Level 2: Learning

AI Literacy for the Modern Workforce · Kirkpatrick Evaluation Framework

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).

It is the only tier of this framework with empirical evidence behind it, and that evidence is qualified, so this document is careful to separate what the data supports from what it does not. The instrument is a ten-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 ten 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 you have to recognize the construct rather than pattern-match the phrasing.

As implemented, the correct answer letter shifts on **eight of the ten** construct pairs; only tokenization and output verification keep their position, which limits position-based recall. The two forms do **not** share an identical answer-key distribution (the pre form runs A2/B3/C2/D3, the post form A3/B3/C2/D2), so a fixed guessing strategy is not perfectly neutral across them.

The vocabulary is staged across the two forms. The pre-assessment uses no course terminology, since learners have not yet encountered it; the post-assessment leans on the program's language (tokenization, next-token prediction, the context window, the augmentation-automation spectrum), so reasoning fluently in that vocabulary is itself part of what the post measures. 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 TEN 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	2	Output-verification priority, structured prompting (the Description competency)	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 long inputs produce systematic omissions as attention degrades over distance
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

That matrix is the traceability artifact: no assessment item exists without a path from a documented gap to a measurable objective.

SCORING AND THRESHOLDS

Scoring is binary per item, unweighted across blocks, with learning gain as the simple post-minus-pre delta. The interface frames outcomes positively: with a ten-item instrument a stable score most often reflects a strong baseline reinforced, so a zero delta reads as "solid foundation," not a failure. Beyond the aggregate, the admin view 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 ten items, a single-item swing moves the score ten points, so the thresholds are interpretive ranges rather than pass/fail cutoffs:

POST SCORE	INTERPRETATION	ACTION
8-10	**Strong**	Solid understanding across all four domains. Ready for application; proceed to Level 3 monitoring.
6-7	**Adequate**	Core concepts acquired, specific gaps remain. Review per-construct analysis; targeted follow-up.
3-5	**Developing**	May have engaged superficially or need support. Check whether gaps are concentrated or distributed.

POST SCORE	INTERPRETATION	ACTION
0–2	**Minimal**	At or below chance (2.5/10). Verify engagement metrics before reading as a learning outcome.

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 most common band for full engagement.
+1	**Incremental**	Modest; may reflect a high baseline. Read against the pre score (+1 from 4 differs from +1 from 8).
0	**Solid foundation**	No numerical change; with ten items, often a strong baseline reinforced.

A summative instrument this short is complemented by a formative 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 split (Handa et al., 2025), the productivity-verification gap (Tamkin & McCrory, 2025), the mechanics of fluent fabrication and tokenization (Karpathy, 2025; Anthropic, 2026), and verification priority (Lee et al., CHI 2025). Construct validity is addressed by the parallel-form controls above. Face validity comes from scenarios built around recognizable roles (marketing director, project manager, legal analyst, data analyst) with no abstract or academic framing.

The empirical anchor is my M.Ed. capstone (Ritchot, Western Governors University, June 2025), which piloted the tokenization instruction this course is built on with ten participants in a mixed-methods pre/post design: pre-assessment scores averaged 3.1 out of 5, post-assessment scores averaged 4.5, and nine of ten participants showed positive gains, with triangulation across reflections and a scenario task confirming the knowledge transferred. It validates the instructional design, not this specific ten-item corporate instrument, and it came from a different population, secondary students rather than the mid-career professionals this program targets, so it carries as suggestive support rather than proof. The corporate instrument has no reliability statistics yet, partly because the privacy-by-design choice keeps every response on the participant's own device, and partly because those statistics require a first deployment cohort it has not had. The honest position is that this is a second iteration on a validated foundation, not a first-generation prototype, which is a meaningfully different claim from "trust me, it works."

A WORKED EXAMPLE

The per-construct view is the design's primary advantage over an aggregate pass/fail. A sample learner scores 6/10 before Module 1 and 9/10 after Module 4, a delta of +3:

#	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

#	CONSTRUCT	PRE	POST	CHANGE
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

The aggregate (+3, moderate growth) hides the useful detail: three specific gaps closed and one (structured prompting) persisted, which 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 (the capstone pilot); the specific ten-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 ten items the instrument trades precision for completion; expansion to fifteen is a viable iteration if deployment data shows ceiling effects.*

REFERENCES

Kirkpatrick Partners — *The Kirkpatrick Model*

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

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

Lee, H. et al. — *The Impact of Generative AI on Critical Thinking*, CHI 2025

Karpathy, A. — *Deep Dive into LLMs like ChatGPT* (2025)

Anthropic — *AI Capabilities and Limitations* (2026)

Ritchot, M. — *M.Ed. Capstone*, Western Governors University (June 2025), prior pre/post validation of the tokenization instruction