Journal of Inclusive Methodology and Technology in Learning and Teaching
ISSN 2785-5104
Anno 6 n. 3 (2026)

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

When Engagement Is Not Learning: Cognitive Safety by Design in Generative AI Tutoring

Marco Iannacone*

Independent Researcher, Milan, Italy
* Correspondence: ianna@pippo.com

Abstract: Generative AI can sustain fluent tutoring interactions while performing the reasoning learners are expected to develop. This conceptual-methodological paper develops Cognitive Safety by Design through an implementation example (Proxima ZSP) and a descriptive synthesis of published synthetic evidence. Five requirements connect assistance to inspectable evidence: diagnose before intervening, provide minimum sufficient mediation, preserve learner-owned reasoning, fade assistance and verify learning after assistance. Across 1,100 dialogues from the Synthetic Learner Evaluation Framework (SLEF), an automated reader classified 85.7–87.8% of classifiable sessions assigned low metacognition as high. The share of sessions in which transfer could be classified ranged from 0.9% to 37.7%. These results motivate explicit records of the opportunities available in each dialogue. Opportunity-based evaluation distinguishes an absent opportunity, ambiguous evidence and an interpretable response, preventing missing evidence from automatically becoming a judgment of low capability. The developing L×M×C framework—depth of reasoning (L), expressed metacognition (M), and consolidation and transfer (C)—provides provisional outcome hypotheses, including session and spacing criteria. The contribution makes attribution explicit: each claim about a learner is linked to the assistance received, opportunities available and subsequent independent performance.

Keywords: generative AI; learner agency; cognitive safety; cognitive engagement; scaffolding

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

A learner may converse extensively with an AI tutor, complete the task and report satisfaction while contributing little of the intended reasoning. Engagement has behavioral, emotional and cognitive dimensions (Fredricks et al., 2004). Turn counts and persistence are behavioral indicators and satisfaction an emotional one; all three can remain high while cognitive work shifts to the system. This possibility sharpens the distinction between performance during instruction and durable learning (Soderstrom & Bjork, 2015).

In high-school mathematics, GPT Base and GPT Tutor raised practice grades by 48% and 127%, respectively, relative to controls. Without access, the Base group scored 17% lower; the Tutor group's exam performance was statistically indistinguishable from controls (Bastani et al., 2025). Safeguards mitigated the observed harm without demonstrating an unassisted learning gain. Other designs produced greater immediate gains than in-class active learning (Kestin et al., 2025), while a writing experiment found improved products without significant knowledge or transfer advantages (Fan et al., 2025). Outcomes thus depend on how assistance is designed and whether learning is assessed with or without it.

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

This conceptual-methodological paper combines a literature-based synthesis, an implementation example and a descriptive synthesis of published Synthetic Learner Evaluation Framework (SLEF) results. It makes three contributions: characterizing the attribution problem, in which tutors supply part of the reasoning later read as evidence of the learner's understanding; translating five established scaffolding requirements into an audit structure; and developing opportunity-based evaluation into a tutoring-design requirement. It also connects these requirements to provisional outcome hypotheses in L×M×C: depth of reasoning (L), expressed metacognition (M), and consolidation and transfer (C).

2. The Assistance Dilemma and Related Work

Help can enable progress or encourage shallow processing; withholding it can stimulate reasoning or leave learners blocked (Koedinger & Aleven, 2007). Metacognitive tutoring addresses help-seeking quality (Aleven et al., 2006). Intelligent tutoring systems have a substantial evidence base (VanLehn, 2011).

Scaffolding connects assisted performance to increasing autonomy (Vygotsky, 1978; Wood et al., 1976). Van de Pol et al. (2010) identify contingency, fading and transfer of responsibility as its core. Their intervention study (2014) operationalizes contingency through diagnosing, checking the diagnosis, providing support and checking learning. Section 4 adapts this cycle to tutors that can themselves perform the reasoning.

Problem solving can consume cognitive resources needed for schema acquisition (Sweller, 1988), while productive failure can make struggle before instruction beneficial under suitable conditions (Kapur, 2008). The assistance dilemma concerns when and how to guide. ICAP adds that interactive engagement requires substantive contributions from both partners: a learner repeating the tutor's inferences may remain in an active, rather than constructive or interactive, mode (Chi & Wylie, 2014).

Zainuddin et al. (2026) measure learner-level epistemic vigilance, cognitive agency and failure resilience through self-report; this proposal concerns system requirements documented in interaction records. Yao (2026) connects cognitive delegation to credential validity and university policy; the focus here is assistance within tutoring. Maurya et al. (2025) evaluate tutor responses across eight pedagogical dimensions; the present audit follows the origin of key reasoning steps and evidence after assistance.

A companion ICERI2026 paper (Iannacone, 2026c) presents the earlier L×M×C specification and introduces Cognitive Safety by Design as an evaluation principle. This paper develops that principle into auditable design requirements and opportunity-based evaluation, draws on completed SLEF results, and outlines subsequent revisions to metacognition measurement and consolidation criteria.

3. The Attribution Problem in Generative Tutoring

A tutor's turn can take the form of a question while supplying the solution: “Divide both sides by two—so x = 4, right?” If the learner's “yes” is scored as a correct response, the decisive step—selecting and applying the operation—came from the tutor. The tutor thus helps produce the very evidence from which the learner's understanding is inferred. Repeating a supplied explanation and constructing one independently warrant different interpretations.

Three kinds of success should be distinguished. Interaction success means the exchange went well; task success means learner and system together solved the problem. Evidence of learning requires capability that persists beyond the assisted

interaction, examined through subsequent independent performance and delayed checks (Soderstrom & Bjork, 2015).

Cognitive Safety by Design requires support that enables progress while preserving the reasoning and monitoring targeted by instruction. Offloading can be beneficial (Risko & Gilbert, 2016), and its appropriateness depends on the objective. If learners are practicing how to choose a solution method, suggesting that method removes the choice they should make. If they are learning to evaluate methods, offering several candidates can be appropriate support.

4. Cognitive Safety by Design

Five requirements make this boundary inspectable (Table 1). Diagnose before intervening means basing the next move on what the learner has shown and recording how certain that interpretation is. Provide minimum sufficient mediation means matching support to the evidence and objective, without exhausting weak hints before explaining. Preserve learner-owned reasoning requires identifying who contributes the target inference. Fade assistance returns responsibility when evidence warrants it. Verify learning after assistance checks subsequent performance while recording which support remains.

Table 1. Design requirements, evidence to retain and signals requiring examination.

Requirement	Evidence to retain	Signal to examine
Diagnose before intervening	Learner response, provisional interpretation and uncertainty before support	An ambiguous response treated as a certain diagnosis
Provide minimum sufficient mediation	Link between observed difficulty, learning objective and support selected	Decisive help supplied too early, or repeated hints leaving the learner blocked
Preserve learner-owned reasoning	Origin of decisive steps and the learner contribution following support	Repetition of a supplied step scored as independent reasoning
Fade assistance	Evidence supporting withdrawal and performance after reduced support	Support removed after acknowledgment alone
Verify learning after assistance	Relevant new task under specified support, with delayed checks where needed	Assisted task success treated as durable learning

Each signal flags a decision for review. Records capture observations, interpretations and chosen support at decision time, because later reconstruction can be shaped by hindsight. Probing should stop once evidence is sufficient to act; each additional question costs effort and may increase frustration. Fading applies to support for the operation being learned: a learner with dyslexia can retain text-to-speech while mathematical scaffolding is reduced.

5. Proxima ZSP as an Implementation Context

Proxima ZSP is an experimental maieutic tutor for Italian lower-secondary learners, approximately 11–14 years old. It asks learners to attempt each step first and increases support gradually when they remain blocked. Table 1 identifies evidence this implementation should expose.

For example, a learner transforms 2x + 3 = 11 into 2x = 8, then proposes x = 6. The tutor could first ask "How did you get from 2x = 8 to x = 6?" to diagnose the difficulty. Depending on the response, "Substitute x = 6 into 2x = 8: what do you obtain?" can test checking behavior. Such monitoring is elicited, not spontaneous. An inverse-operation cue—"Which operation would undo multiplication by two?"—leaves the learner to select and execute the operation.

A direct explanation may be justified if the learner remains blocked. A new equation, such as 3x + 2 = 17, then checks performance without the same cue. This provides a local check of independence; Section 8 proposes spaced recurrence and new contexts to examine consolidation and transfer. These are conditional choices rather than a mandatory sequence of hints.

6. Synthetic Evidence from SLEF

Testing tutor responsiveness to predefined learner states raises a prior measurement question: can those states be recovered from dialogue? SLEF's reference experiment examines this through 1,100 Italian fraction dialogues: 44 profiles × five sessions × five tutoring conditions, with profiles and task assignments shared across conditions (Iannacone, 2026b, 2026d). Each session starts without memory of earlier sessions.

The learner generator is gpt-5.6-terra. A fixed automated reader (the extractor), gemini-3.1-pro-preview under specification v1.0.7, reads completed dialogues blind to assignments and returns a label or INSUFFICIENT_EVIDENCE. Assigned states are the scoring reference. A mismatch can arise from the model playing the learner, the task, the tutor or the extractor. SLEF's dimensions differ from L×M×C's; recovery describes this complete pipeline.

Table 2. Metacognition classification counts by assigned class, and transfer classification coverage.

Condition	Assigned LOW M LOW / HIGH / IE	Assigned HIGH M LOW / HIGH / IE	Transfer coverage
Proxima	11 / 78 / 21	0 / 102 / 8	40/220 (18.2%)
Claude Vanilla	9 / 54 / 47	1 / 74 / 35	20/220 (9.1%)
OpenAI	6 / 41 / 63	1 / 71 / 38	2/220 (0.9%)
Claude Mainstream	11 / 69 / 30	1 / 97 / 12	76/220 (34.5%)
Mistral	10 / 72 / 28	1 / 96 / 13	83/220 (37.7%)

Note. Triplets are LOW / HIGH / INSUFFICIENT_EVIDENCE counts (110 sessions per assigned class); Transfer coverage counts substantive labels over 220 sessions. Proxima and Claude Vanilla use claude-sonnet-4-5, with and without Proxima's instructional apparatus, respectively. OpenAI uses gpt-5.6-sol and Mistral uses mistral-large-2512, each with a minimal generic tutor wrapper.

Claude Mainstream uses claude-sonnet-5 without Proxima's apparatus. Reasoning settings and output budgets also differ. Sources: Iannacone (2026b, Appendix A) and Iannacone (2026d).

Coverage is the share of sessions receiving a label rather than an abstention; mismatch is the proportion of those labels that differ from the assigned state. For assigned LOW metacognition, coverage ranges from 42.7% to 80.9%, while 85.7–87.8% of labels are HIGH. For assigned HIGH, 98.6–100% of labels are HIGH. Agreement between assigned and extracted labels is low (Cohen's κ = .12–.14, calculated on classified sessions). Summing Table 2 gives 314 LOW→HIGH mismatches and four HIGH→LOW. If a tutoring policy treated such HIGH classifications as evidence of independent monitoring, it could withdraw metacognitive support prematurely. This motivates checking independence before fading, as flagged in Table 1.

Proxima and Claude Vanilla share a base model. LOW coverage is 80.9% versus 57.3%, but LOW→HIGH mismatch remains high: 87.6% versus 85.7%. More classifiable sessions therefore coexist with similarly poor recovery of LOW states.

Transfer coverage ranges from 0.9% to 37.7%. Proxima's 18.2% is below Claude Mainstream's 34.5% and Mistral's 37.7%, prompting examination of its transfer opportunities under requirement 5. Table 2 compares synthetic classification outcomes, not educational effectiveness. SLEF's second phase is being designed to distinguish profile enactment, task opportunities and extractor interpretation through independent annotation and alternative extractors. Scarlatos et al. (2026) report limited fidelity for prompting-based student simulations, supporting checks of whether assigned profiles are actually expressed.

7. Opportunity-Based Evaluation

Missing evidence should not automatically imply low capability. Evidence-centered design connects a task model that creates relevant situations to an evidence model that interprets responses (Mislevy et al., 2003). Learning Factors Analysis relates performance to opportunities to practice knowledge components (Cen et al., 2006). The present proposal applies a related distinction to opportunities to observe metacognition and transfer in dialogue: unlike a practice step, which yields a right or wrong answer, a dialogue turn can offer practice without revealing anything about monitoring or transfer.

What counts as an opportunity must be defined before responses are coded, so that coders do not infer an opportunity from the response they are judging. Transfer requires a relevant change of context or representation; monitoring requires an occasion to predict, check or revise. The target operation must remain available to the learner. A question immediately followed by the tutor's solution can close that opportunity.

Three cases follow: no relevant opportunity; an opportunity yielding ambiguous or insufficient evidence; and an interpretable response, whether weak or strong. The first calls for an appropriate task or probe if the judgment is needed. The second calls for further evidence or an explicitly unresolved judgment. The third supports a judgment qualified by preceding assistance, including whether monitoring was elicited or spontaneous.

In the equation example in Section 5, if the tutor supplies the operation and result, the exchange no longer provides an opportunity to observe the learner choosing that operation independently. If the learner gives only a result, the method remains unclear. If the learner chooses and explains a method, the response offers interpretable evidence even when it contains an error. The opportunity is specific to the skill being assessed: a

supplied solution may still allow the learner to demonstrate how they check or evaluate it.

This develops SLEF's opportunity-based analysis (Iannacone, 2026b) into a design requirement. Independent annotators should code opportunities, responses and the origin of key steps before inspecting extractor labels. Their annotations test both classifications and whether abstention was warranted. One hypothesis is that tutor prompts elicit reflective-sounding language, such as "I checked my result," which the extractor interprets as high metacognition regardless of the assigned state. Linking prompts, opportunities and responses makes that explanation testable.

8. L×M×C Revision and Validation Roadmap

L×M×C separates signals that can diverge (Iannacone, 2026a): a correct answer may involve little observable reasoning, while an incorrect one may expose a testable hypothesis. Reasoning can also occur without expressed metacognitive awareness. Each signal must be interpreted against preceding assistance and available opportunities.

The proposed M revision combines conversational context with patterns across sessions, records whether reflective statements were prompted or spontaneous, and compares confidence ratings with performance. Coaching and diagnostic prompts are identified separately: responses to either can be informative, but warrant different inferences about independent monitoring.

For consolidation (C2), the working hypothesis requires performance at or above a target L level in at least three valid same-domain sessions, with an initial minimum interval of seven days between qualifying sessions. This revises the ICERI specification's "same L level" wording and five-day mathematics default (Iannacone, 2026c). A session at L3 also contributes to consolidation at L2, so improvement does not interrupt accumulation toward the lower threshold. Session count and spacing remain configurable for sensitivity analyses accounting for assistance and task difficulty. Independent delayed performance tests the criterion; transfer hypotheses (C3) require appropriate new contexts.

An 18-month field program is being designed for 2027, subject to school partnerships and ethical approval. Validation rests on evidence external to the framework's classifications: human coding, delayed performance, transfer and metacognitive calibration, with comparison conditions in which direct or hybrid instruction could outperform maieutic support.

9. Discussion

For engagement analytics, interaction volume measures participation while leaving the origin of cognitive work unresolved. Provenance and opportunity records help distinguish reasoning constructed by learners from steps supplied by the system. The more of the target reasoning, explanation and monitoring the tutor performs, the less informative joint task completion becomes as evidence of independent learning. Longer dialogue may support reasoning or prolong dependence. A brief or non-standard answer from a learner with a specific learning disorder or a learner of Italian as a second language should not automatically be coded as weak reasoning.

Open definitions and data make these judgments contestable. L×M×C exposes its criteria; SLEF releases dialogue data, extractor specifications and analysis code. The proposal would be weakened if coders cannot reliably distinguish supplied from

constructed steps, if C2 fails to predict delayed performance, or if learners supported this way perform worse independently than with appropriate alternatives.

10. Conclusions

Cognitive Safety by Design connects engagement to the cognitive activity that remains with learners. The audit structure links assistance decisions to evidence; opportunity-based evaluation distinguishes missing occasions from low capability. Proxima supplies an implementation context, SLEF shows where measurement must improve, and L×M×C makes outcome hypotheses explicit. Their educational value depends on independent validation and criteria open to revision.

Declarations

Acknowledgments: The author thanks researchers and educators who provided early methodological feedback, including D. Antonozzi (University of Camerino) for consultation on the metacognition-measurement architecture.

Competing interests: The author is the developer of Proxima ZSP, SLEF and L×M×C. These roles are relevant to the design and interpretation of the work.

Funding: The research was conducted independently and without external funding.

Use of generative AI: Generative AI tools (ChatGPT and Claude) were used during manuscript preparation for critical review of drafts, restructuring and language editing, rewording of selected passages, and checks of references, length and descriptive statistics derived from the published SLEF data. AI models used as research instruments in SLEF are described in Section 6. The author designed the frameworks and the study, interpreted the results, reviewed and verified all AI-assisted content, and takes full responsibility for the manuscript.

Ethics statement: The reported findings concern synthetic interactions, with no human participants or student records.

Data availability: The SLEF Reference Evaluation Dataset v1.1.0 contains the reference data, analysis code and extractor prompt (Iannacone, 2026d); learner-generation prompts and orchestration code are not included. The L×M×C preprint is public (Iannacone, 2026a); the described revision remains in development.

References

Aleven, V., McLaren, B. M., Roll, I., & Koedinger, K. R. (2006). Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence in Education, 16*(2), 101–128. https://doi.org/10.3233/IRG-2006-16(2)02

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences, 122*(26), e2422633122. https://doi.org/10.1073/pnas.2422633122

Cen, H., Koedinger, K. R., & Junker, B. (2006). Learning factors analysis—A general method for cognitive model evaluation and improvement. In M. Ikeda, K. D. Ashley, & T.-W. Chan (Eds.), *Intelligent tutoring systems* (pp. 164–175). Springer. https://doi.org/10.1007/11774303_17

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (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. https://doi.org/10.1111/bjet.13544

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59–109. https://doi.org/10.3102/00346543074001059

Iannacone, M. (2026a). *A multidimensional framework for evaluating reasoning in AI-mediated maieutic tutoring: Theoretical foundations and operational specification* (Version 1.2.7) [Preprint]. EdArXiv. https://doi.org/10.35542/osf.io/hvx37_v2

Iannacone, M. (2026b). *Can assigned learner states be recovered from synthetic tutoring dialogues? A factorial study with SLEF* [Preprint]. EdArXiv. https://doi.org/10.35542/osf.io/esmx2_v1

Iannacone, M. (2026c). *Preserving learner agency in artificial intelligence-mediated education: An open framework for evaluating learning, metacognition and cognition* [Manuscript prepared for an accepted ICERI2026 contribution].

Iannacone, M. (2026d). *SLEF Reference Evaluation Dataset* (Version 1.1.0) [Data set and software]. Zenodo. https://doi.org/10.5281/zenodo.23068043

Kapur, M. (2008). Productive failure. *Cognition and Instruction, 26*(3), 379–424. https://doi.org/10.1080/07370000802212669

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports, 15*, 17458. https://doi.org/10.1038/s41598-025-97652-6

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with Cognitive Tutors. *Educational Psychology Review, 19*(3), 239–264. https://doi.org/10.1007/s10648-007-9049-0

Maurya, K. K., Srivatsa, K. A., Petukhova, K., & Kochmar, E. (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)* (pp. 1234–1251). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.57

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series, 2003*(1), i–29. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Scarlatos, A., Lee, J., Woodhead, S., & Lan, A. (2026). Simulated students in tutoring dialogues: Substance or illusion? In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 42349–42385). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.acl-long.1960

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science, 10*(2), 176–199. https://doi.org/10.1177/1745691615569000

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher-student interaction: A decade of research. *Educational Psychology Review, 22*(3), 271–296. https://doi.org/10.1007/s10648-010-9127-6

van de Pol, J., Volman, M., Oort, F., & Beishuizen, J. (2014). Teacher scaffolding in small-group work: An intervention study. *Journal of the Learning Sciences, 23*(4), 600–650. https://doi.org/10.1080/10508406.2013.805300

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Yao, K. (2026). *What does the credential still certify? Cognitive stewardship for AI-mediated education* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2607.19988

Zainuddin, Z., Churiyah, M., Xiaoyu, W., & Halili, S. H. (2026). From AI literacy to cognitive AI safety: Rethinking safe AI use in education. *Interactive Technology and Smart Education*. Advance online publication. https://doi.org/10.1108/ITSE-12-2025-0396