RUNTIME SUPERVISION FOR LLM TUTORS UNDER IDENTICAL PROMPTING

Anonymous authors
Paper under double-blind review

ABSTRACT

A correct answer is not always an appropriate tutoring response. Large language model (LLM) tutors must follow pedagogical requirements that depend on the learner's prior attempts and misconceptions. Establishing the value of runtime supervision requires separating its contribution from prompting and checking whether LLM judges, models that read and score transcripts, recognize the failures being measured.

We present a reference architecture and controlled evaluation of bounded, state-dependent supervision. The architecture checks conversation-dependent rules, such as withholding direct answers until a diagnostic question has been asked, and permits one corrective regeneration without substituting fallback responses; persistent violations are released and recorded.

Across four nested conditions, the primary comparison uses the same tutor and identical initial prompts, with supervision enabled or disabled. We evaluate one proprietary and one open-weights model from a single developer on 20 data-science and statistics scenarios in 12 topic families, comprising 1,440 multi-turn dialogues and 800 single-turn stability probes. Assessment covers verifiability, stability, auditability, and pedagogical soundness under the Trusted Educational AI Standard (TEAS), alongside inference cost.

Three LLM judges from different model families assess responses, with pedagogical judgments calibrated using openly licensed teaching exemplars and matched flawed alternatives. We test whether judges detect planted errors without flagging harmless changes; failure under prespecified criteria prevents confirmatory conclusions for the affected dimension. The pedagogy score represents the share of a scenario's teaching requirements met, on a 0 to 3 scale. Instructions raise pedagogical adherence by 42% (+0.786 of 3, 95% interval [0.643, 0.929]). Adding reference context raises verifiability by 55% (+0.763 of 3) and moves auditability from 0.53 to 1.83 of 3. Supervision on top of the identical prompt adds 7% to pedagogy (+0.182 of 3, 95% interval [0.000, 0.364]), inconclusive under the registered rule. Delivery costs rise by factors of 1.39 and 4.14 for the proprietary and open-weights models, respectively, while released citation-check failures fall from 12.7% to 1.8% and from 32.4% to 3.4%. Stability falls by 17% with instructions and 20% for the full system; stability and auditability estimates are descriptive because their judges did not pass the planted-error examination.

1 INTRODUCTION

Tutoring requires more than producing a correct answer: the appropriate response depends on what the learner has already attempted and understood. For example, a learner may confuse the probability of detecting fraud with the probability that a flagged transaction is fraudulent. Explaining Bayes' theorem immediately supplies relevant information, but asking how the learner reached that conclusion first creates an opportunity to identify the misconception. Pedagogical guidance emphasizes eliciting understanding and avoiding premature solutions (Jurenka et al., 2024). For LLM tutors, implementing this guidance raises a practical question: how much can instructions accomplish, and what does a separate model that checks a reply before release add?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We address this question through a reference architecture and a four-condition experiment. The baseline condition, B0, gives the tutor minimal instructions; P adds pedagogical instructions; U adds reference material and the complete response requirements; W adds runtime supervision to the identical U prompt. These conditions distinguish the contributions of instructions, context, and supervision, as well as the full system's difference from the base tutor. The runtime policy allows the tutor one revision after a failed check and releases the resulting reply without substituting a template. Following the resulting conversations captures both corrected behavior and violations that remain.

A second question concerns the measurement itself. We use other LLMs to read tutor transcripts and assess whether specified requirements were met; these evaluators are the *LLM judges*. Their judgments must recognize the failures on which the architectural comparison depends. Judge-dependent orderings in tutoring evaluation (Fan et al., 2026) motivate testing that ability directly. We therefore examine the judges using deliberately inserted errors and harmless edits, and compare their assessments with blind human ratings.

The experiment uses two tutor models from one developer: the proprietary `gpt-5.6-terra` (Terra) and the open-weights `gpt-oss-120b` (OSS). A third model, `gpt-5.6-luna`, labels tutor actions using two separately frozen prompts: one for runtime supervision and one for offline evaluation. Three judge models, `gpt-5.6-sol`, `Kimi-K2.6`, and `DeepSeek-V4-Pro`, assess the released conversations. This judge panel combines their assessments of individual requirements into the scores used for comparison.

Supervised generation and conversation-dependent release policies have established precedents in tutoring (Liffiton et al., 2023; Pal Chowdhury et al., 2024; Kadir, 2026; Pisan, 2026). Here, the reference design, the controlled sequence of comparisons, and the examination of the evaluation work together. We run 1,440 scripted multi-turn dialogues and 800 stability probes across 20 data-science and statistics scenarios in 12 topic families. The same tutor and identical initial prompt in U and W isolate the added runtime package; the other comparisons show what the prompt already contributes.

The pedagogy score represents the share of a scenario's teaching requirements met by the tutor, scaled from 0 to 3, where 3 means all assessed requirements were met. A +0.786 difference is therefore 26 points on a 100-point scale, or a 42% rise over the paired base tutor's score of 1.86. The results separate the layers: instructions raise pedagogy by 42% (+0.786 [0.643, 0.929]), and reference context raises verifiability by 55% (+0.763 [0.586, 0.957]) while taking auditability from 0.53 to 1.83. Runtime supervision adds 7% to pedagogy (+0.182 [0.000, 0.364]). Our registered superiority rule requires an effect of at least 0.20 on the 0–3 scale, a 95% interval excluding zero, a one-sided Holm-adjusted p below .05, and positive changes for both models. The runtime result is inconclusive under that rule. Delivery cost rises by factors of 1.39 and 4.14, while released citation-check failures fall from 12.7% to 1.8% for Terra and from 32.4% to 3.4% for OSS. Stability decreases by 17% with instructions and 20% for the full system; stability and auditability estimates are descriptive because their judges fail the examination.

The contributions are:

1. A reference runtime policy for conversation-dependent tutoring requirements, with one corrective revision and explicit accounting for remaining violations.
2. A controlled ladder of prompt, context, supervision, and total-system comparisons, including supervision enabled or disabled under identical initial prompts.
3. An evaluation procedure that tests its judges on known errors, uses blind human review, and determines which dimensions can support confirmatory conclusions.

The supplementary material provides every prompt verbatim, the licensed teaching excerpts used to calibrate the pedagogy judge, the data behind Table 1, and an offline explorer in which readers can open any of the 1,440 dialogues with its runtime checks, revisions, and judge verdicts.

2 Related Work

Supervision of LLM tutors. CodeHelp checks proposed programming assistance and can request another generation to remove solution code (Liffiton et al., 2023). MWPTutor combines pedagog-

ical states, generated utterances, and guardrails (Pal Chowdhury et al., 2024). Authorization-aware release control selects disclosure contracts using learner state and can replace a failed response with a deterministic fallback (Kadir, 2026). A supervisor architecture also uses correction followed by a fallback when required (Pisan, 2026). These studies establish precedents for state-dependent checks and regeneration. Our policy releases the tutor's own response after at most one revision, and our experiment compares that policy with the same initial prompt throughout branching conversations.

Assessment of automated evaluators. Tutoring judgments can depend on the evaluator: Fan et al. (2026) examines judge-contingent helpfulness signals, and MRBench evaluates automated assessments against dimension-specific human annotations (Maurya et al., 2025). Combining judges from different model families has also been studied beyond education (Verga et al., 2024). Our evaluation joins two tests: controlled errors examine whether the judges recognize targeted failures, and human ratings examine their behavior on sampled tutor outputs. The examination has a direct consequence for reporting, since failure removes the corresponding dimension from confirmatory testing.

Pedagogical requirements. Educator-informed principles emphasize eliciting understanding, adapting assistance, and avoiding premature solutions (Jurenka et al., 2024). The Trusted Educational AI Standard (TEAS) organizes assessment into verifiability, stability, auditability, and pedagogical soundness (Syed, 2026). We turn these dimensions into explicit response requirements and separate scores. The sequence of conditions then tests which parts of the system contribute to each dimension.

3 Runtime Policy and Experimental Design

3.1 Checking a proposed reply before release

The runtime follows the sequence in Figure 1. The tutor first proposes a reply to the learner. A separately prompted LLM classifier labels the action in that reply, such as asking a diagnostic question, explaining, or giving an answer. Fixed rules then check whether the action is allowed given the tutor's previous actions in the conversation. For example, a scenario may require a diagnostic question before a direct answer or an explanation followed by a checking question.

The tutor also generates a structured *audit record*: a private-to-the-system part of its response listing the sources used, the suspected misconception, and the reason for its tutoring action. Under the full prompt, factual claims must be linked to identifiers in the scenario's source list inside this audit record; those identifiers must not appear in the learner-facing text. Deterministic citation and audit checks verify that cited identifiers exist, that at least one listed identifier is cited, and that the audit record is well formed. The program runs citation, audit, and pedagogical checks on the proposed reply.

If any check fails, the tutor receives one combined correction request and generates one revised reply. The program checks that revision again and releases it as the tutor wrote it, with any remaining violations saved alongside it. No template replaces the tutor's text. The learner sees the released tutoring text with the audit portion removed. The next scripted learner turn is selected from the released action, so a correction can change the subsequent conversation.

3.2 Four conditions and branching conversations

B0, the baseline, supplies a minimal tutor role and requests an audit record. P adds the scenario's pedagogical instructions. U adds the reference material, known misconceptions, citation requirements, and complete audit schema. W uses the identical U prompt and enables the runtime policy. The planned comparisons are therefore P–B0 for instructions, U–P for reference context, W–U for supervision, and W–B0 for the complete system. All 560 paired U/W initial-prompt hashes match, including 360 dialogue pairs and 200 probe pairs.

The corpus comprises 20 conceptual, not computational, data-science and statistics scenarios in 12 topic families. Each scenario specifies its knowledge, misconceptions, response requirements, and a branching learner script. Three learner profiles and three repeated runs yield nine dialogues for each scenario, model, and condition, for 1,440 dialogues overall. Five single-turn rephrasings per

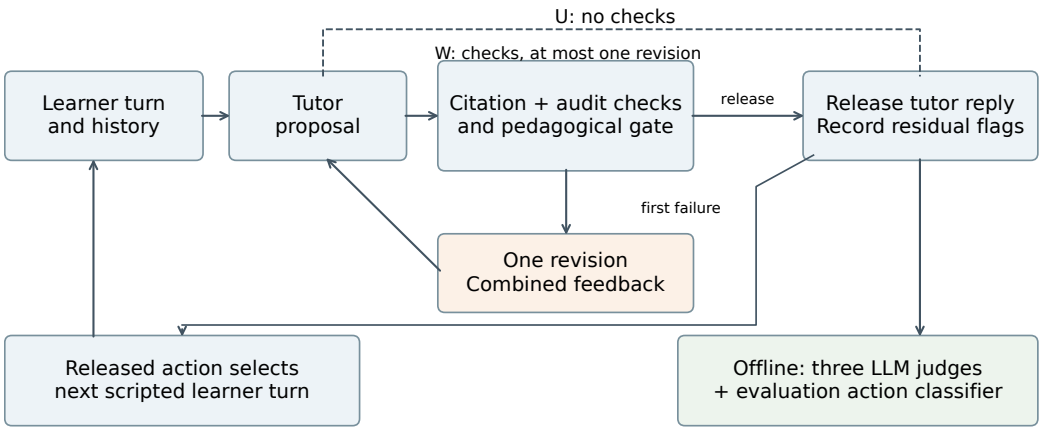

Figure 1: From proposal to release. The same initial prompt supplies a tutor proposal in U and W. U releases it without checks. W checks citations, audit structure, and the permitted pedagogical action; a first failure returns combined feedback for one revision. The revision is checked and released even if violations remain. The released reply determines the next scripted learner turn and is assessed separately by the offline evaluators.

combination yield 800 stability probes. Scenarios were funded in blocks of five under a registered cost-based stopping rule, and four blocks ran. Requests use provider-default temperature and the registered reasoning settings; model-role details are in Appendix A.

Matching U and W fixes the initial instructions, learner profile, and repetition index. Subsequent learner turns follow the script's response to the tutor's released action. This makes the comparison a test of supervision across whole conversations, including the trajectories that intervention produces.

The design, prompts, and code were frozen and linked by hashes before the retained runs. Later changes were documented.

4 Measurement and Analysis

4.1 What the scores mean

Each TEAS dimension uses a separate 0–3 scale. *Verifiability* scores the fraction of extracted factual claims supported by scenario knowledge, scaled by three and aggregated across judges by the median. *Auditability* combines the availability of the tutor's audit record with the accuracy of its source links and misconception description. *Stability* measures whether repeated runs and reworded questions preserve the same factual propositions and sequences of tutoring actions; its score takes the lower of factual and action-sequence consistency. *Pedagogical soundness*, shortened to pedagogy below, scores adherence to the scenario's teaching requirements and adaptation between learner profiles. Appendix B gives the formulas and aggregation rules.

Pedagogy has three components: asking before explaining or answering when required, avoiding endorsement of the specified misconception, and adapting assistance between novice and advanced dialogues. Each judge supplies a Pass or Fail for an individual requirement, with a verbatim tutor quotation supporting a Pass. A majority of the three judges determines the requirement's value. When only two judges supply a usable answer, both must agree; otherwise that assessment is missing, meaning no Pass or Fail can be assigned. A majority across repeated dialogues determines the profile value, and a majority across profiles determines the scenario value. The novice/advanced comparison supplies the adaptivity component, and the three binary components sum to the scenario's 0–3 score.

If any judgment needed for a scenario score is missing, that score is dropped rather than computed from fewer dialogues; requiring U and W scores on both models therefore leaves 11 of the 20 scenarios for the pooled pedagogy comparison. Appendix D gives availability for every condition.

The separately prompted evaluation classifier supplies the action labels used in stability assessment. Worked teaching examples inside the pedagogy judge prompt illustrate each assessed requirement (Appendix B).

4.2 Who Judges the Judges

The judges first encounter practice examples with known errors. The protocol permits one prompt-repair round after this practice, followed by freezing the judge prompts. A separate examination then tests them on 520 previously held-out items across six error types. Each item pairs a deliberately inserted error with control material, including a harmless edit. Detection counts only when the required error is recognized without the corresponding control response. To pass, an error test requires at least 80% detection and at most 20% control alarms; an absent response never counts in the judges' favor. A failed test makes the affected dimension descriptive rather than confirmatory.

Two human teaching assistants independently grade a blind sample of 60 conversations and 100 short exchanges used to assess action labels. The grading pages hide model identity, condition, and automated assessments. Recognition rates compare the automated judgments with the human reference, weighting for the sample's unequal selection probabilities. One of the authors adjudicates disagreements, and three initially undecided items use AI-assisted resolutions; the analysis also retains the version leaving those items undecided. The summary of whether all citation pairs are supported is kept separate from the individual pairs to avoid counting the same evidence twice.

4.3 Planned comparisons and statistical decisions

For each contrast, we pair scenario scores within each model, average the two model differences, and then average over scenarios with all required scores. Uncertainty comes from 10,000 bootstrap resamples of topic families, keeping related scenarios together. The reported 95% intervals describe these paired differences; the superiority criteria are specified in Section 1.

Within the ladder, W–U is the confirmatory comparison. P–B0 and U–P are registered planned contrasts outside the confirmatory family. W–B0 is secondary estimation for verifiability, stability, and auditability; its pedagogy estimate is also a registered planned contrast outside that family. The confirmatory family additionally contains a named-model pedagogy comparison. Auditability and stability follow the examination's reporting decision in every contrast.

Two prespecified test families are reported, both fixed before any result was inspected. The original family tests the full system (W–B0) on verifiability, stability, and auditability, together with W–U pedagogy and the named-model pedagogy comparison; the amended family tests supervision (W–U) on all four dimensions and retains the named-model comparison. Each family receives its own Holm adjustment, with a fixed-five sensitivity assigning excluded tests $p = 1$ (Appendix D.1).

5 Results

5.1 What each layer contributes

Table 1 compares scores on the same scenarios and models: the instruction effect is P minus B0, paired within scenario and model, averaged across the two models and then over complete scenarios. The other contrasts use the same procedure for their respective arms. Relative changes divide each paired difference by the lower-arm mean, making the contribution of each layer interpretable against its own starting point.

Table 1: The four planned comparisons of instructions, context, supervision, and the complete system.

Addition / contrast	Role	Pedagogy	Verifiability	Auditability†	Stability†
Instructions: P–B0	P	$+0.786$ $[0.643, 0.929]$ $n = 7,\ k = 6$ +42% | +26 pts	$+0.050$ $[-0.043, 0.129]$ $n = 20,\ k = 12$ +4% | +2 pts	-0.012 $[-0.075, 0.045]$ $n = 20,\ k = 12$ 0.54 to 0.53 | -0.4 pts	-0.363 $[-0.573, -0.142]$ $n = 18,\ k = 11$ -17% | -12 pts
Context: U–P	P	-0.333 $[-0.500, 0.000]$ $n = 3,\ k = 3$ -13% | -11 pts	$+0.763$ $[0.586, 0.957]$ $n = 20,\ k = 12$ +55% | +25 pts	$+1.297$ $[1.196, 1.406]$ $n = 20,\ k = 12$ 0.53 to 1.83 | +43 pts	-0.008 $[-0.138, 0.128]$ $n = 18,\ k = 10$ -0.5% | -0.3 pts
Supervision: W–U	C	$+0.182$ $[0.000, 0.364]$ $n = 11,\ k = 9$ +7% | +6 pts	$+0.038$ $[-0.048, 0.097]$ $n = 20,\ k = 12$ +2% | +1 pts	$+0.120$ $[0.032, 0.201]$ $n = 20,\ k = 12$ 1.83 to 1.95 | +4 pts	-0.009 $[-0.105, 0.118]$ $n = 19,\ k = 11$ -0.5% | -0.3 pts
Full system: W–B0	S/P	$+0.679$ $[0.444, 0.824]$ $n = 14,\ k = 9$ +35% | +23 pts	$+0.851$ $[0.640, 1.077]$ $n = 20,\ k = 12$ +64% | +28 pts	$+1.406$ $[1.284, 1.507]$ $n = 20,\ k = 12$ 0.54 to 1.95 | +47 pts	-0.420 $[-0.643, -0.182]$ $n = 18,\ k = 11$ -20% | -14 pts

Each cell: difference on the 0–3 scale, 95% interval, complete scenarios n and topic families k, then the relative change (difference over the lower-arm mean) and points on a 100-point scale (difference over three). Auditability shows the two paired means instead of a ratio because its baseline is near zero. Roles: P, registered planned contrast outside the confirmatory family; C, confirmatory; S/P, secondary estimate (pedagogy: P). † Descriptive: the judges for these dimensions failed the examination. Each comparison uses its own complete sample.

Instructions raise pedagogy by 42% (+0.786 of 3, 95% interval [0.643, 0.929]), from 1.86 to 2.64 on the paired sample. Reference context raises verifiability by 55% (+0.763 [0.586, 0.957]) and takes auditability from 0.53 to 1.83 (+1.297 [1.196, 1.406]). Adding supervision contributes a further 7% to pedagogy (+0.182 [0.000, 0.364]), with its statistical decision examined below. The U–P pedagogy estimate is -0.333 [-0.500, 0.000], based on three complete scenarios. Relative to B0, the full system raises pedagogy by 35% (+0.679 [0.444, 0.824]) and verifiability by 64% (+0.851 [0.640, 1.077]). Because these paired samples differ, adjacent estimates need not add to the full-system difference.

Stability decreases by 17%, from 2.14 to 1.77, with pedagogical instructions (-0.363 [-0.573, -0.142]), and by 20%, from 2.14 to 1.72, for the full system (-0.420 [-0.643, -0.182]). These descriptive estimates show that the layers raising adherence and source-related scores do not raise measured response consistency.

5.2 What supervision adds under the same prompt

The W–U pedagogy difference is +0.182 [0.000, 0.364] over eleven scenarios in nine families (Figure 2). Three of the four superiority requirements (Section 1) fail: the effect is below 0.20, the interval touches zero, and the amended-family adjusted $p = .09919$ exceeds .05. Both model contributions are positive on this common subset, +0.273 for Terra and +0.091 for OSS. Under the original family, adjusted $p = .04960$ still leaves the effect-size and interval requirements unmet. The decision remains inconclusive in both families and the fixed-five sensitivities.

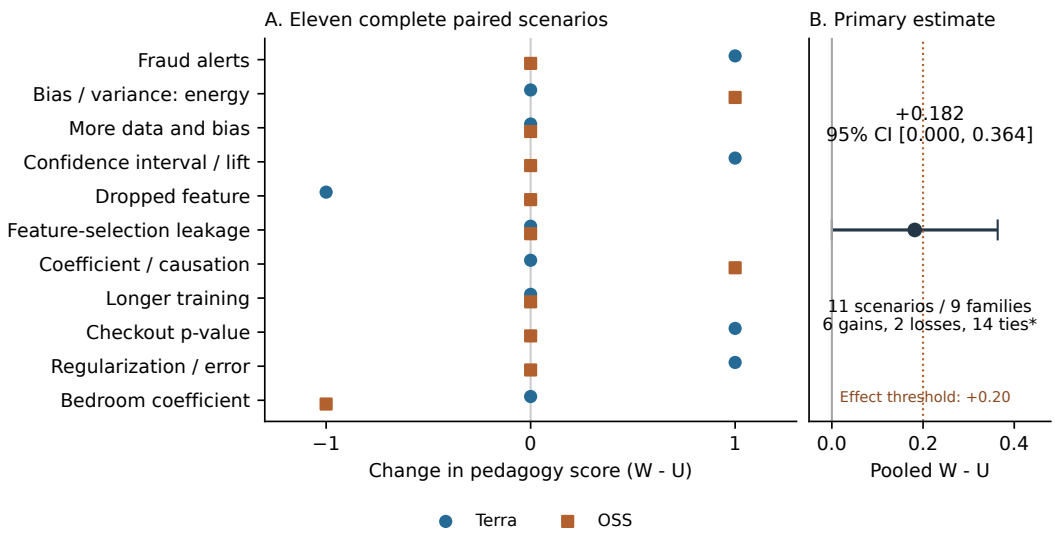

Figure 2: Change in pedagogy score on the 0–3 scale under supervision. Panel A has one marker per model per scenario: circles denote Terra and squares denote OSS. Panel B averages the paired differences and shows its 95% family-bootstrap interval. The asterisk identifies six gains, two losses, and fourteen ties among the 22 model/scenario contributions, rather than independent statistical clusters. The dotted line marks the effect-size requirement.

Prompt sufficiency requires U adherence of at least 2.0 and W–U equivalence, defined as a 90% interval wholly within $[-0.20, 0.20]$, on each model separately. U's available-score means exceed 2.0 for both models. Terra's sixteen paired scenarios give [0.000, 0.400], outside the equivalence margin, whereas OSS's thirteen give [0.000, 0.176], within it. Joint prompt sufficiency is therefore not established. These model-specific comparisons use their own available pairs.

Verifiability W–U is +0.038 $[-0.048, 0.097]$, with a 90% interval $[-0.032, 0.088]$ within the numerical equivalence margin. The planned check for differences in replies containing no factual claims is unavailable for two W scores, so this remains a comparison conditional on extracted claims. Appendix D.1 provides the other analysis family and sensitivities.

5.3 The judges' examination and human comparison

The judges pass the factual-claim, citation-support, and pedagogy error tests (Table 2). These admit verifiability and pedagogy to confirmatory analysis. Auditability and both stability tests fail. The examination therefore distinguishes usable confirmatory dimensions from measurements that can only be described.

Table 2: Judge examination and human comparison. Control alarms follow the specified treatment of missing responses; the strategy value is a conservative bound. Human recognition is the percentage of human Pass or Fail labels recognized by the automated assessment, weighted for sampling. Raw failure counts show the evidence available for each percentage.

Planted error	Detected/items	Control alarms/items	Exam rule
Factual claim	79/83	2/83	Pass
Citation support	77/80	0/80	Pass
Pedagogy	88/96	0/96	Pass
Epistemic stability	63/96	5/96	Fail
Audit note	51/96	0/96	Fail
Strategy stability	36/69	36/69 (bound)	Fail
Human-rated unit	**Pass recognition**	**Fail recognition**	**Raw detected failures**
Pedagogy assertion	86.0%	98.2%	14/16
Citation-support pair	82.6%	35.7%	1/3
Audit note	82.9%	100.0%	14/14

The judges recognize fourteen of sixteen human-identified pedagogy failures and all fourteen audit-note failures; weighted Pass recognition is 86% and 83%, respectively. They recognize one of three citation-pair failures. With so few natural failures, the protocol assigns insufficient validation rather than certifying these recognition rates, and no human-based score correction is applied. The evaluation action classifier matches 76/100 human labels; the original graders match each other on 73/100. Among 27 sampled turns with runtime labels, action labels agree in 14/27, while substituting the human label gives the same rule decision in 27/27; only one is a flag case, so the latter fraction provides little information about sensitivity to violations.

5.4 Released replies and the cost of correction

Supervision reduces the fraction of released turns failing the citation-ID check from 12.7% to 1.8% for Terra and from 32.4% to 3.4% for OSS (Table 3). These counts concern the required hidden source identifiers and audit structure. W revises 115 of 557 Terra turns and 245 of 553 OSS turns. Initial pedagogy flags clear on 80/81 and 75/141 turns, respectively; one additional OSS turn acquires a gate flag after a revision triggered by another check. One Terra reply and 67 OSS replies are released with remaining gate flags, as specified by the policy.

Table 3: Released-turn checks and delivery resources for 180 dialogues per condition and model. Failure rates use released turns as the denominator. Delivery includes tutor and runtime classifier calls; evaluation costs are reported separately in the text.

	Terra U	Terra W	OSS U	OSS W
Released turns	529	557	518	553
Citation check failures	67 (12.7%)	10 (1.8%)	168 (32.4%)	19 (3.4%)
Audit-structure failures	67 (12.7%)	10 (1.8%)	20 (3.9%)	10 (1.8%)
Tutor generations	529	672	518	798
Delivery calls	529	1,344	518	1,596
USD/dialogue	.03204	.04469	.002263	.009366
Total delivery USD	5.77	8.04	0.41	1.69
Additional W cost, USD		2.28		1.28
W/U cost ratio		1.39		4.14

Using configured token prices, the retained experiment costs $926.64, including $897.36 for judge and offline evaluation-classifier calls. The whole project, including development and instrument tuning, costs $2,538.55. At these prices, applying this evaluation setup to a corpus of the same size and dialogue lengths would have an evaluation budget of about $897.36, separate from tutor delivery and instrument development. Over 180 dialogues per model, supervision adds $2.28 for Terra and $1.28 for OSS; per-dialogue delivery rises from $0.03204 to $0.04469 and from $0.002263 to $0.009366, giving factors of 1.39 and 4.14.

Individual conversations explain how the score relates to intervention. In the fraud example, W releases a diagnostic question after its initial proposal was flagged, and all three judges pass the ordering requirement. Majority aggregation can also leave a scenario score unchanged when some dialogues change. Appendix E follows both mechanisms through worked examples and gives the full ordering-transition counts; the evidence explorer in the supplementary material makes every scenario variation inspectable.

6 Discussion and Conclusion

The ladder gives builders a sequence of distinct design choices. Instructions raise pedagogy by 42% (+0.786 of 3), and reference context raises verifiability by 55% (+0.763) while taking auditability from 0.53 to 1.83. The full system raises pedagogy by 35% (+0.679) over the base tutor, with a 20% decrease in measured stability. Auditability and stability remain descriptive. The paired comparisons show which layer contributes to each dimension, rather than attributing the full system's behavior to supervision alone.

Runtime supervision adds 7% to the paired U pedagogy score (+0.182), inconclusive under the registered rule, while reducing released citation-ID failures from 12.7% to 1.8% and from 32.4% to 3.4%. Its delivery-cost multipliers are 1.39 for Terra and 4.14 for OSS. The changed audit and citation output shows an operational effect of checking and correction alongside the uncertain pedagogy increment. Because the final reply remains the tutor's own text, the experiment also exposes the violations that a single revision leaves in place.

The examination of the evaluation yields a substantive result: two of the four dimensions remain eligible for confirmatory analysis. Testing known failures before interpreting scores separates an architectural estimate from the instrument's ability to support it. Human ratings add evidence from naturally generated replies, while the examination supplies controlled opportunities for failure. This procedure makes the strength of each conclusion depend on the measurement that produced it.

6.1 Limitations

The present evidence concerns conceptual data science and statistics, scripted learners, and the two named tutor models. It measures adherence to teaching requirements; studying helpfulness and learning calls for interaction with students. The diagnostic rule uses the occurrence of a question as its prerequisite, and valid source identifiers establish a structural link whose semantic support is assessed separately. Runtime and evaluation classifiers share a model, and one judge shares the tutors' developer; the analysis without that judge has no complete paired pedagogy sample. The human reference includes author adjudication and AI-assisted resolutions, with factual-claim extraction coverage left for an independent enumeration study. Complete-scenario analysis and majority aggregation define which changes enter the score, and missing assessments may be systematic. The supervision effect covers classification, feedback, and extra generation together; separating their contributions requires a compute-matched comparison. Clearing a runtime flag describes the controller's decision, and the release of remaining violations warrants further assessment with students. The tutor, classifier, and judge prompts, the licensed teaching excerpts, the evidence tables, and the offline evidence explorer accompany the paper as supplementary material; the scenarios, runtime rules, and analysis code are released on acceptance.

6.2 Implications for builders

For builders, the design supports testing the prompt first, then assessing what runtime checks add against that same prompt, with both absolute costs and residual violations visible. The evaluation should face its own failure tests before its scores determine the architectural conclusion. A following study can use this reference design to test supervision with actual learners, measuring whether the observed changes in tutor behavior translate into learning.

AI USE STATEMENT

Generative AI helped refine the research questions and framing, drafted scenario content that the authors reviewed and edited, implemented and reviewed the runtime, evaluation, and analysis code, assisted literature discovery, gave feedback on the experimental design, assisted in interpreting results, and drafted and revised the manuscript. Model-based reviews informed corrections to the evaluation instrument. As part of the experiment itself, LLMs generated all tutor replies, all automated assessments, and the planted-error examination items, as described in the methods. Generative AI was not used to formulate or prove mathematical claims, to translate, or to clean or reformat datasets. We have reviewed all AI-assisted work, including the code, the numbers against the retained analysis outputs, and every cited source. Two human teaching assistants supplied the blind ratings; one of the authors adjudicated disagreements, adopting AI suggestions for three initially unresolved items, which a sensitivity analysis leaves unresolved. The authors are responsible for the contents of this submission.

ETHICS STATEMENT

The experimental learners are scripted. Two teaching assistants were paid for grading, with model identity and experimental condition hidden. The teaching sources are named in Appendix B; per-excerpt attribution and licenses are included in the supplementary material. The reference policy deliberately releases residual violations, which must be considered when designing subsequent studies with students.

REPRODUCIBILITY STATEMENT

The appendices specify the conditions, score construction, judge examination, human assessment, and statistical comparisons. Tables and figures are generated from the retained analysis outputs with input hashes. An offline evidence explorer provides the scenario variations, released conversations, checks, and judge evidence. The analysis plan was deposited externally after execution and before results were inspected.

REFERENCES

Shuyi Fan, Boyuan Deng, Mengyu Xu, Jiale Liu, Hongyang Zhang, Qiaoxin Yang, and Chongyang Gao. Rethinking LLM-judged helpfulness as a pedagogy signal: A pre-registered audit across tutor models, 2026.

Irina Jurenka et al. Towards responsible development of generative AI for education: An evaluation-driven approach, 2024.

Nizam Kadir. Auditable release control for pedagogical leakage in LLM tutors, 2026.

Mark Liffiton, Brad Sheese, Jaromir Savelka, and Paul Denny. CodeHelp: Using large language models with guardrails for scalable support in programming classes, 2023.

Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In *Proceedings of NAACL 2025, Volume 1: Long Papers*, pp. 1234–1251, 2025. doi: 10.18653/v1/2025.naacl-long.57. URL https://aclanthology.org/2025.naacl-long.57/.

Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. AutoTutor meets large language models: A language model tutor with rich pedagogy and guardrails, 2024. URL https://arxiv.org/abs/2402.09216.

Yusuf Pisan. Teaching a large language model tutor to withhold the answer: A supervisor architecture and an evidence-driven method for tuning socratic behavior, 2026. URL https://arxiv.org/abs/2608.12292.

Abu Syed. TEAS: Trusted educational AI standard — a framework for verifiable, stable, auditable, and pedagogically sound learning systems. In *Deployable AI Workshop at the AAAI Conference on Artificial Intelligence (AAAI-26)*, 2026.

Pat Verga, Sebastian Hofstätter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models, 2024.

APPENDIX

A EXPERIMENTAL MATERIALS AND MODEL ROLES

The retained corpus has twenty scenarios in twelve topic families: Bayes' theorem (four), bias–variance (three), overfitting (two), class imbalance (two), regression coefficients (two), and one each on p-values, confidence intervals, cross-validation, data leakage, gradient descent, regularization and correlation. Fifteen scenarios contain numerical context; none requires the tutor to perform a computation. Each scenario has a fixed source registry, knowledge nodes, misconception inventory, pedagogical contract, assertions and learner script. Scenarios were funded in blocks of five under the registered cost-based stopping rule; four blocks ran. Table A2 lists every scenario and its available U/W pedagogy scores.

Table A1: Model roles. The provider prefix in retained identifiers is `azure_ai/`. Separate classifier prompts do not imply separate models.

Role	Retained model
Tutor, proprietary	`gpt-5.6-terra`
Tutor, open weights	`gpt-oss-120b`
Judge panel	`gpt-5.6-sol`, `Kimi-K2.6`, `DeepSeek-V4-Pro`
Runtime action classifier	`gpt-5.6-luna`, gate prompt
Evaluation action classifier	`gpt-5.6-luna`, evaluation prompt

Requests use provider-default temperature and a registered medium reasoning setting, with provider-specific transport recorded. These are requested settings and recorded bindings, not a claim of equivalent internal inference across models. The experiment records repetition indices and passes seeds where supported; identical indices do not guarantee identical random draws. The launch procedure probes reasoning transport and binds its receipt to configuration and prompt hashes.

Condition contents. B0 elicits audit metadata in addition to a minimal tutor role, as part of its base response requirements. P adds pedagogical policy lines rendered from the contract, without reference knowledge or citation instructions. U includes source materials, knowledge-node identifiers, known misconceptions, pedagogical policy, citation instructions and the audit schema. W uses that exact initial prompt and enables validation, action classification and one combined corrective regeneration. The knowledge and schema additions in U–P are a bundle. Likewise, W–B0 combines prompting, reference material and runtime intervention. Only W–U is the architectural contrast under identical initial prompting.

Runtime sequence. For each learner turn, the tutor proposes a response. The wrapper checks citations, the audit block and the permitted action under the gate's classified history. If all checks pass, it releases the proposal. Otherwise it sends merged failure feedback to the tutor, obtains one revision, rechecks all three components and releases that revision with residual flags. Audit-stripped visible text is used for action classification. The learner state machine selects its next scripted utterance using the released action; termination can therefore follow different paths under U and W. Such a terminal state is a property of the simulation, not observed mastery.

The pedagogical rule vocabulary represents prohibitions, prerequisites and exceptions. A diagnostic event is satisfied by a prior tutor action label. Checking that event implements a temporal requirement; it does not independently check the quality of the diagnosis. Checks on citation identifiers and audit structure are deterministic, whereas the pedagogical action label is an LLM judgment. The gate and the assessment panel have distinct roles: the panel's scores do not supply runtime feedback in this experiment.

B SCORE CONSTRUCTION AND MEASUREMENT UNITS

Verifiability. Each judge extracts factual claims and assesses support against scenario knowledge. Its claim-support rate is scaled by three; the panel takes the median of the available judge rates under

Table A2: Every retained scenario, with U/W pedagogical scores. A dash means the score is unavailable under the frozen completeness rule. The final column identifies the common primary subset. Scenario IDs omit the shared `ds_` prefix.

	Terra		OSS		Primary
Scenario	U	W	U	W	pair
bayes_fraud_alert_002	2.000	3.000	2.000	2.000	Yes
bayes_hiring_flag_004	3.000	3.000	–	2.000	No
bayes_screening_baserate_001	–	3.000	2.000	2.000	No
bayes_spam_inversion_003	3.000	3.000	–	3.000	No
bv_energy_deeper_001	2.000	2.000	2.000	3.000	Yes
bv_more_data_bias_004	3.000	3.000	3.000	3.000	Yes
bv_zero_error_goal_002	–	3.000	3.000	3.000	No
ci_ab_lift_zero_003	2.000	3.000	2.000	2.000	Yes
corr_dropped_feature_004	3.000	2.000	2.000	2.000	Yes
cv_best_fold_model_104	3.000	3.000	–	–	No
gd_batch_noise_003	2.000	2.000	–	2.000	No
imb_auc_prevalence_104	–	2.000	–	2.000	No
imb_moderation_recall_101	3.000	–	–	–	No
leak_feature_select_all_data_103	3.000	3.000	3.000	3.000	Yes
linreg_coef_causal_001	3.000	3.000	2.000	3.000	Yes
overfit_eval_on_train_003	3.000	3.000	–	3.000	No
overfit_train_longer_005	3.000	3.000	3.000	3.000	Yes
pval_checkout_ab_003	2.000	3.000	2.000	2.000	Yes
reg_credit_trainerr_001	2.000	3.000	2.000	2.000	Yes
regcoef_bedrooms_sign_002	3.000	3.000	3.000	2.000	Yes

the registered fallback rules. Unsupported claims, contradictions and non-supporting citation pairs are retained as separate diagnostic atoms. Differences in extracted claim sets mean that the score is conditional on extraction. Explicit source-ID compliance and semantic factual support are separate measurements. The intended guard on differences in replies with no factual claims is unavailable for the W-related contrasts because two W scenario/model cells lack required atoms. An unavailable guard is not a passed guard.

Auditability. The final score is three times audit availability times trace accuracy. Accuracy includes the audit note's match to the scenario misconception, support for its declared citation pairs and resolution of its named sources against the scenario registry. Omitted grounding receives zero under the same strict rule in all arms. Availability is programmatic; semantic accuracy uses judges. As B0 and P do not receive U/W's source registry, complete-system differences on this score also reflect the supplied context and elicitation contract. Audit notes are observable model-generated metadata, not proof of faithful internal reasoning.

Stability. Epistemic consistency compares fixed knowledge-node labels, entailed, contradicted or omitted, across repetitions and rephrasings. Its score is three times the minimum of seed agreement and paraphrase agreement. Strategy consistency uses evaluation-classifier action sequences, retaining repetitions and termination symbols, with similarity given by one minus normalized Levenshtein distance. The pillar takes the minimum of the scaled components. Raw strategy agreement is descriptive; the registered inferential strategy quantity is a separate chance-corrected agreement against a leave-one-scenario-out marginal-policy null. Failure of the held-out stability detectors makes the pillar descriptive here, irrespective of an interval's numerical direction. It measures consistency at the sampled settings, not longitudinal stability under model updates.

Pedagogy. The two single-dialogue assertions concern ordering of help and the declared prohibition on endorsing the specified misconception. Judges assess each fixed assertion with transcript evidence. A majority across three repetitions determines a profile's assertion value; a majority across three profiles determines its scenario value. A separate panel comparison of the seed-zero novice and advanced dialogues supplies the adaptivity assertion. If b_1, b_2, b_3 are these three binary values, the scenario score is $3(b_1 + b_2 + b_3)/3$. If a required dialogue judgment or adaptivity judgment is unresolved, the entire cell is unavailable. This rule was not replaced after seeing its consequences.

Panel aggregation. Fixed categorical assertions use majority voting. With only two available judgments, agreement must be unanimous. Unresolved items remain missing. For quantitative components, the registered fallback accepts nearby counts conservatively and nearby fractional audit accuracies by their mean; disagreement beyond the specified tolerance is unresolved. Thus a three-judge panel does not eliminate abstention or missingness. Verifiability claim rates and audit block accuracies have their own aggregation, and the pillar scores are rebuilt from those components. We do not average three final pedagogy scores.

Teaching examples inside the judge prompt. The pedagogy prompt includes worked positive/negative examples of the requirements it asks the judge to assess. In each pair, a licensed teaching excerpt illustrates a structural teaching property, and an authored alternative on the same topic violates that property. For example, a pair can distinguish eliciting a learner's understanding before explaining from supplying an explanation immediately. The prompt names the assertion IDs associated with each pair and instructs the judge to attend to the order and content of teaching moves rather than warmth or writing style.

The positive pool contains ten excerpts of at most 150 words from eight documents across MIT OpenCourseWare, Harvard CS50, Khan Academy, The Carpentries, and Cross Validated. Six pedagogical negatives form the calibration pairs, including a style-mimicry example whose teaching structure has been removed. A seventh negative concerns verifiability and is excluded from pedagogy calibration. Each pedagogy call contains two pairs selected for the assertions being assessed, with their identifiers and hashes retained. The property definitions were refined against practice outputs and frozen before the held-out examination. These examples teach the judges how to apply the requirements; examination items are separate.

C Judge Examination and Human Validation

C.1 Known-error constructions and execution

The six constructions target factual support (M-V), citation support (M-C), pedagogy (M-P), epistemic consistency (M-SE), audit accuracy (M-A) and action-sequence consistency (M-SS). Mutants are paired with harmless edits and unmodified material. Detection receives credit only under the required error/control pattern, not merely because a judge changes a score after an edit. M-SS additionally uses repeated classification of unmodified material to measure control instability. The final materialization contains 520 items; 56 of 576 projected constructions exhausted their construction attempts and are distinct from detector misses.

Practice and held-out material are separated. A corrected practice examination allowed exactly one logged judge-prompt repair round, followed by freezing. Code refuses a second such round. The held-out examination completed all 520 items with judge prompts and detector thresholds unchanged after freezing. Verdict-bearing outputs remained sealed until authorized analysis.

The primary missing-measurement rule withholds detection credit when required observations are absent and treats missing control observations unfavorably. Two M-P and 26 M-SS items have missing required measurements. For M-SS, the amended primary detection is 36/69 and its control-alarm bound is 36/69. The observed unmodified-control count is 28/69; the bound must not be reported as an observed false-alarm rate. The original missing-measurement treatment gives 51/69 detections and 28/69 unmodified-control alarms. Neither treatment passes the detector. The detector eligibility decisions agree under the original and amended treatments.

C.2 Human sample, adjudication and analysis

Two teaching assistants independently graded sixty conversations for the designated assertions, declared citation pairs and audit notes, and one hundred short exchanges for action labels. Pages excluded tutor identity, condition, machine verdict, score and sampling metadata. The sampling probabilities were retained for inverse-probability weighting. Disagreements were adjudicated by one of the authors while working from the prepared blind interface.

Three initially unresolved Task A items were resolved using AI suggestions adopted by the author. The analysis retains a sensitivity leaving them unresolved. Weighted pass recognition changes from

85.948% to 85.981% for pedagogy and from 83.995% to 82.933% for audit notes; failure recognition and validation designations do not change. These values are taken from the retained machine-readable adopted and unresolved-reference outputs. The sensitivity isolates the effect of those three resolutions on the reported rates.

One human item asked whether citations were visible in the reply, while the machine score asked whether the cited source supported the claim. These are different questions, so all 60 such items were excluded before comparison, regardless of how they had been answered. The derived "all cited claims are supported" summary is kept separate from its constituent claim–source pairs. The resulting analysis contains 120 pedagogy assertion units (104 Pass, 16 Fail), 219 declared citation pairs (216 Pass, 3 Fail), and 151 audit notes (137 Pass, 14 Fail), all within the sampled conversations.

The retained sample contains too few failures for the protocol's natural-data validation requirements. The human comparison uses the existing sample without a supplementary draw and retains the insufficient-validation designation. The independent factual-claim enumeration in fifteen conversations was not performed, so coverage of automated claim extraction remains unchecked. Neither absence is filled by the planted-error examination.

Table A3: Design-weighted human agreement with 95% family-cluster bootstrap intervals. Raw recognition fractions differ because inclusion probabilities are unequal. Boundary intervals reflect the observed sample, not certainty about population detection.

Unit	Pass recognition [95%]	Fail recognition [95%]	Raw Fail recognition
Pedagogy	86.0 [75.6, 95.4]	98.2 [92.7, 100.0]	14/16
Citation pair	82.6 [78.2, 89.2]	35.7 [0.0, 50.0]	1/3
Audit note	82.9 [76.4, 89.0]	100.0 [100.0, 100.0]	14/14

The citation failure interval uses 8,947 defined replicates of 10,000; the others have no reference failure unit. For audit failures, all observed units agree and a resampling interval degenerates. An ordinary exact binomial interval for 14/14 is approximately [0.768, 1.000]; it is not a design-weighted replacement, but illustrates why the observed perfect fraction does not establish perfect detection. The separately retained joint human/architectural bootstrap gives a pedagogy-effect interval [0.000, 0.375]. It does not replace the primary registered interval.

The evaluation action classifier matches 76/100 final reference labels, with an exact interval of approximately [66.4%, 84.0%]. By reference class, it identifies 59/61 direct answers, 11/15 diagnostic questions, 0/6 scaffolding turns, 3/13 explanation-then-check turns and 3/5 other actions. Original TA–TA agreement is 73/100. The 27 available gate-labeled turns provide 14 matching action labels and 27 matching rule decisions when human labels are substituted into the recorded history. Since only one reference decision is a flag, this is a limited check of final-turn decisions, not an established false-release rate for all initial attempts or histories.

D Complete-Pair Availability and Inferential Results

All 1,440 dialogues and 800 probes remain in the retained corpus. Missing score availability is distinct from failed execution or scenario exclusion. Of 42 unresolved dialogue pedagogy assessments, 37 combine one absent judge with disagreement between the other two; five lack two judges. Across individual dialogue pedagogy judgments, absences number 121 for DeepSeek, 25 for Kimi and two for Sol. Table A4 gives the propagation under the frozen rule. All 160 cross-profile adaptivity judgments resolve, so they do not cause the primary loss.

D.1 Both analysis families

Section 4 specifies the two prespecified test families. Both were fixed before any result was inspected, and both are reported with their own Holm adjustment. The tables show surviving tests and descriptive dimensions. "Terra–OSS" pools B0/U/W within the named-model comparison. The column p_5 retains all five positions, assigning excluded tests $p = 1$.

The named-model pedagogy difference is +0.250, interval [0.033, 0.476], over eight scenarios in seven families. It meets the positive-effect rule in both families. The original W–B0 Verifiability

Table A4: Unresolved dialogue assessments out of 180, followed by computable scenario scores out of twenty. Availability differs by arm.

	B0	P	U	W
Terra unresolved dialogues	0	7	3	1
Terra available scenarios	20	14	17	19
OSS unresolved dialogues	4	17	7	3
OSS available scenarios	16	9	13	18

Table A5: Amended family: instrument-eligible tests. n is complete scenarios and k topic families. Decisions also require the effect, interval and model-direction conditions.

Contrast	Measure	Difference	95% interval	n	k	p_{Holm}	p_5
W–U	Verifiability	0.038	[-0.048, 0.097]	20	12	0.171	0.513
W–U	Pedagogy	0.182	[0.000, 0.364]	11	9	0.099	0.198
Terra–OSS	Pedagogy	0.250	[0.033, 0.476]	8	7	0.025	0.041

difference is +0.851, interval [0.640, 1.077], and meets that family's numerical positive-effect rule. Extraction coverage and the unavailable no-claims guard qualify interpretation, and the contrast is not the runtime effect. In the amended plan W–B0 is an estimation-only secondary contrast.

The primary pedagogy interval is obtained by resampling topic families, retaining paired model/arm differences within sampled families. A scenario-bootstrap sensitivity gives $[-0.045, 0.409]$ and remains inconclusive. The one-sided tail quantity counts bootstrap deltas at or below zero with the registered finite-replicate correction. It is then adjusted within the specified family; it is not a substitute for the declared effect threshold or interval rule. The bootstrap-spread-based 80% power diagnostic is approximately 0.240 for the primary pedagogy comparison, exceeding the 0.20 threshold. This is a scale diagnostic from the realized sample, not a prospective power guarantee.

Prompt sufficiency and sensitivities. U's pedagogy means are 2.647 across seventeen available Terra scenarios and 2.385 across thirteen OSS scenarios. The separate W–U equivalence comparisons use sixteen and thirteen pairs, respectively; their intervals are reported in the main text. Subtracting unpaired arm means would not reproduce the paired estimand. Excluding three designated development families gives +0.200 with interval [0.000, 0.389], still inconclusive. Removing the OpenAI-family judge yields no complete pedagogy pairs under two-judge unanimity and downstream completeness requirements; it is an unavailable sensitivity, not evidence of robustness. Verifiability in that exclusion analysis has difference +0.072 over eighteen scenarios and remains within the equivalence margin.

An exploratory numeric-context split gives +0.278 [0.100, 0.438] in nine complete scenarios containing numerical context and -0.250 $[-0.500, 0.000]$ in two without it. These small, unequal strata do not establish a moderator. The scenario-level correlation between pedagogy changes and response-length changes is -0.079 across the eleven primary pairs; this is not a length-adjusted causal analysis. No such diagnostic replaces the registered effect.

E Worked Traces and Resource Accounting

A diagnostic question before explanation. For example, in `ds_bayes_fraud_alert_002`, Terra, novice profile, repetition zero, the learner confuses a 95% fraud catch rate with the probability that a flagged transaction is fraudulent. The strict scenario contract requires a diagnostic question before explanation. U explains the conditional-probability distinction and then asks a question. The branch classifier labels the response a diagnostic question, but Sol and Kimi fail the ordering assertion; DeepSeek passes it using the later question. This illustrates disagreement about sequence rather than the presence of any question.

For W, the recorded initial attempt is classified as explaining then checking and fails the citation, audit and gate checks. One combined retry follows. The released reply first asks how the learner inferred flag reliability from catch rate; all three judges pass ordering. Citation and audit checks

Table A6: Original family: instrument-eligible tests. W–B0 combines prompt, reference context and runtime machinery.

Contrast	Measure	Difference	95% interval	n	k	p_{Holm}	p_5
W–B0	Verifiability	0.851	[0.640, 1.077]	20	12	$< .001$	$< .001$
W–U	Pedagogy	0.182	[0.000, 0.364]	11	9	0.050	0.149
Terra–OSS	Pedagogy	0.250	[0.033, 0.476]	8	7	0.017	0.033

Table A7: Descriptive dimensions excluded by detector failure. Numerical intervals do not override instrument eligibility.

Contrast	Measure	Difference	95% interval	Scenarios
W–U	Auditability	+0.120	[0.032, 0.201]	20
W–U	Stability	-0.009	$[-0.105, 0.118]$	19
W–B0	Auditability	+1.406	[1.284, 1.507]	20
W–B0	Stability	-0.420	$[-0.643, -0.182]$	18

still fail on that released turn, which is released and recorded according to policy. The rejected W proposal's text was not retained, only its hash and check record. U is a separate generation, not W's rejected proposal. This case cannot isolate which part of the combined feedback caused the changed response.

Across that scenario's nine Terra dialogues, U has zero ordering passes and W eight. The other scenario components pass in both arms, giving scores two and three. OSS remains at two in both arms, so the pooled scenario contribution is +0.5. This is a selected explanatory example, not the average treatment effect or evidence of learning. Negative scenario contributions are also retained: Terra drops from three to two on the correlation/dropped-feature scenario, and OSS drops from three to two on the regression-coefficient/bedrooms scenario. The score changes alone do not identify their causes.

Aggregation and dialogue transitions. The primary subset has six positive, two negative and fourteen zero model/scenario changes. Majority aggregation can mask either improvement or deterioration that does not cross a boundary. The OSS bias–variance/more-data example in the main text demonstrates the former. The full matched ordering transition table is 76 Fail-to-Pass, 43 Pass-to-Fail, 150 Pass-to-Pass, 78 Fail-to-Fail and thirteen unavailable pairs. All 360 pairs are represented. They compare independent generated trajectories with matched experimental indices, not an initially rejected reply with its own revision.

Costs. Main-text delivery costs include tutor generation and the runtime classifier but exclude research judges and the evaluation classifier. Across the 180 U and 180 W dialogues per model, estimated delivery costs are \$5.767043 and \$8.043822 for Terra, and \$0.407307 and \$1.685898 for OSS. Stability probes are separate: each U/W arm has one hundred probes, with costs \$0.710825/\$0.922967 for Terra and \$0.061597/\$0.228596 for OSS. Probe records have generation counts but no explicit retry counter; derived extra generations should not be read as recorded retry counts.

Completing the 160 adaptivity assessments required 480 initial judge calls and five format retries, costing an estimated \$7.189086. Including research evaluation, the retained experiment's ledger totals \$926.636271; the project-wide ledger totals \$2,538.547150. The totals separate the retained experiment from wider project development and validation expenditure.

F Supplementary Material and Evidence Explorer

The supplementary material contains four parts. First, every prompt used in the study, verbatim: the tutor prompts for the four conditions, the runtime and evaluation action-classifier prompts, and the judge prompts for each dimension. Second, the licensed teaching excerpts and authored flawed alternatives that appear as worked examples inside the pedagogy judge prompt, with the source,

license, and access date of every excerpt. Third, the data behind Table 1: for every contrast and dimension, the scenarios used, the paired lower-arm and higher-arm means, the difference and interval, the relative change, and the change in points on a 100-point scale. Fourth, the offline evidence explorer. The scenario files, the runtime rule engine, the analysis code, and the raw judge responses are released on acceptance.

The explorer is a single HTML file that runs in a browser without a server or network connection. It contains all 1,440 dialogues and 800 probes. Selecting a scenario, model, condition, learner profile, and repetition shows the released conversation, the audit metadata, the runtime and evaluation action labels, each check attempt and revision, the panel verdicts, and the score aggregation. A comparison view aligns experimental selections, which need not have identical dialogue histories. Missing responses and unavailable scores are shown explicitly. The explorer also exposes the learner branch structure and the verified initial prompts.

The explorer does not contain the text of rejected tutor proposals, which was not retained, only their hashes and check records. Generated audit notes are model output, not private reasoning traces. Every record carries the identifiers and hashes needed to follow an example to its source.