

WHITE PAPER

Proving AI ROI: The Enterprise Measurement Playbook

Why 60% of enterprises generate no material value from AI — and what the 5% who win are doing differently.

EXECUTIVE SUMMARY

Enterprises are tripling their AI spend while two-thirds generate no measurable return. The gap between AI leaders and laggards is not the model they chose or the vendor they signed — it is the discipline with which they measure. This playbook gives CFOs and Heads of AI the framework to close that gap.

Key findings:

Only 6% of enterprises achieve meaningful EBIT impact from AI, while 60% generate no material value despite record spending.^{2,3} The organizations that break through share one practice: they treat measurement as infrastructure, not a post-hoc reporting task. This paper presents the three measurement blind spots that keep enterprises in pilot purgatory, and a 90-day playbook for CFOs and Heads of AI to move from visibility to scale.

The Scale of the Bet

When your board approved the AI budget last year, they were joining the largest technology investment cycle in history. Global AI spending is forecast to reach \$2.59 trillion in 2026 — a 47% increase year-over-year — with enterprise-directed AI spend alone hitting \$37 billion in 2025, more than triple the \$11.5 billion spent the year before.¹ That number does not include the AI capability baked into the SaaS tools your teams already pay for, or the cloud compute charges accumulating as autonomous agents run in the background.

At the individual tool level, the math compounds fast. A 500-developer organization running GitHub Copilot Enterprise pays roughly \$39 per user per month in licensing alone — approximately \$234,000 annually before you account for the GitHub Enterprise Cloud subscription required to support it. Add Cursor licenses for the power users, Anthropic API access for the teams building custom agents, and the ad hoc ChatGPT Team accounts that finance approved because someone asked nicely, and a mid-sized engineering organization is easily clearing \$500,000 per year on AI coding tools — often with no unified view of what those tools are returning in cycle time reduction, defect rate improvement, or developer throughput.

The question every CFO and Head of AI is now facing is not whether to invest in AI. That decision was made. The question is whether the organization can answer — with data, not inference — what the investment is actually worth.

\$2.59T

global AI spending in 2026, up 47%
year-over-year (Gartner, May 2026)¹

3×

growth in enterprise-directed AI
spend from 2024 to 2025 (\$11.5B →
\$37B)¹

\$500K+

typical annual AI tool spend for a
500-developer enterprise — before
SaaS add-ons

The Measurement Gap

Against this spending backdrop, the returns are concentrated among a remarkably small group of organizations. McKinsey's State of AI 2025, surveying enterprise leaders across industries, found that only 6% of organizations achieve a 5% or greater EBIT contribution from AI.² Only 39% report any enterprise-level EBIT impact at all. The remaining 61% have deployed AI at scale with no measurable effect on their income statement — and many of them have done so while increasing headcount and licensing spend in the same period.

BCG's Widening AI Value Gap, drawn from a 2025 survey of 1,250 global organizations, puts the divergence in starker terms: 60% of enterprises are generating "hardly any material value" from AI, while just 5% are creating substantial value at scale.³ The gap between those two groups is widening, not closing. BCG's AI leaders achieve 1.7 times the revenue growth, 1.6 times the EBIT margin improvement, and 3.6 times the total shareholder return of laggards over the same period.³ The compounding effect of early measurement discipline is already visible in the numbers.

What explains the difference? It is not model quality, vendor selection, or the size of the AI team. BCG's analysis consistently points to one separating factor: the organizations generating outsize returns treat measurement as a core operational discipline. They define ROI metrics before deployment, not after. They track costs at the workflow level, not the invoice level. And they have a unified view across all AI categories — not a different dashboard for every tool they've licensed.

The measurement gap in numbers: NVIDIA's State of AI 2026 survey of 3,200 enterprise leaders found that 30% of organizations still cannot quantify AI's ROI — even as 88% report that AI has increased revenue somewhere in the business.⁴ "Somewhere" is not a measurement framework. It is a board question waiting to happen.

The 30% who cannot measure ROI are not outliers. They represent the middle of the enterprise AI adoption curve — organizations that have moved past the exploratory phase, are spending real money, and are being asked to justify that spending to their boards. The tools to answer that question exist. The measurement infrastructure to use them does not yet, for most.

6%

of enterprises achieve 5%+ EBIT contribution from AI (McKinsey, 2025)²

60%

generate no material value from AI despite record investment (BCG, Sep 2025, n=1,250)³

30%

of enterprise leaders cannot quantify AI ROI at all (NVIDIA, n=3,200)⁴

Three Measurement Blind Spots

Enterprises do not fall into the 60% laggard category because they chose the wrong AI tools. They fall there because they have no measurement layer beneath those tools. Three structural blind spots account for the majority of the gap.

Blind Spot 1: Coding Tools — Productivity Theater

AI coding tools are the most widely deployed category of enterprise AI and the most consistently measured in the wrong way. Most organizations track acceptance rate — the percentage of AI-suggested code that developers accept — as their primary signal of tool value. Acceptance rate is a usage metric, not a productivity metric. A developer who accepts 70% of suggestions from a mediocre tool and produces no faster pull requests is generating zero business value at full license cost.

The right signal is cycle time delta: how much faster are AI-assisted pull requests moving from first commit to merge, compared to non-AI-assisted work by the same developers on similar task types? McKinsey's analysis of enterprise software organizations found that top performers achieve 16–30% productivity gains from AI coding tools — but only when they are measuring the right outcomes and using those measurements to drive adoption discipline.⁷ Organizations with 80–100% developer adoption, sustained by evidence-based coaching rather than mandate, see productivity gains exceeding 110%.⁷ The measurement creates the outcome; it does not just track it.

A second coding tool blind spot is cost attribution. Most organizations see AI coding tool spend as a single line item — total annual contract value divided by developer headcount. They cannot see cost-per-PR, cost-per-merged-line, or spend by team, project, or cost center. Without that granularity, they cannot identify which teams are getting strong ROI (and should have their licenses protected or expanded) and which have idle seats (and should have their licenses reclaimed). The license reclaim opportunity alone typically generates enough savings to fund the measurement platform that found it.

Blind Spot 2: Assistive AI — The 78% Problem

While most organizations are focused on the AI tools they approved, their employees are using a substantially larger set. Microsoft WorkLab's 2025 research found that 78% of employees have brought their own AI tools to work without employer approval.⁶ A separate analysis suggests that 98% of organizations have unsanctioned AI tools actively in use across their workforce — yet only 25% have comprehensive visibility into what those tools are or what data flows through them.⁶

This creates two compounding problems. The first is productivity measurement error: the ROI calculation for your approved AI assistant portfolio is systematically understated, because a material fraction of AI-generated productivity is attributed to tools that do not appear in your reporting. Employees using unapproved tools to get their jobs done are generating business value that never shows up in the ROI case for your managed stack.

The second is data exposure risk. Employees who need AI assistance and cannot get it through approved channels will find it through unapproved ones — and the tools they choose may not meet your data handling requirements, security standards, or regulatory obligations. PII, source code, legal materials, and confidential

deal data regularly flow into consumer-grade AI interfaces. The first time most CISOs learn about this is in a board question or an incident report, not a proactive measurement dashboard.

Blind Spot 3: Agentic AI — The Invisible Cost Center

Agentic AI — autonomous workflows that execute multi-step tasks without direct human involvement — represents the fastest-growing and least-instrumented segment of enterprise AI spend. BCG estimates that agents already account for 17% of the total AI value being created in enterprise organizations, a figure expected to reach 29% by 2028.³ They are also, for most organizations, the segment with the least measurement infrastructure.

Agent costs accumulate in ways that break the intuitions developed around SaaS licensing. There is no per-seat fee. Costs scale with execution volume, LLM token consumption, tool calls, and compute. A single agent workflow running frequently across a large team can accumulate tens of thousands of dollars in cloud costs that do not surface until the monthly invoice — with no breakdown by workflow, team, or business outcome. Without execution-level analytics, the CFO cannot evaluate whether the spend is generating a return, and the Head of AI cannot identify which workflows are cost-effective and which should be rebuilt or retired.

The Pilot Purgatory Problem

The three measurement blind spots described above do not just make ROI calculation difficult. They are the primary cause of AI pilot failure. Gartner estimates that more than half of generative AI projects are abandoned after the proof-of-concept phase — the single most common reason being that organizations cannot demonstrate clear business value from the pilot to justify investment in production deployment.⁵ RAND's analysis of enterprise AI projects puts the overall failure rate at 80.3% when measured against the business value promised at project initiation.⁵

Pilot purgatory is not a technology problem. It is a measurement problem. Projects that cannot demonstrate ROI at the pilot stage fail to clear the internal hurdles required to reach production — not because they are not working, but because no one built the instrumentation to prove that they are. The value was there; the evidence was not.

What winners do differently: BCG's research on AI leaders finds one consistent practice separating the 5% generating substantial value from the 60% generating none: they define measurement criteria before deployment, not after. They tie every AI initiative to a specific, quantifiable business outcome — cycle time, cost per transaction, revenue per employee — and they build the instrumentation to track it from week one, not from the next quarterly review.³

The BCG winner/laggard data makes the business case for measurement infrastructure plain. Organizations in the top AI performance tier achieve 1.7 times the revenue growth and 3.6 times the total shareholder return of their laggard peers over the same period.³ The measurement is not just a reporting exercise. It is the mechanism by which AI value compounds — because what gets measured gets managed, and what gets managed gets scaled.

50%+

of GenAI pilots abandoned after POC — most common
reason: no measurable business value (Gartner)⁵

3.6×

total shareholder return advantage for AI measurement
leaders vs. laggards (BCG, Sep 2025)³

The Measurement Playbook

The path from measurement gap to measurement discipline is not a multi-year transformation. Organizations that move quickly through the following three phases — Visibility, Baseline, and Scale — consistently find that the first 90 days generate enough evidence to fund the next 12 months of AI investment decisions.

WEEKS 1–4

Phase 1: Visibility — Map What You Actually Have

Before you can measure ROI, you need a complete and accurate picture of your AI footprint. This means discovering every tool in active use — not just what was approved, but what employees are actually running. For most organizations, the gap between the approved tool list and the actual tool count is a factor of five to ten. Visibility also means establishing cost attribution at the workflow level: which teams are spending what, on which tools, and for which purposes. The visibility phase typically takes two to four weeks and invariably surfaces quick wins — idle licenses to reclaim, duplicate tools to consolidate, shadow AI tools to either approve or block.

DAY 30

Phase 2: Baseline — Establish the Metrics That Matter

Once you can see your full AI footprint, establish the baseline metrics against which ROI will be measured. For coding tools: PR cycle time by developer cohort (AI-assisted vs. non-AI-assisted), cost per merged PR by team and tool, and developer adoption cohort distribution. For assistive AI: time saved per user per tool, prompt quality scoring, and data risk indicators by tool category. For agentic AI: execution success rate, cost per execution by workflow, and KPI contribution per run. Thirty days of baseline data is enough to see where the signal is strong and where investment is producing noise. These baselines become the internal benchmarks against which every future AI decision is evaluated.

DAY 90

Phase 3: Scale / Fix / Kill — The Investment Decision

With 90 days of baseline data, the scale/fix/kill decision becomes straightforward. Scale: tools and workflows where ROI exceeds the cost threshold, adoption is high, and the productivity delta is measurable — expand licenses, deepen integration, standardize across teams. Fix: tools where usage is high but ROI is unclear — adjust how they are deployed, invest in user coaching, or redefine the KPIs. Kill: tools where usage has not materialized or cost-per-outcome is unacceptable — reclaim licenses, reallocate budget to the programs that are working. The 90-day milestone is typically the first point at which the CFO can walk into a board meeting and answer the AI ROI question with evidence.

Role Takeaways

Closing the AI measurement gap requires alignment across three roles. Each has a distinct stake in the outcome and a distinct set of questions the measurement layer needs to answer.

CFO

Demand workflow-level cost attribution before approving new AI tools — not invoice-level

Use the 90-day baseline period to establish internal benchmarks before the next budget cycle

Treat shadow AI discovery as a cost audit, not just a security audit — idle license reclaim funds the measurement platform

Require vendor ROI claims to be independently verified against your own workflow data, not the vendor's case studies

HEAD OF AI

Build measurement criteria into every AI initiative before deployment begins, not at the 60-day check-in

Establish a cross-pillar dashboard that gives the board a unified view of AI value — not three separate tool reports

Use adoption cohort data (power, casual, idle) to direct coaching resources where they move the ROI needle fastest

Treat the visibility audit as the foundation of your AI strategy — you cannot govern what you cannot see

VP ENGINEERING

Measure cycle time delta, not acceptance rate — it's the signal that survives the board question

Establish per-team and per-project cost attribution for AI coding tools before budgets compound

Use agent execution analytics to catch cost overruns before the monthly cloud invoice — not after

Share productivity data openly with developers — teams with visibility into their own AI ROI consistently out-adopt teams without it

Conclusion

The organizations separating themselves from the 60% who generate no material AI value are not buying better models. They are building better measurement. They define ROI criteria before deployment. They track costs at the workflow level. They surface shadow AI as a signal, not just a risk. And they synthesize all of it into a single view that the board can interrogate without a week of analyst work in between.

That measurement infrastructure is not a luxury for organizations at scale. It is the prerequisite for getting to scale. The 90-day playbook in this paper is the path from AI spend to AI evidence — and from AI evidence to the board confidence that funds the next round of investment. The enterprises that build this capability now will have a compounding measurement advantage that grows with every AI initiative they run. The ones that wait will keep answering the same board question with the same "somewhere in the business" hedge until someone asks for the receipts.

Olakai's Enterprise AI Analytics and Governance Platform gives CFOs and Heads of AI the unified measurement layer their current tool stack cannot provide: Coding IQ for developer productivity and AI spend governance, Assistive IQ for copilot ROI and shadow AI detection, and Agent IQ for agentic workflow analytics and governance — with Kai synthesizing all three into the board-ready answers your executives need. To see it in your environment, visit olakai.ai and request a Magic Link.

About Olakai. Olakai is the Enterprise AI Analytics and Governance Platform — purpose-built to help organizations measure ROI, control costs, and govern risk across every AI tool, every agent, and every team. Coding IQ tracks developer productivity and AI spend governance. Assistive IQ monitors copilot usage, surfaces shadow AI, and enforces content policy. Agent IQ provides real-time observability into agentic workflows. Kai synthesizes all three into conversational intelligence your board can interrogate in plain English. Vendor-neutral. Deployment-flexible. Built for the CFO and Head of AI who need evidence, not estimates. olakai.ai

References

1. Gartner. "Gartner Forecasts Worldwide AI Spending to Grow 47 Percent in 2026." Press Release, May 2026. gartner.com
2. McKinsey & Company. "The State of AI 2025: Agents, Innovation, and Transformation." McKinsey Global Institute, 2025. mckinsey.com
3. Boston Consulting Group. "The Widening AI Value Gap." BCG Global Survey, September 2025 (n=1,250 global organizations). bcg.com
4. NVIDIA Corporation. "State of AI Report 2026." Survey of 3,200 enterprise leaders across industries, 2026. nvidia.com
5. Gartner. "Why Half of GenAI Projects Fail." Gartner Research, 2025. gartner.com
6. Microsoft Corporation. "2025 Work Trend Index." Microsoft WorkLab, 2025. microsoft.com/worklab
7. McKinsey & Company. "The AI Revolution in Software Development." McKinsey Technology, 2025. mckinsey.com