

Level 2: Learning

AI Literacy for the Modern Workforce

Within the Kirkpatrick model, Level 2 measures whether participants acquired the intended knowledge, as distinct from whether they liked the program (Level 1), changed their behavior (Level 3), or produced organizational results (Level 4)¹. It is the only tier of this framework with empirical evidence behind it, and that evidence is qualified, so this document separates what the data supports from what it does not. The instrument is a twelve-item, scenario-based pre/post assessment embedded in the program as a mandatory gate in every platform mode, with no skip option, because without paired pre/post data the level cannot function.

A parallel-form design that resists gaming

The pre and post assessments measure the same twelve constructs through different workplace scenarios. This parallel-form construction controls the primary threat to a pre/post design, test-retest memory: a participant cannot transfer a recalled answer when the scenario, the framing, and the distractors have all changed. The scenario format also defeats the reverse-engineering a motivated test-taker uses on recall quizzes. A stem that asks what you should do in a specific, unfamiliar situation, with distractors drawn from documented misconceptions, is far harder to game than a definitional item, because the test-taker has to recognize the construct rather than pattern-match the phrasing. The distractor rule corrects a defect found in the capstone pilot. Several pilot participants reported that the correct answers were 'too obvious', and the redesign called for more plausible distractors without added cognitive load².

As implemented, the correct answer letter shifts on ten of the twelve construct pairs; only tokenization and output verification keep their position, which limits position-based recall. Both forms share one answer-key distribution (A3/B3/C3/D3), so a fixed guessing strategy scores the same on each. Option length is balanced too: the correct answer is the longest option on two of the twelve items on each form, fewer than the three that chance would give, so picking the longest answer earns no advantage. An automated item check flags any drift in either property, and any post stem that names the concept it tests.

The vocabulary is staged across the two forms. The pre-assessment uses no program terminology, since learners have not yet encountered it. The post-assessment options use the program's language (tokenization, next-token prediction, the context window, the augmentation-automation spectrum), but no post stem names the concept it tests, and several distractors use the same vocabulary wrongly. As on the pre form, a participant has to recognize which mechanism the situation involves rather than match a term in the stem to a term in an option. Transfer to real work depends on that recognition step.

Items are scored binary, with consequence-based feedback on each option (the realistic workplace result of that reasoning) rather than a bare "correct/incorrect," which turns the post-assessment's final screen into a last learning moment.

What the twelve constructs cover

The items span four blocks that follow the instructional progression, and every item traces backward to a documented capability gap and forward to one of the program's seventeen performance objectives.

Block	Items	Domain	Coverage
Usage Patterns	2	How professionals use AI vs. how they report using it; productivity-risk tradeoffs	Modules 1–2
Failure Modes	3	Fluent fabrication, boundary-task failure, context-window limits	Modules 2–3
Mechanics	3	Prediction vs. retrieval, tokenization, architectural capability diagnosis	Module 3
Evaluation	4	Output-verification priority, structured prompting (Description), task decomposition (Delegation), disclosure and accountability (Diligence)	Module 4

#	Construct	Block	Measurement focus
1	Augmentation vs. automation	Usage	Recognizes the gap between self-reported collaborative use and observed single-turn directive use
2	Productivity-verification gap	Usage	Identifies that high time savings and high verification burden coexist
3	Fluent fabrication	Failure	Understands that specific, confident citations can be generated rather than retrieved
4	Boundary task failure	Failure	Knows precise enumeration and character-level tasks fail because models generate plausible answers
5	Context window limits	Failure	Recognizes that material in the middle of a long input is the likeliest to be missed, even when the whole input fits
6	Prediction vs. retrieval	Mechanics	Understands that output is statistically generated text, not retrieved from identified sources
7	Tokenization	Mechanics	Understands that sub-word tokenization makes character-level and non-Latin tasks fail systematically
8	Capability diagnosis	Mechanics	Attributes specific failures to specific architectural properties rather than treating errors as random
9	Verification priority	Evaluation	Identifies which claims in a deliverable carry the highest fabrication risk and consequence
10	Structured prompting	Evaluation	Knows that structure (context, format, constraints) drives output quality, not tone or effort
11	Task decomposition	Evaluation	Splits a deliverable into parts to delegate, draft with AI, or keep, by how checkable each part is and what context it needs
12	Disclosure and accountability	Evaluation	Writes an AI-use note that says which parts AI produced, what was checked, and who answers for the result

That matrix is the traceability artifact: no assessment item exists without a path from a documented gap to a measurable objective. Constructs 11 and 12 were added in the second version of the instrument so that each of the four 4D competencies has at least one construct measured before and after the program; the first version had no item on task decomposition or disclosure. Every response records the version it was given against, so data from revised items is never pooled with data from the items they replaced.

Scoring and thresholds

Scoring is binary per item, unweighted across blocks, with learning gain as the simple post-minus-pre delta. Because a short form has a ceiling, the results screen also reports normalized gain, the share of available improvement achieved: $(\text{post} - \text{pre}) \div (\text{items} - \text{pre})^3$. A learner who moves from 8 to 10 and one who moves from 4 to 6 both gain +2, but the first closed half of the remaining gap and the second a quarter of it. Labels read against the pre score: a zero change from a pre score of 80% or more reads as a strong baseline held, while a zero change from a lower pre score points the learner to the constructs still missed. A participant whose baseline was taken on the earlier ten-item form is compared on the ten constructs answered both times.

Beyond the aggregate, the analytics dashboard exposes a per-construct breakdown, so a participant who improved overall but missed verification priority on both forms is flagged for targeted follow-up rather than averaged into a single number.

At twelve items, a single-item swing moves the score about eight points, so the thresholds are interpretive ranges, stated as percentages, rather than pass/fail cutoffs:

Post score	Interpretation	Action
80% or more (10–12)	Strong	Solid understanding across all four domains. Ready for application; proceed to Level 3 monitoring.
60–79% (8–9)	Adequate	Core concepts acquired, specific gaps remain. Review per-construct analysis; targeted follow-up.
30–59% (4–7)	Developing	May have engaged superficially or need support. Check whether gaps are concentrated or distributed.
Below 30% (0–3)	Minimal	At or below chance (3 of 12). Verify engagement before reading as a learning outcome; the dashboard flags answers given in under ten seconds, a fixed-threshold version of the response-time screen for rapid guessing ⁴ .

Delta	Label	Reading
+4 or more	Strong growth	Substantial gain; confirm it is distributed across blocks, not concentrated in one.
+2 to +3	Moderate growth	Meaningful improvement; the band expected for full engagement, to be confirmed with the first cohort.
+1	Incremental	Modest; may reflect a high baseline. Read against the pre score (+1 from 4 differs from +1 from 8).
0	Held or stalled	No numerical change. From a pre score of 80% or more, a strong baseline held; from a lower pre score, review the constructs missed on both forms.

An end-of-program instrument this short is complemented by an in-program layer: sixteen scenario-based knowledge checks (four per module) and three interpretation checks in Module 1's data narrative. These add temporal resolution the pre/post delta cannot reach: they show *when* a concept was acquired, not only whether.

Validity, and the capstone behind it

Content validity rests on traceability: each construct derives from a specific finding in the research corpus, including the augmentation-automation split⁵, the productivity-verification gap⁶, the mechanics of fluent fabrication and tokenization^{7,8}, and consequence-driven verification (the potential for negative outcomes was the most common reason workers gave for scrutinizing AI output, 116 of 319)⁹. Face validity comes from scenarios built around recognizable roles (project manager, legal analyst, HR specialist, compliance officer, sales lead) with no abstract or academic framing.

Parallel forms control one threat, test-retest memory; they are not evidence that the items measure their constructs. That evidence is planned in three steps. Before deployment, three to five learners think aloud through the post form, which shows whether they reach the key by reasoning about the construct or by matching wording. After the first cohort, each item gets a difficulty index, a discrimination index (its correlation with the rest of the form), and a count of how often each distractor is chosen; an item with discrimination below about 0.2, or a distractor almost no one picks, is revised. Post scores on verification priority are also compared with P10 triage accuracy, which should agree if both measure the same capability.

The empirical anchor is my M.Ed. capstone (Ritchot, Western Governors University, June 2025), which tested the mechanics-first AI literacy instruction this program is built on in a small pilot: one module, ten participants, and a mixed-methods pre/post design¹⁰. Pre-assessment scores averaged 3.1 out of 5 and post-assessment scores averaged 4.5¹¹. Eight of ten participants showed positive gains; one held flat at the scale ceiling and one declined by a point¹². Triangulation found the same recognition of token-based limits in discussions, reflections, and a scenario task, which the capstone reads as evidence the knowledge was retained and applied¹³.

The capstone names its own limits: ten volunteers, a single iteration, and a three-week asynchronous window¹⁴. It is evidence for the instructional design, not for this specific twelve-item corporate instrument, and it came from a different population than the mid-career professionals this program targets, so it counts as suggestive support rather than proof.

The corporate instrument has no reliability statistics yet, because those statistics require a first deployment cohort it has not had. The privacy-by-design choice keeps every response on the participant's own device, so cohort data reaches the organization only through the participant: the results screen gives each participant a results code (item responses, instrument version, and response times, with no name) to paste into the organization's own Level 1 survey under a coded identifier the organization assigns.

Two validated scales used in Lee et al.'s survey are candidates for the validity work that can run alongside the reliability statistics: Kember et al.'s Reflective Thinking Inventory, which predicted critical-thinking enactment ($\beta = 0.52$), and a six-item Propensity to Trust Technology scale adapted to generative AI, which did not¹⁵. Administering both with the pre-assessment would test whether the twelve items measure something beyond a general disposition to reflect. The instrument is a second iteration on a piloted foundation, not a first-generation prototype, and it is not yet proven.

A worked example

The per-construct view is the design's primary advantage over an aggregate pass/fail. A sample learner scores 8/12 before Module 1 and 11/12 after Module 4, a delta of +3 that closes three-quarters of the remaining gap. The analytics dashboard's completed demo learner carries this record:

#	Construct	Pre	Post	Change
1	Augmentation vs. automation	Correct	Correct	Maintained
2	Productivity-verification	Incorrect	Correct	Improved (key gap addressed)
3	Fluent fabrication	Correct	Correct	Maintained
4	Boundary task failure	Incorrect	Correct	Improved (key gap addressed)
5	Context window limits	Correct	Correct	Maintained
6	Prediction vs. retrieval	Incorrect	Correct	Critical misconception corrected
7	Tokenization	Correct	Correct	Maintained
8	Capability diagnosis	Correct	Correct	Maintained
9	Verification priority	Correct	Correct	Maintained
10	Structured prompting	Incorrect	Incorrect	Persistent gap
11	Task decomposition	Correct	Correct	Maintained
12	Disclosure and accountability	Correct	Correct	Maintained

The aggregate (+3, moderate growth) hides the useful detail: three specific gaps closed, and structured prompting persisted. That persistent gap routes the learner back to Module 4's Description material rather than declaring success on the headline number.¹

¹ The instructional design has prior empirical support from the capstone pilot; the specific twelve-item instrument awaits its first deployment cohort for psychometric reliability data. A single-cohort pre/post design without a control group also cannot fully isolate program effects from maturation or concurrent learning, a standard limitation in corporate L&D. The parallel-form design mitigates the largest confound (test-retest memory) but does not eliminate every threat to internal validity. At twelve items the instrument trades precision for completion; expansion to fifteen is a viable iteration if deployment data shows ceiling effects.

Notes

1. Kirkpatrick Partners, “The Kirkpatrick Model,” accessed October 2026, <https://www.kirkpatrickpartners.com/the-kirkpatrick-model/>
2. Michael Ritchot, “Token-Level Abstraction in Large Language Models: A Design-Based Inquiry into AI Comprehension Pathways,” master’s capstone, Western Governors University, June 2025, p. 41, <https://ritchot.me/docs/capstone.pdf>

3. Richard R. Hake, “Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses,” *American Journal of Physics*, 66(1), 64–74, January 1998, <https://doi.org/10.1119/1.18809>
4. Steven L. Wise and Xiaojing Kong, “Response Time Effort: A New Measure of Examinee Motivation in Computer-Based Tests,” *Applied Measurement in Education*, 18(2), 163–183, 2005, https://doi.org/10.1207/s15324818ame1802_2
5. 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>
6. Alex Tamkin and Peter McCrory, “Estimating AI Productivity Gains from Claude Conversations,” Anthropic, November 2025, <https://www.anthropic.com/research/estimating-productivity-gains>
7. Andrej Karpathy, “Deep Dive into LLMs like ChatGPT,” YouTube, February 2025, <https://www.youtube.com/watch?v=7xTGNNLPyMI>
8. Anthropic, “AI Capabilities and Limitations (course),” Anthropic Academy, 2026, <https://anthropic.skilljar.com/ai-capabilities-and-limitations>
9. 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. 10, <https://doi.org/10.1145/3706598.3713778>
10. Michael Ritchot, “Token-Level Abstraction in Large Language Models: A Design-Based Inquiry into AI Comprehension Pathways,” master’s capstone, Western Governors University, June 2025, pp. 30, 36, <https://ritchot.me/docs/capstone.pdf>
11. Michael Ritchot, “Token-Level Abstraction in Large Language Models: A Design-Based Inquiry into AI Comprehension Pathways,” master’s capstone, Western Governors University, June 2025, pp. 2, 38, <https://ritchot.me/docs/capstone.pdf>
12. Michael Ritchot, “Token-Level Abstraction in Large Language Models: A Design-Based Inquiry into AI Comprehension Pathways,” master’s capstone, Western Governors University, June 2025, table, p. 38, <https://ritchot.me/docs/capstone.pdf>
13. Michael Ritchot, “Token-Level Abstraction in Large Language Models: A Design-Based Inquiry into AI Comprehension Pathways,” master’s capstone, Western Governors University, June 2025, p. 39, <https://ritchot.me/docs/capstone.pdf>
14. Michael Ritchot, “Token-Level Abstraction in Large Language Models: A Design-Based Inquiry into AI Comprehension Pathways,” master’s capstone, Western Governors University, June 2025, p. 44, <https://ritchot.me/docs/capstone.pdf>
15. 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. 5, 9, <https://doi.org/10.1145/3706598.3713778>

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