Level 4: Results

AI Literacy for the Modern Workforce · Kirkpatrick Evaluation Framework

Within the Kirkpatrick model, 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.

It is 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 structural equivalent of estimating a gym's health benefit by counting badge swipes at the door, since the turnstile 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: savings range from 20% to 95% by task (Tamkin & McCrory, 2025), and value depends on whether an interaction was augmentation or automation and how it was structured (Handa et al., 2025). The cost of ignoring this arrives as an invoice: Uber's COO reported burning the company's entire annual AI budget in four months without a clean line to user value, and within days Fortune declared the metric dead. This framework measures outcomes downstream of competent use, and treats adoption as an input condition, not a result.

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 usage telemetry, not outcomes. ChatGPT Enterprise reports active users, message volumes, and top tools; Claude Enterprise reports conversation and message counts, projects, and seat utilization; Gemini for Workspace reports active users and per-app feature usage across Gmail and Docs. Every one of those is an adoption metric, the same volume signal this framework rejects as a results measure. The conversation content that could support a quality judgment sits behind a compliance or eDiscovery interface built for litigation, 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. This is why baseline establishment is a deployment prerequisite rather than an afterthought: the provider hands the organization an adoption dashboard, and the outcomes the framework actually needs are the organization's to instrument.

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.

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

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	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

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 (WEF Future of Jobs Report, 2025).

ISOLATING THE PROGRAM'S CONTRIBUTION

Attribution is the central methodological challenge, and the framework is honest that every technique produces an estimate, not a proof. It 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 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 is not invented for this project. It applies the Phillips ROI Methodology, the established framework that extends Kirkpatrick with a fifth level, financial ROI, and is the standard approach to costing training in corporate L&D (Training Industry). 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. What this project supplies is not a new formula but defensible inputs, each drawn from the research corpus rather than a vendor benchmark.

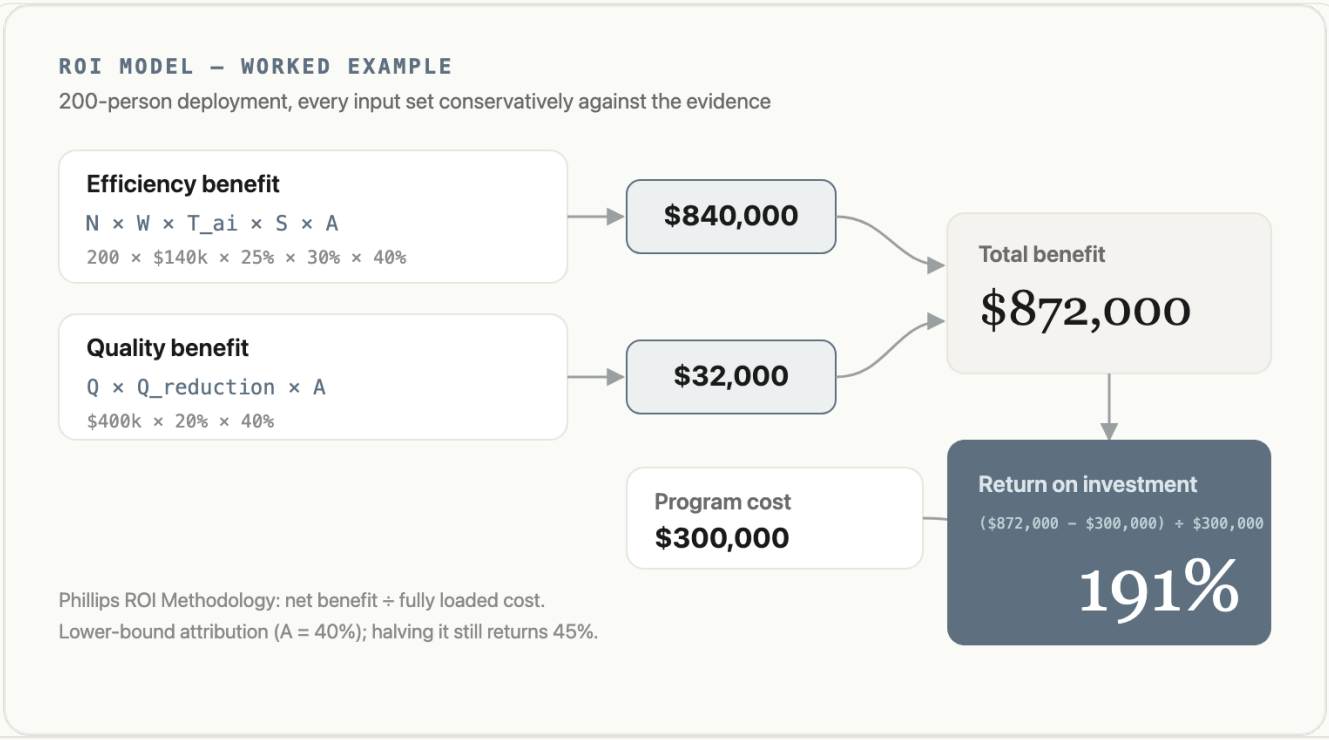
The model calculates two independently attributed benefit streams against total program cost:

Efficiency Benefit	=	`N × W × T_ai × S × A`
Quality Benefit	=	`Q × Q_reduction × A`
ROI	=	**`(Efficiency + Quality − C_program) / C_program × 100%`**

VARIABLE	MEANING	SOURCE
`N`	Participants	Deployment plan
`W`	Fully loaded annual compensation	HR/Finance
`T_ai`	Share of work hours on AI-eligible tasks	Task analysis / Handa et al.
`S`	Time savings per AI-eligible hour, competently used	Measurement / Tamkin & McCrory (14–56%)
`A`	Attribution factor	Isolation methodology
`C_program`	Total program cost	Budget
`Q`	Annual cost of AI-related quality failures	QA/incident extrapolation
`Q_reduction`	Reduction in those failures post-program	KPI Q1

The worked example runs a 200-person deployment at a mid-sized professional-services firm, with every input set conservatively against the documented evidence:

VARIABLE	VALUE	RATIONALE
N	200	Departmental rollout
W	$140,000	Mid-career base ~$100K × 1.4 fully loaded (benefits ~30% per BLS, plus payroll taxes and overhead)
T_ai	25%	Against 57% AI-touched (Handa et al.); 25% = practical, not merely possible
S	30%	Against an 81% in-conversation median; 30% acknowledges full-cycle verification overhead
A	40%	Lower-bound; assumes 60% of improvement is other factors
C_program	$300,000	Amortized development (~735 hrs, Chapman), delivery, ~6 hrs participant time across the course's 37 sections, admin, the full four-level evaluation, and contingency; anchored to ATD's 2024 L&D cost benchmarks
Q	$400,000	Annual AI-related quality-failure cost across 200 employees
Q_reduction	20%	Conservative one-fifth reduction



ROI model worked example · Phillips ROI Methodology · figures match the markdown source

Efficiency Benefit = 200 × $140,000 × 0.25 × 0.30 × 0.40 = **$840,000**. Quality Benefit = $400,000 × 0.20 × 0.40 = **$32,000**. Total $872,000 against $300,000 yields a **191% ROI**. The model is most sensitive to attribution, time savings, and AI-eligible task share; halving attribution still returns a positive 45%. The figure is not a claim about this program's realized return. It demonstrates that results measurement was built into the architecture from the start, ready for an organization to repopulate with its own baselines.

TIMELINE, AND THE FULL CHAIN

MILESTONE	ACTIVITY
Pre-program	Establish KPI baselines (≥3 months historical) 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. That is why the framework was designed as an integrated system from the outset, not four instruments bolted together after the fact.[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.*

REFERENCES

Kirkpatrick Partners — *The Kirkpatrick Model*

Phillips, J. J. / ROI Institute — *The ROI Methodology* (the five-level model extending Kirkpatrick to financial ROI; the formula, isolation step, and conservative Guiding Principles this model follows)

Fortune — *Uber's AI budget burn* and *Tokenmaxxing is over* (May 2026)

Tamkin, A. & McCrory, P. — *Estimating AI Productivity Gains from Claude Conversations* (Nov 2025)

Handa, K. et al. / Anthropic — *Anthropic Economic Index* (augmentation-automation split)

Anthropic Interviewer — *What 1,250 Professionals Told Us About Working with AI* (Dec 2025) (69% concealment baseline)

World Economic Forum — *Future of Jobs Report 2025* (Jan 2025)

Association for Talent Development — *2024 State of the Industry* (per-employee and per-learning-hour L&D cost benchmarks); the worked example also uses the standard 1.25–1.4× fully-loaded employee-cost multiplier and the Chapman Alliance content-development-hours benchmark

OpenAI — *ChatGPT Enterprise workspace analytics*; **Anthropic** — *Claude Enterprise usage analytics*; **Google** — *Gemini for Workspace usage reports*