

Level 4: Results

AI Literacy for the Modern Workforce

Within the Kirkpatrick model¹, Level 4 measures whether the behavioral changes documented at Level 3 produced measurable organizational outcomes. It is the tier that connects individual competency to business results, and therefore the tier executives care about. This document is an evaluation blueprint, not a results report: the program is a portfolio piece with no organizational deployment, so no Level 4 data exists. What follows is the complete methodology, the KPI framework, the isolation techniques, and a worked ROI example that shows how the model operates once an organization supplies its own data.

The attribution chain

Level 1 establishes that the program was perceived as relevant. Level 2 establishes that it produced knowledge gains. Level 3 establishes that the knowledge moved into workplace behavior. Level 4 observes that the organizational metrics those behaviors target then improved. The argument becomes credible only when the links connect: the program was relevant, produced understanding, changed behavior, and the outcomes those behaviors drive moved in the expected direction. Strip out Level 3 and any Level 4 result is exposed to the obvious question: how do you know the program caused it?

Why adoption metrics are not results

The dominant failure mode at this tier is the volume-ROI calculation: deploy AI tools to N employees, measure that X% activated and Y% use them weekly, multiply by a vendor-estimated time saving, and report a number with a dollar sign. It is the equivalent of estimating a gym's health benefit by counting badge swipes at the door: the swipe count says nothing about whether anyone trained. The weakness is the assumption that each unit of usage yields a uniform unit of value, which the research contradicts: estimated in-conversation savings run from about 20% to about 95% across named tasks, with most tasks between 50% and 95%², and value depends on whether an interaction was augmentation or automation and how it was structured³. Ignoring this has a direct cost: Uber's COO reported burning the company's entire annual AI budget in four months without a clean line to user value⁴, and within days Fortune declared the metric dead⁵.

This framework measures outcomes downstream of competent use, and treats adoption as an input condition, not a result. The vendor research agrees. OpenAI's enterprise study concludes that rapid adoption should not be equated with immediate productivity transformation⁶. It also finds that, among adopting public companies, larger firms record lower measured use per employee, a pattern the authors say may reflect a mechanical or organizational scaling effect⁷.

The same gap between adoption and results appears in organizations worldwide. In McKinsey's 2026 global survey, nearly nine in ten respondents report regular AI use in their organization and 80% report personal productivity gains. Yet 37% report any positive effect on their organization's earnings, a figure flat on the year despite more organizations scaling. Respondents do cite other organization-wide benefits, most notably innovation, competitive differentiation, and customer and employee satisfaction, and even those reporting no earnings effect say AI has reduced costs in specific business functions⁸.

The measurement problem is structural. Charting the Atlanta Fed's December 2025 survey of US CFOs, analyst Benedict Evans sorts the reported AI impacts into two kinds. Productivity-type gains are easy to deploy and hard to measure. Cost savings and new revenue are the reverse: hard to deploy, easy to measure (Baslandze et al.)⁹. Usage researchers make a

related point from the supply side: the unit that matters is cost per successful outcome, not cost per token¹⁰.

What is measurable at enterprise scale

A deploying organization will ask where this data comes from, and the answer cuts against the easy assumption that the AI vendor supplies it. What the major providers expose to administrators is mostly usage telemetry, not outcomes. OpenAI's own research on that telemetry measures three things per organization-week: messages, active users, and output tokens¹¹. Its authors state that the task results, drawn from classified message content, do not measure downstream work products, productivity effects, or changes in organizational routines⁶. ChatGPT Enterprise reports seats, weekly active users, messages per active user, and usage by project and skill^{12,13}; Claude Enterprise reports active members against assigned seats, chats, projects, and artifacts¹⁴; Gemini for Workspace reports active users and usage per app and per feature¹⁵. Every one of those is an adoption metric, the same volume signal this framework rejects as a results measure.

OpenAI and Anthropic now add impact measures, and they do not change that conclusion. Both run user surveys: OpenAI sends workspace users a fixed impact survey¹⁶, and Anthropic lets administrators ask users short in-product questions about what they get out of Claude¹⁷. Anthropic also shows an estimated time saved and, for its coding tool, productivity estimates whose inputs the administrator can adjust to the organization's own assumptions¹⁸. A survey answer is self-report, the kind of evidence Module 2 shows drifting from observed behavior. An estimate built on assumed inputs is a projection, not a measurement. Neither records whether the work got better. The conversation content that could support a quality judgment is not in the analytics at all: OpenAI's guide states that workspace analytics exposes no message text and sends administrators who need raw records to its Compliance API¹⁹, an interface built for compliance, not performance evaluation.

The consequence is that the KPIs that matter do not come from the AI provider at all. Efficiency and quality measures come from the organization's own systems: the project tracker, version history, QA and incident logs, and review workflows. Some, such as a clean count of AI-caused errors, require new instrumentation, because most organizations do not yet tag errors by cause. Baseline establishment is therefore a deployment prerequisite: the provider hands the organization an adoption dashboard, and the outcomes the framework needs are the organization's to instrument. That instrumentation is one of the practices separating organizations that report significant AI returns from the rest. They are twice as likely to have defined processes to quantify the impact of their AI initiatives²⁰.

The KPI framework

KPIs fall into four categories, each tracing back to specific Level 3 behaviors. Organizations select from them based on which outcomes matter and which they already have the infrastructure to measure. Employers say the outcomes they most expect from training investment are enhanced productivity (77%), competitiveness (70%), and talent retention (65%)²¹. Those are expectations, not measured results, which is why this framework measures outcomes. The efficiency and quality KPIs speak to productivity; this framework does not measure competitiveness or retention directly.

None of the KPIs assumes a profit motive. In a university, hospital, or public agency, E1–E3 count staff time returned to core work and Q1 counts errors avoided. The ROI Institute has also published a guide to the same method for government training programs²².

Efficiency gains (read together: a faster draft that needs more rounds is not a gain)

KPI	Definition	Source	L3 trace
E1 Time-to-first-draft	Time from assignment to first draft for AI-eligible tasks	PM data (ticket/doc timestamps)	B1, B3, B4
E2 Revision cycles	Average revision rounds before acceptance	Review/approval workflow logs	B5, B6, B8
E3 Rework rate	Share of deliverables needing substantive post-delivery correction	QA / client-feedback systems	B2, B5, B8

Quality improvements (the highest-consequence category: one prevented error in a regulatory filing can outweigh a cohort's efficiency gains; inaccuracy is also among the AI risks organizations most often report working to mitigate, at 52% of high performers and 53% of others²³)

KPI	Definition	Source	L3 trace
Q1 AI-related error rate	Documented errors from unverified AI output (fabricated citations, bad statistics)	Incident reports, QA logs (needs AI-error tagging)	B2, B5, B6
Q2 Appropriate delegation rate	Share of AI-assisted tasks within the model's reliable range	Quarterly sample audits (new infrastructure)	B1, B2, B7

Adoption maturation (cultural change, the organizational analog of the Level 3 behavior change)

KPI	Definition	Source	L3 trace
A1 AI transparency rate	Share of the team openly discussing AI practices	Derived from B9 aggregated by team	B9, B7
A2 Concealment reduction	Drop from the 69% concealment baseline ²⁴	Anonymous pulse survey at 6 months	B9, B12

Capability development (the strategically decisive category)

KPI	Definition	Source	L3 trace
C1 Peer coaching incidence	Documented instances of participants coaching colleagues	Derived from B11 at 90 days	B11, B9
C2 Time-to-competency for new tools	Onboarding time to a new AI tool, alumni vs. non-participants	Opportunistic, when a new tool is introduced	B1, B3, B5 (tool-agnostic)

C2 is the argument that justifies AI-literacy investment over tool-specific onboarding: if alumni onboard to each new tool faster, the program produced transferable literacy, which compounds as organizations adopt more tools through 2030. Employer plans point that way: 86% expect AI to transform their business by 2030, 77% plan to reskill existing staff to work alongside AI, and 62% plan to hire people who already can²⁵. AI is also reaching more parts of the business: 56% of respondents say their organizations use AI in three or more functions in 2026, up from 51%²⁶. The tool count is already high. Each platform shift has multiplied the tools in use, and a typical large US enterprise now runs 400 to 500 SaaS applications (Productiv)²⁷.

Isolating the program's contribution

Attribution is the central methodological challenge, and every technique produces an estimate, not a proof. The framework uses several in combination, in decreasing order of rigor: a comparison-group design where deployment is phased (the later group is a natural control); trend-line analysis against 6–12 months of pre-program baseline; and independently collected manager and participant estimates of the share of improvement attributable to the program, each adjusted downward for optimism. The phased design is a difference-in-differences comparison, the same structure Anthropic's labor-market analysis uses to separate AI effects from the business cycle: a change that appears in both groups is not attributed to the intervention. That analysis also reports the smallest effect its design could detect, about 1 percentage point of unemployment, and this evaluation should state its own up front²⁸. The result is reported as a range, and the ROI model uses the lower bound by default. If the return is positive at the lowest defensible attribution, the case is strong; if it is positive only at the upper bound, it should be communicated as speculative.

The ROI model

The model applies the Phillips ROI Methodology²⁹, the established framework that extends Kirkpatrick with a fifth level, financial ROI³⁰. The formula below is Phillips' own, net program benefits over fully loaded cost expressed as a percentage; the isolation step above is Phillips' "isolate the effects of the program"; and the conservative inputs follow his Guiding Principles, which call for a fully loaded cost, the most conservative estimate among alternatives, and attribution adjusted downward for error. This project supplies defensible inputs. Most are drawn from the research corpus rather than a vendor benchmark; fully loaded compensation (W) and the attribution factor (A) are the adopting organization's own inputs, which the worked example below fills with stated assumptions.

None of the corpus sources supplies a fully loaded compensation figure or an attribution factor. MGI's average-pay figures for its occupation groups exclude benefits³¹. McKinsey's EBIT attribution is a respondent's judgment, not an audited share³². W and A remain assumptions an adopting organization must replace.

Two outside benchmarks give a deploying organization reference points. In US data for June 2026, benefits, including legally required ones, made up 30.0% of private-industry compensation, which puts total compensation at about 1.43 times wages and salaries³³. The worked example's 1.4 multiplier for W matches that ratio and so leaves out overhead. For program cost, ATD's 2024 per-employee L&D cost benchmarks³⁴ are a reference point for a deploying organization's own figure.

Two other Guiding Principles shape the model. It counts one year of benefits only, the rule for short-term programs. It reports the delegation, adoption, and capability KPIs (Q2, A1, A2, C1, C2) as intangible benefits rather than assigning them a dollar value³⁵.

THE ROI MODEL · TERM BY TERM

Efficiency benefit = $N \times W \times T_{ai} \times S \times A$

Quality benefit = $Q \times Q_reduction \times A$

Quality-failure cost $Q = N \times q$

Program cost $C_program = C_fixed + N \times c$

ROI = $(\text{Total benefit} - C_program) \div C_program \times 100$

- N

participants in the cohort (default 200)
- T_{ai}

share of work hours on AI-eligible tasks (default 25%)
- A

share of improvement credited to the program (default 40%)
- c

participant and manager time per participant (default \$525)
- Q_reduction

reduction in those failures (default 20%)
- W

fully loaded annual cost per participant (default \$140,000)
- S

time savings on those tasks with competent use (default 30%)
- C_fixed

build, delivery, evaluation administration, and contingency (default \$235,000)
- q

annual cost of AI-related quality failures per participant (default \$2,000)

Two independently attributed benefit streams, set against total program cost. Program cost and quality-failure cost both scale with N. The calculator below starts at these conservative defaults.

INTERACTIVE · ROI MODEL

ROI calculator

The defaults below are the conservative inputs from the static model, each with its rationale; together they reproduce the document’s worked example. Change any input and the chain recomputes.

N

Participants

200

Mid-sized cohort for a departmental rollout.

T_{ai}

AI-eligible task share

25

%

MGI puts agents, its term for all machines that automate nonphysical work, AI-based or not, at 44% of US work hours as technical potential. Its midpoint scenario automates about 27% of all US work hours by 2030³⁶; 25% sits just below that midpoint. Handa et al. find about 57% of occupations use AI on at least a tenth of their tasks³⁷; their 57% conversation split is a different statistic and is not used here.

W

Fully loaded annual cost

\$140000

Mid-career base ~\$100K × 1.4 fully loaded (BLS puts benefits, legally required ones included, at 30% of private-industry compensation, about 1.43 times wages; overhead not included)³³.

S

Time savings, competent use

30

%

Inside the 14–56% range that controlled trials report; well below the 81% in-conversation median, which leaves out the time spent refining output³⁸.

A Attribution factor
 %

Lower-bound; assumes 60% of improvement is from other factors.

c_{fixed} Fixed program cost

The program's resource plan prices a five-person build at \$190,000–\$258,000, including evaluation administration and 15% contingency; the default sits in the upper half of that range.

c Time cost per participant

About 6.3 hours of participant time (the 230–315-minute program, the Level 1 survey, and three Level 3 self-assessments and check-ins) at $W \div 2,080$, plus about 1 hour of manager time at ~\$89 an hour. The default prices this at the default W.

q AI-quality-failure cost per participant

Placeholder (~\$2,000 per participant a year); no cited source prices AI-related errors. Replace with incident data: errors × hours to fix × loaded rate.

Q_{reduction} Failure reduction
 %

Conservative one-fifth reduction.

EFFICIENCY BENEFIT

$$N \times W \times T_{ai} \times S \times A$$
\$840,000
QUALITY BENEFIT

$$Q \times Q_{reduction} \times A$$
\$32,000
**Total annual benefit**

– Program cost (C_{program})

\$872,000
\$340,000
**RETURN ON INVESTMENT · ILLUSTRATIVE**
156%

$$ROI = (Total\ benefit - Program\ cost) \div Program\ cost \times 100$$

At the current attribution (A = 40%), ROI is 156%; halving A to 20% would give 28%. Break-even at these inputs: A of 15.6% or S of 11.0%.

Formula and conservative-input discipline follow the Phillips ROI Methodology: net program benefits over fully loaded cost, with attribution adjusted downward for error. The defaults are illustrative inputs, not a measured return: no Level 4 data exists until an organization deploys the program and supplies its own baselines.

Timeline, and the full chain

Milestone	Activity
Pre-program	Establish KPI baselines (6–12 months of history where systems hold it, 3 months at minimum) and the A2 concealment baseline
90 days	Level 3 complete; begin watching Level 4 KPIs for direction (no ROI yet)
6 months	First Level 4 point: measure KPIs, run estimation surveys, calculate a preliminary ROI range
12 months	Full evaluation: trend-line analysis, final ROI with confidence range, integrated L1–L4 narrative

Populated, the chain reads as a single argument: participants rated the program relevant and named intended changes (L1), gained measurable knowledge across all four domains (L2), changed specific behaviors that managers and self-reports both confirmed (L3), and the organizational metrics those behaviors target improved (L4). Each level hands the next its evidence; remove any one and the argument has a gap. For that reason the framework was designed as an integrated system from the outset.¹

¹ No Level 4 data exists because the program has not been deployed at organizational scale. Causal certainty is unattainable outside a randomized controlled trial, which is rarely feasible organizationally; the combined isolation techniques produce a defensible estimate reported as a range, not a proof. The model requires pre-program baselines for every tracked KPI, so baseline establishment is a prerequisite for deployment, and some outcomes (cultural transparency, time-to-competency for future tools) may take longer than twelve months to fully materialize.

Notes

1. Kirkpatrick Partners, “The Kirkpatrick Model,” accessed October 2026, <https://www.kirkpatrickpartners.com/the-kirkpatrick-model/>
2. Alex Tamkin and Peter McCrory, “Estimating AI Productivity Gains from Claude Conversations,” Anthropic, November 2025, <https://www.anthropic.com/research/estimating-productivity-gains>
3. Kunal Handa et al., “Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations,” Anthropic, February 2025, <https://arxiv.org/abs/2503.04761>
4. Jake Angelo, “Uber burned through its entire 2026 AI budget in four months. Now its COO is questioning whether it’s worth it,” Fortune, May 2026, <https://fortune.com/2026/05/26/uber-coo-ai-spending-tokens-claude-code/>
5. Jeremy Kahn, “Tokenmaxxing is over. That’s because it never measured what really counts to see ROI from AI,” Fortune, May 2026, <https://fortune.com/2026/05/28/tokenmaxxing-is-dead-companies-didnt-get-the-roi-from-ai-they-wanted-to-see/>
6. Aaron Chatterji et al., “How Organizations Use AI: Evidence from ChatGPT,” OpenAI working paper, August 2026, p. 23, <https://arxiv.org/abs/2608.12236>
7. Aaron Chatterji et al., “How Organizations Use AI: Evidence from ChatGPT,” OpenAI working paper, August 2026, pp. 13–14, <https://arxiv.org/abs/2608.12236>
8. Dan Tinkoff, Lieven Van der Veken, and Michael Chui, with Tara Balakrishnan, “The state of AI in 2026: On the road to ROI,” McKinsey & Company, August 2026, pp. 5, 12, 14, <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>
9. Benedict Evans, “AI eats the world (May 2026 edition),” May 2026, slide 35 (Atlanta Fed, Baslandze et al., December 2025), <https://www.ben-evans.com/s/2026-Spring-AI.pdf>
10. Malika Aubakirova et al., “OpenRouter State of AI: An Empirical 100 Trillion Token Study with OpenRouter,” OpenRouter and Andreessen Horowitz, December 2025, PDF p. 54, <https://openrouter.ai/state-of-ai>
11. Aaron Chatterji et al., “How Organizations Use AI: Evidence from ChatGPT,” OpenAI working paper, August 2026, p. 7, <https://arxiv.org/abs/2608.12236>
12. OpenAI, “ChatGPT Enterprise user analytics guide,” OpenAI Academy, accessed October 2026, <https://academy.openai.com/public/clubs/admins-6o6xf/resources/chatgpt-enterprise-user-analytics-guide>

13. OpenAI, “User analytics for ChatGPT Enterprise and Edu,” OpenAI Help Center, accessed October 2026, <https://help.openai.com/en/articles/10875114-user-analytics-for-chatgpt-enterprise-and-edu>
14. Anthropic, “View usage analytics for Team and Enterprise plans,” Claude Help Center, accessed October 2026, <https://support.claude.com/en/articles/12883420-view-usage-analytics-for-team-and-enterprise-plans>
15. Google, “Review Gemini usage in your organization,” Google Workspace Admin Help, accessed October 2026, <https://knowledge.workspace.google.com/admin/generative-ai/review-gemini-usage-in-your-organization>
16. OpenAI, “ChatGPT Enterprise user analytics guide,” OpenAI Academy, accessed October 2026, Admin FAQ, “Can admins customize the impact survey?”, <https://academy.openai.com/public/clubs/admins-606xf/resources/chatgpt-enterprise-user-analytics-guide>
17. Anthropic, “View usage analytics for Team and Enterprise plans,” Claude Help Center, accessed October 2026, section “Surveys”, <https://support.claude.com/en/articles/12883420-view-usage-analytics-for-team-and-enterprise-plans>
18. Anthropic, “View usage analytics for Team and Enterprise plans,” Claude Help Center, accessed October 2026, sections “What are the results?” (estimated time saved) and “Claude Code”, subsection “Value” (adjustable inputs), <https://support.claude.com/en/articles/12883420-view-usage-analytics-for-team-and-enterprise-plans>
19. OpenAI, “ChatGPT Enterprise user analytics guide,” OpenAI Academy, accessed October 2026, Admin FAQ, “Does workspace analytics replace the Compliance API?”, <https://academy.openai.com/public/clubs/admins-606xf/resources/chatgpt-enterprise-user-analytics-guide>
20. Dan Tinkoff, Lieven Van der Veken, and Michael Chui, with Tara Balakrishnan, “The state of AI in 2026: On the road to ROI,” McKinsey & Company, August 2026, p. 19, <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>
21. World Economic Forum, “The Future of Jobs Report 2025,” January 2025, p. 48, https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf
22. Jack J. Phillips and Patricia Pulliam Phillips, “The Bottom Line on ROI: A One-Day Workshop Describing How to Calculate the ROI in Government Training Programs,” ROI Institute, 2009, <https://www.calhr.ca.gov/wp-content/uploads/sites/361/2025/07/how-to-calculate-roi-in-government-training-programs.pdf>
23. Dan Tinkoff, Lieven Van der Veken, and Michael Chui, with Tara Balakrishnan, “The state of AI in 2026: On the road to ROI,” McKinsey & Company, August 2026, p. 23, <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>
24. Kunal Handa et al., “Introducing Anthropic Interviewer: What 1,250 Professionals Told Us About Working with AI,” Anthropic, December 2025, research post, “AI’s impact in the general workforce” (general-workforce sample, N = 1,000: 69% mentioned social stigma, a theme count from interview transcripts; fact-checker quotation), <https://www.anthropic.com/research/anthropic-interviewer>
25. World Economic Forum, “The Future of Jobs Report 2025,” January 2025, pp. 62–63, Figure 4.14, https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf
26. Dan Tinkoff, Lieven Van der Veken, and Michael Chui, with Tara Balakrishnan, “The state of AI in 2026: On the road to ROI,” McKinsey & Company, August 2026, p. 5, <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>
27. Benedict Evans, “AI eats the world (November 2025 edition),” November 2025, slide 10 (Productiv), <https://www.ben-evans.com/s/2025-Autumn-AI.pdf>
28. Maxim Massenkoff and Peter McCrory, “Labor Market Impacts of AI: A New Measure and Early Evidence,” Anthropic, March 2026, pp. 11–12, <https://www.anthropic.com/research/labor-market-impacts>
29. ROI Institute, “ROI Methodology,” accessed October 2026, <https://roiinstitute.net/roi-methodology/>
30. Training Industry, “Phillips ROI Methodology (glossary entry),” accessed October 2026, <https://trainingindustry.com/glossary/phillips-roi-methodology/>
31. Lareina Yee et al., “Agents, robots, and us: Skill partnerships in the age of AI,” McKinsey Global Institute, November 2025, p. 57, <https://www.mckinsey.com/mgi/our-research/agents-robots-and-us-skill-partnerships-in-the-age-of-ai>
32. Dan Tinkoff, Lieven Van der Veken, and Michael Chui, with Tara Balakrishnan, “The state of AI in 2026: On the road to ROI,” McKinsey & Company, August 2026, p. 18, <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>
33. U.S. Bureau of Labor Statistics, “Employer Costs for Employee Compensation — June 2026,” September 2026, summary, private industry workers (wages and salaries \$32.82, 70.0%; benefits \$14.07, 30.0% of total compensation of \$46.89 per hour worked), <https://www.bls.gov/news.release/ecec.nr0.htm>
34. Association for Talent Development, “ATD Research: L&D Professionals Are Optimistic About TD’s Value Within Organizations (2024 State of the Industry),” 2024, <https://www.td.org/content/press-release/atd-research-l-and-d-professionals-are-optimistic-about-td-s-value-within-organizations>
35. Jack J. Phillips and Patricia Pulliam Phillips, “The Bottom Line on ROI: A One-Day Workshop Describing How to Calculate the ROI in Government Training Programs,” ROI Institute, 2009, p. 15, <https://www.calhr.ca.gov/wp-content/uploads/sites/361/2025/07/how-to-calculate-roi-in-government-training-programs.pdf>
36. Lareina Yee et al., “Agents, robots, and us: Skill partnerships in the age of AI,” McKinsey Global Institute, November 2025, pp. 11–12, 37, <https://www.mckinsey.com/mgi/our-research/agents-robots-and-us-skill-partnerships-in-the-age-of-ai>
37. Kunal Handa et al., “Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations,” Anthropic, February 2025, p. 13, <https://arxiv.org/abs/2503.04761>
38. Alex Tamkin and Peter McCrory, “Estimating AI Productivity Gains from Claude Conversations,” Anthropic, November 2025, p. 13, <https://www.anthropic.com/research/estimating-productivity-gains>