TRACE3

2026
THE INVISIBLE BUDGET LINE
WHITEPAPER

Prepared By

Madhu Joshi
Senior Solution Architect

Brian Moresco
Senior Manager, Trace3 Digital Consulting

Colby Buchanan
Cloud FinOps Consultant

Chelsea Robertson
Senior Cloud FinOps Consultant

Date Issued:
07/08/26

Document Version:
Final

Table of Contents

1. Executive Summary 1
2. The Pattern We Have Seen Before 2
3. The AI Tooling Cost Landscape in 2026 4
4. The Model Tier Escalation Problem 5
5. What the Data Shows 7
6. The Six Governance Gaps 12
7. A Framework for AI Spend Visibility and Control 16
8. How Trace3 Approaches This 18
9. Three First Steps for Finance Leaders 19
10. Conclusion 19
11. References 21
12. Glossary 22

1. Executive Summary

Most large enterprises have already doubled their AI spend, often without any formal decision to do so. No contract was renegotiated. No budget was approved. No signature was required. The cost simply moved as engineers defaulted to each new top-tier model and vendors shifted billing from seats to consumption and it slid below a noticeable threshold.

This is what separates AI tooling from every other cost line a finance team already governs: it grows through accumulated default, not decision. By the time the board asks for total cost of ownership, the number is larger than anyone forecast and no one in the organization can explain how it got there.

Four facts define where things stand today:

- **AI tooling spend is material and growing.** A 50-person engineering team running frontier AI coding tools can spend $5,000 to $15,000 per month in tooling costs alone, before factoring in any cloud infrastructure costs. Individual developers at peak agentic usage can generate $2,000 per month in token billing. [1][2]
- **Most organizations cannot calculate the ROI.** MIT Project NANDA studied 300-plus enterprise AI initiatives and found that 95% of organizations saw zero measurable return on their investment. Not because the tools do not work. Because there is no framework tracking what was spent against what was produced. [3]
- **Vendor pricing escalates without anyone "raising" prices.** AI model vendors often keep existing tiers unchanged while launching a new top tier at double the previous flagship price each cycle. A team that defaulted to the best available model **in early** 2026 **can end up** paying 2x per token without any explicit sign-off.
- **The savings are available but unused.** Prompt caching, batch processing, and model routing can reduce effective AI API spend by 50 to 95 percent. Most enterprise teams aren't taking advantage of these savings and optimization opportunities. [4]

Organizations that built cloud cost governance before the board started asking questions had a lasting structural advantage. AI tooling is the same opportunity, on a faster timeline.

This paper covers the mechanisms driving invisible AI spend, the six governance gaps most organizations have not addressed, and a practical framework for taking control before the renewal cycle forces the conversation.

2. The Pattern We Have Seen Before

In 2012 and 2013, cloud infrastructure was a developer productivity story. Engineering teams moved workloads to AWS and Azure because it was faster and more flexible than buying hardware. Finance teams approved it. The costs were low; the argument was compelling.

By 2015, cloud spend was a significant and frequently surprising budget item across every sector. The spend was real. The value was real. The visibility was not. Costs lived across teams and project budgets, measured in units finance teams had no natural way to read: compute hours, storage gigabytes, data transfer. The FinOps discipline emerged to bridge this gap and translate cloud functions into explainable, measurable, and projectable metrics.

During this adoption phase for cloud, the cost and productivity statistics that eventually came out were sobering. Flexera found 27% of cloud spend is wasted. Forrester found 94% of enterprises overspend in the cloud, typically 30% above original projections. [5] By 2017, RightScale estimated annualized waste across AWS, Azure, and GCP at $10 billion. [6]

Like its cloud predecessor, AI tooling follows the same path albeit at an accelerated pace. The adoption curve that took cloud three to four years to climb, AI developer tooling has climbed in 18 months. Three structural similarities are worthy of emphasis.

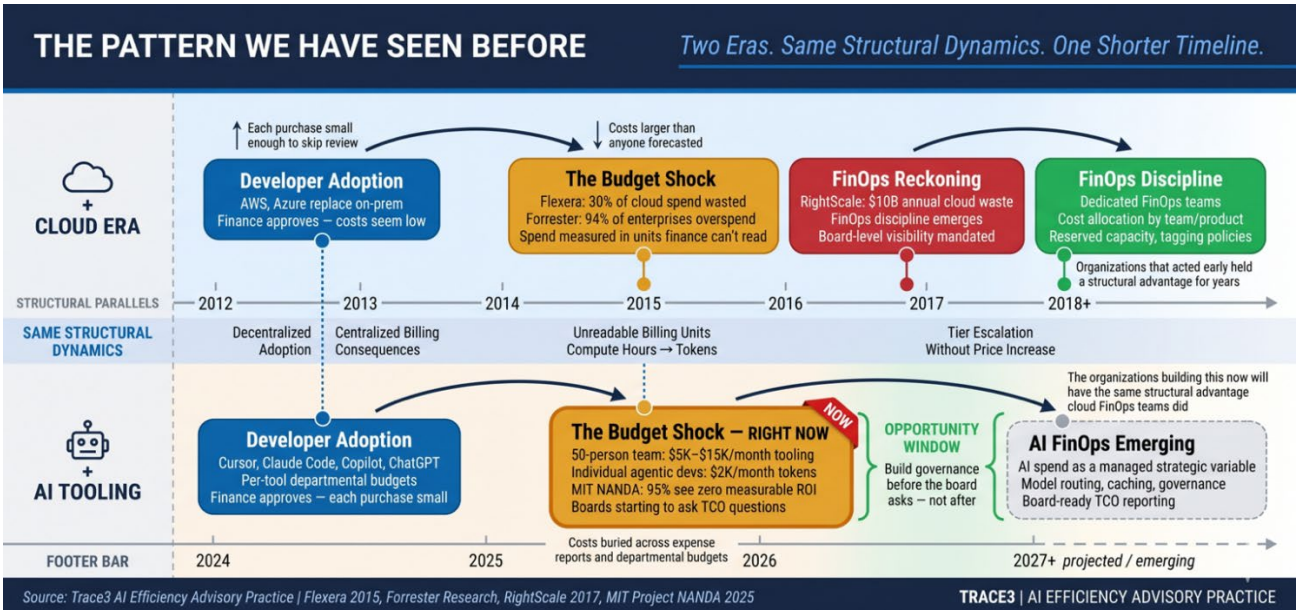

Figure 1: AI tooling follows the same adoption-to-crisis-to-governance arc as cloud infrastructure - on a compressed timeline.

Decentralized Adoption, Centralized Billing Consequences

Engineers adopt AI tools through departmental budgets or expense reimbursement. Cursor, Claude Code, GitHub Copilot, ChatGPT Enterprise. Each purchase is small enough to skip procurement review. Aggregated across a full engineering organization, the number is not

small. DX Research benchmarks total AI coding tool spend at $200 to $600 per developer per month for teams mixing inline and agentic tools. [1] For a 200-person engineering team, that is $40,000 to $120,000 a month, spread invisibly across dozens of cost centers.

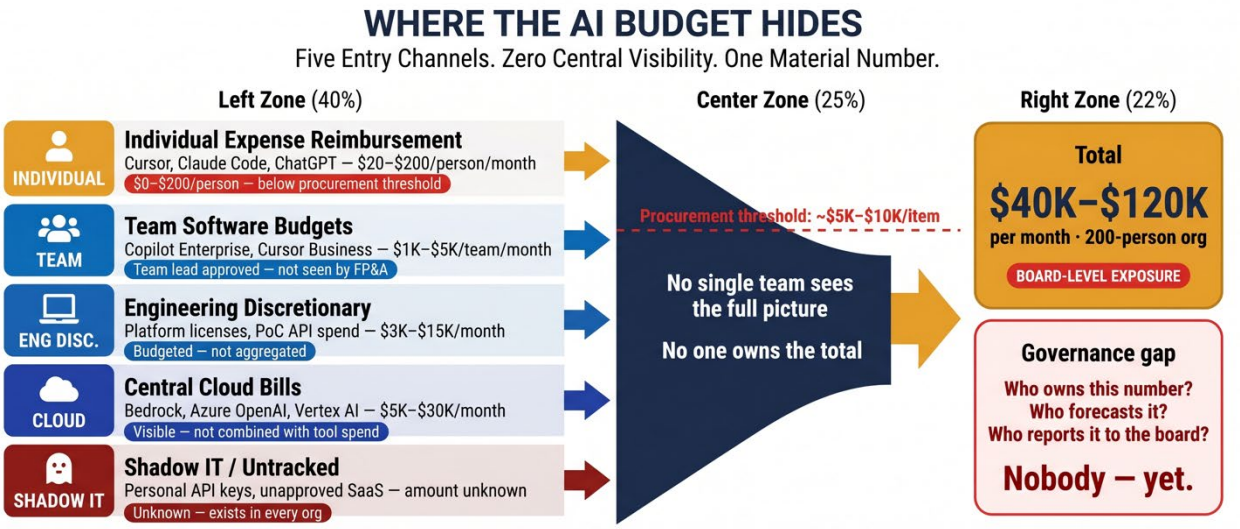

Figure 2: Five entry channels. Zero aggregated visibility. The same $40,000–$120,000 per month that would require board approval as a single line item arrives below procurement threshold, scattered across cost centers.

Consumption Models That Finance Teams Cannot Read

Cloud vendors charge in compute-hours and API calls. AI vendors charge in tokens. A token is roughly three-quarters of a word. A complex coding request can consume tens of thousands of tokens. Finance teams have no natural unit of comparison nor historical baseline to these hyper-variable cost drivers, nor do they have sufficient context to reasonably forecast and manage overruns.

The shift from fixed-request to consumption-based billing is already underway across every major platform. Cursor moved from a fixed 500-request model to credit-based billing in June 2025. GitHub Copilot transitioned to usage-based AI Credits in June 2026. Claude Code meters directly at API token rates. [7][8] None of these transitions came with clear advance notice of the financial impact.

Pricing That Escalates Without a Price Increase

Established AI tiers have been largely stable on price. The portfolio has not. Vendors often add a new top tier product at double the previous flagship price with each product cycle, while keeping existing tier prices flat. For PE-backed organizations where EBITDA margin discipline is a board-level priority, this can be a dangerous precedent. The vendor did not raise prices. The engineering team did not change behavior. The spend doubled, with nothing in the procurement record to explain it.

3. The AI Tooling Cost Landscape in 2026

The pricing architecture of the tools most used by engineering teams is worth understanding before any governance work begins.

The Cursor Case Study

Cursor's June 2025 pricing change is the clearest documented example of how usage-based billing creates FP&A exposure without warning.

On June 16, 2025, Cursor moved from a fixed allotment of 500 premium requests per month to a credit-based system tied to AI model API costs. [7] At the same $20 monthly price, a Pro plan now yields roughly 200 to 225 premium requests using frontier models. Less than half the previous allotment. Heavy users can exhaust credits in days. Pay-as-you-go overages apply automatically unless the user has manually enabled a cap. [9]

Cursor issued a public apology on July 4, 2025, and offered refunds for unexpected charges. [7] The lesson is not specific to Cursor. A single pricing change, poorly communicated, can increase effective costs by 20x or more for high-usage teams. One team of five reported spending over $4,600 in six weeks in 2026, double their entire prior year spend, without any change in usage patterns. [10]

Three FP&A risks come out of this model:

- The subscription cost is predictable. The total cost is not due to product tiering.
- The heaviest users, often the most senior engineers, are the most likely to generate overages.
- Spending caps are opt-in. The default is uncapped billing.

The Real Cost of Developer AI at Scale

In June 2026, Microsoft's Experiences 6and Devices division, which covers Windows, Microsoft 365, Outlook, Teams, and Surface, reportedly ordered thousands of engineers off Claude Code, citing token billing that reached approximately $2,000 per engineer per month and had exhausted the division's annual AI budget in months. [2] The replacement was GitHub Copilot Enterprise at a flat $39 per seat. The CFO arithmetic was straightforward.

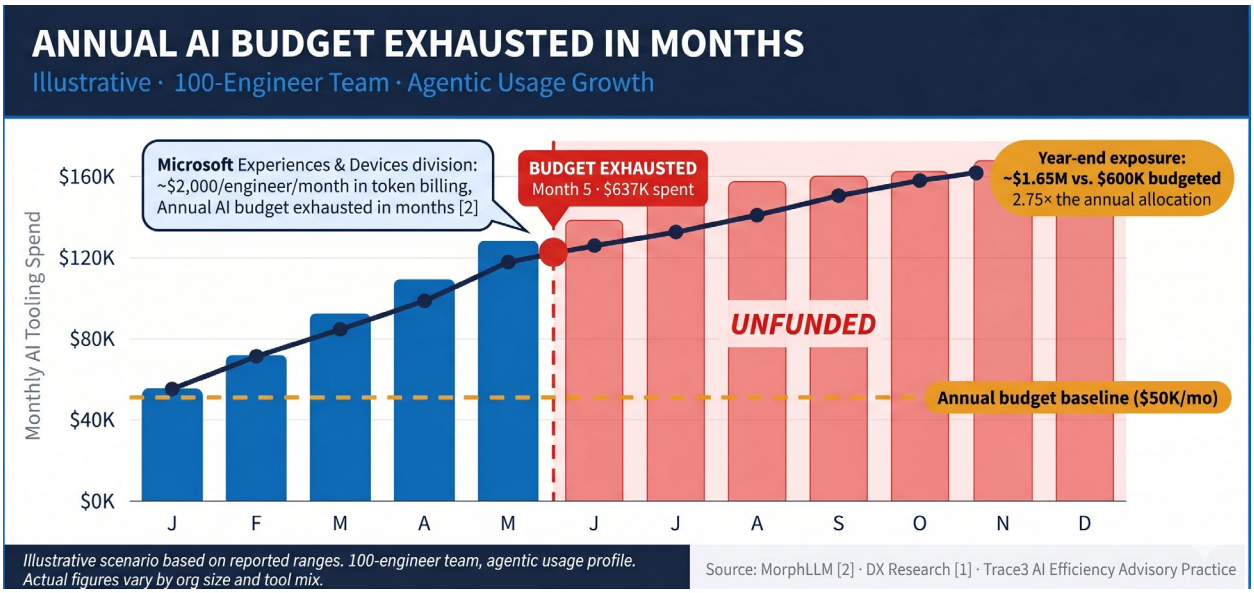

Figure 3: Illustrative 100-engineer team. Annual AI budget at $50K/month baseline exhausted by Month 5. Seven months of spend remain unfunded.

> *"When 25 developers each have different usage patterns and model preferences, the actual monthly bill can range from $500 to $5,000-plus on the same Teams plan."*
>
> CloudZero, Cursor AI Pricing Analysis, 2026 [11]

For globally distributed engineering teams, these risks compound. Usage patterns vary by time zone and team culture. Oversight is harder to centralize. Without a single owner of AI spend, budget variance shows up in retrospect rather than in the forecast.

4. The Model Tier Escalation Problem

The pricing structure of AI model providers follows a pattern most finance teams have not yet mapped. Within any given tier, prices are stable across generations. Anthropic's Sonnet tier has held at $3 per million input tokens and $15 per million output tokens across multiple successive model releases. [4] The tier itself does not move. What moves is the portfolio around it.

Current Anthropic pricing per million tokens: [4]

Model	Input (per M tokens)	Output (per M tokens)
Haiku 3 (legacy)	$0.25	$1.25
Haiku 4.5	$1.00	$5.00
Sonnet 4.6	$3.00	$15.00
Opus 4.8	$5.00	$25.00
Fable 5	**$10.00**	**$50.00**

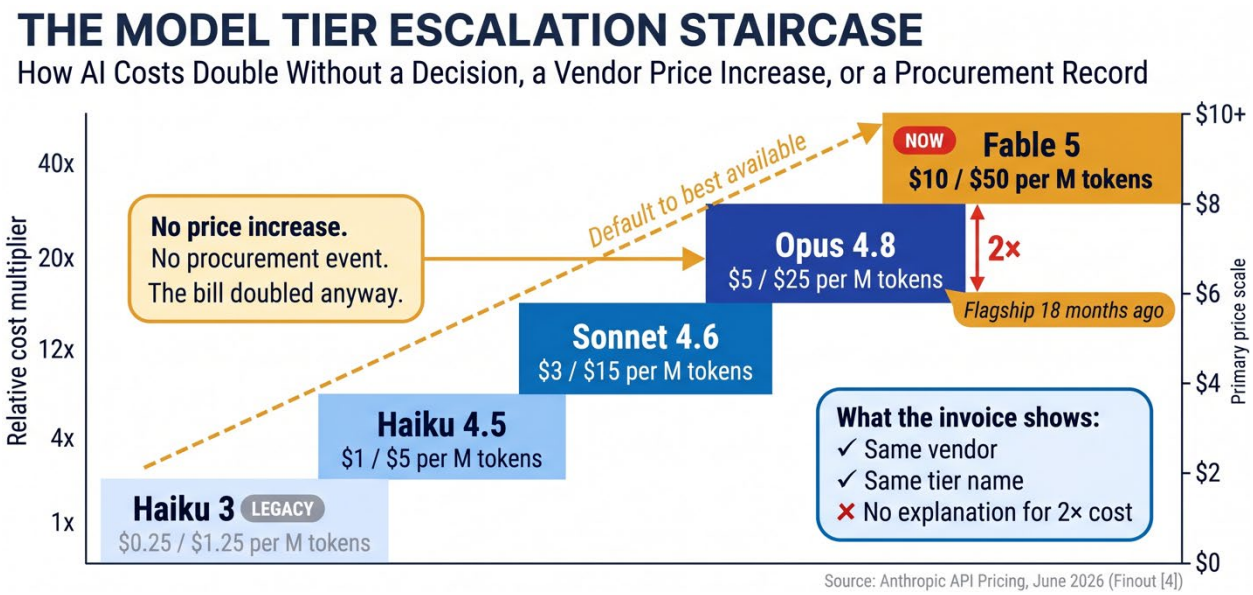

Figure 4: The tier price holds. The portfolio moves around it. Engineers defaulting to the best available model in 2026 pay 2× what projections built on Opus anticipated.

Each product cycle, a new top tier launches at roughly double the price of the previous flagship. Opus was the premium model earlier in 2026. Fable 5, released June 9, 2026, is priced at exactly 2x Opus. [4][12] Opus is now the mid-tier option. An engineering team that built cost projections around Opus and continued defaulting to the best available model has doubled its per-token spend. No deliberate decision was made. Nothing appears in the procurement record.

A working developer summarized the mechanism in a recent forum thread: "They never raise prices. Sonnet has been $3 per million tokens for four generations. They just keep adding a new tier on top at double the price, and yesterday's flagship quietly becomes the mid option." That is an accurate description of the financial dynamic, and most enterprise finance teams have not yet recognized it.

The Tokenizer Trap

Cost increases can also invisibly occur within a tier. The Opus 4.7 release introduced a tokenizer change that increased token consumption for identical text by up to 35% compared to the previous version. [4] An organization that did not change usage patterns paid 35% more the month after the update. No price change. No procurement flag. No notification. The increase would only show up in variance analysis if someone was tracking spend at the model version level, which almost no finance team does.

What This Looks Like in Practice

An engineering organization in early 2025 benchmarks AI costs using Opus, the flagship at the time, and builds projections accordingly. By mid-2026, Fable 5 has launched and engineers default to it as the best available model. AI spend on this factor alone has doubled. The CFO reviews the renewal, sees the same vendor and the same tier name in the billing description, and finds a bill 2x the projection. No decision in the organization's history explains it.

5. What the Data Shows

The findings in this section come from a sample size subset of Trace3's anonymized usage dataset from our portfolio of enterprise clients. We analyzed real production consumption, not vendor benchmarks or hypothetical scenarios, to understand how AI spend is accumulating in live environments. The data points to three connected sources of exposure: teams are shifting demand toward more expensive frontier models, many workloads still run at full rack rate, and supporting infrastructure costs often sit outside the AI budget line. Together, those patterns show how AI spend can grow materially even when no single team appears to be making a major financial decision.

Market Sector Concentration

Our customer sample shows faster AI adoption among consumer-facing companies, especially in customer service and nondeterministic marketing use cases, while financial services and healthcare organizations are moving more cautiously under stricter guardrails. However, cost inefficiencies and waste are pervasive across industries, with between 30% and 50% of AI cloud spend evaporating into idle resource commitments and over-

provisioned infrastructure. [5] The drivers of overspend are nuanced by industry: the concentration of overspend in the retail and consumer goods sectors are often focused on pilot to production efforts, while financial services have high spending commitments on proprietary LLM architectures, and healthcare and life science investments in AI are often elevated by regulatory and security enhancements. No industry is immune to AI driven cost and utilization inefficiencies.

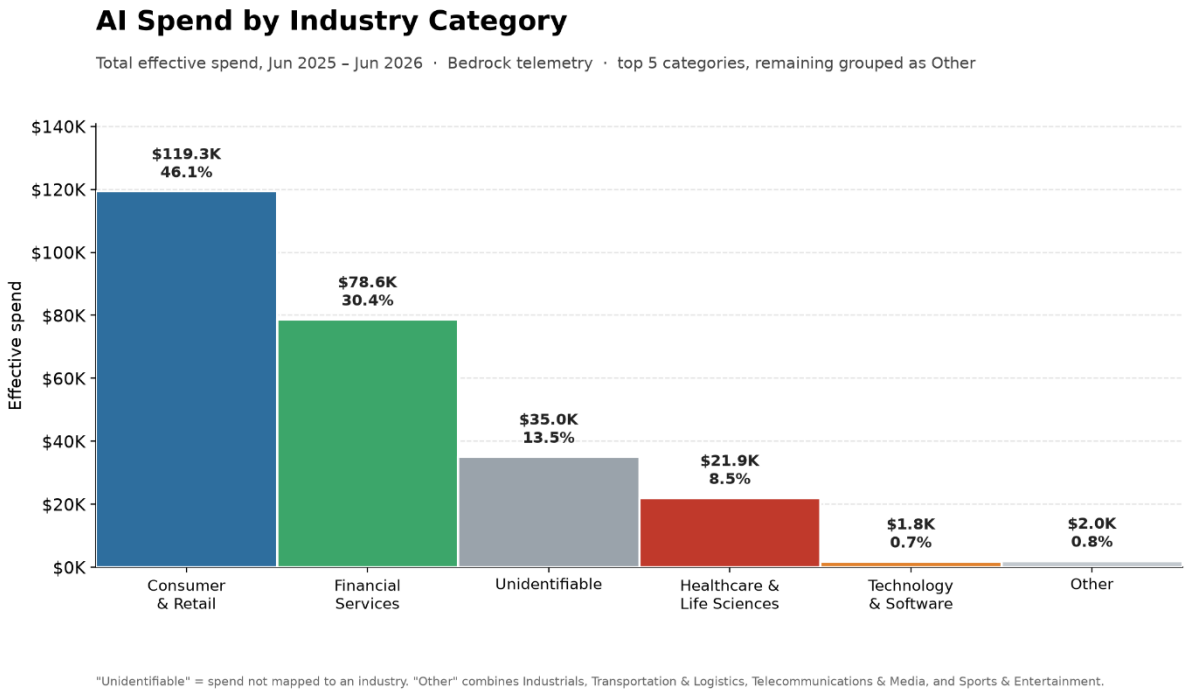

Figure 5: Tracks AWS AI spend over the trailing twelve months categorized by industry.

Frontier Model Concentration

As new frontier models come to market, organizations are not simply adding AI capacity; they are moving existing workloads into the most capable and expensive tiers. In 2025, mid-tier "workhorse" models carried most demand. By mid-2026, frontier models had become the dominant consumption category, changing the cost profile of AI usage even when application volume remained steady.

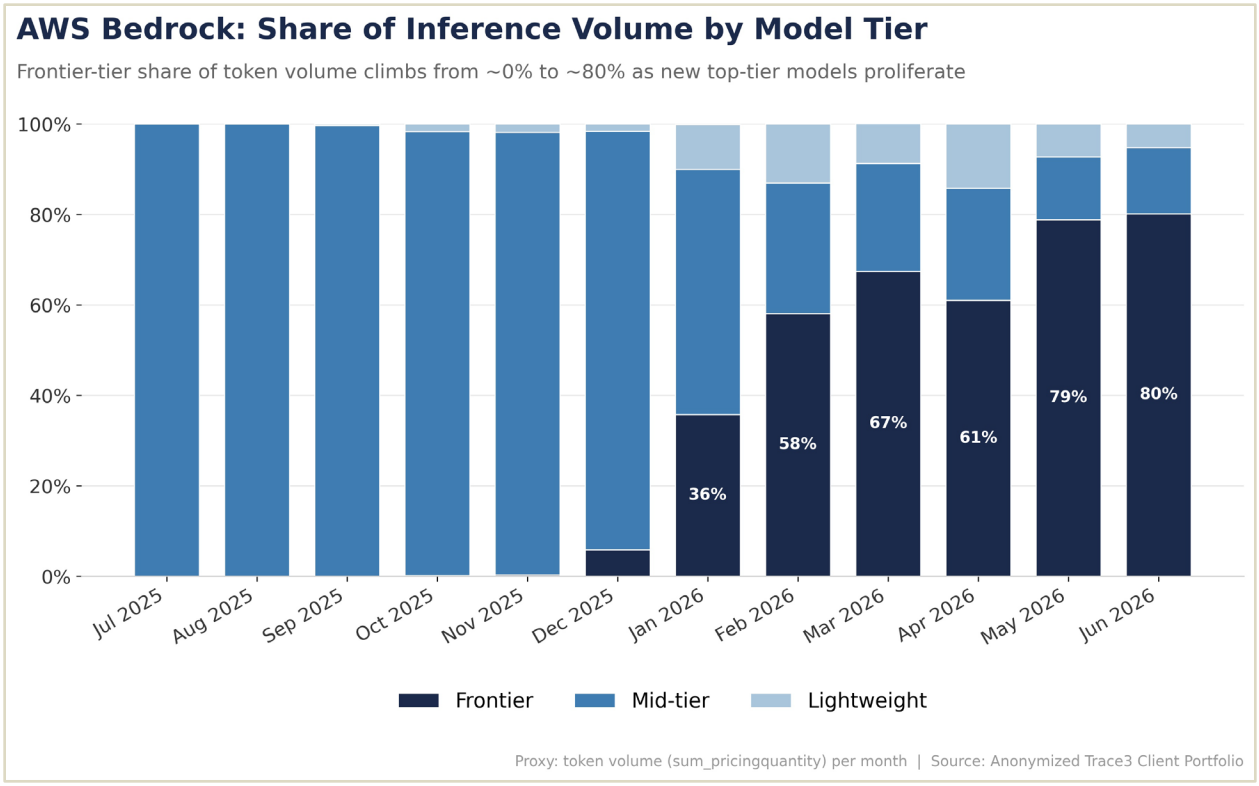

Figure 6: Tracks AWS Bedrock inference volume across frontier, mid-tier, and lightweight models over the trailing twelve months. Frontier-tier share rises from effectively zero in late 2025 to 80% by June 2026, nearly inverting the prior mid-tier-dominant mix. The inflection point arrives in early 2026, as new top-tier models become broadly available and teams default to the most capable option.

* Frontier = Opus / Fable / Gemini Pro / Nova Premier / large reasoning OSS. Mid-tier = Sonnet / Gemini Flash / 70B-class. Lightweight = Haiku / Flash-Lite / Nano / Micro / small OSS.

Rack Rate Exposure

While model mix explains where demand is moving, rack rate exposure explains how much of that demand is being billed inefficiently. In many environments, the largest source of avoidable AI overspend is not the model itself; it is paying full list price for work that could be served at a lower effective rate with no loss of quality. Prompt caching and batch processing are the clearest examples. Both can materially reduce cost, yet both are often underused because teams focus first on model selection rather than how the model is called.

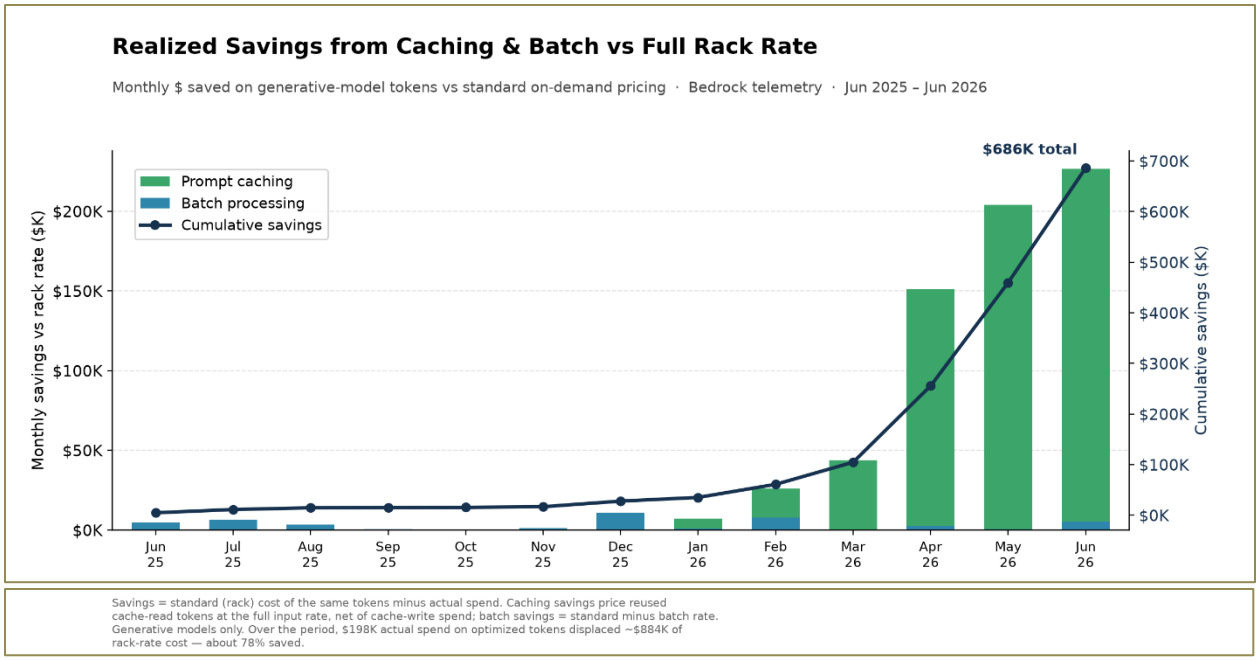

Figure 7: Shows the cost impact of ~78% savings for a client moving eligible Bedrock calls from full on-demand pricing to optimized execution patterns. It compares the standard cost of processing the same workload against the lower effective cost achieved through prompt caching and batch processing.

Batch processing is also highly effective for asynchronous workloads, but it requires teams to redesign how requests are queued and executed. The graph above showcases the meaningful savings that can be gleaned within workloads through effective pre-production planning.

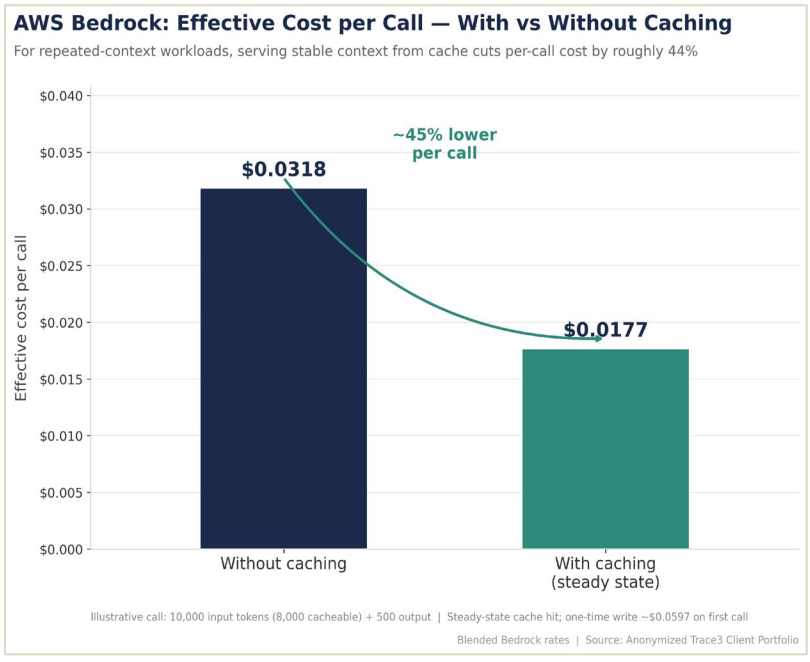

Figure 8: The workload does not change, but the unit economics improve materially when reusable context is cached instead of processed again on every request. These results represent a blended rate across all Bedrock models used in the usage portfolio, and per call effective pricing will vary by model.

Without caching, every input token is billed at the full on-demand rate, producing an effective cost of roughly $0.032 per call. With caching at steady state, the 8,000 reusable context tokens are read from cache at approximately $0.47 per million tokens instead of $2.24 per million, reducing the effective cost to roughly $0.018 per call, a 44.3% reduction within the same usage pattern.

The calculation assumes 80% of the input is stable, reusable context, which is common in RAG and agentic workloads where the same instructions, codebase context, and/or reference material is passed repeatedly. Trace3's portfolio data supports that assumption: in observed cached workloads, cached-read tokens outnumber fresh input tokens by roughly 15:1.

The Hidden Infrastructure Layer

The AI budget line extends beyond token costs. Enterprise AI deployments increasingly require supporting data infrastructure that does not appear on AI vendor invoices.

- Vector databases that store proprietary code and documentation for tools like GitHub Copilot Enterprise and custom retrieval implementations
- Data pipelines that feed organizational knowledge into AI context windows
- Log storage for AI interaction data, often required for compliance and audit purposes
- Data transit costs when regional data residency requirements route queries through specific endpoints at premium rates

These costs almost never appear in initial AI tooling budgets and are rarely attributed to AI spend in cost center reporting.

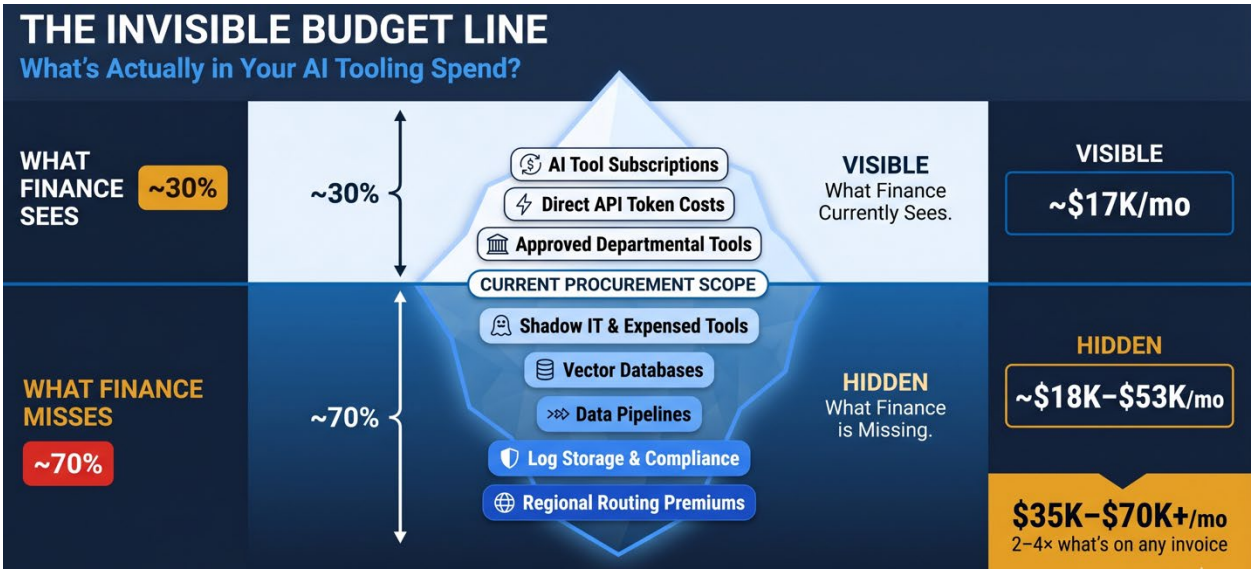

Figure 9: The AI budget line extends beyond vendor invoices. The hidden infrastructure layer grows as deployments mature and almost never appear in the initial cost estimate.

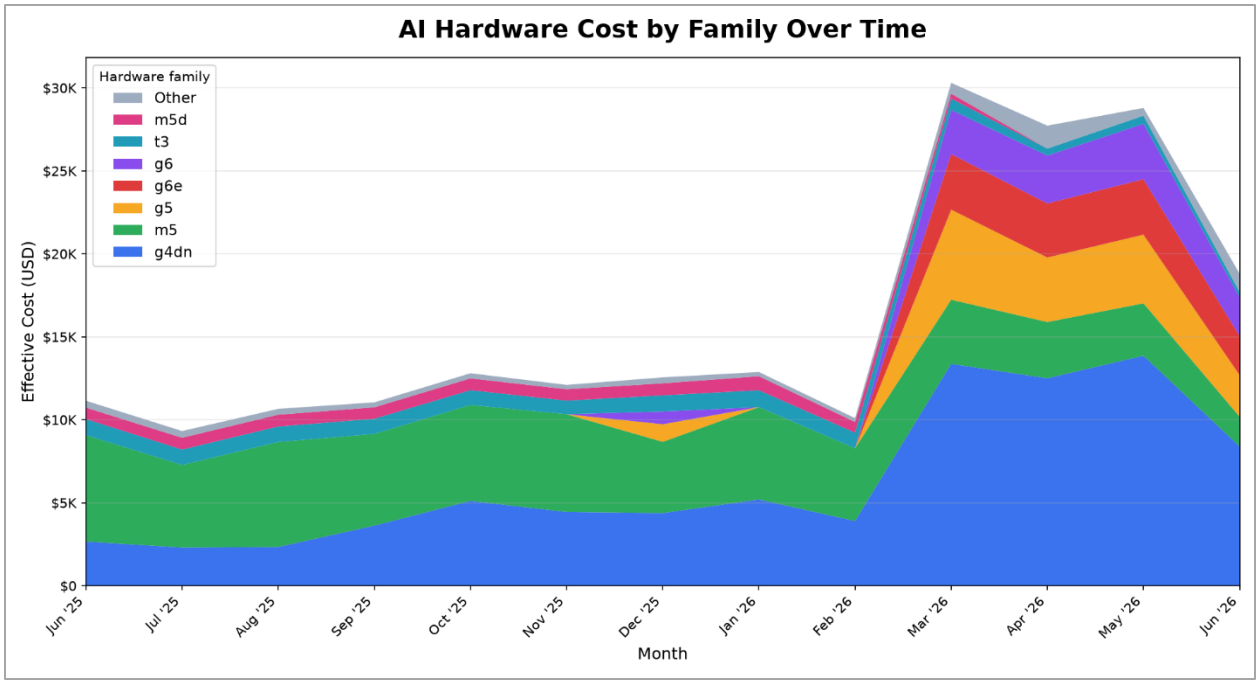

Figure 10: Early spend may appear concentrated in general-purpose compute, but the mix shifts as teams add GPU-backed instances, higher-memory families, and specialized acceleration to support inference, embedding, retrieval, and model-serving patterns. The result is a second cost curve that grows alongside token consumption but is often owned, tagged, and forecast separately from the AI tools that created the demand.

** Hardware family: the AWS EC2 instance family — a group of instance types built on the same underlying hardware and optimized for a similar workload. g-series (g4dn, g5, g6, g6e) = GPU / accelerated • m-series (m5, m5d, m4) = general purpose • c-series = compute-optimized • r-series = memory-optimized • t-series (t3, t2) = burstable. 'Other' groups the 9 lowest-cost families*

6. The Six Governance Gaps

Most organizations have no framework for governing AI tooling costs. Six gaps appear consistently in organizations that have not yet addressed this.

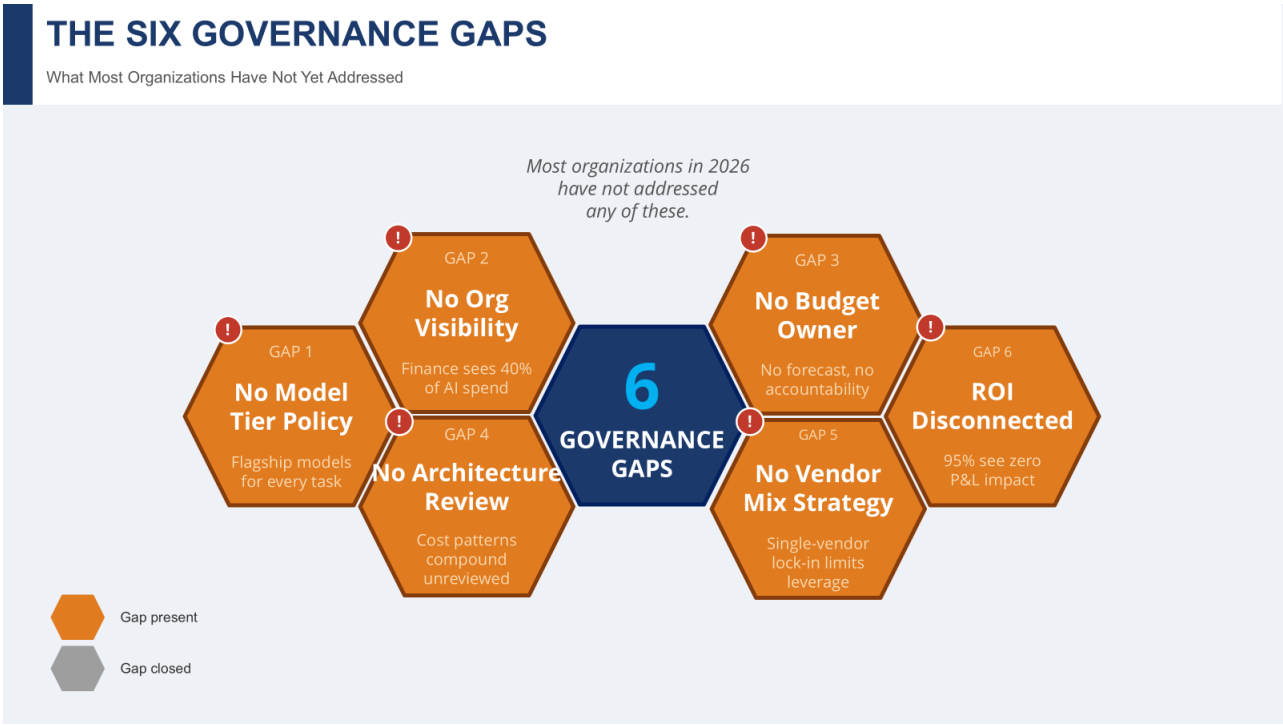

Figure 11: A self-assessment framework. Most organizations operating in 2026 will recognize several or all these gaps.

Gap 1: No Model Tier Policy

Engineers choose AI models based on availability and capability. There is no organizational guidance on when a frontier model is justified versus when a mid-tier model is sufficient for the task. The result is routine work running on flagship models, at flagship prices, every cycle.

Gap 2: No Organizational Visibility

Individual team leads may have a view of their own Cursor or Copilot usage. Finance teams rarely have an aggregated view across tools, teams, and vendors. No one holds a complete picture of what the organization is spending, which models are in use, or where costs are concentrated. Gartner found that organizations typically have visibility into only 40% of the SaaS applications in use. [5] AI tooling is the same problem, at faster growth rates.

Gap 3: No Budget Owner

AI tools enter the organization through departmental software budgets, engineering discretionary spend, and individual expense reimbursement. No single person is accountable for the total. Without an established owner of cost management, there is no forecast, no variance analysis, and no feedback loop for improvement.

Gap 4: No Architecture Review

The patterns that drive unnecessary cost, including synchronous calls where batch processing would work, missing prompt caching, oversized context windows, and frontier model usage on routine tasks, require an engineering review to identify and fix. Most organizations have not done this review; and consequently fail to discover that meaningful cost reduction is available without reducing engineering output.

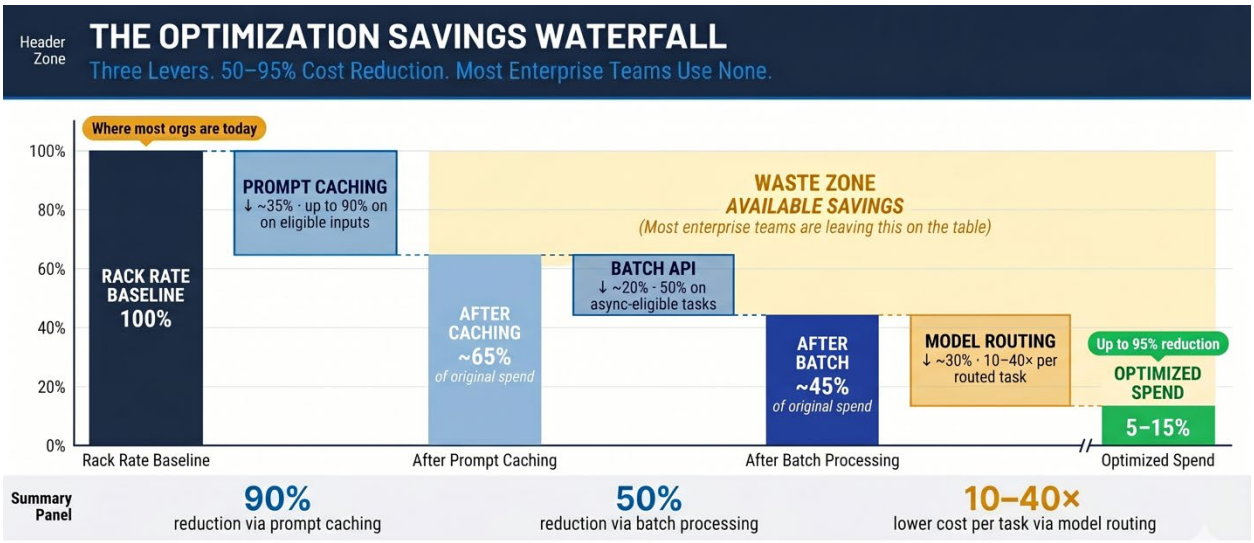

Figure 12: Three optimization levers available today. Combined, they reduce effective AI API spend by 50 to 95 percent on eligible workloads. Most enterprise teams use none.

Gap 5: No Vendor Mix Strategy

AI model pricing varies significantly across providers for comparable capabilities on specific task types. Organizations that run everything through one vendor are leaving cost optimization on the table. Single-vendor concentration also limits negotiating leverage at renewal.

Gap 6: The ROI Disconnect

MIT Project NANDA studied 300-plus enterprise AI initiatives, interviewed 150 executives, and surveyed 350 employees. Despite $30 to $40 billion in enterprise investment, 95% of organizations saw zero measurable P&L impact. [3] IBM's 2025 CEO Study of 2,000 executives across 33 countries found that only 1 in 4 AI initiatives delivered the expected return. [11]

When costs are buried in expense reports and consumption data is disconnected from business outcomes, Finance struggles to defend the investment to leadership or the board. Thus, a strategic capability investment comes to be regarded as a cash and resource drain rather than a value-added innovation. That is the reframing a PE-backed board will apply if this Total-Cost-Of-Ownership (TCO) question cannot be answered with clarity and evidence.

Trace3 Perspective: What This Data Means

The patterns above are not an abstract market trend. They are the conditions that create the six governance gaps most organizations have not yet closed, and they are accumulating in production environments right now, regardless of whether anyone is watching.

Frontier model concentration is the mechanism behind Gaps 1 and 2. When engineers default to the most capable model and no policy guides that choice, cost shifts upward without a decision ever being made. Within our client portfolio, the frontier-tier share of Bedrock inference volume moved from effectively zero in late 2025 to 80 percent by mid-2026. That is not a gradual drift. It is a structural inversion, and in most organizations, it happened below every threshold that would have triggered a budget conversation.

Rack rate exposure is Gap 4 made visible. The savings from prompt caching and batch processing are real, readily available, and require no reduction in engineering output to capture. Our portfolio data shows a 44% reduction in effective per-call cost for workloads where caching is implemented, and a 78% reduction when batch processing is added for eligible asynchronous work. Most enterprise teams are using neither. The gap is not technical. It is organizational: no one has been assigned to close it.

The hidden infrastructure layer explains why Gap 3 matters. Vector databases, data pipelines, regional routing costs, and compliance log storage do not appear on AI vendor invoices. They accumulate in adjacent cost centers, in compute, data, and cloud infrastructure, where they are not attributed to AI spend and not visible to anyone tracking the AI budget line. Without a single owner accountable for tracking and managing the holistic financial impacts of AI workloads, these sublimated costs are often undiscoverable until something forces the exercise.

Gap 5 limits an organization's ability to optimize AI workload costs downstream. Organizations running a single-vendor AI stack have already surrendered negotiating leverage at the one moment it matters most: renewal. By the time the CFO sees the full number, the options are narrower than they would have been twelve months earlier.

Gap 6 is where the MIT Project NANDA data described above lands the hardest. When costs are distributed across cost centers and outcomes are not measured against full spend, finance cannot construct an ROI narrative even if the tools are delivering value. Again, 95% of the organizations in the study saw zero measurable P&L impact, not because AI tools do not work, but because no framework existed to connect what was spent to what was produced. A leadership body that cannot get clarity as to this question of true ROI will become progressively more reticent in continuing to fund AI opportunities.

The data detailed above is not a warning to immediately throttle back on AI related commitments. It is simply evidence a) these gaps exist within current investment conditions; b) they are costing organizations real money today; and c) they are entirely addressable. Organizations that close these gaps before the board asks enter the conversation with control, evidence, and a plan. Those that wait are pulled into the discussion reactively, where the conversation is harder because it begins from a position of deficit rather than preparedness.

7. A Framework for AI Spend Visibility and Control

The following three-phase framework is designed for organizations at the early stages of building AI spend discipline. In 2026, that is most organizations.

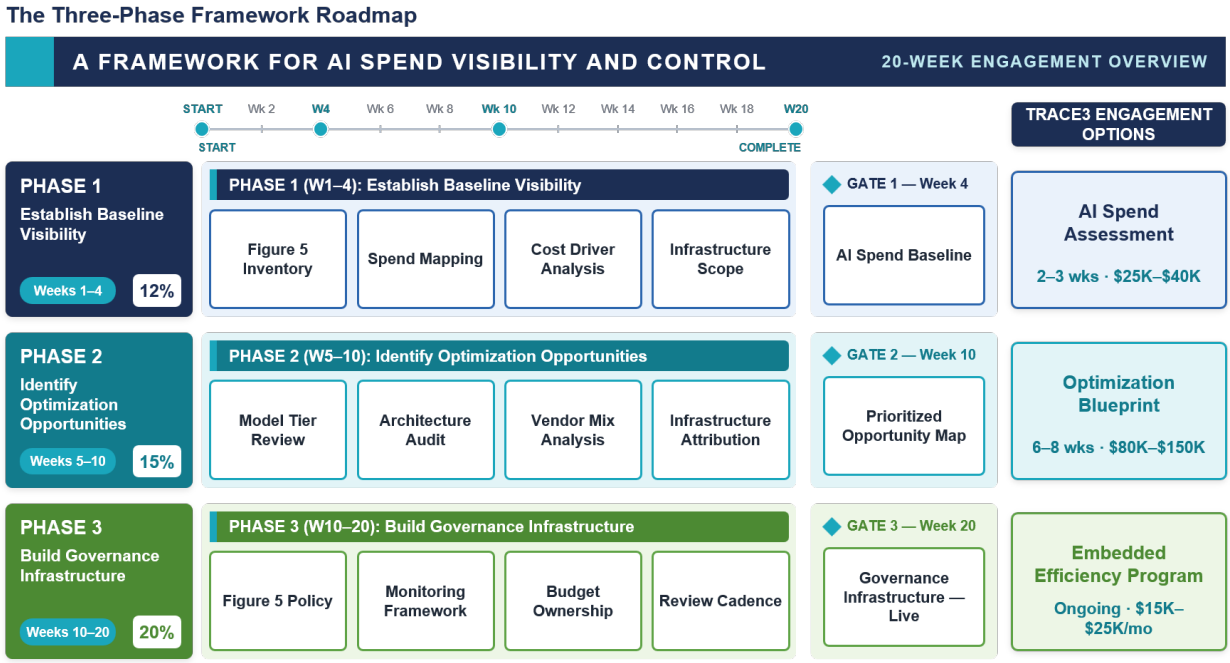

Figure 13: A 20-week path from invisible AI spend to governance infrastructure capable of answering board-level TCO questions.

Phase 1: Establish Baseline Visibility (Weeks 1 to 4)

You cannot optimize what you cannot see. This phase assembles the complete picture of AI tooling spend across the organization.

- Inventory all AI tools in use, including tools purchased through departmental budgets, individual expense reimbursement, or shadow IT
- Map spend to the vendor, model tier, team, and use case level
- Identify the top cost drivers by team, tool, and model tier

- Include supporting infrastructure costs in scope from the start: vector databases, data pipelines, regional routing

The output is an AI spend baseline. Most finance teams that complete this exercise find the number is larger and more diffuse than anyone expected. That baseline becomes the foundation for board-level TCO reporting and every optimization conversation that follows.

Phase 2: Identify Optimization Opportunities (Weeks 5 to 10)

With a baseline in place, a technical assessment identifies where spend is disproportionate to value.

- Model tier appropriateness: are frontier models being used for tasks that mid-tier models handle equally well?
- Architecture optimization: are prompt caching, batch processing, and async workflows being used where applicable?
- Vendor mix: is the organization over-concentrated in a single provider?
- Infrastructure attribution: are supporting data costs correctly allocated to AI spend?

The output is a prioritized opportunity map with projected savings against each recommendation, structured for FP&A reporting rather than engineering documentation.

Phase 3: Build Governance Infrastructure (Weeks 10 to 20)

Optimization without governance is a one-time fix. Sustainable AI cost management requires institutional infrastructure.

- A model tier policy that gives engineering teams clear guidance on appropriate model selection by task and workload type
- A monitoring and alerting framework that gives finance and engineering shared, real-time visibility into spend
- A defined budget owner and accountability structure for AI tooling as a distinct cost category
- A review cadence that updates recommendations as the model landscape shifts, which it does roughly quarterly

The output of Phase 3 is an organization that can answer TCO questions with data, defend AI investment with ROI evidence, and absorb the next round of vendor tier escalation without a budget crisis.

8. How Trace3 Approaches This

Governing AI spend is not a procurement exercise. Putting hard caps on engineering teams slows down the work those tools were bought to accelerate. It also does not fix the underlying problem. It just puts a lid on it.

Trace3 approaches this as strategic investment analysis. We connect technical telemetry to financial forecasting and monitoring, so organizations can route tasks to cost-appropriate models, optimize the supporting infrastructure spend, and treat AI as a managed strategic capability rather than an unmonitored line item.

The distinction becomes critical when leadership asks the questions that determine continued investment: What are we spending? What value is it producing? Are the unit economics improving over time? A procurement-led approach can answer only the first question, usually in the form of a reduction target or renewal savings figure. A strategic investment analysis answers the full set. It connects spend to output, shows whether efficiency is improving quarter over quarter, and gives finance a defensible ROI narrative. For a PE-backed organization facing investor reporting, margin scrutiny, or a potential liquidity event, that difference is material: one approach explains what was cut; the other explains why the investment deserves to continue.

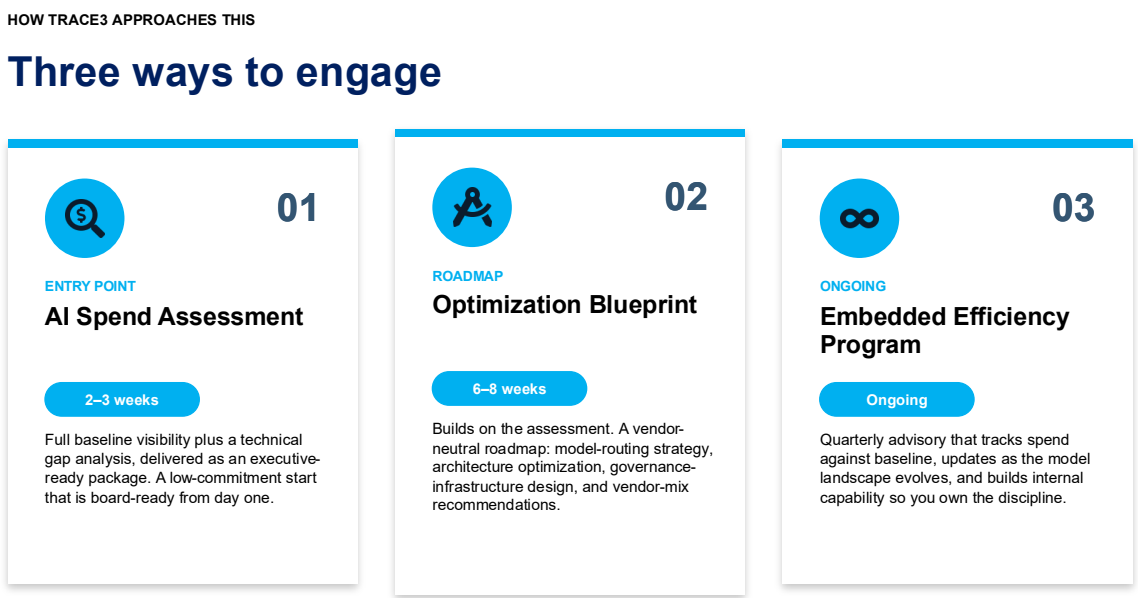

Figure 14: Trace3's three-tier engagement model for AI spend governance. A progressive engagement structure that builds from a low-commitment diagnostic to a full optimization roadmap, and on to an ongoing advisory retainer.

Our differentiation is evidence. Through our work with enterprise clients on Amazon Bedrock, Azure AI, and Google Vertex AI, we have direct visibility into how AI costs accumulate in production environments. That evidence is the foundation of our recommendations, not a generic governance framework.

9. Three First Steps for Finance Leaders

For a VP of Finance at an organization with a significant global engineering function and active AI tooling deployment, these three actions have the highest leverage.

Figure 15: Three first-step actions for FP&A leaders. A prioritized sequence of the highest-leverage moves a finance organization can take to bring AI tooling spend under governance, beginning with visibility, moving to accountability, and culminating in a renewal-timed assessment.

10. Conclusion

The organizations that built cloud cost governance in 2013 and 2014 had a structural advantage over those that built it in 2017 in response to a crisis. The advantage was both financial and organizational. They built internal fluency, tooling, and accountability structures that let them manage cloud costs as a strategic variable. When the board asked questions, they had answers backed by data.

AI tooling costs present the same opportunity on a shorter timeline. For PE-backed organizations under capital efficiency mandates, the stakes are higher than they were in the

cloud era. Ungoverned AI spend is discoverable by investors, auditors, and acquirers. It compresses EBITDA margins. It is also entirely preventable.

The question is not whether AI tooling spend will require board-level visibility. It will. The question is whether your organization builds that capability before the board asks, or after.

Ready to Take the First Step?

If the engagement model above matches where your organization is today, whether you are beginning with a diagnostic, building toward a full optimization roadmap, or looking for an ongoing advisory partner, Trace3's AI Efficiency Advisory team is ready to start the conversation. Contact your Trace3 account executive or reach out directly to discuss an AI Spend Assessment for your organization.

11. References

[1] Bruneaux, Taylor, *"AI Coding Assistant Pricing and ROI Guide 2026,"* getdx.com, June 12, 2026. [Online]. Available: https://getdx.com/blog/ai-coding-assistant-pricing/

[2] Morph, "*Microsoft Cancels Claude Code After Token Costs Blow Budget*," morphllm.com, June 18, 2026. [Online]. Available: https://www.morphllm.com/ai-coding-costs

[3] MIT Project NANDA, "*The GenAI Divide: State of AI in Business 2025*," MIT Media Lab's NANDA Initiative, July 2025, (as cited in Kahn, Jeremy, "*An MIT report that 95% of AI pilots fail spooked investors. But it's the reason why those pilots failed that should make the C-suite anxious,*" Fortune, August 21, 2025). [Online]. Available: https://fortune.com/2025/08/21/an-mit-report-that-95-of-ai-pilots-fail-spooked-investors-but-the-reason-why-those-pilots-failed-is-what-should-make-the-c-suite-anxious/

[4] Finout Writing Team, *"Anthropic API Pricing in 2026: Complete Guide,"* finout.io, June 6, 2026. [Online]. Available: https://www.finout.io/blog/anthropic-api-pricing

[5] nOps/Gartner, "*25+ Stunning FinOps Statistics That Show the True State of Cloud Spending,"* nops.io, December 11, 2025. [Online]. Available: https://www.nops.io/blog/23-stunning-finops-statistics/

[6] RightScale/GlobeNewswire, "*RightScale Estimates Companies to Waste More Than $10 Billion in Cloud Spending Over the Next Year,"* globenewswire.com, November 13, 2017. [Online]. Available: https://www.globenewswire.com/news-release/2017/11/13/1208218/0/en/RightScale-Estimates-Companies-to-Waste-More-Than-10-Billion-in-Cloud-Spending-Over-the-Next-Year.html

[7] Cursor, "*Clarifying Our Pricing,"* cursor.com, July 4, 2025. [Online]. Available: https://cursor.com/blog/june-2025-pricing

[8] Developers Digest, "*AI Coding Tools Pricing: The June 2026 Reality Check*" developersdigest.tech, July 1, 2026. [Online]. Available: https://www.developersdigest.tech/blog/ai-coding-tools-pricing-june-2026

[9] NxCode Team, "*Cursor AI Pricing 2026: Free vs Pro vs Business,"* nxcode.com, March, 2026. [Online]. Available: https://www.nxcode.io/resources/news/cursor-ai-pricing-plans-guide-2026

[10] Jimeng, "*When Cursor Silently Raised Their Price by Over 20x,"* medium.com, February 17, 2026. [Online]. Available: https://medium.com/@jimeng_57761/when-cursor-silently-raised-their-price-by-over-20-and-more-what-is-the-message-the-users-are-6af93385f362

[11] IBM Institute for Business Value, "*IBM CEO Study 2025: The CEO's Clarity Crisis*," ibm.com, May 18, 2026. [Online]. Available: https://www.cloudzero.com/blog/cursor-ai-pricing/

[12] Lens, Brian, "*Claude Fable 5 and Mythos 5: Pricing, API Costs, and Benchmark Comparison vs Opus 4.8*," arcastgroup.com, February 17, 2026. [Online]. Available: https://www.arcastgroup.com/insights/the-roi-of-ai-75-of-projects-fail.-build-a-business-case-that-works

12. Glossary

For readers less familiar with AI infrastructure terms:

Token: The unit AI vendors use to measure and bill text processing. Roughly 0.75 words, or 4 characters. A typical paragraph contains 100 to 150 tokens. Complex AI coding tasks can consume 50,000 to 200,000 tokens per interaction.

Prompt Caching: Storing repeated context, such as a system instruction or codebase summary, so the AI does not re-process it on every request. Can reduce input costs by up to 90% on eligible workloads.

Batch API Processing: Grouping AI requests for asynchronous processing rather than real-time responses. Typically cuts costs by 50%. Appropriate for code review, documentation generation, and analysis tasks that do not require an instant response.

Model Routing: Directing different task types to different AI model tiers based on complexity and cost requirements. Routing a simple classification task to a lightweight model rather than a frontier model can reduce cost per task by 10 to 40 times.

RAG (Retrieval-Augmented Generation): A technique where an AI model retrieves relevant documents or code from a database before generating a response. Allows AI to work with proprietary organizational knowledge. Requires supporting infrastructure including vector databases.

Agentic Workflow: An AI workflow where the model autonomously executes multi-step tasks, such as browsing files, writing code, running tests, and iterating, without human input at each step. Token consumption compounds significantly compared to single-turn interactions.

Vector Database: A database that stores and retrieves information by semantic similarity. Used to hold organizational knowledge such as code, documentation, and policies for AI retrieval. Common examples include Pinecone, Weaviate, and pgvector.

EBITDA: Earnings Before Interest, Taxes, Depreciation, and Amortization. The primary metric PE-backed boards use to assess operational profitability and valuation. Ungoverned AI tooling costs compress EBITDA margins directly.

About Trace3

Trace3 is an IT solutions and services company helping organizations accelerate innovation through technology. The AI Efficiency Advisory practice advises enterprise clients on AI cost governance, model optimization, and building the internal capability to manage AI spend as a strategic investment. Contact your Trace3 account executive to learn more.

Copyright 2026 Trace3. All rights reserved.