the vulnerability was never in the code — it was in the trust boundary between instructions and content.

Wilson's framing is the one worth pinning to a wall in your risk committee: incident data is backward-looking. It tells you what got observed, classified, and reported. It doesn't tell you what's dangerous in the systems you're deploying this quarter.

Where this gets expensive fast

The gap runs in the other direction too — and the entries where it runs widest are the ones enterprises are funding least.

Misinformation sits at #13 in the expert vote and #2 in the incident record — the widest disagreement in the whole comparison. The authors don't call it an expert error; they note the incident corpus is heavy with deepfake and AI-generated disinformation cases that describe harm produced by AI rather than a vulnerability inside an LLM. Different failure mode, same taxonomy bucket.

Persistent memory poisoning (expert #4) and MCP tool interface exploitation (expert #7) are the two newest categories, and the incident record can't place either one — the confidence interval on both spans most of the taxonomy. That's not because they're rare. It's because they're new enough that reporting hasn't caught up. Real 2026 CVEs already exist for both: a KQL-injection flaw in the Azure Data Explorer MCP Server (CVSS 8.3) that lets a prompt-injected agent run arbitrary queries, and an unauthenticated MCP bridge in an agent harness (CVSS 10.0) that exposed API keys and let an attacker poison the learning store.

Lambros's line on this is the one to bring into a budget conversation: a count of zero incidents measures your blindness, not your safety. Nobody files an advisory for a procurement agent quietly told once that invoices under a threshold clear without a second signature — because nobody knows it happened. Every approval after that looks like the process working.

The BFSI angle

That procurement example isn't hypothetical for lending and credit operations — it's the exact shape of exposure that shows up in decisioning workflows once an agent starts remembering exceptions instead of just executing rules. It's also precisely why control boundaries need to be architectural, not promptable. A rule written into a system prompt is a suggestion to the model. It is not an enforceable control, and it will not hold up to an examiner, an auditor, or an attacker.

What actually helps, starting Monday

Stop treating the OWASP Top 10 as a spend queue. It's a coverage map built from 29 votes and an incident corpus its own authors call weak. Build your priority order from your own exposure — production reach, what a failure touches, and which controls you've actually tested against your live system, not a hypothetical one.

Put the authorization gate outside the model. An agent can propose a high-impact action — a DNS change, a payment override, a limit exception — but it should not hold the authority to execute it unsupervised. That's the control Wilson names first, and the tradeoff is explicit: you lose some autonomous improvisation in exchange for keeping a human or a hard-coded boundary between "the model suggested this" and "this happened."

Log what the system actually did, not just what it output. The prompt, the retrieved context, the tool calls, the arguments passed, and — critically — the model's confidence score. A model executing on a poisoned instruction doesn't behave erratically. It behaves certain. Most monitoring treats certainty as the sign of a healthy system, which is exactly backwards when the certainty itself is the symptom.

Fund agent memory and MCP tool boundaries on architecture, not advisory volume. These are the entries with the thinnest incident record and the highest severity CVEs already on file. Wiring in the boundary now, while these systems are still being built, is a rounding error. Retrofitting it in two years is a re-architecture.

The question for your next board meeting

If your current AI risk ranking rests on a 29-person practitioner vote and an incident record its own authors say disagrees with that vote at a near-certain rate — what, exactly, is deciding which controls get funded next year?

Sources: Lambros & Wilson, "Prompt Injection: Rank #1 or #12?" (arXiv, Aug 18, 2026); OWASP Top 10 for LLM Applications 2026 (published Aug 4, 2026); CrowdStrike 2026 Global Threat Report; Ivanti 2026 State of Cybersecurity.

Key takeaways

- A low CVE count for prompt injection measures scanner blindness, not safety — the attack lives in the trust boundary between instructions and content, not in product defects.
- OWASP's own comparison shows a Cohen's kappa of 0.20 between expert rankings and 6,639 real-world incidents: the two rankings are effectively uncorrelated.
- The widest gaps are in categories enterprises fund least — misinformation, persistent memory poisoning, and MCP tool exploitation — because incident data lags behind emerging attack surfaces.
- Architectural controls beat prompt-level rules: authorization gates outside the model, full execution logging, and hardened memory/tool boundaries are the practical starting points.
- Use the OWASP Top 10 as a coverage map, not a spend queue; fund controls based on your production exposure and tested controls, not advisory position alone.

Read the full piece at beyondtheblackbox.in/blog/prompt-injection-enterprise-ai-security-cve-dashboard

AI GOVERNANCE

The EU AI Act Just Got Teeth — And It's Proof CLEAR Was Right About Vendor Risk

On August 2, 2026, EU enforcement powers over general-purpose AI model providers went live, with fines up to 3% of global turnover. For institutions that never sold a single product in Europe, this is not a compliance footnote — it's the first real-world instance of the exact failure mode the CLEAR framework's lineage and resilience checks were designed to catch.

August 2026 · 7 min read · Prashant Bhardwaj

Part Three of this series described a Friday-to-Monday scenario: a foundation model provider pushes an update over the weekend, and an underwriting co-pilot gives a materially different answer to a near-identical case on Monday morning, with no internal change request behind it. That scenario was framed as a routine operating condition — the ordinary cost of building on a model an institution does not own. August 2, 2026 added a second, sharper version of the same risk, and this one carries legal teeth.

That is the date the European Commission's AI Office became formally entitled to enforce obligations on providers of general-purpose AI models — the foundation models underneath most institutions' credit decisioning co-pilots, document reconciliation tools, and customer-facing chat interfaces. The obligations themselves — technical documentation, risk mitigation for systemic-risk models, incident reporting — have technically applied since August 2025. What changed on this date is that the AI Office can now request a provider's technical documentation, run its own evaluations of a model, order corrective measures, and — the part that should get a technology team's attention — restrict, withdraw, or recall a model from the EU market entirely, backed by fines of up to €15 million or 3% of global annual turnover, whichever is higher.

An institution in Mumbai or Manila with no EU customers might reasonably assume this is someone else's regulation. It isn't, for a simple reason CLEAR framework's lineage check already forces an institution to confront: the foundation model an underwriting co-pilot runs on is not owned by the institution, and the provider of that model now answers to a regulator with the power to force a model change on a timeline the institution does not control. A GPAI provider ordered to restrict or modify a model to satisfy the AI Office does not ship a regional variant for customers outside Europe. It ships a change. That change reaches every deployment of that model, everywhere, on the provider's schedule — not because the institution did anything differently, but because the institution's vendor operates in a jurisdiction with enforcement power now, and the institution's jurisdiction doesn't.

This is precisely the gap lineage was built to close. Part Three asked whether an institution can identify the exact model version, prompt version, and knowledge source behind any given output. Before August 2026, the primary reason a model version might change without notice was routine vendor maintenance — the ordinary churn of a fast-moving industry. Now there is a second reason, and it is regulatory: a provider under an AI Office corrective order, or one that pre-emptively adjusts a model's behaviour to avoid one. Both look identical from inside an institution's system logs — an output that shifted for a reason nobody internally requested. An institution without lineage tracking has no way to tell a routine model refresh from a regulatorily-forced one, and no way to answer a domestic regulator's question — "why did this credit decision change?" — with anything better than "the vendor changed something."

Resilience is the second CLEAR pillar this enforcement regime tests directly, and it tests the part institutions are least likely to have actually built. A documented rollback to a known-good model version, an automatic route to human review when confidence drops, a kill switch that works in minutes — these were framed in Part Three around outages and silent degradation. A GPAI provider ordered by the AI Office to restrict a model's availability is a scenario with the same shape and a harder deadline. If a provider suspends or materially alters a model in response to enforcement action, an institution's fallback plan is the only thing standing between that event and a credit decisioning workflow going dark, or worse, continuing to run on a model whose behaviour has quietly shifted underneath it. Institutions that treated resilience as a checkbox for the CLEAR audit now have a live reason to treat it as an operating requirement.

There is a temptation to read this as a story about Europe, or about GPAI providers, and set it aside as someone else's regulatory burden. The more useful read is narrower and more uncomfortable: this is the first documented instance of a foundation model provider facing real legal consequences for what it ships, from a regulator with actual enforcement power. It will not be the last, and India's own AI governance conversation — still largely advisory — is watching this precedent closely. An institution that has already built lineage tracking and resilience into its credit decisioning AI system is not affected by any single enforcement action. An institution that treated those two checks as documentation exercises for an internal audit finds out, the first time a vendor changes something it didn't ask for, that the checkbox and the capability were never the same thing.

Key takeaways

- EU AI Act enforcement powers over GPAI providers became active August 2, 2026, with fines up to €15 million or 3% of global turnover, and powers to force model restriction, withdrawal, or recall.
- A model change ordered by a regulator and a routine vendor update look identical inside an institution's own systems — lineage tracking is the only way to tell them apart and answer a domestic regulator's questions.
- Institutions outside the EU are still exposed: a GPAI provider's model change reaches every deployment globally, on the provider's timeline, regardless of where the institution operates.
- Resilience — documented rollback, automatic human-review fallback, a working kill switch — moves from a CLEAR audit checkbox to a live operating requirement once a vendor's model can be forced to change on short notice.
- This is a preview of where India's own AI governance conversation is likely headed as enforcement-backed frameworks prove out elsewhere.

Read the full piece at beyondtheblackbox.in/blog/eu-ai-act-enforcement-clear-framework-lesson

AI GOVERNANCE

How to Choose an AI Vendor Without Marrying One

The bake-off picks a model. The contract picks a gravity well. Enterprises that stay solvent treat the lab as a supplier and the control plane as the asset.

September 2026 · 6 min read · Prashant Bhardwaj

Eighty-five percent of enterprise leaders plan to increase AI spend in 2026, after already committing an average of 12% of functional budgets to it in 2025. Only 22% have gotten AI to actually work across more than one business unit. Eleven percent, per Gartner's September 2026 survey of 1,300+ functional leaders, cannot say what their own function spent on AI last year. Put those three numbers next to each other and the story isn't "adoption is slow." It's that budget, use case, and vendor are moving faster than anyone is tracking them — and a chunk of the enterprise doesn't even own the ledger showing where the money went.

That's not a maturity problem procurement can out-negotiate its way past. It's a control-plane problem, and it starts earlier than the RFP. Most vendor selections are still a prompt bake-off with a legal appendix stapled on: forty prompts, a scoring rubric, a winner circled in a slide deck. The model behind that winner will change at least once before the contract term is up. What won't change is which cloud it runs on, which identity system gates it, where the logs live, and whether the "agent" the vendor demoed is actually something your team owns or a runtime borrowed from someone else's platform.

Gartner's own read on the market makes the split explicit. In its December 2025 assessment of nearly 30 AI technology races, three different companies won three different contests: Microsoft was named the company to beat in enterprisewide AI, on the strength of its partner ecosystem and control of everyday work surfaces. Google took enterprise agentic platforms, built around its model stack and infrastructure. OpenAI led the large language model race outright, still the frontrunner on raw model capability. Three winners, three different questions. A CIO who reads "OpenAI is winning" as "we should buy OpenAI" has answered a model question with a platform decision — and those are not the same purchase.

Call it gravity instead of vendor. You're not choosing the smartest model in a bake-off; you're choosing which cloud, which identity system, which data path, and how expensive it becomes to leave. Enterprises already on Microsoft find OpenAI arriving by default through Copilot and Azure. Enterprises on AWS or multi-cloud infrastructure often route Anthropic through Bedrock, frequently alongside Google's Vertex. Enterprises anchored in Google Cloud get Gemini as the path of least resistance through Workspace and BigQuery. None of those defaults are wrong. What's wrong is treating the demo as the decision and discovering the gravity well only once the exit is expensive.

Four questions do more to protect an enterprise than another round of prompt comparisons — a companion set to the four questions a board should ask about any AI ROI claim, asked one stage earlier, before the contract is signed rather than after. First: what's the riskiest action this system is allowed to take without a human approving it first? If the honest answer is "summarize a document," almost any credible lab clears the bar. If the answer is "file a claim, approve a limit, or pay an invoice," the vendor's safety story and your own approval architecture both start to matter far more than which model wrote the best paragraph in the demo. Second: where does the prompt, the retrieved document, and the resulting log actually sit, and does the vendor train on any of it? Residency and subprocessor terms deserve a paragraph in the evaluation, not a footnote added the week before signature — this matters especially for organizations operating across India, the EU, or other jurisdictions with active data-residency rules. Third: if the model changes in ninety days, what breaks? If the honest answer is "our entire tool layer," the enterprise has already married the API, not rented the model — an interoperable tool layer is the closest thing procurement has to a prenup. Fourth: when the system is wrong, what can be handed to an auditor? A service-level agreement is not an audit trail. An owner, a policy version, a tool allowlist, and an evaluation set the vendor doesn't control are — the same accountability discipline the CLEAR framework argues for at the model layer, applied one level up, to the vendor relationship itself.

None of this argues for picking one lab. Paying two frontier labs at once is increasingly the default rather than the exception, and it's often cheaper than a forced rewrite later. The pattern worth building toward: let the system of engagement — drafting, coding assistance, everyday copilots — run on more than one model, routed by task rather than brand loyalty. Keep the system of action — anything that writes, approves, or pays — behind a single control plane that your own team owns: identity, tool allowlist, logging, and evaluation, regardless of which lab is answering the prompt underneath. Two labs without that shared plane is just two bills and two audit gaps. Two labs with it is genuine leverage the next time a renewal conversation starts.

This is also where Gartner's spend numbers loop back around. The 11% of organizations that can't say what they spent on AI last year almost certainly can't say which vendor a given dollar routed through either — the two blind spots share a root cause. Fixing it isn't a bigger dashboard project for 2027. It's naming one owner per AI line item now, tied to a business outcome and reviewed on the same cadence as cloud spend, so that renewal season isn't the first time anyone adds up what's actually running where. Gartner's own data backs the case for doing this early: organizations that track return by category — productivity, risk, revenue — and are willing to kill underperforming bets reported positive returns on 81% of their initiatives; organizations that don't track it that way couldn't even state the return on nearly a third of theirs.

The same survey surfaces a second, quieter warning worth folding into the same review: the AI use cases enterprises pursue most — cybersecurity threat detection, IT service-desk automation, code generation — are not the ones delivering the strongest returns — asset and cost optimization, synthetic data generation, and, again, code generation. Popularity and payoff are not the same signal, and a vendor evaluation that only asks "can you do the popular thing well" is optimizing for the wrong list.

None of this requires naming a winner, and doing so would defeat the point. The lab that fits this year's gravity may not fit next year's, and it shouldn't need to for the enterprise to stay solvent. What has to survive the next model release, the next renewal, and the next name change is the layer underneath: one owner, one control plane, one exit clause that doesn't require rebuilding the tool stack from scratch. Sign the lab as a supplier. Don't marry it.

For how this applies to building an agent stack on a mid-market budget, see the 2027 agent stack for mid-market enterprises.

Key takeaways

- Gartner's September 2026 survey: 85% of leaders plan to increase AI spend in 2026, but only 22% have scaled AI past one business unit, and 11% can't say what their function spent in 2025.
- High performers who track ROI by category (productivity, risk, revenue) report positive returns on 81% of AI initiatives; low performers can't state the return on nearly a third of theirs.
- Gartner's December 2025 "Companies to Beat" analysis names three different leaders for three different races — Microsoft (enterprisewide AI), Google (enterprise agentic platforms), OpenAI (LLM providers) — proof that platform and model are separate purchasing decisions.
- Four questions beat a bake-off: what's the riskiest unsupervised action, where does the data actually sit, what breaks if the model swaps in 90 days, and what can be handed to an auditor when the system is wrong.
- The safest architecture is two labs and one control plane: let engagement tools be multi-model, but route anything that writes, approves, or pays through identity, tooling, and logging your own team owns.

Read the full piece at beyondtheblackbox.in/blog/choose-ai-vendor-without-marrying-one

A clearer AI agenda starts with a sharper conversation.

Bring the governance, spend, or vendor question your leadership team cannot resolve in a slide deck.

Request a strategy conversation >

beyondtheblackbox.in/contact