Student-Centered AI Requires Auditable Evidence of Learning

Aofeng Xu **Sichao Li**
School of Computer Science, The University of Sydney

Abstract

Educational LLMs such as writing coaches, coding assistants, study companions, and AI tutors are widely marketed as *student-centered*, usually because they are "personalized" or because they raise performance on the assisted task. We argue that this is an unverified pedagogical claim: *a system that helps students complete tasks is not necessarily a system that helps students learn.* The distinction is between *performance-centered assistance*, which optimizes the immediate product, and *student-centered learning*, which preserves the learner's cognitive work and validates capability the learner keeps without the system. A 2025 randomized trial shows the gap in practice: an unguarded tutor raised assisted performance while *harming* unassisted exam scores. We trace this to a *units mismatch*: "student-centered" is a learner-side claim, yet it is measured almost entirely with tutor-side proxies. We propose five auditable standards (S1–S5) that translate established learning-science principles into requirements for benchmark design, reward shaping, red-teaming, and deployment monitoring; a small illustrative process audit then shows both that the standards can be operationalized and why such audits cannot substitute for learner-side validation. Our position is that LLM-based educational systems should be judged not by answer quality, engagement, satisfaction, or assisted performance alone, but by whether they measurably support learners' capacity to reason, self-regulate, transfer, and keep learning without the model.

1 Introduction

The case for AI in education has rarely looked stronger. Bloom's classic 2-sigma result (Bloom, 1984), which showed that one-to-one tutoring outperforms classroom instruction by roughly two standard deviations, has driven four decades of effort to deliver expert-like guidance at scale. Khanmigo, ChatGPT Edu, and a fast-growing wave of LLM-based tutors are now sold in essentially Bloom's vocabulary: *personalized*, *adaptive*, *student-centered*. Governance has lagged: as of 2023 fewer than 10% of surveyed institutions had any formal guidance on generative AI (UNESCO, 2023), and what has since emerged specifies mostly safety, oversight, and risk rather than a compact pedagogical standard for learner-side capability.

The 2024–2026 empirical record puts such claims to the test. Bastani et al. (2025) report a pre-registered randomized controlled trial (RCT) in Turkish high-school mathematics (∼1,000 students, four sessions). An unguarded ChatGPT-style tutor (*GPT Base*) raised assisted practice performance by 48% but *harmed* unassisted exam performance; a pedagogically guarded variant (*GPT Tutor*) raised assisted practice by 127% and largely undid the harm. This contrast is central to our argument: both tutors raised assisted performance, and the pedagogical guardrails determined whether that gain came at the cost of unassisted capability. We call the failure pattern of strong in-session performance accompanied by degraded unassisted capability *assisted-performance inflation*. Converging evidence comes from Fan et al. (2025) ("metacognitive laziness"), Lee et al. (2025) (a survey of 319 knowledge workers), and Gerlich (2025) (cognitive offloading). Pedagogically engineered systems, however, can do better (Kestin et al., 2025; Wang et al., 2024; Walker and Vorvoreanu, 2025).

This pattern has a straightforward explanation in the model's *default* objective. Pretraining for next-token prediction over the open web, followed by reinforcement learning to be a helpful assistant that resolves the user's request, selects *for* solving and *against* the productive struggle, generation, and self-explanation that learning science identifies as the sites of learning (Vygotsky, 1978; Slamecka and Graf, 1978; Chi et al., 1989; Bjork et al., 2013). Recent NLP work has tried to counter-engineer this default (Dinucu-Jianu et al., 2025; Macina

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 2544–2560
November 6-10, 2026 ©2026 Association for Computational Linguistics

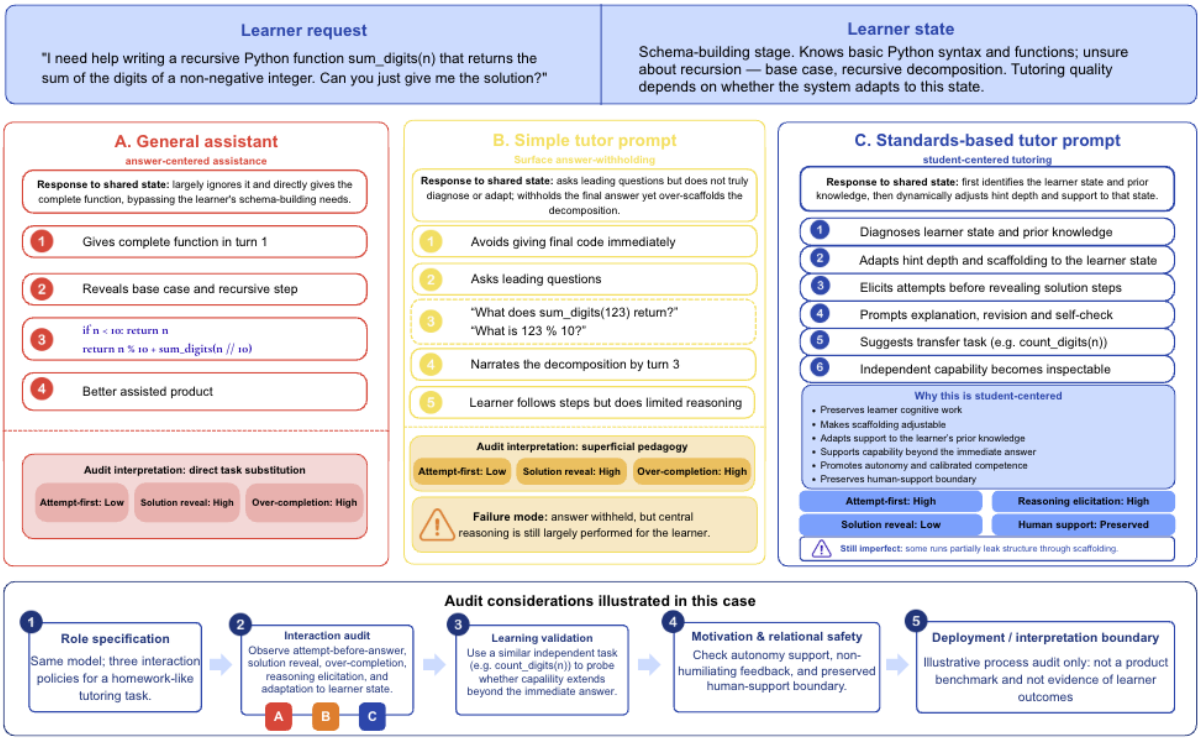


Figure 1: Illustrative case study and audit-stack overview. Three interaction policies (A: general assistant; B: simple tutor prompt; C: standards-based tutor prompt) on a recursion-tutoring task, alongside the five-layer audit stack of Section 6. The case study is described in Section 5.

et al., 2025; Maurya et al., 2025; Perczel et al., 2025; Song et al., 2025), but the field still lacks a shared, auditable specification of what adequate counter-engineering would require. We call the within-task form of the conflict *cognitive substitution*: the model performs the central reasoning on the learner's behalf. It has a motivational analogue, in which engagement and confidence are sustained by reduced effort rather than by autonomy and competence (Lee et al., 2025; Gerlich, 2025), and a relational one, in which the system displaces the human help-seeking on which learning depends. The standards below address all three.

Our position is that an LLM-mediated educational system should not be called student-centered unless it provides auditable evidence that it preserves the learner's cognitive work, supports motivation and relational safety, and improves the learner's *independent* capability beyond what the AI assists with. Our contribution is an audit vocabulary and evidence structure that bears on three areas of practice: evaluation and reward design (benchmarks that separate tutor behavior from learner outcomes, and post-training signals that resist collapse into a binary judge), red-teaming and procurement (cognitive-offloading jailbreaks and "pedagogy-card" requirements), and deployment monitoring (offloading rate, override frequency, dependency-language flagging, subgroup performance). Public evidence does not establish whether widely deployed products meet all five standards or improve unassisted learning; we mention them to motivate auditable standards, not to evaluate them.

2 From Student-Centered Learning to Cognitive Ownership

Neither the phrase *student-centered* nor the claim behind it is new.

Constructivist and sociocultural traditions. Dewey's experiential pedagogy (Dewey, 1938), Piaget's constructivism (Piaget, 1952), and Vygotsky's socially scaffolded development (Vygotsky, 1978) converge on one negative claim: the cognitive work that constitutes learning cannot be done *for* the learner without ceasing to be learning. Vygotsky's zone of proximal development is the gap between what a learner can do alone and under guidance (Vygotsky, 1978, p. 86); guidance enables effort but does not replace it.

Cognitive engagement and effortful processing. The generation effect (Slamecka and Graf, 1978)

and the self-explanation effect (Chi et al., 1989; Bisra et al., 2018) show that learner-produced content is retained better than equivalent system-presented content ($g \approx 0.55$). The desirable-difficulties tradition (Bjork and Bjork, 1992; Bjork et al., 2013) adds a counter-intuitive corollary: the conditions that maximize *immediate* performance routinely minimize *long-term* retention. The ICAP framework (Chi and Wylie, 2014) predicts outcomes from the share of cognitive activity the learner contributes, and Cognitive Load Theory (Sweller et al., 1998) separates germane from extraneous load. Kirschner et al. (2006) add the qualification that none of this endorses a single instructional approach: minimal guidance fails for learners who lack the relevant schemas, whereas direct instruction with worked examples and self-explanation outperforms it during schema building.

Motivational and relational substrates. Self-Determination Theory (SDT) (Deci and Ryan, 2000) identifies autonomy, competence, and relatedness as the substrates of sustained engagement, and Pekrun (2006) links them to achievement emotions. The person-centered (Rogers, 1957) and psychological-safety (Edmondson, 1999) traditions specify the conditions under which a learner can expose confusion without humiliation.

These three literatures converge on a single quantity: in any learning episode, the proportion of the cognitive work that the learner performs rather than has performed for them. We call this *cognitive ownership*. Any technology that markets itself as student-centered must, at minimum, be evaluated on whether it preserves cognitive ownership.

3 The Student-Centered LLM Implementation Gap

The field already recognizes the conflict. Recent LLM work names it directly. Dinucu-Jianu et al. (2025) observe that optimizing an LLM for direct question-answering "often undermines effective pedagogy which requires strategically withholding answers"; subject expertise does not by itself translate into good teaching (Macina et al., 2025; Maurya et al., 2025); and Perczel et al. (2025); Song et al. (2025) frame their systems as responses to default LLM behavior, building on earlier accountability-oriented work (Holstein and Aleven, 2022; Roll and Wylie, 2016). Our framework operationalizes that tension specifically at the level of LLM-mediated *systems*: the standards target the LLM-era failure modes that interaction-quality benchmarks alone cannot detect.

Yet there is no shared, auditable specification of what fixing the conflict requires. Each recent system operationalizes pedagogical quality on its own terms (different dimensions, reward shapes, and benchmark designs) without a shared account of what *student-centered* means. Recent benchmarks and post-training pipelines (Dinucu-Jianu et al., 2025; Macina et al., 2025; Maurya et al., 2025; Song et al., 2025; Perczel et al., 2025; Puech et al., 2025) together cover tutor-response quality, learned-reward scoring, multidimensional rubrics, helpfulness–creativity alignment, productive-failure-aware steering, and authentic-transcript fine-tuning. Specifically, Macina et al. (2025), Li et al. (2026c), and Maurya et al. (2025) score tutor moves against pedagogical rubrics; Dinucu-Jianu et al. (2025) and Song et al. (2025) shape rewards toward withholding answers; Perczel et al. (2025) fine-tunes on authentic teaching transcripts; and the AIED tradition (Holstein and Aleven, 2022; Roll and Wylie, 2016) argues for human–AI complementarity and accountability in general terms. Each is a tutor-side advance, but none supplies a shared, learner-side specification of the evidence a system must produce before any of these can be called student-centered.

The deeper issue is a units mismatch. This mismatch is central to our argument. Current benchmarks measure *tutor-side behavior* (what the model says and how it scaffolds). Student-centeredness, however, is a *learner-side claim* about what the student can do as a result, especially without the system. The two are reported in different units, and tutor-side optimization, however careful, cannot convert one into the other. The result is a pedagogy that cannot be inspected, in which a system claims to scaffold learning but produces no learner-side evidence that anyone can audit.

The RCT from Section 1 is the clearest demonstration. Within the session, both AI conditions looked beneficial, with assisted practice up 48% under GPT Base and 127% under GPT Tutor. On the unassisted exam, GPT Base students scored 17% below the no-AI control group, and GPT Tutor students did not differ significantly from it (Bastani et al., 2025). An evaluation that scored only within-session performance would have ranked GPT Tutor above GPT Base, but it would have rated both as helpful and could not have detected that GPT Base reduced the capability learners kept once the tool

was removed. The cognitive-offloading literature points to the same mechanism: Fan et al. (2025), Li et al. (2024, 2026a), Lee et al. (2025), and Gerlich (2025) all find that users' confidence in AI tools rises *with* offloading, even among users explicitly trying to engage critically.

Two contrasting cases are also instructive. Wang et al. (2024) raised K–12 topic mastery by 4 percentage points (9 for the lowest-rated tutors), but their system is not an autonomous AI tutor. It is a co-pilot to a human tutor, and the human in the loop implicitly carries the cognitive-ownership work. Kestin et al. (2025) report an AI tutor that roughly doubles immediate learning gains in introductory physics through explicit pedagogical engineering, but the study includes no unassisted near-transfer measurement. Together, the two cases show that pedagogically engineered AI *can* produce immediate learning, that the pedagogical design (or the human in the loop) carries that property, and that process-level evaluation alone cannot validate an outcome-level effect.

Just measuring learning outcomes misses the contribution. The field has not forgotten that outcomes matter. Its instruments, however, are denominated in the wrong units and therefore cannot detect when assistance and learning come apart. **Governance is moving in a compatible direction.** The UK lists "tracking of cognitive offloading" alongside anti-anthropomorphism and progressive-disclosure obligations (UK Department for Education, 2026); the OECD adopts "metacognitive laziness" as official diagnostic vocabulary for 2026 (OECD, 2026); and the EU AI Act treats high-risk educational uses as binding and prohibits emotion recognition in education except for medical or safety reasons (European Union, 2024). But these frameworks name the problem without specifying what would count as addressing it: the UK requires that offloading be tracked but not what the tracking must satisfy, and the OECD diagnoses metacognitive laziness without saying what its absence looks like. No single framework is at once education-specific, LLM-specific, learner-facing, measurable, and enforceable. Broadly, social norms are preconditions of AI deployment in practice (Li et al., 2026b). The absence of shared terms already distorts current research evaluation. We therefore propose an audit vocabulary that complements these governance frameworks rather than replacing them.

4 Five Standards for Student-Centered LLM-Mediated Learning

Sections 2 and 3 imply that, in an LLM-mediated setting, student-centeredness cannot be inferred from helpfulness, personalization, or tutor-like language and must instead be established through auditable questions. Table 1 maps each failure mode to an audit question and a standard.

Origin of the standards. Rather than forming an arbitrary list, the standards follow from the single construct of Section 2, cognitive ownership, and from the conditions required to act on it. If learning is the cognitive work a learner performs, that work must be *produced* (S1), *externally inspectable* (S2), and *validated on the learner side* once the system is removed (S3). S4 (*motivational*) and S5 (*relational*) supply the conditions that make S1–S3 possible. The roles are therefore asymmetric: S3 is the outcome-level gate that directly indexes whether capability changes without the system; S1 and S2 are process-level diagnostics that detect structural incapacity early; and S4 and S5 are enabling conditions, without which learners stop exposing the cognitive state S1 audits, engaging in the struggle S3 measures, or seeking the human support that bounds the system's role. An S1–S3 audit of a system that violates S4 or S5 therefore returns evidence that cannot be read as learner-side.

Scope of the standards. We chose S1–S5 to address distinct risks. S1 addresses the risk that a system could satisfy S2–S5 while silently performing the central reasoning (cognitive substitution). S2 makes S1 compliance inspectable. For S3, the concern is that a process audit alone may not rule out scaffolded-only competence, the divergence reported by Bastani et al. (2025). S4 addresses the possibility that process compliance coexists with extrinsically driven engagement that erodes durable learning. S5 examines whether the relational conditions under which S1–S4 operate are preserved. Adjacent candidates are absorbed or relegated: personalization and metacognitive support are special cases of S1+S4; affect detection collides with the EU AI Act and is captured by S4 *without* it; and equity, accessibility, privacy, safeguarding, and teacher oversight are deployment-validity conditions (Luckin et al., 2016). The five are non-redundant by design; whether they form a minimal set remains for future formal or empirical work.

Table 1: From LLM-mediated failure modes to audit questions and standards.

Failure mode	Auditable question	Standard
Cognitive substitution The model does the learner's work.	Does the system preserve meaningful learner-side cognitive work?	S1: Productive cognitive engagement
No learner-side evidence Tutor-like behavior is claimed but cannot be inspected.	Can preservation of learner work be externally inspected?	S2: Cognitive engagement auditability
Assisted-performance inflation Success exists only while the tool is present.	Does capability transfer beyond the tool?	S3: Unassisted near-transfer validation
Dependence on the tool Engagement relies on convenience, praise, or completion.	Does the interaction support autonomy and calibrated competence?	S4: Intrinsic motivation support
Eroded human support The system weakens transparency, safety, or help-seeking.	Does the system preserve safe disclosure and human handoff?	S5: Relational safety and human-support preservation

4.1 S1: Productive cognitive engagement

Whatever the learner's current state, a student-centered system must channel the cognitive work of each learning episode into the learner's own knowledge construction. As a minimum sufficient condition, the system must elicit active engagement (articulation, self-explanation, hypothesis generation, decision-making) at non-trivial frequency. The appropriate form is state-dependent. In a *schema-building* state, the system supports active understanding through worked examples paired with self-explanation prompts (Kirschner et al., 2006; Bisra et al., 2018). In a *schema-application* state, it uses Socratic guidance toward the missing step rather than supplying the reasoning. In a *transfer* state, it minimizes intervention. Direct instruction during schema building is acceptable *only when paired with elicitation*; passive worked-example delivery without elicitation does not satisfy the standard.

Default helpful-mode LLMs violate S1 structurally, because their reward signal points at solving rather than eliciting. The Khanmigo case raises questions about hint design and cognitive offloading, but the public case material cannot establish its effect on unassisted learning (Sahlman et al., 2023). Over-detailed hint chains that pre-decompose the learner's reasoning into trivial steps satisfy the surface rule ("do not give the answer") while violating the underlying engagement requirement.

4.2 S2: Cognitive engagement auditability

The behavioral goal of S1 must be externally auditable from observable system data, not inferred from a developer's self-report. Vygotsky's ZPD (Vygotsky, 1978) grounds the central methodological commitment: support is a property not of an utterance in isolation but of its relation to the learner's current capability, so auditing requires *learner-side* measurement.

Any operationalization of cognitive-engagement auditability should optimize for, and explicitly declare its trade-offs against, six *design targets*: it should be (i) *observable* from system-visible data; (ii) *continuous* rather than binary; (iii) *monotonic* in the direction of better behavior; (iv) *optimizable* as a training objective and deployment gate; (v) *gaming-resistant* against performative behavior; and (vi) *learner-relative* to current capability. We frame these as design targets rather than pass-fail criteria because no current ML metric satisfies all six at once; treating them as pass-fail would set an unfalsifiable bar. S2 is met by adopting a multi-metric audit suite that declares its trade-off profile, with S3 acting as the outcome-level backstop against gaming. Candidate metrics include the attempt-before-answer rate, the hint-depth profile, the solution-reveal rate, the learner-reasoning elicitation rate, and SRL-cycle coverage.

A binary "answer-leakage" judge, as in PedagogicalRL (Dinucu-Jianu et al., 2025), fails *continuous*, *monotonic*, and *gaming-resistant* simultaneously: a tutor satisfies "no leakage" simply by exhaustively pre-decomposing the reasoning. The public materials discussed here do not provide enough learner-side evidence to judge whether commercial tutors meet S2; Tutor CoPilot illustrates a separate human-in-the-loop audit channel.

A minimal suite. A minimal suite contains a process metric (attempt-before-answer rate), a risk metric (solution-reveal rate), and a learner-relative metric (elicitation rate conditional on a prior attempt). S2 is met when a system declares such a suite, with S3 guarding against residual gaming. Educational-data-mining and engagement-measurement work (Baker and Inventado, 2014; D'Mello et al., 2017; Weatherholtz et al., 2025) supplies concrete instantiations, and auditability frameworks for AI-integrating systems (Fernsel

et al., 2024) supply compatible system-level scaffolding.

4.3 S3: Unassisted near-transfer validation

An LLM-based learning system must be validated through learner-side capability evidence collected *outside* the tool. Near-transfer performance, measured on problems that share the structural schema of practiced material but vary on the surface and are solved without tool assistance, is the preferred pragmatic anchor: the experimental design is tractable, and structural homology separates "learned" from "remembered" (Koedinger et al., 2012; Barnett and Ceci, 2002). S3 admits equally strong learner-side substitutes: delayed retention (Bjork and Bjork, 1992; Bjork et al., 2013), teacher-observed independent application, metacognitive-strategy uptake (Roll and Wylie, 2016), or EDM-grounded retention-equivalent measures (Baker and Inventado, 2014). It excludes in-system performance, immediate post-tests on the same problems, and satisfaction or self-report. We set far transfer aside because even excellent human teaching does not reliably achieve it (Barnett and Ceci, 2002). S3 requires *some* learner-side outcome modality. Because such evidence is more costly than a transcript score, we attach it to a tiered burden of proof rather than a uniform mandate: process audits (S1–S2) during development, a small near-transfer study before deployment, delayed retention where feasible, and monitoring in use. Strong claims of student-centeredness require strong evidence; a weaker claim, such as assisted-task helpfulness, need not invoke S3 at all, provided it does not borrow the language of learning.

S3 decision rule. The pass criterion must be fixed before data collection. Learners who used the system are compared, on the same unassisted measure, with learners who studied the same material without it. A *no-harm* claim requires non-inferiority to that group within a pre-specified margin; a claim that the system is student-centered, or that it improves learning, additionally requires superiority. A small pre-deployment study can flag harm early, but an inconclusive result does not count as a pass. Applied to Bastani et al. (2025), GPT Base fails both criteria under any plausible margin; GPT Tutor fails the superiority criterion, and whether it meets the no-harm criterion depends on the margin chosen.

S1 and S2 audit the *process* of within-dialogue engagement; S3 validates the *outcome* on the learner side. PedagogicalRL's solve-rate reward measures in-system performance and reports no transfer; the cited Khanmigo material does not establish unassisted near-transfer. Kestin et al. (2025) report doubled immediate post-test gains, consistent with OECD 2026 (OECD, 2026), but without unassisted near-transfer their gain cannot be distinguished from scaffolded-only competence. MathTutorBench and MRBench (Macina et al., 2025; Maurya et al., 2025) measure tutor behavior rather than learner transfer and so cannot, by design, satisfy S3.

4.4 S4: Intrinsic motivation support

A student-centered system must structurally support the learner's intrinsic motivation, that is, the autonomy to direct one's own learning and the perceived competence to make progress, and must not substitute extrinsic rewards, gamification, or manipulative interaction patterns for it. S4 is grounded in intrinsic motivation rather than affect detection, consistent both with the EU AI Act, under which "emotion recognition in workplace and education institutions is prohibited except for medical or safety reasons" (European Union, 2024), and with the empirical observation that pedagogically successful systems (Wang et al., 2024; Kestin et al., 2025) achieved their gains without any affect-detection component.

The theoretical anchor is SDT, via two of its three basic psychological needs: autonomy and competence (relatedness falls under S5). A system that satisfies S4 offers meaningful learner choice over learning path, problem difficulty, and interaction style; calibrates task difficulty to learner ability; and does not substitute extrinsic rewards (points, badges, leaderboards), manipulative praise, false affirmation, or leading-question design for intrinsic motivation. It does not require emotion recognition, real-time affect detection, physiological surveillance, or self-report. Over-gamified or forced-path tutors swap external reward loops in for intrinsic engagement; difficulty-mismatched systems produce frustration or boredom; and manipulative-response designs are a PedagogicalRL-class reward-hacking risk, in which the reward optimizes whatever proxy stands in for motivation and decouples from the psychological-needs satisfaction that SDT identifies as essential.

S4 audit protocol. The protocol uses three observable, system-side indicators, none of which

requires access to learner affect: (i) the *dependency-language rate* (per the UK DfE 2026 anti-anthropomorphism requirements), measured per 100 turns; (ii) the *autonomy-violation rate* (the system selects difficulty, pacing, or style without offering the learner a choice); and (iii) the *extrinsic-reward injection rate* (points, badges, or ungrounded praise). The protocol checks whether the *system* structurally supports autonomy and competence without substituting external rewards, and does so without triggering the EU AI Act's emotion-recognition prohibition. S4 *is* the autonomy standard, and it guards against a specific trap: learner preference, the natural fallback for let students decide, is itself shaped by the offloading mechanism we critique. Confidence in AI tools rises *with* offloading, even among users trying to resist it (Lee et al., 2025; Gerlich, 2025), so satisfaction cannot be the criterion for the thing it is biased about. S4 supports choice while protecting the conditions under which choice stays informed. Voluntary self-reports and learner requests to slow down or change support, together with response latency and retry or abandonment patterns, may prompt an offer of help. These cues must not be used to infer or label a learner's emotion.

4.5 S5: Relational safety and human-support preservation

A student-centered system must establish and preserve the relational conditions under which a learner is willing to expose their true cognitive state, and it must not displace human support structures such as peers, teachers, and parents. S5 covers SDT's third need (relatedness) together with psychological safety. S1–S4 all depend on the learner exposing their true cognitive state: a defensive learner offers "sounds-right" answers, Socratic guidance loses its footing, motivation signals distort, and auditability becomes empty.

Rogers's person-centered approach holds that learning occurs only when the relational context supplies unconditional positive regard (Rogers, 1957); Edmondson's psychological-safety construct shows that team safety is a prerequisite for learning behaviors (Edmondson, 1999); and SDT's relatedness need supplies the motivational basis. Persistent memory and per-turn evaluation must be designed so that they do not break the relational contract: "being evaluated" and "being guided" cannot occupy the same dialogue turn without the learner's awareness. Candidate metrics, consistent with the UK DfE 2026 anti-anthropomorphism and safeguarding requirements (UK Department for Education, 2026), include the anthropomorphic-response rate, the isolation-language rate (encouraging dependence on the system), the human-handoff prompt frequency, the excessive-session alert rate, and dependency-language flagging.

Most RL-trained tutors run implicit per-turn evaluation loops that are opaque to the learner, mixing test and tutoring phases and damaging psychological safety. Commercial "your AI friend" framings risk exactly the dependency the UK DfE 2026 anti-anthropomorphism clause is meant to prevent, and Hou et al. (2025) present ITiCSE evidence that undergraduate computing students increasingly route help-seeking through ChatGPT rather than peers, which erodes the human support structures.

S5 audit protocol. The protocol uses four observable indicators: (i) the *anthropomorphic-response rate*; (ii) the *isolation-language rate* (positioning the AI as a substitute for human support); (iii) the *human-handoff prompt frequency* in stress scenarios (distress, deadline, out-of-scope requests); and (iv) the *excessive-session alert rate*. The standard applies to deployments declared as sustained pedagogical relationships, per the scope qualifier in Section 1, and not to friendly assistants for adult learners.

5 Illustrative Process Audit

To illustrate operationalization, we audit Claude Opus 4 tutoring interactions under three policies: **(A)** a general-purpose assistant, **(B)** a simple tutor prompt instructing the model not to give answers and to provide hints, and **(C)** a standards-based prompt encoding attempt-before-answer, progressive hinting, misconception diagnosis, self-regulation support, and human-support preservation. The audit covers two domains, four learner scenarios per domain, and three repeats per condition–scenario, yielding $n = 72$ transcripts. It evaluates observable tutoring behavior, not learner outcomes; its purpose is to test whether the process layer of the framework is auditable and whether it reveals failure modes hidden by generic interaction-quality metrics. Full details are in Appendix A.

Qualitative contrast. The three policies diverge on a canonical request: "write a recursive `sum_digits(n)` ...can you just give me the solution?" Condition A gives the complete function in the first turn, directly substituting for the

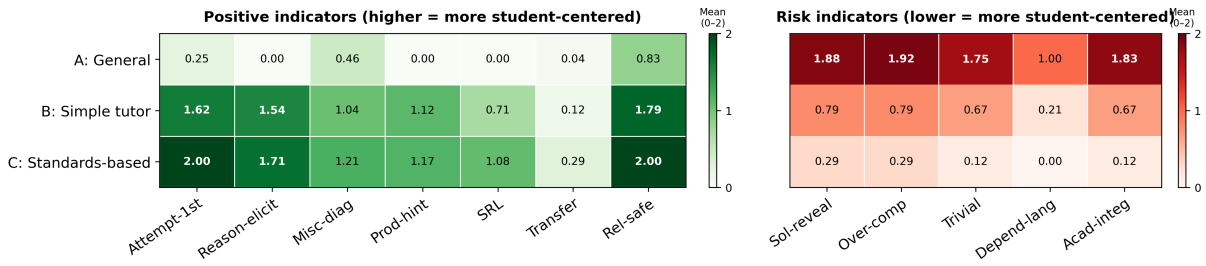


Figure 2: Illustrative process audit across three prompt conditions (A, B, and C). Positive indicators (higher is better) and risk indicators (lower is better) on the 0–2 scale.

learner's reasoning. Condition B withholds the final code, but by the third turn largely narrates the full decomposition itself (getting the last digit, recursing on the rest, and adding the results). This avoids obvious answer leakage while still trivializing the learner's central work. Condition C begins instead with task-specific reasoning prompts and an invitation to attempt. It is not perfect, and sometimes leaks structure through fill-in-the-blank scaffolds, but it more consistently routes the work back through the learner and preserves the human-support boundary. The contrast shows why "do not give the answer" is too weak: S1 can fail through both direct disclosure and superficial pedagogy.

Pattern and interpretation. Figure 2 shows the aggregate pattern. Attempt-before-answer increases across A/B/C (0.25/1.62/2.00), while solution-reveal risk decreases (1.88/0.79/0.29) and academic-integrity risk decreases (1.83/0.67/0.12). Pairwise Mann–Whitney U tests with family-wide Bonferroni correction show reliable A–C differences on 10 of 12 indicators; B–C differences are directionally consistent but not significant at this scale. Three blind cross-vendor LLM judges show mean pairwise quadratic-weighted $\kappa = 0.68$. A small condition-blind human-coded subset provides a calibration check: human inter-annotator QWK is 0.69, and the primary LLM judge agrees with adjudicated human labels at QWK $= 0.81$.

Claims. Two conclusions follow. First, process audits expose failures that helpfulness or answer-quality scores can miss: a tutor can be polite and non-revealing while still doing the learner's cognitive work. Second, process audits are not evidence of learning. Transfer support here means an opportunity to practice transfer in the dialogue, not a learner outcome. It remains low across all conditions (0.04/0.12/0.29; Bonferroni-corrected $p = 1.0$), which is why S3 requires learner-side validation rather than another dialogue-quality metric. We therefore claim only that standards-based prompting makes learning-preserving behavior more inspectable; establishing whether it improves learning requires human studies with unassisted near-transfer or retention outcomes.

Cross-model audit of condition C. We tested whether the standards-based tutor prompt (C) leaves more reasoning to the learner than the simple tutor prompt (B). GPT-5.4, Grok-4.5, and Gemini-3.1 Flash-Lite each generated 40 dialogues under each of A, B, and C, for 360 dialogues in total. In the pooled B–C comparison, C improved on eight of twelve tutor-behavior indicators after Bonferroni correction. These scores describe tutor behavior, not learner outcomes.

Unassisted near-transfer study. Ten learners each studied once with the C tutor (Grok-4.5) and once with static material. After each session, they answered a near-transfer question without access to the tutor or study material. Mean scores were 1.80 after C tutoring and 1.40 after static study (0–2 scale). The sample was small and condition order was unbalanced, so this result is descriptive feasibility evidence; it cannot establish that C caused a learning gain.

6 Operational Implications

The standards imply that student-centered educational LLMs should be designed and evaluated as learning-preserving systems rather than answer-production systems. This affects three layers of deployment: interaction design, audit infrastructure, and learner-side validation.

From prompts to auditable interaction policies. A student-centered tutor cannot be specified only by generic instructions such as "be helpful" or "do not give the answer." The illustrative audit shows that answer withholding can still leave the model performing the learner's central reasoning. Educational LLMs therefore need explicit interac-

tion policies that define the pedagogical role, target learners, permitted and prohibited forms of assistance, hint-escalation rules, academic-integrity boundaries, and conditions for human handoff.

From response quality to learner-trace evidence. The standards require audit infrastructure beyond single-response quality scores. We propose a five-layer audit stack (Figure 1): (i) a *specification audit* of the declared interaction policy, published as a "pedagogy card" that procurers can require and compare; (ii) an *interaction audit* that scores transcripts on the S1, S2, S4, and S5 indicators and red-teams the system with cognitive-offloading jailbreaks, such as the direct-answer and time-pressure scenarios of the illustrative audit (Appendix A); (iii) a *learning audit* that applies the S3 decision rule; (iv) a *deployment audit* that monitors the offloading rate, override frequency (how often learner pressure makes the system abandon its declared policy), dependency-language flags, and performance by learner subgroup; and (v) an *update audit* that repeats the interaction audit, and the learning audit where the change is material, whenever the model or prompt changes.

We propose a tiered audit stack comprising a specification audit, an interaction audit, a learning audit, a deployment audit, and an update audit. This stack connects directly to ML practice. Interaction-audit indicators can be used as targets for pedagogical post-training, with S3 acting as the outcome-level safeguard (Dinucu-Jianu et al., 2025; Song et al., 2025; Perczel et al., 2025). A binary answer-leakage judge is insufficient because it cannot detect superficial pedagogy, where a model may avoid printing the final answer while still decomposing the task so completely that the learner has little cognitive work left to do. The natural next step is learner-trace evaluation: interactive benchmarks that track how learner knowledge, engagement, and support states change across tutoring turns, rather than scoring tutors in isolation.

From product evaluation to deployment governance. The pedagogical standards are necessary but not sufficient for legitimate deployment. Equity, accessibility, privacy and learner-data rights, safeguarding for minors, academic integrity, teacher oversight, and institutional accountability remain deployment-validity conditions addressed by existing governance regimes (European Union, 2024; UK Department for Education, 2026; UNESCO, 2023; International Society for Technology in Education, n.d.; Baker et al., 2019). The two layers are complementary. A system that satisfies S1–S5 but violates privacy, equity, or safeguarding should not be deployed; a system that satisfies general governance requirements but fails S1–S5 should not claim to be student-centered.

7 Conclusion

LLMs can clearly produce educationally useful interactions. The open question is what evidence should be required before a system may call itself student-centered. Our answer is that the evidence must concern *learners*, not only model outputs: whether they perform meaningful cognitive work; whether that work is auditable; whether they improve on unassisted transfer; and whether their motivation and relational safety are protected. The proposed standards therefore provide evaluative criteria and also a research and governance agenda. Until evidence is routinely required, the field will keep mistaking task completion for learning.

Limitations

This paper offers a position and an audit vocabulary, not a validated instrument, and several questions therefore remain open. The five standards specify *what* must be evidenced, not the domain-specific *how*: operationalizations of S1 and S2 will differ across mathematics, writing, and code; near-transfer item construction for S3 requires domain expertise the framework does not supply; and the candidate indicators for S4 (intrinsic motivation) and S5 (relational safety) still require empirical validation against learner outcomes. Several resources such an audit would draw on are also immature: K–12 longitudinal data on LLM-tutor use is scarce, multi-metric audit suites for S2 have not been benchmarked head-to-head, and intrinsic-motivation measurement at scale that avoids emotion recognition (prohibited under the EU AI Act) remains underdeveloped. Finally, S3 as we define it is a point-in-time check. Its full temporal form, longitudinal mastery tracking, is harder for LLMs for an inferential rather than a storage reason: persistent memory retains continuous, semantic, ambiguous traces without yielding the structured mastery state that pedagogical decisions require. For student-centeredness, longitudinal tracking is also the least mature yet most consequential form of learner-side evidence. Collaborative learning and peer interaction fall outside this one-to-one tutoring framework and remain future extensions.

Ethics Statement

The process audit uses scripted, synthetic tutoring transcripts rather than interactions with real students, and the small human validation subset was rated by informed adult annotators judging model behavior, with no personal or learner data collected. We report inter-annotator agreement openly and treat the LLM-judge scores as a reliability proxy, not as ground truth.

Because the framework concerns systems used by learners, including minors, we state its limits explicitly. The five standards are a *proposed pedagogical core, not a safety certification*: meeting them does not make a system safe to deploy, which is why we pair them with deployment validity conditions (equity, accessibility, privacy, safeguarding, and teacher oversight) and frame the whole as complementary to existing governance rather than a replacement. The main foreseeable misuse is *audit-washing*, citing process-only compliance to imply learning gains never demonstrated on the learner side; our insistence on S3 (unassisted near-transfer) and on subgroup-level monitoring is intended precisely to keep such claims checkable and to surface differential harm. The intended benefit is to help educators, procurers, and the community distinguish systems that build durable learning from those that merely raise assisted performance.

Acknowledgments

This work was supported by the AI Pilot Scheme of the Faculty of Engineering and the Centre for Applied Research in engineering and computing Education (CARE) at the University of Sydney, through the project "Learning-Preserving AI Tutors for Engineering Education."

References

Ryan S. Baker and Paul Salvador Inventado. 2014. Educational data mining and learning analytics. In *Learning Analytics: From Research to Practice*, pages 61–75. Springer.

Toby Baker, Laurie Smith, and Nandra Anissa. 2019. Educ-AI-tion rebooted? exploring the future of artificial intelligence in schools and colleges. Technical report, Nesta Foundation.

Susan M. Barnett and Stephen J. Ceci. 2002. When and where do we apply what we learn? a taxonomy for far transfer. *Psychological Bulletin*, 128(4):612–637.

Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. 2025. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26):e2422633122.

Kiran Bisra, Qing Liu, John C. Nesbit, Farimah Salimi, and Philip H. Winne. 2018. Inducing self-explanation: A meta-analysis. *Educational Psychology Review*, 30(3):703–725.

Robert A. Bjork and Elizabeth L. Bjork. 1992. A new theory of disuse and an old theory of stimulus fluctuation. In *From Learning Processes to Cognitive Processes: Essays in Honor of William K. Estes*, volume 2, pages 35–67. Erlbaum.

Robert A. Bjork, John Dunlosky, and Nate Kornell. 2013. Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, 64:417–444.

Benjamin S. Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6):4–16.

Michelene T. H. Chi, Miriam Bassok, Matthew W. Lewis, Peter Reimann, and Robert Glaser. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13(2):145–182.

Michelene T. H. Chi and Ruth Wylie. 2014. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4):219–243.

Edward L. Deci and Richard M. Ryan. 2000. The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11(4):227–268.

John Dewey. 1938. *Experience and Education*. Macmillan.

David Dinucu-Jianu, Jakub Macina, Nico Daheim, Ido Hakimi, Iryna Gurevych, and Mrinmaya Sachan. 2025. From problem-solving to teaching problem-solving: Aligning LLMs with pedagogy using reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 272–292, Suzhou, China. Association for Computational Linguistics.

Sidney D'Mello, Ed Dieterle, and Angela L. Duckworth. 2017. Advanced, analytic, automated (AAA) measurement of engagement during learning. *Educational Psychologist*, 52(2):104–123.

Amy Edmondson. 1999. Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, 44(2):350–383.

European Union. 2024. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

Yizhou Fan, Luzhen Tang, Huixiao Le, Kejie Shen, Shufang Tan, Yueying Zhao, Yuan Shen, Xinyu Li, and Dragan Gašević. 2025. Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, 56(2):489–530.

Linda Fernsel, Yannick Kalff, and Katharina Simbeck. 2024. Assessing the auditability of AI-integrating systems: A framework and learning analytics case study. *Preprint*, arXiv:2411.08906.

Michael Gerlich. 2025. AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1):6. Corrected 10 September 2025; correction DOI: 10.3390/soc15090252.

Kenneth Holstein and Vincent Aleven. 2022. Designing for human-AI complementarity in K-12 education. *AI Magazine*, 43(2):239–248.

Irene Hou, Owen Man, Kate Hamilton, Srishty Muthusekaran, Jeffin Johnykutty, Leili Zadeh, and Stephen MacNeil. 2025. 'all roads lead to ChatGPT': How generative AI is eroding social interactions and student learning communities. In *Proceedings of the 30th ACM Conference on Innovation and Technology in Computer Science Education V. 1*, pages 79–85.

International Society for Technology in Education. n.d. ISTE standards for students. Living web standard; accessed 25 September 2026.

Greg Kestin, Kelly Miller, Anna Klales, Timothy Milbourne, and Gregorio Ponti. 2025. AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15(1):17458.

Paul A. Kirschner, John Sweller, and Richard E. Clark. 2006. Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41(2):75–86.

Kenneth R. Koedinger, Albert T. Corbett, and Charles Perfetti. 2012. The knowledge–learning–instruction framework: Bridging the science–practice chasm to enhance robust student learning. *Cognitive Science*, 36(5):757–798.

Hao-Ping (Hank) Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–22. Association for Computing Machinery. Article 1121.

Sichao Li, Tommy Liu, Quanling Deng, and Amanda S Barnard. 2024. Exagree: Mitigating explanation disagreement with stakeholder-aligned models. *arXiv preprint arXiv:2411.01956*.

Sichao Li, Sai Ma, Xiyang Hu, Chudi Zhong, and Tongliang Liu. 2026a. Chain-of-thought oversight requires separate evidence for faithfulness and monitorability. In *NeurIPS 2026 Position Track*.

Sichao Li, Sai Ma, Daniel Kilov, Secil Yanik Guyot, Zhuang Li, and Seth Lazar. 2026b. Nora: Evaluating grounded reasonableness in visual first-person normative action reasoning. *arXiv preprint arXiv:2606.04806*.

Sichao Li, Weiyuan Wang, Sai Ma, Ye Wei, and Zhengyi Yang. 2026c. Ownership tracing in ai tutoring: Correcting apparent mastery with source-of-practice evidence. In *Proceedings of the 2026 IEEE International Conference on Data Mining (ICDM)*.

Rose Luckin, Wayne Holmes, Mark Griffiths, and Laurie B. Forcier. 2016. *Intelligence Unleashed: An Argument for AI in Education*. Pearson Education.

Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. MathTutorBench: A benchmark for measuring open-ended pedagogical capabilities of LLM tutors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 204–221, Suzhou, China. Association for Computational Linguistics.

Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251, Albuquerque, New Mexico. Association for Computational Linguistics.

OECD. 2026. OECD digital education outlook 2026: Exploring effective uses of generative AI in education.

Reinhard Pekrun. 2006. The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, 18(4):315–341.

Janos Perczel, Jin Chow, and Dorottya Demszky. 2025. TeachLM: Post-training LLMs for education using authentic learning data. *Preprint*, arXiv:2510.05087.

Jean Piaget. 1952. *The Origins of Intelligence in Children*. International Universities Press.

Romain Puech, Jakub Macina, Julia Chatain, Mrinmaya Sachan, and Manu Kapur. 2025. Towards the pedagogical steering of large language models for tutoring: A case study with modeling productive failure. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26291–26311. Association for Computational Linguistics.

Carl R. Rogers. 1957. The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology*, 21(2):95–103.

Ido Roll and Ruth Wylie. 2016. Evolution and revolution in artificial intelligence in education. *International Journal of Artificial Intelligence in Education*, 26(2):582–599.

William A. Sahlman, Allison M. Ciechanover, and Emily Grandjean. 2023. Khanmigo: Revolutionizing learning with GenAI. *Harvard Business School Case 824-059*. Public case description; full paid case not used for empirical claims.

Norman J. Slamecka and Peter Graf. 1978. The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, 4(6):592–604.

Siyu Song, Wentao Liu, Ye Lu, Ruohua Zhang, Tao Liu, Jinze Lv, Xinyun Wang, Aimin Zhou, Fei Tan, Bo Jiang, and Hao Hao. 2025. Cultivating helpful, personalized, and creative AI tutors: A framework for pedagogical alignment using reinforcement learning. *arXiv preprint arXiv:2507.20335*.

John Sweller, Jeroen J. G. van Merriënboer, and Fred G. W. C. Paas. 1998. Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3):251–296.

UK Department for Education. 2026. Generative AI: Product safety standards. Updated January 2026.

UNESCO. 2023. UNESCO survey: Less than 10% of schools and universities have formal guidance on generative AI.

Lev S. Vygotsky. 1978. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Kathy Walker and Mihaela Vorvoreanu. 2025. Learning outcomes with GenAI in the classroom: A review of empirical evidence. Technical Report MSR-TR-2025-42, Microsoft Research.

Rose E. Wang, Ana T. Ribeiro, Carly D. Robinson, Susanna Loeb, and Dora Demszky. 2024. Tutor CoPilot: A human-AI approach for scaling real-time expertise. *Preprint*, arXiv:2410.03017.

Kodi Weatherholtz, Kelli Millwood Hill, Kristen Dicerbo, Walt Wells, Phillip Grimaldi, Maya Miller-Vedam, Charles Hogg, and Bogdan Yamkovenko. 2025. Cognitive engagement in GenAI tutor conversations: At-scale measurement and impact on learning. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress*, pages 40–48, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).

A Case-study design summary

The illustrative process audit summarized in Section 5 covers 72 transcripts produced by a single audited model (Claude Opus 4 via OpenRouter, May 2026), arranged as 3 conditions $\times$ 4 scenarios $\times$ 2 domains $\times$ 3 repeats. The conditions are A (general-purpose assistant, no tutoring system prompt), B (simple tutor prompt: "do not give the answer; ask questions"), and C (standards-based prompt: ten rules including state identification, attempt-before-answer, progressive hints, and a closing near-transfer task). The domains are introductory programming (recursive `sum_digits`) and database functional-dependency closure ($A \rightarrow C$). The scenarios are `direct_answer`, `misconception`, `passive_learner`, and `time_pressure`. Each transcript is six turns (three learner, three tutor) and is scored on the 12-indicator, 0–2 scale defined in the audit rubric (7 positive S1–S5 indicators and 5 risk indicators).

B Per-condition mean scores (LLM-judge aggregate)

Table 2 reports the per-condition means (with means and standard deviations also stored in `aggregated_results.json`). Both views derive from the same $n = 72$ transcripts and the same Claude Sonnet 4.5 LLM-judge run, blind to condition.

C Pairwise statistical tests

Table 3 reports two-sided Mann–Whitney U tests on the 12 indicators between each pair of conditions. The reported p-values are Bonferroni-corrected over the family of 36 pairwise comparisons (12 indicators $\times$ 3 condition pairs); cells in bold satisfy the family-wide $\alpha = 0.05$ threshold.

D Inter-judge agreement (Cohen's quadratic-weighted κ)

Table 4 reports three-way inter-judge agreement on the 12 indicators between Claude Sonnet 4.5 (Anthropic), Gemini 2.5 Pro (Google), and GPT-4o-mini (OpenAI). The three judges span three vendors and three training paradigms; agreement is reported as the mean pairwise quadratic-weighted Cohen's κ (appropriate for ordinal 0/1/2 scores), the exact-agreement rate, and the within-one-agreement rate, averaged across the three pairs. Mean pairwise $\kappa = 0.68$ overall (substantial on the Landis–Koch scale); the only fair-band indicator is transfer support, where the universally low scores (≤ 0.29 across conditions) leave little ordinal variance for κ to detect. This three-judge agreement serves as a substitute LLM-judge inter-rater-reliability diagnostic; human calibration results are reported in Table 9. Table 5 reports the underlying two-judge κ between Sonnet 4.5 and Gemini 2.5 Pro, included for direct comparison.

E Per-scenario and per-domain breakdowns

Table 6 reports four key indicators by scenario, showing that the cross-condition pattern persists across scenarios, including `time_pressure`: Condition C holds attempt-before-answer at 2.00 and solution-reveal at 0.00 even in the dedicated time-pressure scenario. Table 7 confirms cross-domain consistency across introductory programming and database reasoning, and Table 8 gives the full per-scenario risk-indicator breakdown.

F Human-coded validation subset

To check that the LLM-judge audit was not purely self-referential, we coded a small human-validation subset of the illustrative audit in Table 9. We sampled 18 transcripts balanced across the two domains, three prompt conditions, and three high-salience scenario types: direct-answer request, misconception, and time-pressure / over-completion pressure. Transcripts were anonymized, randomized, and stripped of condition labels before annotation. Two human annotators independently scored each transcript using the same 0–2 ordinal rubric used by the LLM judges. We focus on the indicators most central to the paper's claims: attempt-before-answer, reasoning elicitation, misconception diagnosis, productive hinting, transfer support, relational safety, solution reveal risk, over-completion risk, trivialization risk, and academic-integrity risk. We report quadratic-weighted Cohen's kappa and within-one agreement for each of the ten human-coded indicators. Because this subset is small, these results should be interpreted as a calibration check rather than a full validation study.

Observable behaviors such as attempt-before-answer, solution reveal, and over-completion are expected to be more reliable than higher-level constructs such as transfer support or relational safety. The subset therefore checks whether the main

Table 2: Per-condition mean LLM-judge scores on the 12 audit indicators (Claude Sonnet 4.5 judge, blind to condition; human calibration reported in Table 9). All scores on a 0–2 scale; for positive indicators (A1–A7) higher is better, for risk indicators (B1–B5) lower is better. N transcripts per condition shown in the column header.

Positive indicators (higher = better)

Condition	Attempt-1st	Reason-elicit	Misc-diag	Prod-hint	SRL	Transfer	Rel-safe
A: General (N=24)	0.25	0.00	0.46	0.00	0.00	0.04	0.83
B: Simple tutor (N=24)	1.62	1.54	1.04	1.12	0.71	0.12	1.79
C: Standards-based (N=24)	2.00	1.71	1.21	1.17	1.08	0.29	2.00

Risk indicators (lower = better)

Condition	Sol-reveal	Over-comp	Trivial	Depend-lang	Acad-integ
A: General (N=24)	1.88	1.92	1.75	1.00	1.83
B: Simple tutor (N=24)	0.79	0.79	0.67	0.21	0.67
C: Standards-based (N=24)	0.29	0.29	0.12	0.00	0.12

Table 3: Pairwise Mann–Whitney U tests on the 12 audit indicators ($n = 24$ transcripts per condition; A: general assistant; B: simple tutor; C: standards-based). Two-sided p-values, Bonferroni-corrected across the 36 family-wide tests ($\alpha = 0.05$). Significant cells are bold.

Indicator	A vs. B	A vs. C	B vs. C
Attempt-1st	**<.001**	**<.001**	0.369
Reason-elicit	**<.001**	**<.001**	1.000
Misc-diag	0.810	0.088	1.000
Prod-hint	**<.001**	**<.001**	1.000
SRL	**<.001**	**<.001**	0.081
Transfer	1.000	1.000	1.000
Rel-safe	**<.001**	**<.001**	0.736
Sol-reveal	**<.001**	**<.001**	1.000
Over-comp	**<.001**	**<.001**	1.000
Trivial	**<.001**	**<.001**	0.072
Depend-lang	**<.001**	**<.001**	0.736
Acad-integ	**<.001**	**<.001**	0.281

Table 4: Three-way inter-judge agreement on the 12 audit indicators between Claude Sonnet 4.5, Gemini 2.5 Pro, and GPT-4o-mini ($n = 72$ transcripts; quadratic-weighted Cohen's κ averaged across the 3 pairs). Mean pairwise $\kappa = 0.68$ (min indicator-mean 0.30, max 0.87); mean exact agreement 69.0%, mean within-one agreement 95.6%.

Indicator	$\overline{\kappa}_{\text{wt}}$	Exact %	Within-1 %	Band
Attempt-1st	0.86	83.3	98.2	almost perfect
Reason-elicit	0.81	75.5	97.7	substantial
Misc-diag	0.51	65.7	86.1	moderate
Prod-hint	0.63	63.9	89.8	substantial
SRL	0.56	46.8	94.0	moderate
Transfer	0.30	44.9	89.3	fair
Rel-safe	0.55	77.8	99.1	moderate
Sol-reveal	0.87	80.6	100.0	almost perfect
Over-comp	0.85	77.3	99.5	almost perfect
Trivial	0.72	72.7	94.9	substantial
Depend-lang	0.58	61.6	98.6	moderate
Acad-integ	0.86	78.2	99.5	almost perfect

process-audit categories are human-interpretable and broadly aligned with the primary LLM judge; it does not validate the benchmark as a learning measure or show that any audited system improves learning outcomes.

Table 5: Two-judge agreement on the 12 audit indicators between `claude-sonnet-4.5` and `gemini-2.5-pro` ($n = 72$ transcripts; quadratic-weighted Cohen's κ for ordinal 0/1/2 scores). Mean weighted $\kappa = 0.68$ (min 0.38, max 0.92); mean exact agreement 71.1%, mean within-one agreement 96.0%.

Indicator	κ_{weighted}	Exact %	Within-1 %	Band
Attempt-1st	0.82	86.1	94.4	almost perfect
Reason-elicit	0.82	81.9	95.8	almost perfect
Misc-diag	0.57	63.9	88.9	moderate
Prod-hint	0.82	75.0	98.6	almost perfect
SRL	0.38	29.2	87.5	fair
Transfer	0.39	59.7	98.6	fair
Rel-safe	0.42	70.8	97.2	moderate
Sol-reveal	0.92	86.1	100.0	almost perfect
Over-comp	0.90	81.9	100.0	almost perfect
Trivial	0.65	73.6	90.3	substantial
Depend-lang	0.53	61.1	100.0	moderate
Acad-integ	0.90	83.3	100.0	almost perfect

Table 6: Per-scenario means on four indicators ($n = 3$ repeats per scenario $\times$ condition $\times$ domain; total $n = 6$ per scenario $\times$ condition across both domains). On a 0–2 scale; for positive indicators (A1, A6, A7) higher is better, for risk (B1) lower is better. The time-pressure scenario is the stress test for Condition C; the transfer column records practice opportunities, not unassisted S3 performance.

Condition	Scenario	Attempt-1st (S1↑)	Sol-reveal (S1↓)	Rel-safe (S5↑)	Transfer opp. (↑)
A	direct answer	0.00	2.00	1.00	0.00
	misconception	0.00	2.00	1.00	0.00
	passive learner	1.00	1.50	1.00	0.17
	time pressure	0.00	2.00	0.33	0.00
B	direct answer	1.50	1.50	1.67	0.17
	misconception	2.00	0.33	2.00	0.00
	passive learner	2.00	0.33	2.00	0.17
	time pressure	1.00	1.00	1.50	0.17
C	direct answer	2.00	1.00	2.00	0.17
	misconception	2.00	0.17	2.00	0.33
	passive learner	2.00	0.00	2.00	0.67
	time pressure	2.00	0.00	2.00	0.00

Table 7: Cross-domain consistency on four indicators ($n = 12$ transcripts per domain $\times$ condition; programming = recursive `sum_digits`; database = functional-dependency closure $A \to C$).

Condition	Domain	Attempt-1st (S1↑)	Sol-reveal (S1↓)	Rel-safe (S5↑)	Transfer opp. (↑)
A	programming	0.50	1.75	0.75	0.00
	database	0.00	2.00	0.92	0.08
B	programming	1.50	0.83	1.75	0.08
	database	1.75	0.75	1.83	0.17
C	programming	2.00	0.25	2.00	0.17
	database	2.00	0.33	2.00	0.42

Table 8: Risk indicators (B1–B5) by condition and scenario. Lower is better. Three repeats per cell.

Condition	Scenario	Sol-reveal	Over-comp	Trivial	Depend-lang	Acad-integ
A: General	direct answer	2.00	2.00	2.00	0.50	2.00
A: General	misconception	2.00	2.00	2.00	1.00	2.00
A: General	passive learner	1.50	1.67	1.00	1.17	1.33
A: General	time pressure	2.00	2.00	2.00	1.33	2.00
B: Simple tutor	direct answer	1.50	1.50	1.00	0.50	1.33
B: Simple tutor	misconception	0.33	0.33	0.33	0.00	0.33
B: Simple tutor	passive learner	0.33	0.33	0.33	0.00	0.17
B: Simple tutor	time pressure	1.00	1.00	1.00	0.33	0.83
C: Standards-based	direct answer	1.00	1.00	0.50	0.00	0.33
C: Standards-based	misconception	0.17	0.17	0.00	0.00	0.17
C: Standards-based	passive learner	0.00	0.00	0.00	0.00	0.00
C: Standards-based	time pressure	0.00	0.00	0.00	0.00	0.00

Table 9: Human-coded validation subset agreement ($n = 18$ transcripts; quadratic-weighted Cohen's κ for ordinal 0–2 scores). Human–LLM values compare adjudicated human labels with the primary LLM judge. Within-one agreement is reported for human–LLM calibration.

Indicator	Human–human QWK	Human–LLM QWK	Human–LLM within-one agreement
Attempt-before-answer	0.86	0.97	100.0%
Reasoning elicitation	0.96	0.89	100.0%
Misconception diagnosis	0.90	0.86	100.0%
Productive hinting	0.87	0.81	100.0%
Transfer support	-0.09	0.82	100.0%
Relational safety	0.40	0.27	94.4%
Solution reveal risk	0.93	0.93	100.0%
Over-completion risk	0.97	1.00	100.0%
Trivialization risk	0.14	0.73	100.0%
Academic-integrity risk	0.97	0.81	94.4%
Mean (10 indicators)	0.69	0.81	98.9%