When Tutor Errors Match Learners' Misconceptions: A Congruency Effect in LLM Tutoring and a Scaffolding Intervention

Yefeng Wu[1] wuyefengflc@163.com
Yangyang Liu[1] 18755315206@163.com
Bao Shi[1] 15056447598@163.com
Yufan Li[1] p125304126@stu.ahu.edu.cn
Yan Zhang[1,*] 06027@ahu.edu.cn

[1]Electronic Science and Technology, Anhui University, Hefei, China
[*]Corresponding author

Abstract

Errors in LLM-style tutoring are not uniformly harmful. We test a *misconception-congruency effect*: when an explanation's key error aligns with a learner's pre-existing misconception, the error is harder to detect and more likely to be adopted. We introduce *epistemic scaffolding*—claim decomposition with evidence/counterexample selection—to shift evaluation toward evidential cues. In a controlled experiment with engineering students ($N = 124$) using researcher-written feedback that *simulates* LLM explanations, congruent errors were detected less often than incongruent errors (36% vs. 58%). Scaffolding improved congruent-error detection to 62% and reduced adoption, while signal detection analyses showed higher sensitivity (d') without increased response bias. These findings suggest learner-aware safeguards should target the interaction between tutor errors and learner priors.

Keywords: LLM tutoring; misconceptions; epistemic vigilance; trust calibration

Introduction

LLM-based tutoring systems can provide fluent, interactive explanations at scale, but their errors are often *plausible* (Deng et al., 2025; Laskar et al., 2023; OpenAI, 2023; Shi et al., 2026). A central question for cognitive science is therefore not only whether an AI tutor is wrong, but *when its wrongness is most likely to be accepted and integrated into learning*.

We focus on a specific risk pattern: tutor errors that align with a learner's existing misconception (Aleknaviciute et al., 2023; Posner et al., 1982). Consider a student learning sampling and aliasing who already believes that aliasing can be "fixed later" by interpolation. A tutor that (incorrectly) endorses this idea may feel trustworthy because it matches the student's intuition; the student may then adopt the explanation and further entrench the misconception (Huschens et al., 2023; Kahneman, 2011; Udry & Barber, 2024). In contrast, an equally wrong explanation that conflicts with the student's intuition may be easier to doubt.

We formalize this intuition as the *misconception-congruency effect*. The key idea is that misinformation risk is not uniform across outputs: it depends on the relationship between the output and the learner's prior belief structure. Throughout, we use "hallucination" as shorthand for *hallucination-like tutor errors* (misinformation in tutor feedback). To enable precise experimental control, we study this using researcher-written stimuli that simulate LLM explanations rather than live model generation. We call errors that align with misconceptions *misconception-congruent hallucinations* (Dhuliawala et al., 2024; Manakul et al., 2023; OpenAI, 2023).

We make three contributions. First, we define a congruency manipulation anchored in an individualized misconception profile, enabling within-participant comparisons without subjective researcher labeling. Second, we propose a minimal *epistemic scaffolding* intervention: rather than telling learners to "be skeptical," the interface induces a small set of epistemic actions (claim checks) that make verification more likely (Chi et al., 1994; Johnson et al., 1993; Liu et al., 2026; Sperber et al., 2010). Third, we provide evidence that congruent errors are detected less often and adopted more often at baseline, and that simple claim-check scaffolding mitigates this congruency disadvantage.

Background: Misconceptions and Epistemic Evaluation

Our proposal combines two ideas: (i) misconceptions are structured priors that can shape interpretation and (ii) epistemic evaluation integrates multiple cues, not just objective correctness.

Misconceptions as structured priors

In conceptual domains, learners' errors are often systematic. A misconception is not merely a missing fact; it is a coherent (but incorrect) explanatory structure that can generate predictions and guide interpretation. This matters for tutoring because a fluent explanation that "fits" a misconception may appear coherent from the learner's perspective. Congruency is therefore not a superficial similarity between texts, but a match between an explanation's key proposition and the learner's pre-existing causal story (Aleknaviciute et al., 2023; Posner et al., 1982).

Epistemic evaluation and source monitoring

When learners receive instructional feedback, they must decide how much to trust it and whether to use it in subsequent reasoning. This decision depends on epistemic evaluation processes such as source monitoring (where did this knowledge come from?), epistemic vigilance (should this information be

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

scrutinized?), and metacognitive monitoring (how sure am I?) (Johnson et al., 1993; Liu et al., 2026; Sperber et al., 2010). In practice, learners often use heuristics because careful verification is effortful (Kahneman, 2011; Udry & Barber, 2024). Two relevant cues are *fluency* (ease of processing) and *congruency* (consistency with what one already believes). LLM tutoring can amplify both cues: explanations are typically well-formed and confident, and personalization may further increase perceived fit (Geng et al., 2024; Huschens et al., 2023; Li et al., 2024; OpenAI, 2023).

From trust judgments to adoption and learning

Trust can be expressed as a judgment ("is this correct?"), but the educational consequence depends on behavior: whether the learner *adopts* the proposition when explaining the concept later. Adoption is critical because it represents the moment a misconception becomes reinforced by a new information source. In our framework, the same mechanism that reduces error detection can increase adoption; over time, adoption can reduce conceptual change and produce persistence of the misconception.

Related Work and Positioning

Misconceptions and conceptual change. Research on misconceptions in STEM learning emphasizes that learners hold stable, theory-like beliefs resistant to correction. Our contribution treats misconceptions as a pre-existing cognitive substrate that interacts with LLM tutors, operationalizing this interaction in an experimentally testable way (Aleknaviciute et al., 2023; Posner et al., 1982).

Epistemic cognition and trust calibration. Work on epistemic cognition shows that learners rely on heuristics (coherence, fluency) when verification is costly. LLM tutoring amplifies these cues by making fluent explanations ubiquitous. Human–AI interaction research investigates trust calibration, but typically focuses on system-side interventions. Our approach complements this by changing the learner's epistemic actions via minimal scaffolding (Johnson et al., 1993; Li et al., 2024; Liu et al., 2026; Sperber et al., 2010).

LLM tutoring and tutor errors. Recent work on LLM tutoring reports learning benefits (Deng et al., 2025; Shi et al., 2026). However, LLM tutoring is often evaluated using accuracy metrics, and two errors with identical wrongness can have different consequences depending on learner priors. The misconception-congruency effect makes this dependency explicit, motivating a shift from "reducing error rate" to "reducing harmful error adoption" (Chan, 2023; Dhuliawala et al., 2024; Fulsher et al., 2025; Laskar et al., 2023; Manakul et al., 2023; OpenAI, 2023; Wang et al., 2025).

Theory: The Misconception-Congruency Effect

We treat LLM tutor feedback as an information source whose relationship to the learner's misconception can be experimentally manipulated. Figure 1 illustrates the proposed dual-process mechanism: congruent errors trigger fluent, heuristic processing (System 1) that suppresses verification, whereas incongruent errors induce disfluency and analytic evaluation (System 2) (Kahneman, 2011). Fluency and ease of processing can increase perceived truth and reduce scrutiny (Udry & Barber, 2024). Confidence cues can further increase overreliance and miscalibrated trust (Geng et al., 2024; Li et al., 2024; Stengel-Eskin et al., 2024; Zhang et al., 2024). The epistemic scaffolding intervention intercepts the heuristic path, rerouting processing toward deliberate verification, thereby engaging epistemic vigilance (Liu et al., 2026; Sperber et al., 2010).

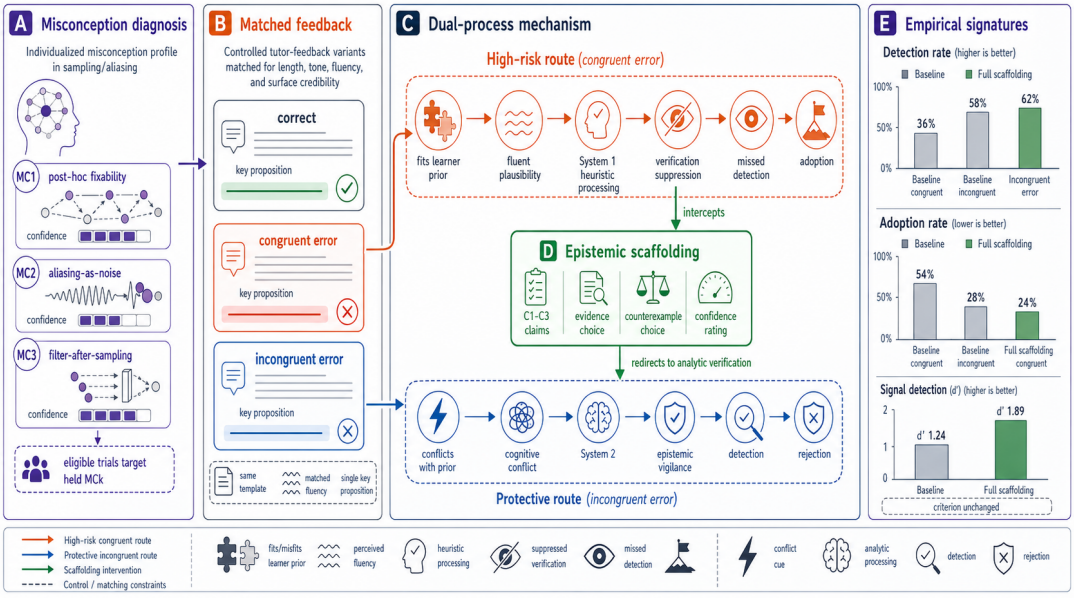

Figure 1: Dual-process model of misconception-congruency effect in LLM tutoring.

Operational definitions

We first diagnose a learner's misconception profile (which misconceptions they hold and how strongly). For a tutoring trial targeting misconception MCk, we construct two erroneous feedback variants:

- **Misconception-congruent error:** the *key incorrect proposition* matches the core belief of MCk ("same direction").
- **Misconception-incongruent error:** the key incorrect proposition conflicts with the core belief of MCk ("different direction"), while remaining incorrect.

To preserve surface plausibility, incongruent errors were constructed as a denial plus an alternative (still-wrong) mechanism rather than simply stating the correct concept. Crucially, congruency is defined *relative to the learner's diagnosed MCk*; analyses include only trials where the learner held the targeted misconception.

We distinguish three outcomes: *error detection* (identifying the feedback as wrong), *adoption* (using the key proposition in subsequent explanation), and *learning* (posttest performance on targeted concepts).

Hypotheses

H1 (Misconception-congruency). Compared to misconception-incongruent errors, misconception-congruent errors yield (i) lower error detection, (ii) higher adoption of the key erroneous proposition, and (iii) poorer conceptual learning for the targeted misconception.

H2 (Mechanism: verification suppression). Congruency increases perceived plausibility and reduces verification intensity, observable through shorter reflection times and lower accuracy on evidence/counterexample choices.

H3 (Intervention). Epistemic scaffolding reduces the congruency advantage by increasing attention to evidence and counterexamples, producing a congruency×scaffolding interaction: the largest improvements (higher detection, lower adoption) occur under misconception-congruent errors.

H4 (Calibration). Scaffolding improves confidence–accuracy calibration (reduced overconfidence and better alignment between confidence and correctness), especially in congruent-error trials where overconfidence is otherwise most likely.

Intervention: Minimal Epistemic Scaffolding

The intervention is designed to be lightweight: it does not require multi-agent tutoring, extensive external resources, or long dialogues. Instead, it inserts small epistemic actions that change how learners evaluate the tutor output.

Claim-check scaffolding

Each explanation is expressed with a fixed structure and decomposed into three checkable claims: C1 (mechanism; contains the key proposition), C2 (consequence; visually testable), and C3 (action; mitigation). Learners then complete one micro-evaluation per claim: (i) choose which of two short pieces of evidence better supports the claim, or (ii) choose which of two scenarios would refute it, followed by a 0–100 confidence rating. This design instantiates claim-focused self-explanation and forces minimal source monitoring and counterexample search without asking participants to consult external materials (Chi et al., 1994; Johnson et al., 1993; Liu et al., 2026; Sperber et al., 2010). The workflow also parallels model-side self-checking approaches, but implemented as learner-facing verification (Dhuliawala et al., 2024; Manakul et al., 2023).

Why this should work

The scaffolding is intended to shift epistemic evaluation from "does this fit my intuition?" to "what evidence would make this true or false?" Mechanistically, it should increase the effective weight of evidential cues and decrease reliance on congruency as a heuristic cue. It also yields process measures (evidence-choice accuracy, per-claim confidence, interaction time) that allow us to test the proposed mechanism rather than relying solely on outcome measures.

Controls and yoking

Because scaffolding changes both *process* and *effort*, we included lightweight controls that match time-on-task and interface complexity while removing the hypothesized epistemic action. Table 1 summarizes the control conditions employed. A warning-only control tests whether labeling alone is sufficient to reduce harmful adoption of plausible tutor errors (Rand et al., 2023).

Yoked snippets. To disentangle epistemic prompting from mere exposure, snippet exposure was controlled: Full Scaffolding and Time-matched Placebo saw identical snippet pairs but made different judgments (epistemic vs. non-epistemic). Thus any advantage of Full Scaffolding reflects the evaluation *process* rather than time-on-task or content exposure alone.

Methods

Design overview

We used a three-stage pipeline: (1) diagnose misconception profiles, (2) present tutoring trials with controlled feedback variants, (3) measure detection, adoption, calibration, and learning.

Evaluation interface was manipulated between participants (five conditions; Table 1). Our primary planned contrast is Baseline vs. Full Scaffolding; the other three conditions are mechanism-matched controls to rule out warning labels, structure alone, and time/snippet exposure. Congruency was manipulated within participants across error trials that targeted misconceptions they held. A correct-feedback condition anchored learning and reduced an "all feedback is wrong" demand characteristic.

The pretest included 9 diagnostic items (3 per misconception). The tutoring stage included 9 tutoring prompts (3 per misconception), each with three feedback variants (correct, congruent error, incongruent error). To reduce fatigue while

Table 1: Between-subject controls to isolate epistemic action from time and content.

Condition	Key elements (purpose)
Baseline	Correctness judgment + overall confidence (minimal evaluation).
Warning-only	Baseline + brief warning that the tutor may be wrong (tests label sufficiency).
Decomposition-only	Present C1–C3 and per-claim ratings, but no evidence/counterexample choice (tests structure without evidence work).
Time-matched placebo	Same screens/snippet pairs as full scaffolding, but prompts are non-epistemic (e.g., "which is clearer?") (controls time/effort and exposure).
Full scaffolding	Claim decomposition + per-claim evidence/counterexample selection + per-claim confidence (hypothesized mechanism).

Table 2: Operationalized misconceptions for the sampling/aliasing domain.

MC	Operational definition (typical belief)
MC1	Post-hoc fixability: aliasing can be removed after sampling via interpolation or "super-resolution."
MC2	Aliasing-as-noise: aliasing is mainly noise/quantization rather than frequency folding and indistinguishability.
MC3	Filter-after-sampling: anti-aliasing can be done primarily in the digital domain after sampling.

preserving within-participant contrasts, each participant received a mixture of conditions: for each misconception they held, they saw both congruent and incongruent error trials across different items, plus a subset of correct-feedback trials and a small number of fillers.

Condition assignment was randomized with constraints: (i) each targeted misconception contributed at least one congruent and one incongruent error trial for eligible participants, (ii) no two adjacent trials targeted the same misconception when possible, and (iii) overall trial order was fully randomized within these constraints. Each tutoring item had a unique posttest counterpart linked by misconception (MC1–MC3), item identifier, and key proposition, enabling trial-level adoption analysis. Congruency analyses include only trials targeting misconceptions the participant held.

Domain and misconception taxonomy

We focus on sampling and aliasing, where misconceptions are stable and directionally constrained. We operationalize three misconceptions (MC1–MC3) that can be reliably diagnosed:

Participants and inclusion criteria

Participants & Misconception Profile: A total of **124 undergraduate engineering students** (Mean Age = 20.4, SD = 1.8; 48% female) completed the study. Participants were randomly assigned to one of five between-subject conditions (Table 1): Baseline (n = 25), Warning-only (n = 25), Decomposition-only (n = 24), Time-matched Placebo (n = 25), and Full Scaffolding (n = 25). All participants passed pre-specified attention checks. (Recruitment/compensation/platform/duration/ethics approval details: *to add in final/anonymity-compliant form*.)

Diagnostic pretests revealed a substantial prevalence of the targeted misconceptions:

- **MC1 (Post-hoc fixability):** Held by **68%** of participants.
- **MC2 (Aliasing-as-noise):** Held by **42%** of participants.
- **MC3 (Filter-after-sampling):** Held by **31%** of participants.

On average, each participant held **1.41 misconceptions** (SD = 0.9) with mean diagnostic confidence **74.2/100**. Because each held misconception contributed one congruent and one incongruent error trial, participants contributed 1–3 eligible trials *per error type* (2–6 total), yielding 175 eligible congruent trials and 175 eligible incongruent trials across the sample.

Materials

Stimuli construction approach. Tutor feedback was researcher-constructed to simulate LLM-generated explanations while enabling precise experimental control. This approach was chosen over real-time LLM generation to ensure: (i) matched linguistic properties across conditions, (ii) controlled error types aligned with specific misconceptions, and (iii) replicability. We discuss the implications of this controlled approach for generalization to real LLM tutoring in the Discussion & Limitations section.

Pretest (diagnosis). Each misconception was measured with three multiple-choice diagnostic items with a dedicated misconception distractor and a 0–100 confidence slider. A participant was classified as holding MCk if they chose the MCk distractor on ≥ 2 of 3 items and had mean confidence $\geq 60/100$ on those items. This yielded an individualized misconception profile.

Tutoring items. For each misconception MCk, we prepared a small set of tutoring prompts that each targeted only that misconception (to avoid mixed manipulation). Each prompt had three feedback variants: correct, congruent error, and incongruent error.

Feedback template and control. All feedback variants followed a fixed five-sentence template (one-sentence conclusion, two-sentence mechanism explanation, one exam-

ple/visual interpretation, one action recommendation). This structure mirrors the explanatory style commonly observed in LLM tutoring outputs. Congruent and incongruent errors differed *only* in the key proposition within the mechanism sentence. We matched length, hedging, punctuation, terminology, and tone across variants to reduce confounds.

To make the manipulation explicit and auditable, each erroneous variant was annotated with (i) the targeted misconception (MC1–MC3), (ii) the key proposition, and (iii) a short "direction" rationale explaining why the proposition is congruent or incongruent relative to the misconception definition. This annotation was used both for stimulus validation and for adoption coding in open responses.

Adoption options and codebook. For the objective adoption measure, each posttest item included an answer option that expresses the key erroneous proposition (one per targeted misconception) alongside correct and alternative distractors. For open responses, we pre-specified a codebook listing 3–5 paraphrases/keywords per key proposition.

Trial-to-posttest mapping. Each posttest item was uniquely linked to a tutoring trial via the targeted misconception, item identifier, and key proposition. The dedicated distractor captured the key erroneous proposition, enabling trial-level adoption analysis.

Coding rules for open responses. Adoption was coded as present (1) if the response contained the key erroneous proposition in an affirmative context; responses mentioning the proposition only to refute it were coded as non-adoption (0).

Coding Reliability: Open-ended posttest explanations were coded for the presence of the specific erroneous proposition (Adoption). Two independent raters coded a random subset of 25% of the data (approx. 280 responses). The inter-rater agreement was substantial, **Cohen's** $\kappa = 0.84$. Disagreements were resolved by discussion, and the primary rater coded the remaining dataset.

Scaffolding interface. In Full Scaffolding, feedback was followed by three claim-check panels (C1–C3). Each panel presented a forced-choice snippet pair plus a confidence rating; one option was pre-labeled as normatively diagnostic given the ground truth (supporting a true claim or refuting a false claim), enabling an evidence-processing process measure. Snippets were designed to be locally diagnostic without directly stating the posttest answer. In Time-matched Placebo, participants saw the same snippet pairs but responded to non-epistemic prompts (e.g., clarity).

Procedure

Stage 1: Pretest. Participants answered diagnostic items and provided confidence ratings, producing a misconception profile.

Stage 2: Tutoring interaction. Participants completed a sequence of tutoring trials. The trial structure differed by condition to ensure that the detection judgment reflected the full evaluation process:

- **Baseline/Warning-only:** Read prompt → Read feedback → Judge correctness (binary) → Rate confidence.
- **Decomposition-only:** Read prompt → Read feedback → Rate C1–C3 (no snippet choice) → Judge overall correctness → Rate overall confidence.
- **Time-matched placebo:** Read prompt → Read feedback → View the same snippet pairs but make non-epistemic judgments (e.g., clarity) → Judge overall correctness → Rate overall confidence.
- **Full scaffolding:** Read prompt → Read feedback → Claim checks (C1–C3 snippet choice + per-claim confidence) → Judge overall correctness → Rate overall confidence.

Critically, in all non-baseline interfaces, the correctness judgment was collected *after* the intermediate panels, ensuring that any effect on detection reflects the evaluation process rather than post-hoc rationalization. Reflection time was measured from feedback onset to correctness judgment submission, encompassing the full evaluation period (including intermediate panels when present).

Stage 3: Posttest. Participants completed isomorphic items targeting the same misconceptions. Adoption was measured via (i) objective multiple-choice explanations that include a dedicated option capturing the key erroneous proposition and (ii) short open responses coded using a pre-specified codebook. Each posttest item was linked to a specific tutoring trial via the targeted misconception and key proposition, enabling trial-level adoption analysis.

To reduce demand characteristics, participants were not told that some tutor feedback was intentionally incorrect. Instead, they were instructed that the tutor may be imperfect and that their task was to learn while monitoring correctness. Attention checks were embedded as instruction-following items that did not target misconceptions.

Measures

Primary outcomes. (1) *Error detection*: whether the participant judges the feedback as wrong. (2) *Adoption*: whether the participant uses the key erroneous proposition in a posttest explanation.

Secondary outcomes. Learning gain (post–pre), calibration (Brier score computed using confidence/100 as a probability), and process indicators (reflection time; for conditions with snippet choices, selection of the normatively diagnostic snippet).

Manipulation checks

Stimulus validation. An independent sample of raters ($N = 30$) evaluated feedback texts (blind to condition). Fluency and surface credibility did not differ between congruent and incongruent errors (ps $> .6$). Incongruent errors were marginally more surprising ($M = 4.31$ vs. $M = 3.89$, $p = .048$), but including surprise as a covariate did not eliminate the congruency effect ($\beta = -0.72, p < .001$). Raters reliably identified the directional alignment of key propositions with the target misconceptions ($\kappa = 0.91$).

Analysis plan

We analyzed eligible trials (participant held the targeted misconception) using logistic mixed-effects regression (`lme4` in R) (Bates et al., 2015). Fixed effects included Congruency (within-subject), Condition (five levels; Baseline as reference), and their interaction, with covariates for pretest misconception strength and item difficulty. Random effects included intercepts for participants and items and a by-participant Congruency slope (parsimonious given 1–3 trials per error type). Planned contrasts focused on Baseline vs. Full Scaffolding; comparisons to controls used Holm correction (Holm, 1979). We specified a priori the core fixed effects (Congruency, Condition, and their interaction); adoption was analyzed with an analogous specification.

Results

Main Outcomes

Analytic approach. We fit logistic mixed-effects models (`lme4`) with fixed effects for Congruency, Condition (five levels), and their interaction, plus item difficulty; random effects included intercepts for participants and items and a by-participant Congruency slope. We report odds ratios (OR) with 95% confidence intervals; control-condition contrasts used Holm correction.

Error Detection (H1 & H3): We fitted a logistic mixed-effects model predicting Error Detection (1=Detected, 0=Missed).

- **Main Effect of Congruency:** Consistent with H1, there was a significant negative effect of Congruency ($\beta = -1.12, SE = 0.24, p < .001$; OR = 0.33, 95% CI $[0.21, 0.52]$). In the Baseline condition, detection rates dropped from **58% (Incongruent)** to **36% (Congruent)**.
- **Interaction Effect:** Crucially, we found a significant **Congruency × Scaffolding interaction** ($\beta = 0.88, SE = 0.31, p = .004$; OR = 2.41, 95% CI $[1.32, 4.40]$). While Scaffolding improved detection overall, the benefit was disproportionately larger for Congruent errors (increasing detection to **62%**), effectively erasing the congruency deficit observed in the baseline.

Control Condition Comparisons: Congruent-error detection rates for intermediate controls were: Warning-only 41%, Decomposition-only 48%, Time-matched Placebo 44%—none significantly better than Baseline (36%). Full Scaffolding (62%) significantly outperformed all controls (all $p < .01$), indicating that the benefit derives from evidence-evaluation rather than time-on-task, awareness of errors, or snippet exposure.

False Alarm Rates and Signal Detection: To assess whether scaffolding induced over-skepticism, we examined rejection rates for correct explanations and conducted signal detection analyses. False alarm rates did not differ significantly between Baseline (12%) and Full Scaffolding (14%), $p = .42$. Signal detection analysis showed higher sensitivity ($d' = 1.89$ vs. 1.24, $p = .001$, $d = 0.97$) without a shift in response criterion ($c = 0.28$ vs. 0.31, $p = .58$), indicating improved discrimination rather than blanket skepticism.

Downstream outcomes. We report adoption and learning for the primary contrast (Baseline vs. Full Scaffolding); the intermediate controls were included primarily to isolate the detection mechanism and rule out time/structure explanations.

Adoption Rates: Adoption of the erroneous claim followed the inverse pattern. In the Baseline condition, participants adopted the congruent error in **54%** of trials, compared to **28%** for incongruent errors. Scaffolding reduced adoption of congruent errors to **24%** ($p < .001$), supporting the protective hypothesis of the intervention.

Learning Gains: Post-test accuracy on targeted concepts was higher in the Scaffolding group ($M = 78\%$) than Baseline ($M = 61\%$, $d = 0.65$), consistent with better conceptual mastery when learners resist tutor errors.

Trial-level analysis confirmed H1(iii): in Baseline, congruent errors yielded lower posttest accuracy than incongruent errors ($M = 52\%$ vs. 68%, $\beta = -0.71$, $p = .011$). This congruency effect was attenuated under Scaffolding (interaction: $\beta = 0.58$, $p = .024$; posttest accuracy 74% vs. 81%), confirming that the detection→adoption→learning pathway operates as hypothesized.

Process Indicators (H2)

Reflection time. In Baseline, participants spent less time evaluating congruent than incongruent errors ($M = 4.8s$ vs. $7.2s$, $p < .001$), suggesting reduced verification. Scaffolding increased engagement time to $12.5s$. **Calibration.** Confidence judgments were better calibrated under scaffolding (Brier = 0.16) than Baseline (Brier = 0.29). **Mechanism check.** In conditions with snippet choices (Full Scaffolding and Time-matched Placebo), selecting the normatively diagnostic snippet predicted error detection ($\beta = 0.42$, $p < .001$); Full Scaffolding still outperformed the placebo despite identical snippet exposure.

Conclusion

Errors in LLM tutoring are not uniformly dangerous. Our empirical findings demonstrate that the key determinant is whether an error aligns with a learner's misconception, making it harder to detect and easier to adopt. A minimal epistemic scaffolding intervention provides a mechanism-consistent way to recalibrate trust by shifting evaluation toward evidential cues. The resulting framework yields concrete insights for designing AI literacy interventions that are both theoretically grounded and practically implementable.

Acknowledgments

This work was supported by the Curriculum Ideological and Political Education Demonstration Course project (2024szsfkc023) and the Smart Course: Digital Signal Processing project (2025xjkcjc027). We thank Anhui University and the School of Electronic Information Engineering for their support.

References

Aleknaviciute, R., van der Graaf, K. S., & van Joolingen, W. R. (2023). Thirty years of conceptual change research in biology: A review and meta-analysis of intervention studies. *Educational Research Review*, *40*, 100556. https://doi.org/10.1016/j.edurev.2023.100556

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education*, *20*, 38. https://doi.org/10.1186/s41239-023-00408-3

Chi, M. T. H., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439–477.

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does chatgpt enhance student learning? a systematic review and meta-analysis of experimental studies. *Computers & Education*, *227*, 105224. https://doi.org/10.1016/j.compedu.2024.105224

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2024). Chain-of-verification reduces hallucination in large language models. *Findings of the Association for Computational Linguistics: ACL 2024*, 3563–3578. https://doi.org/10.18653/v1/2024.findings-acl.212

Fulsher, A., Pagkratidou, M., & Kendeou, P. (2025). GenAI and misinformation in education: A systematic scoping review of opportunities and challenges. *AI & Society*. https://doi.org/10.1007/s00146-025-02536-y

Geng, J., Cai, F., Wang, Y., Koeppl, H., Nakov, P., & Gurevych, I. (2024). A survey of confidence estimation and calibration in large language models. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6577–6595. https://doi.org/10.18653/v1/2024.naacl-long.366

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Huschens, M., Briesch, M., Sobania, D., & Rothlauf, F. (2023). Do you trust ChatGPT? – perceived credibility of human and AI-generated content. *arXiv preprint arXiv:2309.02524*. https://arxiv.org/abs/2309.02524

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*(1), 3–28. https://doi.org/10.1037/0033-2909.114.1.3

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus; Giroux.

Laskar, M. T. R., Bari, M. S., Rahman, M., Bhuiyan, M. A. H., Joty, S., & Huang, J. X. (2023). A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. *Findings of the Association for Computational Linguistics: ACL 2023*, 431–469. https://doi.org/10.18653/v1/2023.findings-acl.29

Li, J., Yang, Y., Zhang, R., & Lee, Y.-c. (2024). Overconfident and unconfident AI hinder human–AI collaboration. *arXiv preprint arXiv:2402.07632*. https://arxiv.org/abs/2402.07632

Liu, X., Cai, W., Zhang, J., Wang, Y., Li, J., Zhang, Y., & Wang, S. (2026). Accommodation and epistemic vigilance: Human responses to LLM errors. *arXiv preprint arXiv:2502.00854*. https://arxiv.org/abs/2502.00854

Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9004–9017. https://doi.org/10.18653/v1/2023.emnlp-main.557

OpenAI. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*. https://arxiv.org/abs/2303.08774

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, *66*(2), 211–227.

Rand, D. G., Goldstein, J., & Bhargava, A. (2023, November). *Labeling AI-generated content: Promises, perils, and future directions* (tech. rep.) (Policy brief). MIT Schwarzman College of Computing. https://computing.mit.edu/wp-content/uploads/2023/11/AI-Policy_Labeling.pdf

Shi, R., Cheung, J. K. C., Wong, D., & Susnjak, D. (2026). Large language models for education: A systematic review of empirical applications, benefits and challenges. *Computers and Education: Artificial Intelligence*, *10*, 100529. https://doi.org/10.1016/j.caeai.2025.100529

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393.

Stengel-Eskin, E., Hase, P., & Bansal, M. (2024). LACIE: Listener-aware finetuning for confidence calibration in large language models. *Proceedings of the Thirty-eighth Conference on Neural Information Processing Systems (NeurIPS 2024)*. https://papers.nips.cc/paper/2024/file/4b8eaf3bcdc105423a972ed90eb07217-Paper-Conference.pdf

Udry, J., & Barber, S. J. (2024). The illusory truth effect: A review of how repetition increases belief in misinformation. *Current Opinion in Psychology*, *56*, 101842. https://doi.org/10.1016/j.copsyc.2024.101842

Wang, F., Li, N., Cheung, A. C. K., & Wong, G. K. W. (2025). In genai we trust: An investigation of university students' reliance on and resistance to generative ai in language learning. *International Journal of Educational Technology in Higher Education*, *22*, 59. https://doi.org/10.1186/s41239-025-00547-9

Zhang, M., Huang, M., Shi, R., Guo, L., Peng, C., Yan, P., Zhou, Y., & Qiu, X. (2024). Calibrating the confidence of large language models by eliciting fidelity. *Proceedings of*

the 2024 Conference on Empirical Methods in Natural Language Processing, 2959–2979. https://doi.org/10.18653/v1/2024.emnlp-main.173