

Structured AI Tutoring vs. Cognitive Offloading in Undergraduate STEM

A citation-lineage and evidence map centered on Kestin et al. (2025)

Research question

Research question: Does using AI tools for problem-solving improve undergraduate STEM students' learning and long-term retention, or does it create a cognitive-offloading effect that reduces skill-building - and does the outcome depend on how the AI is used, such as answer-giving versus Socratic tutoring?



Conceptual illustration: the same AI interface can sit on a learning path that preserves cognitive work or a shortcut path that displaces it; the empirical literature increasingly suggests the difference is pedagogical design and user behavior, not AI access alone.

Seed paper	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. <i>Scientific Reports</i> , 15, 17458. doi:10.1038/s41598-025-97652-6
Scope	20 related papers selected for conceptual lineage and research-question relevance: 10 confirmed antecedents cited by the seed + 10 recent descendants/adjacent studies. The seed itself is shown separately and is not counted in the 20-paper limit.
Literature check	Recent-source verification performed through 18 August 2026. Direct citation/extension relationships are labeled separately from inferred thematic relationships.

Executive synthesis

Bottom line

The evidence does not support a simple "AI helps" or "AI harms" conclusion. Structured AI that keeps the learner cognitively active - prompting attempts, providing targeted explanations or hints, asking Socratic questions, and withholding complete solutions when appropriate - can improve learning. By contrast, open-ended answer-giving can raise assisted task performance while reducing independent performance, persistence, study time, or retention once the tool is removed.

What the seed establishes. Kestin et al. randomized 194 undergraduates in introductory physics to a carefully engineered GPT-4 tutor versus a strong active-learning classroom condition. The AI condition produced substantially higher immediate post-test performance in less median time and higher engagement/motivation, demonstrating that an AI tutor can outperform current best-practice instruction when its interaction is tightly scaffolded.

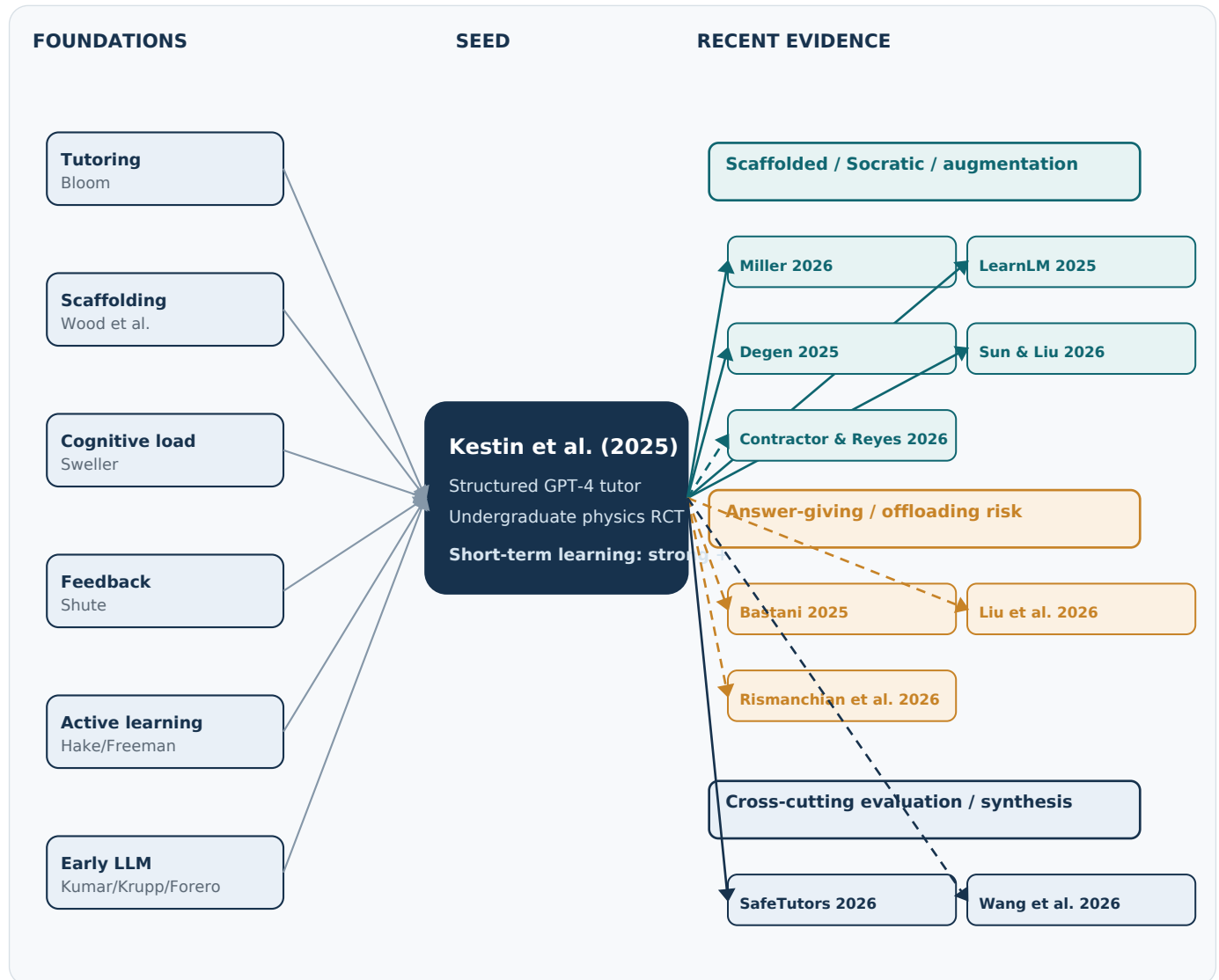
What the seed does not establish. The intervention spans two lessons and measures immediate post-tests; the authors explicitly identify retention, longer course integration, higher-order synthesis, and the decision of when to give an answer versus guide reflection as open research directions. That limitation is decisive for your question.

What subsequent evidence adds. A direct extension (Miller, 2026) shows that pedagogical system prompts change engagement and cognitive load. Later/parallel experiments sharpen the mechanism: guarded or Socratic tutoring preserves cognitive work, while unrestricted answer access can create dependence. Contractor and Reyes (2026) add one-week undergraduate evidence in which explanation-seeking "augmentation" users retain gains better than text-generating "automation" users.

Where cognitive offloading becomes visible. Bastani et al. (2025) observe a reversal from higher assisted performance to lower unassisted performance under unrestricted GPT-4; Liu et al. (2026) find reduced independent performance and persistence after brief AI exposure; Rismanchian et al. (2026) report less time on AI-susceptible college-math problems and weaker later proctored retention in a large longitudinal dataset.

Current answer to the research question. The most defensible conclusion is conditional: AI can improve learning when it behaves like a tutor and is used as an augmentation; it can reduce skill-building when it behaves like an answer engine and is used as a replacement for reasoning. What remains under-tested is the exact combination you care most about: a long-duration, undergraduate STEM RCT that randomizes answer-giving versus Socratic scaffolding and measures delayed, unaided transfer and retention.

Relationship map



Solid arrows indicate a verified citation or explicit extension relationship. Dashed arrows indicate thematic adjacency selected for the research question; those papers were not confirmed as direct citations of the seed in the accessible full text searched.

1. The seed paper as the pivot

Seed: 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. doi:10.1038/s41598-025-97652-6.

Dimension	What the seed paper actually shows	Implication for your question
Population / context	194 eligible students in Harvard Physical Sciences 2, an introductory physics course for life sciences; crossover design across two lessons.	Strong undergraduate STEM relevance, but one institution and a narrow learning window.
Comparator	Research-based in-class active learning, not passive lecture.	The effect is not merely "AI beats lecture"; the tutor outperformed an already effective pedagogy.
AI design	GPT-4 tutor with instructor-crafted prompts, sequential scaffolding, brief responses, step-by-step reference solutions, targeted feedback, and self-pacing.	The intervention is closer to an engineered Socratic/scaffolded tutor than to unrestricted ChatGPT.
Outcome	Immediate post-test learning; median AI post-score 4.5 versus 3.5 for active learning, with large adjusted effects; AI students also reported greater engagement and motivation.	Strong evidence for immediate acquisition and user experience, not durable retention.
Critical gap	The authors call for studies of spacing/retention, extended course use, higher-order synthesis, and decisions about when the tutor should openly provide answers versus guide reflection.	Your question is a direct continuation of the seed paper's stated limitations and future-work agenda.

Interpretive caution

Because the seed tutor explicitly discourages giving away the full solution and embeds scaffolding, it should not be generalized to "ChatGPT use" in the abstract. The treatment is a particular pedagogy delivered through an LLM. That distinction explains why later studies of unrestricted answer access can produce opposite effects without actually contradicting Kestin et al.

2. Foundational papers the seed builds on

All ten papers below are confirmed antecedents: they appear in the seed paper's reference list and are used to motivate its tutoring design, active-learning comparator, learning/perception measures, or contrast with unguided LLM use.

CONFIRMED ANTECEDENT	1. Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. <i>Educational Researcher</i>, 13(6), 4-16. doi:10.3102/0013189X013006004 Relationship to seed: Directly cited by Kestin et al. as the benchmark for the effectiveness - and scalability problem - of expert one-to-one tutoring. One-sentence summary: Bloom framed individualized tutoring as a high-performing instructional benchmark, motivating the seed paper's attempt to reproduce tutor-like personalization at scale with AI. Research-question fit: Foundational to the tutoring side of the question, but it predates generative AI and does not address offloading.
CONFIRMED ANTECEDENT	2. Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem-solving. <i>Journal of Child Psychology and Psychiatry</i>, 17(2), 89-100. Relationship to seed: Directly cited by the seed paper as the basis for scaffolding content and sequencing support during problem solving. One-sentence summary: This classic tutoring study formalized scaffolding as contingent support that helps learners perform beyond current capability while preserving learner participation in the solution process. Research-question fit: Provides the clearest pre-AI rationale for Socratic or hint-based assistance rather than simply supplying answers.
CONFIRMED ANTECEDENT	3. Sweller, J. (2011). Cognitive load theory. In J. P. Mestre & B. H. Ross (Eds.), <i>The Psychology of Learning and Motivation: Cognition in Education</i> (pp. 37-76). Elsevier Academic Press. Relationship to seed: Directly cited and explicitly implemented in the AI tutor design through brief, targeted exchanges intended to manage working-memory demands. One-sentence summary: Cognitive load theory argues that instruction should reduce extraneous load and manage task complexity so limited working memory can be devoted to schema construction. Research-question fit: Central to distinguishing productive offloading of irrelevant burden from harmful offloading of the reasoning practice itself.
CONFIRMED ANTECEDENT	4. Shute, V. J. (2008). Focus on formative feedback. <i>Review of Educational Research</i>, 78(1), 153-189. doi:10.3102/0034654307313795 Relationship to seed: Directly cited by Kestin et al. to justify targeted, timely feedback as a core design principle of the tutor. One-sentence summary: Shute synthesizes evidence that formative feedback is most useful when it is timely, specific, non-evaluative, and oriented toward closing the gap between current and desired performance. Research-question fit: Supports AI as a feedback partner that responds to student thinking rather than a solution generator.

CONFIRMED ANTECEDENT	5. Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. American Journal of Physics, 66(1), 64-74. doi:10.1119/1.18809 Relationship to seed: Directly cited as evidence that interactive engagement outperforms traditional lecture in introductory physics, helping define the strong control condition the AI tutor must beat. One-sentence summary: Across thousands of introductory physics students, interactive-engagement methods produced substantially larger conceptual gains than traditional instruction. Research-question fit: Establishes active problem solving as the benchmark against which AI-supported learning should be judged.
CONFIRMED ANTECEDENT	6. Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. Proceedings of the National Academy of Sciences, 111(23), 8410-8415. doi:10.1073/pnas.1319030111 Relationship to seed: Directly cited to establish active learning as an evidence-based STEM standard rather than a weak lecture-only comparator. One-sentence summary: This meta-analysis of 225 STEM studies found that active learning improves exam performance and lowers failure rates relative to traditional lecturing. Research-question fit: Makes the seed RCT more consequential: its comparison is against an established effective pedagogy, not passive instruction.
CONFIRMED ANTECEDENT	7. Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. Proceedings of the National Academy of Sciences, 116(39), 19251-19257. doi:10.1073/pnas.1821936116 Relationship to seed: Directly cited and authored partly by the same team; it motivates the seed paper's separation of objective learning gains from students' subjective experience. One-sentence summary: Students in active classrooms learned more even when they felt they learned less, demonstrating that perceived fluency or enjoyment can diverge from actual learning. Research-question fit: Important for AI studies because a frictionless, confidence-inducing tool can feel helpful while reducing independent skill formation.
CONFIRMED ANTECEDENT	8. Kumar, H., Rothschild, D. M., Goldstein, D. G., & Hofman, J. M. (2023). Math education with large language models: Peril or promise? SSRN Electronic Journal. doi:10.2139/ssrn.4641653 Relationship to seed: Directly cited by Kestin et al. as early mixed evidence on LLM-supported learning and as a precursor to testing how instructional structure changes outcomes. One-sentence summary: In a preregistered experiment with about 1,200 participants, LLM explanations improved later math performance over answers alone, with the largest gains when learners attempted problems before consulting the explanation. Research-question fit: One of the most direct antecedents for the answer-giving versus explanation/attempt-first dimension in the research question.

CONFIRMED ANTECEDENT	9. Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K. E., Lukowicz, P., Kuhn, J., Küchemann, S., & Karolus, J. (2024). Unreflected acceptance - Investigating the negative consequences of ChatGPT-assisted problem solving in physics education. In HHAi 2024: Hybrid Human AI Systems for the Social Good (pp. 199-212). doi:10.3233/FAIA240195 Relationship to seed: Directly contrasted with the seed results as evidence that unrestricted ChatGPT use can elicit shallow interaction and uncritical acceptance. One-sentence summary: Physics-background participants frequently over-trusted incorrect ChatGPT-supported solutions and used copy-and-paste prompting far more than search-engine users, indicating limited reflection during problem solving. Research-question fit: Directly supports the cognitive-offloading concern in higher-education STEM problem solving.
CONFIRMED ANTECEDENT	10. Forero, M. G., & Herrera-Suárez, H. J. (2023). ChatGPT in the classroom: Boon or bane for physics students' academic performance? arXiv preprint arXiv:2312.02422. arXiv:2312.02422 Relationship to seed: Directly contrasted by Kestin et al. as a prior physics study in which relatively unstructured ChatGPT integration was associated with poorer academic performance. One-sentence summary: A second-semester engineering-physics cohort encouraged to use ChatGPT earned lower exam grades than a prior control cohort, while many students simultaneously reported convenience and concerns about reduced critical or independent thinking. Research-question fit: A close pre-seed warning that unguided AI access may harm learning even when students perceive it as useful.

3. Recent papers that extend, cite, or directly test the seed paper's thesis

The first five are verified citation-lineage papers. The final five are deliberately included as adjacent evidence because they address the exact causal mechanisms in your question even though I did not confirm a direct citation to Kestin et al. in the accessible full text.

CONFIRMED EXTENSION	1. Miller, K. (2026). Prompt Matters: How Pedagogical Engineering Shapes Behavior and Engagement with AI Tutors. Technology, Knowledge and Learning. doi:10.1007/s10758-026-09983-6 Relationship to seed: The paper explicitly states that it uses the same framework as Kestin et al. and acknowledges the work as an extension of the earlier experiment. One-sentence summary: In a randomized study (N=125), an enhanced tutor prompt incorporating active learning, cognitive-load management, and growth-mindset support produced more student-AI interaction, higher engagement and motivation, and less cognitive overload than a minimal prompt, while both groups learned. Research-question fit: Directly isolates prompt-level pedagogy as a mechanism, although it does not establish long-term retention.
CONFIRMED DESCENDANT	2. LearnLM Team, Google & Eedi. (2025). AI tutoring can safely and effectively support students: An exploratory RCT in UK classrooms. arXiv preprint arXiv:2512.23633. arXiv:2512.23633 Relationship to seed: The paper directly cites Kestin et al. in its review of evidence for effective generative-AI tutoring. One-sentence summary: In an exploratory RCT with 165 secondary-school students, human-supervised LearnLM tutoring matched human tutoring and improved success on novel subsequent-topic problems by 5.5 percentage points, with tutors specifically praising its Socratic questions. Research-question fit: Strong evidence for guided dialogue and transfer, but the population is K-12 rather than undergraduates.
CONFIRMED DESCENDANT	3. Hazra, R., Ghuku, B., Marchenko, I., Tokarieva, Y., Layek, S., Banerjee, S., Stoyanovich, J., & Pechenizkiy, M. (2026). SafeTutors: Benchmarking Pedagogical Safety in AI Tutoring Systems. arXiv preprint arXiv:2603.17373. arXiv:2603.17373 Relationship to seed: The benchmark paper directly cites Kestin et al. while reframing educational AI safety around preserving learning effort and scaffolding, not merely avoiding harmful content. One-sentence summary: Across mathematics, physics, and chemistry, SafeTutors finds widespread pedagogical harms such as answer over-disclosure and collapsed scaffolding, with multi-turn failure rates rising sharply in its evaluation scenarios. Research-question fit: Does not measure student learning directly, but operationalizes the exact failure mode behind answer-giving and cognitive offloading.

CONFIRMED PREPRINT-LINEAGE	4. Degen, P.-B., & Asanov, I. (2025). Beyond Automation: Socratic AI, Epistemic Agency, and the Implications of the Emergence of Orchestrated Multi-Agent Learning Architectures. arXiv preprint arXiv:2508.05116. arXiv:2508.05116 Relationship to seed: The paper cites the 2024 Kestin et al. preprint version and extends the design logic toward explicitly Socratic, dialogue-driven AI support. One-sentence summary: In a controlled experiment with 65 pre-service teacher students, a Socratic tutor was rated as providing more support for critical, independent, and reflective thinking than an uninstructed chatbot. Research-question fit: Directly tests Socratic versus generic chatbot interaction, but outcomes are largely self-reported rather than delayed performance measures.
CONFIRMED DESCENDANT	5. Sun, Y., & Liu, F. (2026). The impact of an AI Digital Teacher on human-AI collaborative learning in higher education. Smart Learning Environments, 13, 30. doi:10.1186/s40561-026-00454-0 Relationship to seed: The authors directly cite Kestin et al. as evidence of strong gains in structured physics and frame their study as testing whether such benefits generalize to a complex non-STEM domain. One-sentence summary: A semester-long RCT with 84 undergraduates found that adding a bespoke AI teacher to traditional literature instruction improved objective tests and analytical essays while reducing extraneous and increasing germane cognitive load. Research-question fit: Useful evidence for duration and cognitive-load mechanisms, but not a STEM problem-solving test.
INFERRED / PARALLEL	6. 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. doi:10.1073/pnas.2422633122 Relationship to seed: I did not confirm a direct citation to the seed paper in accessible full text; it is included because it is the closest causal test of unrestricted answer-giving versus a learning-protective tutor design. One-sentence summary: In a field experiment with nearly 1,000 high-school math students, unrestricted GPT-4 raised assisted performance but led to 17% lower unassisted grades after access was removed, whereas a guarded GPT Tutor largely mitigated that learning penalty. Research-question fit: Probably the strongest current causal evidence that design and answer access determine whether AI becomes a crutch.
INFERRED / ADJACENT	7. Contractor, Z., & Reyes, G. (2026). Experimental Evidence on the Learning Impact of Generative AI. arXiv preprint arXiv:2607.08849. arXiv:2607.08849 Relationship to seed: No direct citation to Kestin et al. was confirmed in the accessible version; it is selected because it directly fills the seed paper's retention and usage-mode gap. One-sentence summary: In a randomized experiment with 211 undergraduates, AI access improved unaided knowledge immediately and about one week later, but delayed gains were concentrated among augmentation users who sought explanations, while automation users' short-run gains vanished once AI was removed. Research-question fit: The clearest undergraduate evidence so far that augmentation versus automation predicts durable benefit, though the task is not a standard STEM course problem set.

INFERRED / ADJACENT	8. Liu, G., Christian, B., Dumbalska, T., Bakker, M. A., & Dubey, R. (2026). AI Assistance Reduces Persistence and Hurts Independent Performance. arXiv preprint arXiv:2604.04721. arXiv:2604.04721 Relationship to seed: The accessible full text contains no Kestin citation; it is included as mechanism-level causal evidence that immediate AI assistance can undermine independent skill and persistence. One-sentence summary: Across randomized experiments totaling 1,222 participants, AI assistance improved performance while available but reduced later unassisted performance and increased giving up after only brief exposure, especially when users obtained direct answers. Research-question fit: Strong causal evidence for a cognitive-offloading/persistence mechanism, with less direct correspondence to authentic undergraduate STEM courses.
INFERRED / ADJACENT	9. Rismanchian, S., Uzun, H., Matayoshi, J., Cosyn, E., & Kurd-Misto, E. (2026). Faster Completion, Less Learning: Generative AI Reduced Study Time on Math Problems and the Knowledge They Build. arXiv preprint arXiv:2605.21629. arXiv:2605.21629 Relationship to seed: The paper does not cite Kestin et al. in the accessible version; it is included because it tests long-run college-math behavior and delayed proctored retention at scale. One-sentence summary: Using 3.2 million ALEKS learning interactions over a decade, the authors estimate a 26.9% post-ChatGPT decline in time spent on AI-susceptible college math problems and a 25% cumulative decline in the odds of correct responses on later proctored retention items. Research-question fit: Highly relevant to durable STEM learning, but the quasi-experimental design infers AI use from task susceptibility rather than observing prompts directly.
INFERRED / SYNTHESIS	10. Wang, G., Wang, W., Yang, D., & Ren, J. (2026). Generative AI, Cognitive Offloading, and Learner Agency in Higher Education: A Scoping Review. <i>Behavioral Sciences</i>, 16(7), 1150. doi:10.3390/bs16071150 Relationship to seed: I did not find a direct Kestin citation in the accessible article; it is included as a current higher-education synthesis of the augmentation-versus-replacement mechanism central to the seed paper's design logic. One-sentence summary: Synthesizing 123 higher-education studies, the review concludes that scaffolded, self-regulated, augmentation-oriented AI use is associated with learner agency, whereas replacement-oriented use clusters with offloading, overreliance, dependence, and weakened judgment. Research-question fit: A useful map of the broader literature, but its conclusions are configurative rather than causal and not STEM-specific.

4. What the 20-paper map says about your research question

Sub-question	Best current reading	Most probative papers	Confidence / gap
Does AI improve immediate problem-solving performance?	Often yes, especially when explanations, feedback, or tutoring are available; assisted performance is not the same as learning.	Kestin 2025; Kumar 2023; Bastani 2025; Miller 2026	High confidence for immediate performance; interpretation must separate assisted output from unaided knowledge.
Does AI improve durable, unaided learning?	Conditionally. One-week undergraduate gains can persist under augmentation, but unrestricted solution access can reduce later unassisted performance or retention.	Contractor & Reyes 2026; Bastani 2025; Liu et al. 2026; Rismanchian et al. 2026	Moderate confidence; direct long-duration undergraduate STEM RCTs remain scarce.
Is there a cognitive-offloading / deskilling effect?	Yes under some usage patterns. Evidence appears as copying/over-trust, reduced persistence, reduced time-on-task, and weaker unassisted performance.	Krupp 2024; Bastani 2025; Liu et al. 2026; Rismanchian et al. 2026	Moderate-to-high for mechanism; ecological and causal strength varies by study.
Does answer-giving versus Socratic/scaffolded tutoring matter?	Very likely yes. Attempt-before-help, pedagogical prompts, guardrails, and Socratic questions consistently point toward greater engagement, reflection, transfer, or preserved independent ability.	Kumar 2023; Miller 2026; LearnLM 2025; Degen 2025; Bastani 2025; SafeTutors 2026	Strong convergent evidence, but few studies randomize the exact same model into answer-giving versus Socratic modes with delayed STEM tests.
Can we claim long-term retention benefits in undergraduate STEM?	Not yet. The seed is short-term; the strongest delayed evidence either uses non-STEM tasks, younger students, broad participant samples, or quasi-experimental college math data.	Kestin 2025; Contractor & Reyes 2026; Rismanchian et al. 2026	Low-to-moderate; this is the central research gap.

The emerging mechanism

A useful working model for a new study

AI as cognitive amplifier: student attempts first -> AI diagnoses or asks a question -> targeted hint/explanation -> student produces the next reasoning step -> delayed unaided retrieval/transfer.

AI as cognitive substitute: student delegates the reasoning step -> AI returns a complete solution -> student verifies superficially or copies -> immediate task success rises while practice opportunities, persistence, and durable schema construction fall.

A rigorous next RCT would hold content, model, time, and interface constant while randomizing only the assistance policy: (A) full answer on request, (B) explanation after an independent attempt, and (C) Socratic/hint-based tutoring that withholds full solutions until predefined thresholds. Outcomes should include immediate mastery, unaided transfer, delayed retention at one week and several weeks, time-on-task, persistence, hint/answer-seeking behavior, and process traces that distinguish productive self-explanation from cognitive surrender.

The most informative primary endpoint for your exact question is not homework completion or post-test performance while AI is present; it is delayed unaided performance on novel or isomorphic problems after AI removal. A secondary endpoint should capture whether students continue attempting difficult problems before seeking an answer.

5. Selection method, relationship audit, and limitations

Selection rule. I prioritized papers that are closest to the target construct: undergraduate or STEM problem solving; objective learning or retention; manipulation or observation of AI assistance style; scaffolding/Socratic interaction; and evidence of cognitive offloading, overreliance, or independent performance after AI removal. Canonical papers were retained only when they directly explain the seed tutor's design.

Relationship verification. "Confirmed antecedent" means the work is cited by the uploaded seed paper. "Confirmed extension/descendant" means the newer paper directly cites the published seed or its 2024 preprint, or explicitly calls itself an extension. "Inferred / adjacent" means the paper was selected because it tests the same causal mechanism, but a direct citation relationship was not confirmed in the accessible full text searched.

Publication-status caveat. Several of the most directly relevant 2026 studies are arXiv preprints. Their results should be treated as provisional until peer review and replication. Conversely, peer-reviewed status does not remove limitations in population, intervention duration, or causal identification.

Bibliometric caveat. This report is a targeted evidence map, not an exhaustive citation census. The seed article has accumulated a fast-moving citing literature, and some recent papers may be indexed unevenly across databases. The ten recent selections are therefore optimized for your research question, not for raw citation count.

Recent paper	Relationship status used here	Verification basis
Miller, K et al.	CONFIRMED EXTENSION	Direct statement/citation located in accessible full text.
LearnLM Team, Google & Eedi et al.	CONFIRMED DESCENDANT	Direct statement/citation located in accessible full text.
Hazra, R et al.	CONFIRMED DESCENDANT	Direct statement/citation located in accessible full text.
Degen, P et al.	CONFIRMED PREPRINT-LINEAGE	Direct statement/citation located in accessible full text.
Sun, Y et al.	CONFIRMED DESCENDANT	Direct statement/citation located in accessible full text.
Bastani, H et al.	INFERRED / PARALLEL	No direct seed citation confirmed in accessible full text; thematic relationship is an inference.
Contractor, Z et al.	INFERRED / ADJACENT	No direct seed citation confirmed in accessible full text; thematic relationship is an inference.
Liu, G et al.	INFERRED / ADJACENT	No direct seed citation confirmed in accessible full text; thematic relationship is an inference.
Rismanchian, S et al.	INFERRED / ADJACENT	No direct seed citation confirmed in accessible full text; thematic relationship is an inference.
Wang, G et al.	INFERRED / SYNTHESIS	No direct seed citation confirmed in accessible full text; thematic relationship is an inference.

Primary source trail for recent papers

- Miller (2026): Springer article and notes explicitly identify the same AI tutoring framework as Kestin et al. and call the study an extension.
- LearnLM Team (2025): arXiv full text directly lists Kestin et al. (2025) in its references and discusses mixed evidence on guarded versus unguarded AI tutoring.

- Sun & Liu (2026): Springer/Smart Learning Environments article directly cites Kestin et al. when motivating the gap beyond structured physics.
- SafeTutors (2026): arXiv full text directly cites Kestin et al. and defines answer over-disclosure and lost scaffolding as pedagogical safety risks.
- Degen & Asanov (2025): arXiv PDF directly cites the 2024 Research Square preprint version of the Kestin study.
- Bastani et al. (2025), Contractor & Reyes (2026), Liu et al. (2026), Rismanchian et al. (2026), and Wang et al. (2026): included for topical/causal relevance; no direct Kestin citation was confirmed in the accessible full text searched.

6. Compact bibliography: the 20 selected related papers

Seed paper is listed on the cover and in Section 1; the 20 entries below are the selected antecedent and recent papers requested.

- Bloom, B. S. (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. doi:10.3102/0013189X013006004
- 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.
- Sweller, J. (2011). Cognitive load theory. In J. P. Mestre & B. H. Ross (Eds.), *The Psychology of Learning and Motivation: Cognition in Education* (pp. 37-76). Elsevier Academic Press.
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153-189. doi:10.3102/0034654307313795
- Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1), 64-74. doi:10.1119/1.18809
- Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410-8415. doi:10.1073/pnas.1319030111
- Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences*, 116(39), 19251-19257. doi:10.1073/pnas.1821936116
- Kumar, H., Rothschild, D. M., Goldstein, D. G., & Hofman, J. M. (2023). Math education with large language models: Peril or promise? *SSRN Electronic Journal*. doi:10.2139/ssrn.4641653
- Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K. E., Lukowicz, P., Kuhn, J., Küchemann, S., & Karolus, J. (2024). Unreflected acceptance - Investigating the negative consequences of ChatGPT-assisted problem solving in physics education. In *HHAI 2024: Hybrid Human AI Systems for the Social Good* (pp. 199-212). doi:10.3233/FAIA240195
- Forero, M. G., & Herrera-Suárez, H. J. (2023). ChatGPT in the classroom: Boon or bane for physics students' academic performance? arXiv preprint arXiv:2312.02422. arXiv:2312.02422
- Miller, K. (2026). Prompt Matters: How Pedagogical Engineering Shapes Behavior and Engagement with AI Tutors. *Technology, Knowledge and Learning*. doi:10.1007/s10758-026-09983-6
- LearnLM Team, Google & Eedi. (2025). AI tutoring can safely and effectively support students: An exploratory RCT in UK classrooms. arXiv preprint arXiv:2512.23633. arXiv:2512.23633
- Hazra, R., Ghuku, B., Marchenko, I., Tokarieva, Y., Layek, S., Banerjee, S., Stoyanovich, J., & Pechenizkiy, M. (2026). SafeTutors: Benchmarking Pedagogical Safety in AI Tutoring Systems. arXiv preprint arXiv:2603.17373. arXiv:2603.17373
- Degen, P.-B., & Asanov, I. (2025). Beyond Automation: Socratic AI, Epistemic Agency, and the Implications of the Emergence of Orchestrated Multi-Agent Learning Architectures. arXiv preprint arXiv:2508.05116. arXiv:2508.05116
- Sun, Y., & Liu, F. (2026). The impact of an AI Digital Teacher on human-AI collaborative learning in higher education. *Smart Learning Environments*, 13, 30. doi:10.1186/s40561-026-00454-0
- 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. doi:10.1073/pnas.2422633122
- Contractor, Z., & Reyes, G. (2026). Experimental Evidence on the Learning Impact of Generative AI. arXiv preprint arXiv:2607.08849. arXiv:2607.08849
- Liu, G., Christian, B., Dumbalska, T., Bakker, M. A., & Dubey, R. (2026). AI Assistance Reduces Persistence and Hurts Independent Performance. arXiv preprint arXiv:2604.04721. arXiv:2604.04721
- Rismanchian, S., Uzun, H., Matayoshi, J., Cosyn, E., & Kurd-Misto, E. (2026). Faster Completion, Less Learning: Generative AI Reduced Study Time on Math Problems and the Knowledge They Build. arXiv preprint arXiv:2605.21629. arXiv:2605.21629
- Wang, G., Wang, W., Yang, D., & Ren, J. (2026). Generative AI, Cognitive Offloading, and Learner Agency in Higher Education: A Scoping Review. *Behavioral Sciences*, 16(7), 1150. doi:10.3390/bs16071150