

Level 3: Behavior

AI Literacy for the Modern Workforce

Within the Kirkpatrick model¹, Level 3 measures whether participants apply what they learned to their on-the-job behavior. It is the tier that distinguishes knowing from doing, and therefore the tier that determines whether the program produced transfer. It is also the tier most AI training skips. The rest of this document is the instrument itself: the behavioral indicators, the two-instrument design, the rating scale, the interval schedule, and the analysis. The design measures what a participant does with AI, not how much they use it, sampled at 30, 60, and 90 days.

The transfer problem this level exists to catch

The L&D literature documents a persistent gap between learning and application. Philippa Hardman's synthesis of the transfer research estimates that only 10 to 20% of learning investment produces measurable on-the-job behavior change, against a global L&D spend in the hundreds of billions of dollars annually².¹ Peer-reviewed work questions that number but not the gap. Ford, Yelon and Billington argue that the often-quoted figure of 10% transfer has no empirical basis and call it a delusion³. What training professionals report instead is decay. Those at 150 organizations estimated that 62% of employees apply what they learned immediately after training, 44% after six months, and 34% after a year⁴. In a later survey, organizations that evaluated training more often reported higher transfer, and of Kirkpatrick's four levels only evaluation of behavior and results was linked to it⁵. An evaluation that stops at Level 2 cannot see this gap at all; it certifies that knowledge was acquired and assumes the rest. This instrument is built to both measure the gap and create the conditions to close it: it adds indicators at 60 and 90 days that target the specific barriers (peer influence, manager reinforcement, organizational environment) where transfer tends to stall.

Why usage volume is not behavioral change

Through early 2026, a growing number of organizations measured AI adoption through usage volume: token consumption, login frequency, leaderboards ranking employees by how much they used a tool. The approach conflates activity with competency. OpenAI's own researchers make a similar point. Their 2026 enterprise study cautions that message volume is a measure of usage intensity, not a complete measure of economic importance⁶. It adds that its task results, drawn from classified message content, do not measure downstream work products or productivity effects⁷.

Volume measurement also carries a price. By mid-2026, one in five respondents reported that their organization was limiting AI use because of operating costs, tokens included. CFOs had begun speaking of "tokenomics" because token volumes for reasoning-heavy work grew faster than per-token prices fell⁸. A volume metric therefore rewards the behavior that is both riskiest and most expensive. A participant who generates thousands of tokens through single-turn, underspecified prompts and accepts first-generation output without verification registers as a high-usage employee on every volume metric, while demonstrating precisely the behaviors this program identifies as problematic: poor delegation, poor description, absent discernment, absent diligence.

The research the program is built on shows why volume cannot proxy value. Most tasks show estimated in-conversation savings between 50% and 95%, and named examples run from about 20% for checking diagnostic images to about 95% for compiling information from reports⁹. Of sampled Claude.ai conversations, 57% are augmentation and 43% are automation¹⁰.

The value of an interaction therefore depends entirely on what was delegated and how it was structured. The most competent user often shows lower consumption, because they decompose the task, prompt well enough to need fewer iterations, and route unreliable work back to their own hands.

One caution on the 57%: it is a share of conversations. It is not the 57% of US work hours that McKinsey Global Institute calls technically automatable, a feasibility figure rather than a usage one¹¹.

Token volume is also an unstable measure. When per-token prices fall, consumption can rise. In the OpenRouter data, as cost per token dropped for the low-priced, capable models the authors call efficient giants, those models got integrated everywhere and total consumption soared, with longer contexts and more iterations. The authors present this as some evidence of a Jevons effect and do not test it formally¹². Analyst Benedict Evans puts inference efficiency gains at 50 to 100 times a year¹³. A team whose token count doubled may have done twice the work. It may also have done the same work at a lower price per token with more retries. Volume cannot tell the difference.

The market reached the same conclusion. By late May 2026, Fortune declared "tokenmaxxing" over as a measure of AI return¹⁴, after Uber's COO reported burning the company's entire annual AI budget in four months without a clean line to user value¹⁵ and Amazon deprecated an internal usage leaderboard (a senior vice president telling staff to "please don't use AI just for the sake of using AI")^{16,17}. At Meta, an employee-built leaderboard that ranked staff by token use came down two days after the press reported it; Meta said the employee removed it without a company request¹⁸. These episodes confirm a design principle that predated them: an organization that incentivizes volume rewards the highest-risk behavior, uncritical high-volume delegation with no verification layer.

Two instruments, anchored to evidence

A traditional observation checklist assumes a co-located manager who can watch the work. That assumption fails for distributed knowledge work, where structured prompting, citation verification, and task decomposition all happen invisibly at a workstation. The instrument addresses this with an evidence-informed parallel design:

- **Manager evidence review**, completed by the direct manager at each interval, rates each indicator from three sources: review of work artifacts (deliverables, documentation, shared prompt logs), a structured 15-to-20-minute check-in conversation, and any direct observation available. The manager is not assumed to have witnessed each behavior firsthand.
- **Participant self-assessment**, completed at each interval, rates the same indicators, but each rating must carry a specific recent example (situation, behavior, outcome). Fabricating a convincing, detailed example is harder than checking a frequency box, which is the primary control against self-report inflation. The design borrows from the one peer-reviewed survey of AI-assisted critical thinking in the corpus, which asked for observable actions rather than the mental state, on the ground that critical thinking 'as a pure mental phenomenon is difficult for people to self-observe,' and paired each yes/no with a free-text justification¹⁹. The same study names the residual risk: participants sometimes conflated a task feeling easier with having thought about it less, often when the output satisfied them²⁰. The manager evidence review exists to catch that conflation.

Where the two agree, the change is well-evidenced. Where they diverge, the gap is the most useful signal in the instrument.

The rating scale

Both instruments use a four-point frequency scale with no neutral midpoint, which forces a directional call between "occasionally" and "consistently." The scale tops out at modeling for others, because peer modeling is how one person's competence becomes a team's.

Rating	Label	Definition
1	Not yet observed	Not seen since completion. Not a negative judgment: may indicate insufficient opportunity, organizational barriers, or that the behavior needs more time.
2	Occasionally	Seen in some situations but not consistent; demonstrated when prompted or when conditions are favorable.
3	Consistently	Observed regularly across relevant work, applied as a standard part of the workflow without prompting.
4	Modeling for others	Applies it consistently and actively demonstrates it to colleagues, explains the reasoning, or coaches. Expected mainly at 60 and 90 days.

The twelve behavioral indicators

The indicators map to [the program's 4D competency framework](#) (Delegation, Description, Discernment, Diligence) and to the program's action map. Four of the core eight track a named sub-category in the source framework: B1 to Goal and Task Awareness, B2 to Platform Awareness, B7 to Transparency Diligence, and B8 to Deployment Diligence²¹. Anthropic's AI Fluency: Framework & Foundations course (2025), which teaches the framework, uses the same names except for the first, which it calls Problem Awareness, the name the program's action map uses²².

Each indicator is phrased as a concrete, observable action, never a mindset: "verifies factual claims against independent sources before incorporating them" is measurable; "values accuracy" is not. Each is also classified by evidence type, which sets how much weight the manager review can carry. **Artifact** evidence is a reviewable work product and carries high confidence; **conversation** evidence is elicited through the check-in and carries moderate confidence; **self-report** evidence is visible mainly to the participant and leans on the specificity requirement. The eight core indicators run at all three intervals; transfer indicators are added at 60 days, organizational-influence indicators at 90.

#	Competency	Indicator	Evidence
B1	Delegation	Before using AI for a task, defines which components need human judgment and which are appropriate for AI.	Conversation
B2	Delegation	Identifies when a task falls outside the model's reliable range and adjusts (less delegation, more verification, or manual work).	Self-report
B3	Description	Provides structured context, format, and quality criteria when prompting, rather than single-turn underspecified requests.	Artifact
B4	Description	Iterates on outputs through multi-turn refinement rather than accepting or discarding first-generation results.	Artifact

#	Competency	Indicator	Evidence
B5	Discernment	Verifies factual claims, citations, and statistics against independent sources before incorporating them.	Artifact
B6	Discernment	Evaluates the reasoning in an output, not just the product, checking for gaps, unsupported assumptions, and circular logic.	Conversation
B7	Diligence	Documents AI's role in a deliverable, distinguishing AI-generated, AI-assisted, and human-authored components.	Artifact
B8	Diligence	Takes full accountability for accuracy before sharing, applying the review standard used for a junior colleague's work.	Self-report
B9	Transfer (60d)	Discusses AI practices openly with colleagues, using the 4D vocabulary to describe how they work.	Conversation
B10	Transfer (60d)	Has changed the AI approach on at least one regular task based on the program (more involvement where apt, less where risky).	Conversation
B11	Influence (90d)	Has helped a colleague improve their AI practices, through coaching, sharing materials, or demonstrating techniques.	Conversation
B12	Influence (90d)	Has raised a concern about output quality, data handling, or delegation appropriateness in a team setting.	Conversation

B1 to B10 are the workplace forms of action-map behaviors: B1 observes D1, B2 observes D3 and D8, B3 observes D4 and D5, B4 observes D6, B5 observes D7, B6 observes D9, B7 observes D10, B8 observes D11, and B9 observes D12. D2, the call between augmenting a task and keeping it human, shows up in B1 and B10. B11 and B12 have no counterpart in the action map. They measure influence on the team, which the business goal's team-level gains depend on.

Three of the twelve (B2, B8, and partly B1) are self-report by nature, because they describe cognitive processes that leave no artifact; the specificity requirement is their primary control. Grading confidence by evidence type keeps that limit visible: organizations should weight artifact-anchored claims heavily and temper self-report claims accordingly, rather than treating all twelve as equally observable.

Progressive intervals and convergence

Interval	Indicators	Focus
30 days	B1–B8 (core)	Individual baseline: are the most actionable practices taking hold?
60 days	B1–B10	Transfer: is the participant adapting their workflow and discussing practices openly?
90 days	B1–B12	Influence: is the change propagating to peers and team practice?

The core indicators produce a three-point trend line. The expected pattern is progressive adoption. An indicator that rises from 30 to 60 days but falls by 90 signals a behavior adopted and then crowded out, a different problem from one never adopted. The 90-day interval is therefore a sustainability check rather than a delayed re-measurement.

The 60-day open-discussion indicator targets the concealment dynamic directly: the Anthropic Interviewer study found 69% of professionals conceal their AI use at work²³, and a program that builds private skill while ignoring the social conditions for visible use produces no change an organization can measure. The 2026 State of AI survey shows the same gap at scale, though it does not measure its social causes. Eighty percent of respondents report individual productivity gains, while 37% report any positive earnings effect at their organization. Its authors conclude that the organizations likely to convert one into the other are those that transform their businesses, not just adopt AI tools²⁴.

The parallel design produces a manager rating and a self rating for each indicator, and the four agreement patterns each carry distinct meaning. Both high is the target. Manager-high/self-low usually means the participant has internalized the practice to the point it feels unremarkable. Manager-low/self-high is the most actionable, because each of its possible causes points to a different fix: the behavior may be happening invisibly, a legibility problem; it may be overstated, a self-report problem; or the manager may not recognize it, an orientation problem. Both low, cross-referenced against the Level 2 data, separates a transfer failure (knowledge present, behavior absent) from an upstream knowledge gap.

Limitations, and alternatives

The instrument has more than one weak point. The most obvious is the manager: it only works if managers spend real attention on it, reading the artifacts and the written examples rather than speed-running the rating scale, and a manager who treats it as one more compliance checkbox produces confident noise. The participant side is just as fragile. Self-assessment depends on honest reporting, and the 69% concealment finding cuts straight against it: where AI use carries social stigma, people under-report what they actually do, and a well-meaning L&D team reading the aggregate may have no idea how much pressure sits on the ground beneath the numbers. Anthropic's own internal study reached the same conclusion from the other direction and recommends reading self-reports only alongside usage data²⁵. The data can look clean while the culture it came from is not.

The instrument also fails without visible warning when it is added to already-full days. If the manager review and the self-assessment become one more task on a manager's or an employee's plate, with no protected time to reflect and engage honestly, both default to whatever rating clears the form fastest. Two further risks compound these: evaluation apprehension, where a participant shapes a self-report to look good to the person who rates them, especially if the check-in reads as appraisal rather than development; and uneven evidence, since the artifact-anchored indicators assume an organization where deliverables and prompt logs are reviewable, which is not every organization.

None of this is fatal, but it does mean the framework should not be deployed naively. Some mitigation is structural: protect real time for the check-in, frame it as development rather than appraisal, and lean on more than one measurement surface. To that end the framework offers alternatives a deploying organization can layer in or substitute: sampled artifact audits, where an L&D reviewer or quality lead scores a random sample of AI-assisted deliverables against the same indicators; peer review, where colleagues close enough to the work can see it; and self-assessment with spot-checks, where specific examples are verified against real artifacts rather than taken on faith. No single surface is bias-free, but each fails differently, and the more independent the surfaces, the less any one of them decides the outcome.²

¹ The transfer estimates and the global L&D-spend figure are widely cited but should be read as directional, since both depend on how "behavior change" and "L&D spend" are defined across organizations. The Saks estimates come from training professionals' reports, not direct measurement. Saks and Belcourt's study is twenty years old, Ford, Yelon and Billington's fifteen, and Saks and Burke's fourteen, so read them as the shape of the problem, not a rate to plan against.

² This is an evaluation instrument design, not a results report. No cohort has run these instruments; the program is an independent portfolio project not yet deployed at organizational scale. The indicators, rating scale, evidence-type classifications, interval design, and convergence protocol are complete and ready for a deploying organization to administer through its own HR or performance-management infrastructure. Attribution also remains inherently limited: behavioral change at 30, 60, or 90 days cannot be definitively separated from concurrent factors such as policy changes or independent learning.

Notes

1. Kirkpatrick Partners, “The Kirkpatrick Model,” accessed October 2026, <https://www.kirkpatrickpartners.com/the-kirkpatrick-model/>
2. Philippa Hardman, “The Course Is Dying as the Unit of Learning,” Substack, April 2026, <https://drphilippahardman.substack.com/p/the-course-is-dying-as-the-unit-of>
3. J. Kevin Ford, Stephen L. Yelon, and Abigail Q. Billington, “How Much Is Transferred from Training to the Job? The 10% Delusion as a Catalyst for Thinking about Transfer,” *Performance Improvement Quarterly*, 24(2), 7–24, 2011, <https://doi.org/10.1002/piq.20108>
4. Alan M. Saks and Monica Belcourt, “An Investigation of Training Activities and Transfer of Training in Organizations,” *Human Resource Management*, 45(4), 629–648, 2006, <https://doi.org/10.1002/hrm.20135>
5. Alan M. Saks and Lisa A. Burke, “An Investigation into the Relationship between Training Evaluation and the Transfer of Training,” *International Journal of Training and Development*, 16(2), 118–127, 2012, <https://doi.org/10.1111/j.1468-2419.2011.00397.x>
6. Aaron Chatterji et al., “How Organizations Use AI: Evidence from ChatGPT,” OpenAI working paper, August 2026, p. 18, <https://arxiv.org/abs/2608.12236>
7. Aaron Chatterji et al., “How Organizations Use AI: Evidence from ChatGPT,” OpenAI working paper, August 2026, p. 23, <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. 15, 17, <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>
9. Alex Tamkin and Peter McCrory, “Estimating AI Productivity Gains from Claude Conversations,” Anthropic, November 2025, <https://www.anthropic.com/research/estimating-productivity-gains>
10. Kunal Handa et al., “Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations,” Anthropic, February 2025, p. 3, <https://arxiv.org/abs/2503.04761>
11. Lareina Yee et al., “Agents, robots, and us: Skill partnerships in the age of AI,” McKinsey Global Institute, November 2025, p. 4, <https://www.mckinsey.com/mgi/our-research/agents-robots-and-us-skill-partnerships-in-the-age-of-ai>
12. Malika Aubakirova et al., “OpenRouter State of AI: An Empirical 100 Trillion Token Study with OpenRouter,” OpenRouter and Andreessen Horowitz, December 2025, pp. 48, 53–54, <https://openrouter.ai/state-of-ai>
13. Benedict Evans, “AI eats the world (May 2026 edition),” May 2026, slide 20, <https://www.ben-evans.com/s/2026-Spring-AI.pdf>
14. 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/>
15. 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/>
16. Financial Times, “Amazon shuts down its internal AI-use leaderboard,” *Financial Times* (confirmed by Business Insider), May 2026, <https://www.ft.com/content/b1a62a7f-6df5-4c90-94ce-64ce9c9961b6>
17. Brent D. Griffiths, “Amazon says it shut down a token leaderboard: ‘Don’t use AI just to use AI,’” *Business Insider*, May 2026, <https://www.businessinsider.com/amazon-ai-leaderboard-tokenmaxxing-2026-5>
18. Jacqueline Munis, “A Meta employee created a dashboard so coworkers can compete to be the company’s No. 1 AI token user—and Zuckerberg doesn’t even rank in the top 250,” *Fortune*, April 2026, <https://fortune.com/2026/04/09/meta-killed-employee-ai-token-dashboard/>
19. Hao-Ping (Hank) Lee et al., “The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers,” *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 2, 5, <https://doi.org/10.1145/3706598.3713778>
20. Hao-Ping (Hank) Lee et al., “The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers,” *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, p. 15 (participants conflated task effort with critical-thinking effort; self-confidence may be miscalibrated), <https://doi.org/10.1145/3706598.3713778>
21. Rick Dakan and Joseph Feller, “Framework for AI Fluency: Practical Overview Document (Version 1.1),” Ringling College of Art and Design; University College Cork, January 2025, pp. 4, 6, <https://ringling.libguides.com/ai/framework>
22. Rick Dakan, Joseph Feller, and Anthropic, “AI Fluency: Framework & Foundations (course),” Anthropic Academy, 2025, <https://anthropic.skilljar.com/ai-fluency-framework-foundations>
23. 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>

24. 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. 14, 29, <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>
25. Saffron Huang et al., “How AI Is Transforming Work at Anthropic,” Anthropic, December 2025, research post, “Usage and productivity”: METR study, experienced developers on familiar codebases overestimated their AI productivity boost; Appendix, “Limitations”: selection, social desirability, recency bias; read alongside usage data, <https://anthropic.com/research/how-ai-is-transforming-work-at-anthropic/>

© Michael Ritchot. Code licensed under the MIT License. Instructional content licensed under CC BY-NC-SA 4.0.