Exploring the learning engineering flower with agentic simulation

William Gray-Roncal[1] and Jodi Lis[2]

[1] Johns Hopkins University, School of Education; wgr@jhu.edu
[2] IEEE ICICLE; jodillis@gmail.com

Abstract. We use agentic simulation to examine how expertise contributes to the design and revision of small learning tools. Four dynamic design-team conversations crossed two fictional cases, a feedback handoff and a shared-equipment workshop, with a generalist and a full flower team. A facilitator routed questions and reviews as the conversation developed. Frozen artifacts were tested with prescribed initial and revised responses under easy, medium and hard browser conditions, followed by separate model reviews of process and result under an expert-role prompt. The retained baseline evidence contains 4 frozen candidates, 12 scenario records and 16 accepted reviews. The conversations made concrete repairs traceable to their authors, including an unsupported self-review claim, draft replacement and a lost case fact. Specialist participation and version-specific approval remained uneven. Passing browser checks did not establish response-specific feedback or learning. A planned user change added a copyable practice record in the same conversation. Its direct browser check was separate from four fresh version reviews. The study illustrates inspectable design contributions and limits of the simulation; one conversation per cell cannot establish a condition effect or human learning outcome.

Keywords: learning engineering; team competencies; agentic simulation; design evaluation

Introduction

The learning engineering flower places a team's shared understanding at the center of context-dependent areas of expertise. Its petals are not an exhaustive prescription; the challenge can require additional expertise and changing contributions (Ali et al., 2025). This raises an empirical question about influence: when does a contribution change a learning tool, and what evidence shows that the change matters? A role's presence alone does not answer it.

Agent pipelines can already plan, generate and review interactive learning resources (Gröner et al., 2026). We examine the conversation that produces an executable candidate, including questions, incomplete review and repair. Explicit models can expose assumptions and generate questions without predicting human behavior (Epstein, 2008). Our purpose is to identify inspectable mechanisms and hypotheses for subsequent study. We ask how the assigned roles participate, which concerns reach implemented changes, and what the resulting tools support under the recorded conditions.

Study design

The Design / Develop phase used two modest fictional challenges. The handoff brief targeted distinguishing an observation from an unsupported motive, asking a useful clarifying question and proposing a feasible next step. The workshop brief targeted reasoning about equipment, assistance, participant provisions and a feasible alternative when circumstances changed. Both required an editable scenario, an initial response, visible feedback, optional revision, skipping and reset. The common interface required scenario text to be handled as

data in self-contained HTML without external assets. Neither case supplied observations from real learners.

Each case was run once with a generalist and once with a full specialist team, producing four design conversations. The generalist role specified broad introductory knowledge and implementation permission; the agent was instructed to do competent work and could acknowledge uncertainty. It did not receive specialist playbooks or authority to approve specialist domains. The full team comprised Learning Sciences; Learning Experience Design; Data Science; Assessment, Measurement, Evaluation; Learning Environment Engineering; Software Engineering; Subject-Matter Expertise; and Education and Training Professionals. Data Science was also assigned assessment expertise. These were prompt assignments and authority rules; they did not erase the model's prior knowledge or establish professional competence.

Each configuration included a facilitator that routed questions, requested implementation and recorded readiness, but could not author solutions or supply specialist approval. Pending requests and the conversation determined the next contribution. Specialists were instructed to stay within assigned domains; the runtime validated declared permissions and approval authority, not substantive expertise. Implementations received version numbers, and approvals and objection resolutions remained attributed to their authors and versions. A separate freeze did not convert unreviewed domains into agreement. Shared understanding was represented by the common record and explicit treatment of assumptions and disagreement; no shared cognitive state was measured.

Four initial command-line attempts and one compatibility retry timed out without accepted messages; their records were retained separately. The reported pilot used fresh collaboration-service actors and manual transfer of exact retained prompts and schemas. The actual model was inherited from the session and not independently reported, so the planned command-line label does not identify it. Transport instructions restricted actors to their input and response files; sandbox equivalence and absence of other tools were not independently established. We report no token or compute total.

After format rejections, later dispatches clarified existing and new artifact versions, parent versions, addressed replies and null decision fields outside decision messages. These changes preserved the case, role, rubric and budget definitions. Exact per-call dispatches, rejected and accepted response bytes, and provenance were retained; the host edited no responses or scores.

A status check found the version 3 process-review worker interrupted before a response file existed. Archived transport-only follow-up requested status and resumption of the same pending review with its retained prompt and schema; the case, rubric and budget were unchanged. The resumed review counted once, and its returned response was not edited.

Context handling required recorded amendments. Three frozen runs received a post-freeze increase in the runner's input allowance from 180,000 to 600,000 characters for complete review records. The workshop full-team run reached the original cap during design and used a separately archived continuation amendment. That continuation preserved the configuration and all 39 preceding events, and retained the refused 217,067-character prompt byte for byte before resuming its addressed turn. The cap is a runner character limit, not a measurement of the model's context window. Source snapshots pin the engine version applicable to each run; the study did not use one unchanged binary throughout. Supplying the complete record does not establish equal model attention to every part.

Observation and evaluation

The Deploy / Test phase used the same three fixtures for each candidate in a case. Easy combined a 1280 by 800 desktop viewport with a clear task. Medium used 390 by 844 pixels, with conflicting handoff records or a missing workshop kit. Hard used 360 by 740 pixels while offline, with uncertain handoff evidence and unavailable coverage, or interrupted equipment and assistance. These bundled conditions do not isolate a difficulty effect. The 12 baseline scenario records represent four design conversations, with three observations per candidate.

Browser actions applied the case, entered a deliberately problematic response, requested feedback, opened revision, supplied a replacement response, resubmitted by keyboard, skipped and reset. Six checks examined case application, nonempty feedback, editable revision, skip state and response clearing. Observations retained rendered output, errors, network requests and operating conditions. This was not a complete accessibility, security or usability assessment. Both response texts came from the fixture; their difference cannot show learning caused by the tool.

Fresh evaluator actors, prompted as experts in learning sciences and systems engineering, received the retained conversation, assigned permissions, artifact and bound browser observations. Each frozen version was scheduled for one process review and three result reviews, one per difficulty; evaluators neither repaired nor executed the artifact. Accepted responses had to satisfy the unchanged evaluation contract and cite exact evidence representations. Raw failures were retained without host score edits. Six process and six result criteria used a 0-4 ordinal scale: 0 indicated observed contradiction or an absent mechanism; 1, a partial or fragile response with material observed gaps; 2, adequate evidence for a limited portion; 3, strong and coherent evidence across recorded conditions; and 4, consistently supported evidence that handled exceptions and uncertainty appropriately. Missing or insufficient evidence remained Unknown. Process, result and the human flourishing, system performance and learning axes were reported separately without totals or ranking.

After initial testing and four review requests or attempts, a planned study request asked the handoff full team to add a concise, copyable record of the applied scenario, response, feedback and next action. This was not field-use feedback. A frozen successor required three same-fixture regressions, four fresh reviews and an explicit feature check; neither the successor nor those regressions were independent replicates or holdouts.

Conversation results

The conversations are between design agents. Deployment observations use scripted browser actions and prescribed responses; no learner dialogue was simulated.

Table 1. Design trajectories at the first freeze

Case and condition	Msgs	Calls / failures	Roles	Built / frozen	Approved domains
Handoff Generalist	3	4 / 1	1	1 / v1	0
Handoff Full team	17	19 / 2	6	2 / v2	1
Workshop Generalist	3	3 / 0	1	1 / v1	0
Workshop Full team	21	22 / 1	5	3 / v3	4

Messages include the facilitator; contributing roles exclude it and evaluators. Calls / failures count collaboration-service design attempts before first freeze. Built counts implemented versions. Approved domains refer only to the frozen version.

Each generalist conversation comprised a facilitator request, one implementation and a readiness response. Their recorded controls operated, but the short trajectories offered little opportunity to observe disagreement or evidence-driven repair. Full-team trajectories included addressed proposals, implementations and focused reviews. Their greater length and revision count cannot establish efficiency or quality differences in this design.

In the handoff full-team conversation, the assessment specialist objected to a status message that treated an unchanged response as evidence that self-review had occurred (m00024). Software Engineering replaced it with an observable event description in version 2 (m00032), after which the objection's author resolved the issue and approved the assessment design (m00035 and m00037). The corrected status reported an unchanged submission and withheld any correctness or improvement judgment. This repair preserved voluntary revision while removing an unsupported inference about the learner's reasoning.

Workshop reviews led to two further repairs. Learning Experience Design found that copying an earlier response could overwrite a nonempty draft (m00018); version 2 restricted copying to empty drafts (m00022). Subject-Matter Expertise then found that its own proposed case change had omitted the independent fact that a volunteer could guide the manual activity (m00029). Version 3 restored that fact and clarified the change notice (m00037). The supplied source established these within-domain repairs; their interactions still required executed observation.

Review coverage remained uneven. Several specialists approved handoff version 1, but those approvals did not transfer to version 2. At first freeze, only the assessment domain approved version 2. Data Science and Education and Training Professionals had made no accepted contribution. Availability, participation and current-version approval therefore described different coverage. An absent approval records a review gap; it does not itself demonstrate product failure.

Observed behavior and expert reviews

The result tables use the retained study summary dated 03 November 2026 at 20:52 UTC; every entry is bound to its recorded artifact version and journal prefix.

All 12 baseline scenario records were present; 12 were marked observed. The specified browser interactions recorded 72 passes across 72 mechanical checks, with 0 runtime errors and 0 network requests. Nonempty feedback could pass a mechanical check while offering little explanation of a response.

Table 2. Process ratings by frozen artifact version

Process criterion	HG v1	HF v2	HF v3	WG v1	WF v3
Human concerns	3	3	4	3	4
Domain boundaries and uncertainty	4	4	4	4	4
Responsive dialogue	2	4	4	3	4
Decision traceability	3	4	4	3	4
Learning reasoning	3	4	4	2	4
Participation	2	3	3	2	3

HG = handoff generalist; HF = handoff full team; WG = workshop generalist; WF = workshop full team. Each column contains one process review; a successor continues the same conversation. Domain boundaries rates recorded authority and uncertainty, not measured expertise or depth of knowledge. Criteria map to human flourishing (human concerns, participation), system performance (domain boundaries, traceability), and learning (dialogue, learning reasoning), each retaining its separate 0-4 rating.

Table 3. Handoff result ratings by difficulty and artifact version

Candidate and scenario	AA	TB	R	CS	FR	CU
Generalist v1 easy	3	3	3	3	3	3
Generalist v1 medium	3	3	3	3	2	3
Generalist v1 hard	3	3	3	3	2	3
Full team v2 easy	3	4	2	3	3	3
Full team v2 medium	3	3	2	3	3	3
Full team v2 hard	3	3	2	3	3	3
Full team v3 easy	3	3	2	3	3	2
Full team v3 medium	3	3	3	3	2	3
Full team v3 hard	3	3	3	3	2	2

AA = agency, access and burden; TB = task behavior; R = recovery and updates; CS = capability support; FR = feedback and revision; CU = credible use. The axes are human flourishing (AA, CU), system performance (TB, R), and learning (CS, FR). Each criterion retains its 0-4 rating; no totals or ranks are computed.

Table 4. Workshop result ratings by difficulty and artifact version

Candidate and scenario	AA	TB	R	CS	FR	CU
Generalist v1 easy	3	2	3	2	2	3
Generalist v1 medium	3	2	3	2	2	3
Generalist v1 hard	3	3	3	2	2	3
Full team v3 easy	3	3	3	3	2	3
Full team v3 medium	3	3	3	3	2	3
Full team v3 hard	3	3	3	3	2	3

Criterion codes, axes and the separate 0-4 scale follow Table 3.

Both baseline full-team process reviews rated responsive dialogue and decision traceability 4, crediting the concern, implementation and originating-reviewer follow-up chains. The corresponding generalist ratings were 2 and 3 for handoff, and 3 and 3 for workshop. These are descriptive judgments about four realized conversations. The full-team participation ratings of 3 retain gaps in practical delivery and operating perspectives; neither the roster nor the longer transcript establishes complete review or a condition effect.

The handoff generalist's phrase cue identified an unsupported motive in the easy input but missed the carelessness claim in the hard input. Its feedback and revision ratings were 3, 2 and 2 across easy, medium and hard. Full-team version 2 received 3 in each fixture, although its substantive self-review prompts also remained the same across the supplied initial and revised responses. Those reviewers credited transparent explanations and honest limits. The ratings do not establish response-specific diagnosis or improvement in a person's reasoning.

Handoff recovery ratings require particular care. The generalist received 3 in all fixtures and full-team version 2 received 2, but the latter rationales cited limited executed coverage rather than observed failures: case replacement preceded a populated attempt, and reset followed skip after the draft was already cleared. The generalist easy review acknowledged the same limitation while assigning 3. The differing scores are model assessments of limited recorded evidence; they do not establish that the full-team tool was less recoverable.

In the easy workshop fixture, the supplied initial response allocated six kits when only three were available. The generalist displayed the facts and generic resource questions without explaining the mismatch. Its feedback and revision ratings were 2 throughout. All three full-team reviews also rated feedback and revision 2: the structured practice retained the applied facts and a revision route, but did not diagnose the response's substantive error. Those reviews rated capability support 3, compared with 2 for the generalist, without claiming attainment. In the hard fixture, the full-team tool supported conditional reflection during the stated assistance interruption while leaving the revised alternative's feasibility unverified. Corrected source-level case handling and successful controls did not establish whether learners could recognize and repair their own planning errors.

Response to the planned user change

The planned user-change request, recorded in event 55, led to nine further authoring messages and handoff version 3 in the same conversation. Assessment review required the applied case, submitted wording and offered guidance to remain associated with the same attempt. Learning Experience Design distinguished that submitted record from an editable draft and made viewing and copying optional. Software Engineering implemented a labeled, read-only summary with native text selection. Assessment and experience reviewers then approved the implemented addition within their respective scopes (m00070 and m00072); the other six domains remained not_reviewed for this version.

The version 3 process review rated human concerns (agency and burden) 4, compared with 3 for version 2, crediting optional summary disclosure, a return-to-draft control and clear last-submission labeling. Dialogue and traceability remained 4: the review linked the assessment concern about matching case, response and guidance to the implementation and scoped follow-up reviews. Participation remained 3 because questions about delivery and available assistance lacked educator or learner input. These judgments concern decisions in one continued design conversation; they do not measure human burden or expertise.

Fresh result reviews rated feedback and revision 3 for easy and 2 for medium and hard: the same general prompts did not diagnose conflicting records or the hard case's unmet coverage prerequisite. The hard credible-use rating of 2 also noted the full review's uncertain fit to a

fictional two-minute constraint; no human completion time was measured. Easy recovery and credible use were 2 because canonical execution coverage was limited, not because a failure was observed. Medium and hard recovery ratings were 3 while also retaining coverage limits. These are model judgments of scoped packets, not an overall ranking or evidence of changed recovery reliability. The separate inspection below does not alter their recorded ratings.

A separate direct browser inspection used hosted HTML whose fingerprint matched frozen version 3. The copied text equaled the complete displayed summary, retaining the applied title, facts, operating conditions, task, submitted response with its line breaks, guidance, next action and practice limitation. Editing the draft preserved the previous submitted record; resubmission replaced it. Blank submission, pending case edits and invalid case application preserved that record. Valid application cleared it, and the next copied summary contained the changed case and response. Skip and reset cleared populated attempts; voluntary return began fresh, and reset restored the starting case. An unchanged submission retained neutral wording without a correctness or improvement judgment. Two inspection expectation mismatches were retained and resolved against the current interface; neither was an observed artifact failure.

The three unchanged regression fixtures produced 18 passes across 18 mechanical checks, with no recorded runtime errors or network requests. The direct inspection was separate from those fixtures and was not included in expert review packets. Its two copied examples contained 2,960 and 3,072 characters. They demonstrate a copyable, context-bound practice record; whether that amount of detail is concise or manageable for intended users remains unmeasured. The addition did not supply diagnostic feedback, human learning evidence or a verified workplace outcome.

Discussion and limitations

The most direct contribution is the link from a specific concern to an implemented change and a named follow-up review. In these conversations, assessment expertise narrowed an inference, interaction design protected a draft, and case expertise repaired a fact introduced by the case author. The value of those changes can be examined in the artifact without assuming that every assigned specialist contributed or that a review label certifies the whole tool. The flower is useful here as a way to organize questions about contribution and coordination. The pilot does not validate its domain assignments as a sufficient or necessary team composition.

The tests also expose a practical design tradeoff. Guided self-review can keep uncertainty visible and support a small, editable tool, while leaving the learner responsible for recognizing a substantive error. A functioning revision route and honest caveats do not demonstrate that a novice can use the guidance well. Follow-up work should examine whether people detect the specific errors exposed by these fixtures, explain a justified revision and adapt to a changed case without the same prompts. Such evidence would address capability support beyond the present browser behavior.

With one conversation per case and condition, work was not experimentally matched: calls, speakers, versions and review coverage varied. Scenario observations and successor versions remained nested within their original conversations. Two cases cannot establish which expertise is generally required, and bundled difficulty conditions do not isolate device, connectivity or task complexity. The transport change and selective context continuation also prevent treating execution as one stable protocol. Their retained records permit inspection of these limitations.

The actors and evaluators were language models, with shared background capabilities that role instructions did not remove. Language-model simulations can reproduce some patterns while also introducing systematic distortions (Aher et al., 2023). Fresh evaluators were separate from the authoring turns, but they were not blinded human judges, independently replicated raters or evidence of population-level reliability. They could infer the condition from the conversation and artifact. Model-judge research identifies verbosity, position and self-enhancement biases and reasoning limits (Zheng et al., 2023); this pilot did not measure or remove those biases. Exact evidence references constrain unsupported judgments but do not make the ratings objective.

Future experiments could replicate conversations across cases and models, vary one expertise assignment while accounting for work opportunities, and seek independent human judgments of the resulting changes. Dede's (2026) component-based research motivates examining which elements contribute under specified conditions; applying that approach to domain instructions and conversational participation remains a methodological proposal. Human participants would be needed to evaluate burden, accessibility in use, understanding, retention and transfer. No human learning, workplace effectiveness or team-performance claim follows from the present simulation.

Conclusion

Dynamic agentic conversations made several design contributions inspectable from proposal through repair. They also showed that configured expertise, participation, current-version approval and observed behavior are distinct records. The study provides a bounded method for asking how those records relate. Its evidence supports concrete questions about design and evaluation, with unresolved coverage and model-based inference kept visible. Establishing which team capabilities improve learning requires replicated studies and evidence with people.

Acknowledgments. The authors acknowledge support from the LENS program.

Data availability. The workspace at https://capabilitymatters.org/agentic-le/dynamic/ provides saved conversations, artifacts, scenario observations and evaluator rationales. The separate direct feature inspection, excluded from the expert packets, is at https://capabilitymatters.org/agentic-le/dynamic/study/user-update-feature-check.json. Code and protocol are available at https://github.com/wrgr/capabilitymatters/tree/main/tools/lifecycle and in docs/lifecycle/dynamic-flower-study.md. The complete invocation archive, including prompts, rejected outputs, source snapshots, amendments and validation receipts, is retained separately offline.

Use of artificial intelligence. Artificial intelligence assisted software development, generation, manuscript preparation, technical observation and artifact assessment. The collaboration-agent model used for the pilot was inherited from the session and was not independently reported.

References

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. Proceedings of Machine Learning Research, 202, 337-371. https://proceedings.mlr.press/v202/aher23a.html

Ali, Z., Morehead, K., Jay, M., Arner, T., Lis, J., & Braitling, B. (2025). Learning engineering is a team sport: Collaborative explorations for team building strategies with short scenarios in various contexts. International Consortium for Innovation and

Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale. https://doi.org/10.59668/2109.21837

Dede, C. (2026). What works when for whom under what conditions: Learning engineering as an enabler of component-based research. Proceedings of the Learning Engineering Research Network Convening (LERN 2026). https://doi.org/10.59668/2551.25412

Epstein, J. M. (2008). Why model? Journal of Artificial Societies and Social Simulation, 11(4), Article 12. https://jasss.soc.surrey.ac.uk/11/4/12.html

Gröner, F., Verma, A., & Bronowitz, J. (2026). Automatically generating interactive learning experiences with an LLM-driven agentic pipeline. Proceedings of the Learning Engineering Research Network Convening (LERN 2026). https://doi.org/10.59668/2551.25387

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems, 36. https://arxiv.org/abs/2306.05685