When and How Does LLM Support Help Learning? Toward Data-Driven Design Principles for AI-Enhanced Educational Interfaces

Eason Chen
Carnegie Mellon University
eason.tw.chen@gmail.com

ABSTRACT

Large language models are rapidly deployed in educational systems, yet fundamental questions remain about when LLM support helps learning and when it may hinder it. This dissertation develops design principles for LLM-enhanced educational interfaces through three studies: (1) an experiment showing that expertise paradoxically reduces AI collaboration effectiveness in educational decision-making; (2) a classroom deployment revealing that conversational chatbot usage is negatively associated with exam performance while structured proof-review is not; and (3) a controlled experiment demonstrating that open-ended self-explanation with LLM feedback significantly enhances transfer in calculus ($p = .030$) while requiring $3.5\times$ fewer practice problems. Drawing on learning analytics and educational data mining perspectives, this work contributes methods for automated assessment of student reasoning ($\kappa = 0.78$ vs. human consensus), behavioral pattern detection from interaction logs, and adaptive feedback design informed by data-driven student models.

Keywords

Large language models, Self-explanation, Automated assessment, Learning analytics, Learner behavior modeling, Human-AI collaboration, Educational data mining

1. PROBLEM AND MOTIVATION

The integration of large language models into education promises scalable, personalized support that was previously possible only through human tutoring or expensive rule-based systems [9]. However, emerging evidence suggests that the *form* of LLM support matters as much as its availability. Unconstrained conversational AI can reduce student-generated reasoning and encourage answer outsourcing [15], while structured interfaces that anchor feedback in student work may better preserve productive struggle [1]. Yet we lack a systematic understanding of *which* design choices matter and *why*.

Self-explanation, the process of articulating the reasoning behind a solution, is one of the most robust learning strategies in educational research [3, 12]. However, it faces longstanding practical barriers: it is time-costly [4], explanation quality varies widely [2], and benefits may disappear when learning time is equated with additional practice [10]. LLMs could provide scalable feedback on student-generated explanations, but whether this preserves the generative processing that makes self-explanation effective remains unclear.

These challenges point to a fundamental gap: while LLMs are being widely adopted in education, we lack principled design guidelines grounded in empirical evidence about which forms of AI support enhance learning and which undermine it. My dissertation addresses this gap through the following research question:

How can LLM-supported self-explanation be designed to maximize learning transfer, and what design principles govern the effectiveness of AI feedback on student reasoning?[1]

2. THEORETICAL FRAMING

I draw on three complementary perspectives.

Self-explanation theory establishes that articulating reasoning promotes learning by helping learners identify gaps and integrate new knowledge [3]. Decades of work in intelligent tutoring systems has shown that prompting students to explain solution steps improves learning, particularly when combined with tutorial dialogue that ensures explanation quality [1, 2]. The open question is whether LLM feedback preserves these generative benefits or short-circuits them.

The ICAP framework [5] predicts that learning outcomes improve as engagement moves from passive to active to constructive to interactive. Self-explanation is constructive; when LLM feedback responds adaptively to student explanations, the interaction may approach the interactive level, where each partner's contribution triggers further knowledge-building. This predicts that open-ended self-explanation with adaptive AI feedback should outperform menu-based selection.

The assistance dilemma [6] and the Knowledge-Learning-Instruction (KLI) framework [7] provide a principled basis for understanding when instructional support helps versus hin-

Eason Chen. When and How Does LLM Support Help Learning? Toward Data-Driven Design Principles for AI-Enhanced Educational Interfaces. In Anthony Botelho, Maria Mercedes T. Rodrigo, Adish Singla, Hiroaki Ogata, Hyojeong So, and Young Hoan Cho (eds.) Proceedings of the 19th International Conference on Educational Data Mining, Seoul, Republic of Korea, June, 2026, pp. 896–900. International Educational Data Mining Society (2026).

https://doi.org/10.5281/zenodo.21040097

[1]To preserve double-blind review, we do not cite the author's own publications in this submission. Full citations will be included in the camera-ready version.

ders. The assistance dilemma captures the tension between providing information (reducing errors but potentially reducing learning) and withholding it (increasing productive struggle). The KLI framework specifies that optimal instruction depends on the type of knowledge: sense-making knowledge benefits from constructive activities such as self-explanation, while routine knowledge benefits from practice [8]. In AI-assisted contexts, this manifests as over-reliance (answer outsourcing) versus under-reliance (dismissing valuable feedback) [14]. Interface design and learner characteristics both moderate these patterns [13, 11].

Together, these predict that effective LLM-supported self-explanation requires interfaces that promote generative processing before feedback, provide quality-ensuring feedback without bypassing reasoning, and calibrate support to the type of knowledge and learner characteristics [8].

3. RESEARCH PROGRESS

My research has progressed from understanding how humans interact with AI in educational contexts (Study 1) to examining how LLM interface design shapes learning outcomes in classrooms (Study 2) to directly testing LLM-supported self-explanation as a mechanism for transfer (Study 3). Each study builds on insights from the previous one, progressively narrowing from broad human-AI interaction patterns to specific design principles for self-explanation feedback.

3.1 Study 1: Expertise and AI Reliance in Educational Decision-Making (Published)

Motivation. Before designing LLM tools for learners, I wanted to understand how people use AI recommendations in educational contexts and how expertise moderates reliance. If we cannot predict who benefits from AI assistance, we cannot design effectively for diverse learners.

Method. A within-subjects study ($N = 95$) examined how experienced tutors (experts) and non-tutors (novices) evaluated tutor-praise responses under three AI-assisted conditions: no AI, AI with textual reasoning explanations, and AI with inline highlighting. The task involved judging the pedagogical appropriateness of tutor praise for student work, requiring subject-matter expertise. I decomposed errors into over-reliance (accepting incorrect AI recommendations) and under-reliance (rejecting correct AI advice), treating these as distinct behavioral indicators rather than collapsing them into a single accuracy metric.

Results. AI assistance significantly improved accuracy over unassisted performance ($p = .009$), but human-AI collaboration consistently underperformed the AI-only baseline (88% accuracy), quantifying the "cost of human oversight." A paradox of expertise emerged: novices outperformed experts when AI was present ($p = .021$) because they followed AI suggestions more closely, while experts frequently overrode correct advice. Textual reasoning reduced under-reliance when AI was correct ($p < .05$) but increased over-reliance when AI was wrong, revealing a fundamental trade-off in explanation design. A repeated-measures ANOVA on time cost showed significant main effects for both highlighting ($p < .001$) and reasoning ($p < .001$), with a significant interaction ($p = .001$): users substituted between explanation types rather than simply adding reading time.

Implications. Expertise moderates AI support effectiveness, and explanation design involves fundamental trade-offs between over-reliance and under-reliance. Through the lens of the KLI framework [7], the expertise paradox can be understood as a knowledge-type mismatch: experts' deep domain knowledge leads them to override AI on the basis of sense-making processes, even when the AI's pattern-matching is more accurate for this specific task. These insights motivated me to examine how interface *constraints*, not just explanations, shape the effectiveness of LLM support for learning.

This study demonstrates a key methodology for learning analytics: decomposing aggregate accuracy into process-level indicators (over-reliance, under-reliance, time cost) that reveal the mechanisms behind human-AI interaction effects. This contrasts with prior work that treats AI assistance as a binary treatment and overlooks the quality of human-AI coordination.

3.2 Study 2: Conversational vs. Structured LLM Support in a Course (Under Review)

Motivation. Study 1 showed that interface design affects AI reliance in a controlled task. Study 2 tests whether these dynamics generalize to authentic classroom settings with real learning goals.

Method. We deployed an LLM-powered tool in an undergraduate discrete mathematics course ($N = 148$) using a staggered-access design. The tool integrated two LLM-powered components: a chatbot for conversational math support and a proof-review tool that provides structured, localized feedback on students' written proof attempts. The proof-review tool requires students to first produce their own proof before receiving feedback, while the chatbot allows open-ended conversation. We analyzed adoption patterns, outcome associations, and mediation pathways using log data combined with course performance records.

Results. Earlier access improved homework scores ($B = 2.71$, $p = .026$) but not exam scores. After controlling for self-efficacy and prior performance, chatbot usage negatively predicted exam performance ($B = -0.030$, $p = .014$), while proof-review showed no such association ($p = .480$). To understand this divergence, we coded chatbot transcripts for answer-seeking behaviors (directly requesting solutions rather than conceptual explanations). Students classified as answer-seekers scored significantly lower ($d = 0.55$), and a serial mediation model revealed that chatbot usage partially mediated the self-efficacy $\rightarrow$ exam pathway.

Implications. The proof-review tool, which requires students to generate their own work first (analogous to self-explanation), showed no negative association, while the open-ended chatbot did. This motivated Study 3: if anchoring LLM feedback in student-generated reasoning is key, then explicitly scaffolding self-explanation with LLM feedback should enhance learning. However, this study is observational; usage-outcome associations may reflect selection effects (struggling students may use the chatbot more). A

controlled experiment was needed to test the causal mechanism.

This study exemplifies learning analytics methodology: mining interaction logs to discover behavioral patterns (answer-seeking) and testing predictive models (regression, mediation) for learning outcomes. The answer-seeking classifier represents a potentially deployable tool for real-time intervention in AI-assisted learning environments.

3.3 Study 3: LLM-Supported Self-Explanation in Calculus (Under Review, Central to Thesis)

Motivation. Studies 1 and 2 converged on the hypothesis that LLM interfaces promoting generative processing should yield better learning outcomes. Study 3 directly tests this through a controlled experiment.

Method. A between-subjects experiment ($N = 92$) compared three conditions in a calculus learning environment: no self-explanation (control, $n = 29$), menu-based self-explanation ($n = 35$), and open-ended self-explanation with LLM feedback ($n = 28$). All participants had a fixed 60-minute practice session. Transfer was measured via "Not Enough Information" (NEI) items requiring metacognitive judgment about whether a problem provides sufficient information. Participants were recruited via Prolific and screened for algebra proficiency without advanced calculus background.

The automated evaluation pipeline scored student explanations against a four-level rubric (0, 0.3, 0.7, 1.0) capturing completeness and conceptual accuracy. Two human graders achieved $\kappa = 0.70$ (quadratic-weighted); agreement between the LLM and human consensus was $\kappa = 0.78$ (substantial). The system provided color-coded quality indicators (red/yellow/green) after each attempt, with corrective explanations after two consecutive low-quality responses.

Results. The open-ended condition significantly outperformed control on NEI transfer questions ($\beta = +11.9$ percentage points, $p = .030$, $d = 0.44$; adjusted $R^2 = .204$), while completing 3.5× fewer practice problems (16.9 vs. 58.9) with comparable overall learning gains (+9.3% vs. +9.4%). Menu-based self-explanation did not differ from control ($p = .343$). This directly addresses the concern that self-explanation benefits disappear when learning time is equated [10].

Notably, open-ended participants completed 3.5× fewer problems yet achieved comparable learning gains, suggesting that the time spent explaining was at least as productive as additional practice. They also spent more time on the post-test (20.6 vs. 14.9 minutes, $p \approx .09$), writing more detailed explanations, a behavioral indicator that the intervention shifted their problem-solving approach toward more generative engagement. For NEI multiple-choice items (identifying insufficient information without requiring explanation), the open-ended condition showed a non-significant advantage ($\beta = +9.3\%$, $p = .183$, $d = 0.26$), suggesting that the self-explanation effect is strongest when assessed through explanation quality rather than answer accuracy alone.

Open questions. The sample was not drawn from a classroom. The transfer effect was specific to NEI items; whether it extends to other transfer forms is unknown. The feedback design was fixed; which components are essential has not been isolated.

This study makes a computational contribution to educational data mining: an LLM-based automated assessment pipeline achieving substantial inter-rater reliability with human graders, enabling scalable feedback provision and data collection. This pipeline could be applied to other domains requiring assessment of open-ended student reasoning.

4. PROPOSED THESIS DIRECTION

Study 3's finding that LLM-supported open-ended self-explanation enhances transfer forms the central contribution of my thesis. The finding that open-ended self-explanation with LLM feedback enhances transfer while requiring dramatically fewer practice problems opens several questions: which feedback components are responsible, whether fixed-format feedback can be improved through personalization, and whether these effects hold in authentic classrooms. My planned work addresses these through three directions:

Feedback mechanism decomposition. Study 3 bundles rubric-based scoring, color-coded indicators, and corrective explanations into a single treatment. Which components drive the transfer effect? I plan factorial experiments that systematically vary feedback type (score only vs. score + hint vs. full explanation) and feedback timing (immediate vs. delayed) to isolate causal effects. The automated evaluation pipeline from Study 3 provides the infrastructure to maintain consistency across conditions while scaling to larger samples. This addresses a gap in the self-explanation literature, where feedback designs are typically studied as monolithic treatments rather than decomposed into their constituent mechanisms.

Adaptive feedback intensity. A recurring pattern across my studies is that many comparisons yielded null effects: explanation styles did not improve accuracy in Study 1, proof-review showed no positive association in Study 2, and menu-based self-explanation did not differ from control in Study 3. One interpretation is that fixed-format support is too rigid: the same feedback may over-assist strong students (reducing productive struggle) while under-assisting weak ones (failing to correct misconceptions). I plan to investigate feedback that adapts to explanation quality in real time: minimal acknowledgment for strong explanations (preserving autonomy), targeted hints for partial ones (scaffolding without giving away the answer), and corrective feedback only after repeated failures (preventing frustration). This directly operationalizes the assistance dilemma [6] through data-driven student models, and the KLI framework [7] provides principled criteria for when sense-making support should be intensified versus withdrawn.

Classroom deployment. Study 3 used Prolific participants in a single session. Validating these effects in an authentic course would strengthen ecological validity and test whether benefits persist over longer periods, including across exam performance. The staggered-access methodology from Study 2 provides a template for quasi-experimental deployment. Longitudinal interaction data would also enable trajectory analysis of explanation quality over time, connecting to the EDM

community's work on early warning systems and detector-based interventions. Building on Study 2's answer-seeking coding, I plan to develop automated classifiers from interaction logs that could trigger just-in-time interventions when students shift toward unproductive usage patterns.

4.1 Research Timeline

My timeline prioritizes the highest-impact extensions first. In the near term (next 6 months), I will conduct the feedback mechanism decomposition experiment, as this directly builds on the existing Study 3 infrastructure and addresses the most pressing open question. Concurrently, I will begin developing the answer-seeking classifier using coded transcripts from Study 2 as training data. In the medium term (6–12 months), I will implement the adaptive feedback system, informed by both the decomposition results and trajectory analysis from Study 3 data. The classroom deployment will follow (12–18 months), integrating the adaptive system into an authentic course using the staggered-access design from Study 2.

5. EXPECTED CONTRIBUTIONS

- **Empirical evidence** that LLM-supported open-ended self-explanation enhances transfer ($p = .030$) with 3.5× fewer problems, directly addressing concerns that self-explanation benefits disappear when learning time is equated with additional practice [10].
- **Automated assessment and behavioral detection** pipelines: an LLM-based explanation evaluator ($\kappa = 0.78$ vs. human consensus) enabling scalable feedback, and an answer-seeking classifier from interaction logs enabling just-in-time interventions. These address key bottlenecks in both self-explanation research and EDM early warning systems.
- **Error decomposition methodology** for human-AI interaction, decomposing outcomes into over-reliance, under-reliance, and time cost as process-level indicators, grounded in the KLI framework's distinction between knowledge types [7]. This provides diagnostic information beyond aggregate accuracy for designers of AI-enhanced educational tools.
- **Design guidelines for LLM integration in education**, synthesizing across all three studies: (1) anchor AI feedback in student-generated work; (2) design for generative processing before assistance; (3) monitor and intervene on unproductive usage patterns; (4) calibrate support intensity to knowledge type, learner characteristics, and explanation quality, operationalizing the assistance dilemma [6] through data-driven student models.

6. FEEDBACK SOUGHT

My dissertation is at the transition between completed empirical work and planned extensions. I would particularly value feedback on:

1. **Thesis scope:** Should I focus narrowly on self-explanation feedback design (decomposing which feedback components drive transfer), or maintain the broader framing that connects interface design, reliance patterns, and self-explanation across all three studies?
2. **Transfer measurement:** My current transfer measure is specific (recognizing insufficient information). What additional measures of deep understanding would strengthen the claim that LLM-supported self-explanation promotes genuine conceptual transfer?
3. **Addressing null effects:** Many comparisons across my studies were non-significant. Is adaptive feedback intensity the most promising direction, or are there other design levers more likely to produce robust learning gains?
4. **Theoretical framing:** Is the ICAP-based argument (that LLM feedback elevates self-explanation from constructive to interactive) well-supported, or are there alternative explanations for the transfer effect that the EDM community would find more compelling?
5. **Ecological validity:** What study designs would most convincingly demonstrate that these effects matter in real classrooms, beyond single-session Prolific experiments?

7. REFERENCES

[1] V. Aleven and K. R. Koedinger. An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science*, 26(2):147–179, 2002.

[2] V. Aleven, K. R. Koedinger, and O. Popescu. The need for tutorial dialog to support self-explanation. 2000.

[3] M. T. Chi, M. Bassok, M. W. Lewis, P. Reimann, and R. Glaser. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13(2):145–182, 1989.

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

[5] M. T. Chi and R. Wylie. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4):219–243, 2014.

[6] K. R. Koedinger and V. Aleven. Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review*, 19:239–264, 2007.

[7] K. R. Koedinger, A. T. Corbett, and C. Perfetti. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, 36(4):757–798, 2012.

[8] K. R. Koedinger, J. C. Stamper, E. A. McLaughlin, and T. Nixon. Testing theories of transfer using error rate learning curves. *Topics in Cognitive Science*, 8(3):589–609, 2016.

[9] X. Li et al. Bringing generative AI to adaptive learning in education. In *Proceedings of AIED 2024*, 2024.

[10] K. L. McEldoon, K. L. Durkin, and B. Rittle-Johnson. Is self-explanation worth the time? a comparison to additional practice. *British Journal of Educational Psychology*, 83(4):615–632, 2013.

[11] K. Morrison et al. The impact of AI explanations on human decision-making. In *Proceedings of CHI 2023*, 2023.

[12] B. Rittle-Johnson, A. M. Loehr, and K. Durkin. Promoting self-explanation to improve mathematics

learning: A meta-analysis and instructional design principles. *ZDM Mathematics Education*, 49:599–611, 2017.

[13] H. Vasconcelos et al. Explanations can reduce overreliance on AI systems during decision-making. In *Proceedings of the ACM on Human-Computer Interaction*, 2023.

[14] O. Vereschak, G. Bailly, and B. Caramiaux. How to evaluate trust in AI-assisted decision making? a survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–39, 2021.

[15] C. Zhai et al. The effects of over-reliance on AI dialogue systems on students' cognitive abilities. *Smart Learning Environments*, 11:28, 2024.