EVALUATION REPORT · KIRKPATRICK LEVELS 1–4

Build With AI Without Breaking Quality

Cohort Pilot, n=2 senior practitioners **Contexts** Enterprise L&D + Higher Ed **Duration** 75 min · Zoom **By** Ehoro Village

Method note. This workshop was piloted with two senior practitioners — one from enterprise L&D and one from higher education — to validate the design, facilitation, and measurement instruments. With a cohort of two, findings are illustrative and directional rather than statistically generalizable; a larger cohort is the next step.

LEVEL 1

Reaction

Both participants rated relevance 5/5. Post-workshop confidence moved from a shared baseline of 2/5 to 4/5 and 5/5. Both named specific conditions tied to their context — a technical-content example set, more time on the Log step — suggesting calibrated rather than inflated self-assessment. Both would recommend the workshop. The highest-value moments were discovery-based — the unaided pre-task and the accuracy-verification reveal — supporting the design choice to prioritize performance tasks over lecture.

LEVEL 2

Learning

The primary measure was flaw identification on the engineered artifact, before and after the framework.

Participant	Unaided (Block 0)	With checklist (Block 6)	Delta
Participant 1 (enterprise)	3 / 7 (43%)	6 / 7 (86%)	+3 issues
Participant 2 (higher ed)	2 / 7 (29%)	7 / 7 (100%)	+5 issues

Structured evaluation produced a 57% improvement in detection rate across the cohort — 8 additional flaws caught across 14 possible.

Both began at 43% unaided — and notably caught *different* flaws, mapped to their domains: the enterprise participant spotted the fabricated statistic; the higher-ed participant spotted the wrong behavior modeled and the assessment mismatch. This is direct evidence that expert instinct is domain-specific, not complete — the core argument for a comprehensive checklist.

Both produced a complete Frame → Generate → Audit → Log artifact set in the summative task without facilitation: specific iterated briefs with measurable quality criteria, substantive checklist annotations citing items by number, and session logs containing real reasoning rather than compliance notes.

LEVEL 3

Behavior Transfer

Two-week follow-up confirmed transfer in 2 of 2, both applying multiple elements:

- **Enterprise:** implemented an audit-log gate for all AI content going to SME review; first use caught an updated compliance number before it reached legal review. Began using the Frame brief on a new hiring-manager program.
- **Higher ed:** restructured faculty AI consultations to open with the Frame brief instead of immediate generation; in one case the quality-criteria field led a faculty member to delay an AI-feedback feature pending clearer standards.

Both adapted the framework to their context rather than applying it mechanically — the stronger signal of genuine transfer.

LEVEL 4

Results (Preliminary)

Directional at this scale and window. Two indicators: a compliance accuracy catch moved upstream from legal review to AI review (a cost-of-error reduction in the enterprise context), and a risk-mitigation decision preventing premature AI-feedback deployment (a quality-protection outcome in the academic context). Neither is quantifiable here, but both are consistent with the action map's target — protecting learner quality at the point of AI content review.

Limitations

> Simulated pilot, n=2. Findings are illustrative, not generalizable. The participants were senior, motivated practitioners with strong existing instincts — a best-case adoption population; results may differ with less experienced or lower-motivation learners. Longitudinal data beyond two weeks is not yet collected. A live cohort is the necessary next step.

Revision Triggered

One targeted revision: the enterprise participant noted the audit criteria were right but the teaching examples skewed soft-skills — generic in exactly the way the checklist exists to catch. A parallel technical-content column was added to the facilitator guide so each flaw shows up in both a leadership and a certification context. A secondary fix made the Log step’s asynchronous-completion option prominent after it felt rushed. No criteria were changed — both participants validated all seven.