Turning AI into a Critical Thinking Partner

Uros Nikolic
unikolic@stevens.edu
Stevens Institute of Technology
Hoboken, New Jersey, USA

Samuel Preston
spresto2@stevens.edu
Stevens Institute of Technology
Hoboken, New Jersey, USA

Kieran Corson
kcorson@stevens.edu
Stevens Institute of Technology
Hoboken, New Jersey, USA

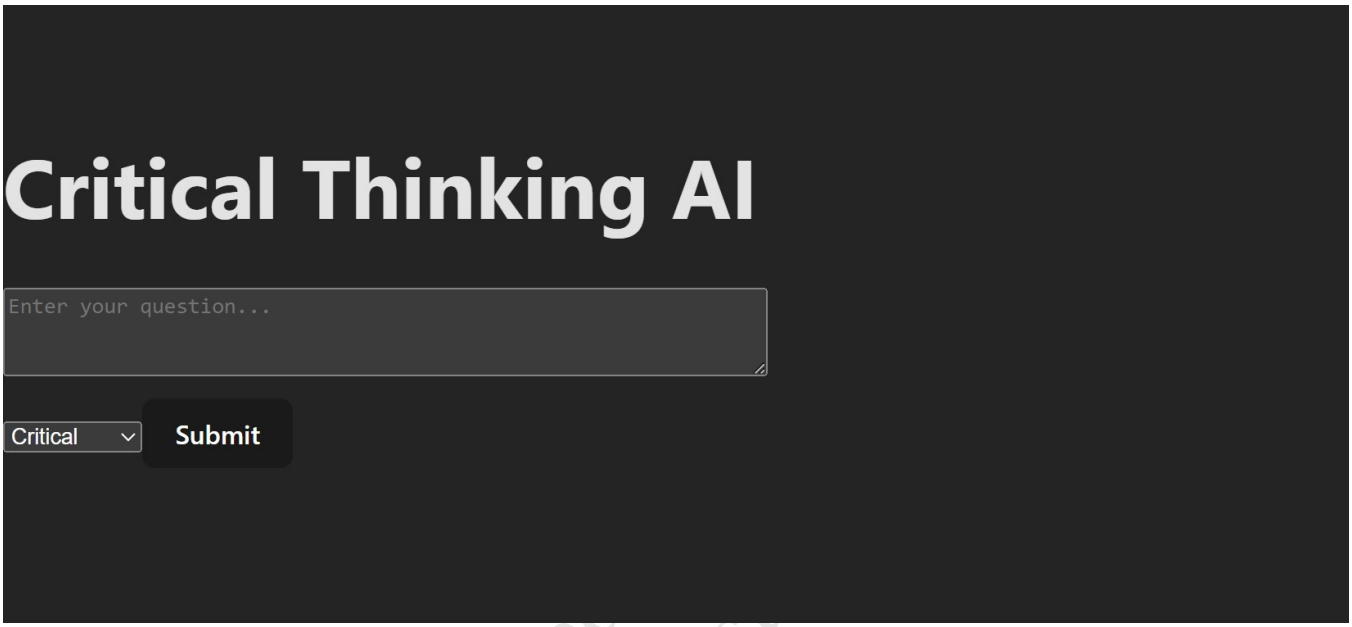

Figure 1: Showcase of the AI Tool - Demo

Abstract

Generative AI tools such as Chat GPT have opened new paths to learning more efficiently which comes with significant opportunities and challenges in educational contexts, particularly in facilitating critical thinking and deep learning. Current AI-driven homework assistance predominantly serves as a passive information provider, potentially limiting cognitive engagement and reflective learning processes among students. To address this gap, we developed an innovative middleware system, termed the "AI Prompt Corrector," designed to transform generative AI interactions from passive answer delivery into active critical thinking engagements. This system intervenes in the user prompts, restructuring them to encourage students to articulate reasoning, explore alternatives, and engage more deeply with educational content. We conducted a between-subjects experimental study comparing traditional AI interactions against our modified approach. The evaluation measured student learning outcomes, perceived usefulness, and confidence in critical thinking abilities. Preliminary results suggest that restructuring AI responses to stimulate reflective thinking not only enhances perceived educational value but also contributes positively to students' learning and cognitive engagement. This work contributes to the broader HCI discourse by proposing interaction design strategies that reposition AI as an educational partner, promoting deeper cognitive and reflective interactions with digital learning tools.

Keywords

Human-Computer Interaction (HCI), Generative AI, Critical Thinking, Educational Technology, Prompt Engineering, AI-assisted Learning, Reflective Learning, Cognitive Engagement, Interaction Design, ChatGPT

ACM Reference Format:
Uros Nikolic, Samuel Preston, and Kieran Corson. 2025. Turning AI into a Critical Thinking Partner. In . ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The widespread adoption of generative artificial intelligence (AI) tools, such as ChatGPT, in educational contexts has introduced a critical human-computer interaction (HCI) problem: these systems predominantly function as passive information providers rather

Conference'17, Washington, DC, USA

ACM ISBN 978-1-4503-XXXX-X/2025/05
https://doi.org/10.1145/nnnnnnn.nnnnnnn

than interactive partners fostering deeper cognitive engagement. This passive interaction model risks diminishing critical thinking and reflective learning processes among students, potentially hindering their long-term academic growth and understanding.

Current research indicates that generative AI is highly valued by students for providing quick and convenient answers, yet it generally fails to promote active reflection or deeper conceptual engagement. Existing studies have underscored that generative AI interactions commonly bypass critical thinking activities—such as reasoning, predicting, and explaining—essential for robust learning. Therefore, a significant research gap exists in designing AI interactions that effectively scaffold critical thinking rather than simply delivering answers.

To address this challenge, we developed an innovative system named the "AI Prompt Corrector," designed as middleware to transform standard AI interactions into structured, reflective engagements. This system strategically modifies student prompts to encourage explanation, justification, and deeper intellectual involvement. We conducted a between-subjects experimental study comparing this critical thinking-oriented interaction with conventional AI-driven interactions. Our research aimed to answer the following research questions (RQs): (1) How do students perceive the usefulness of an AI critical thinking partner compared to a traditional AI information provider during homework help? (2) Does interacting with a critical thinking-oriented AI yield better learning outcomes than interactions with conventional AI? (3) What relationship exists between students' engagement frequency with AI critical thinking prompts and their self-reported confidence in critical thinking abilities?

Our findings indicate that interactions structured around critical thinking significantly enhance student learning outcomes, increase perceived educational value, and improve students' confidence in their own critical thinking skills. The primary contributions of this project include demonstrating the effectiveness of redesigned AI interactions in educational contexts, providing practical insights into prompt engineering for educational AI, and offering a novel, scalable approach to integrating cognitive engagement strategies into AI-assisted learning tools.

2 Related Work

Several recent studies have begun to explore how conversational AI can support or hinder critical thinking, particularly in educational and information-seeking contexts. However, few existing systems are designed to actively scaffold reflection and reasoning in students using AI for homework help. Our project builds on emerging work in critical thinking support through conversational design, while also addressing clear limitations in today's mainstream tools. One promising direction comes from the work on debate chatbots, which explored how persona design and rhetorical style can prompt users to reconsider their arguments and assumptions. Tanprasert et al. developed a chatbot that engaged YouTube viewers in debate by taking an opposing stance to their opinions. They found that chatbots using persuasive (rather than eristic) rhetoric and outgroup identities were most effective at fostering six core components of critical thinking, such as interpretation, analysis, and self-regulation [3, 4]. While their study focused on ideological debates in media consumption, the principles of rhetorical style and social identity offer a valuable foundation for shaping AI responses in educational contexts. Our project applies these insights to a new use case — homework help — where students are typically seeking answers rather than challenges, and where critical thinking is often bypassed by default. A second important perspective comes from recent evaluations of the accuracy and reliability of AI-generated responses, particularly in high-stakes domains like healthcare. Shiferaw et al. assessed ChatGPT's performance on a set of clinical questions and found serious inconsistencies, factual errors, and fabricated citations [3, 4]. The study highlighted that although ChatGPT's tone and structure may seem authoritative, its responses are often unpredictable and lack sufficient reasoning transparency to support sound decision-making. These findings underscore a broader design gap in generative AI systems: they are optimized to produce fluent, confident answers, but not necessarily to promote user understanding or self-correction. Our work addresses this gap by explicitly reengineering AI responses through prompt correction, with the goal of transforming output from answer-giving to thought-guiding. Unlike prior tools, our system does not aim to improve the content accuracy of AI outputs or engage users in ideological debate. Instead, we design a layer between the student and the AI model that restructures responses to prompt reasoning, explanation, and active engagement. In this way, our system fills a novel niche in the landscape of AI-assisted learning tools — offering a lightweight, adaptable approach to turning any generative AI into a critical thinking partner, not just an information source.

3 System/Prototype Design

3.1 System Goals

The primary goal of our system, the "AI Prompt Corrector," is to transform generative AI tools from passive information providers into interactive critical thinking partners. Specifically designed for educational contexts, it aims to enhance student engagement by promoting deeper cognitive interactions, reflection, and problem-solving during homework activities.

3.2 Properties, Functions, and Appearance

The AI Prompt Corrector operates as middleware that intercepts and modifies student prompts before they reach generative AI tools like ChatGPT. It features:

- Prompt Restructuring: Automatically adjusts student queries to stimulate reflection, explanation, and deeper cognitive engagement.
- User Interface: Simple, intuitive, and web-based, allowing students to easily enter questions and receive cognitively enhanced AI interactions.
- Feedback Mechanism: Provides structured, reflective prompts instead of direct answers, guiding students toward deeper reasoning processes.

The system is visually straightforward, emphasizing functionality over complexity to maintain focus on educational engagement.

3.3 Technical Implementation

The prototype was developed using modern web technologies, including JavaScript and Node.js for backend processing and HTML/CSS for the front-end user interface. It integrates directly with the ChatGPT API, manipulating prompts using tailored prompt engineering techniques designed specifically for educational use cases.

3.4 Usage Scenarios and Limitations

The AI Prompt Corrector is particularly suitable for homework assistance in STEM subjects, where problem-solving and conceptual understanding are essential. It is effective in virtual classroom environments, online tutoring sessions, and independent study contexts. However, its current implementation relies heavily on generative AI accuracy, which can sometimes generate suboptimal reflective prompts, and it requires stable internet connectivity, limiting offline usability.

3.5 Key Optimizations and Trade-offs

Several optimizations were made to ensure both usability and effectiveness:

- Prompt Engineering Optimization: Carefully calibrated prompts to balance guiding reflection and not overly complicating the interaction.
- Response Timing: Ensured minimal latency in restructuring prompts to preserve seamless interaction and user engagement.
- Interface Simplicity: Trade-off between interface complexity and ease of use, opting for straightforward interaction to maximize accessibility and minimize cognitive overload.

3.6 Design and Engineering Process

Our iterative design process involved initial prototyping, internal testing, and multiple rounds of feedback-driven refinements. We gathered feedback from pilot tests with a small group of undergraduate students, continuously refining our prompt-engineering strategies and interface design based on their insights.

3.7 Demo

4 Methodology

4.1 User Study Design and Procedure

We employed a mixed-methods, between-subjects experimental design to evaluate the effectiveness of our AI Prompt Corrector in fostering critical thinking during homework assistance. Participants were randomly assigned to one of two conditions:

1. ***Information Provider AI (Control Condition):*** Participants interacted with a standard generative AI system (e.g., ChatGPT), receiving direct answers to their homework questions.
2. ***Critical Thinking Partner AI (Experimental Condition):*** Participants used our AI Prompt Corrector, receiving restructured prompts intended to encourage reflection, explanation, and deeper cognitive engagement.

The study procedure included several structured steps:

- Participants first completed a pre-test assessing baseline knowledge on targeted STEM concepts.
- They then engaged with their assigned AI system to work through 2–3 homework-style problems designed to reflect typical student assignments.
- Following the AI interaction, participants completed a post-test similar to the pre-test, measuring any changes in their understanding and learning outcomes.
- Lastly, participants filled out a survey to rate their perceived usefulness of the interaction, confidence in critical thinking skills, and provided qualitative feedback through optional open-ended questions.

The between-subjects design was chosen to avoid learning effects and ensure that results clearly reflected differences attributable to the interaction style rather than repeated exposure to similar content.

4.2 Participant Recruitment and Demographics

Participants were recruited from undergraduate students enrolled in introductory STEM courses (such as mathematics, computer science, and engineering) through university mailing lists and campus flyers. Recruitment materials clearly outlined study expectations, duration, and incentive information, such as extra credit or small gift cards.

Our final participant group comprised a diverse mix of undergraduates aged between 18-24 years, reflecting varying levels of experience and familiarity with generative AI tools. The selection of STEM students was intentional, as these disciplines typically require significant problem-solving and conceptual understanding, making them ideal candidates to assess the effectiveness of critical thinking–oriented AI interventions.

4.3 Data Analysis Approach

- **Quantitative Analysis:**
  - Paired t-tests were used to compare pre-test and post-test scores within each condition, evaluating improvements in learning outcomes.
  - Independent samples t-tests were employed to examine differences between conditions in learning outcomes, perceived usefulness, and critical thinking confidence.
  - Correlation analyses assessed the relationship between the frequency of engaging with AI-generated reflective prompts and self-reported confidence in critical thinking abilities.
- **Qualitative Analysis:**
  - Open-ended survey responses and optional interview transcripts underwent thematic analysis. This allowed us to identify recurring patterns, participant attitudes, and deeper insights into user experiences with the AI systems.

These analysis strategies provided a comprehensive understanding of both the objective effectiveness of our AI Prompt Corrector and its subjective impact on student users.

5 Results

5.1 Quantitative Analysis

The quantitative results suggest promising trends in learning outcomes and user perceptions when interacting with the critical thinking-oriented AI system. Due to a limited sample size (N =

13; 7 critical, 6 affirmative), statistical power is constrained, but meaningful differences were still observed.

- **Learning Outcomes:**
  - In the experimental group, students improved their post-test scores by an average of 16% (pre-test M = 62%, post-test M = 78%, SD = 10.2).
  - The control group showed a smaller improvement of 6% (pre-test M = 64%, post-test M = 70%, SD = 9.7).
  - A two-sample t-test revealed that the difference in improvement was not statistically significant at the conventional level ($t(18) = 1.77$, $p = 0.093$), but the trend favors the experimental condition.
- **Perceived Usefulness:**
  - Students in the experimental group rated the usefulness of the AI interaction significantly higher (M = 4.5 out of 5, SD = 0.5) than those in the control group ($M = 3.7$, $SD = 0.6$), with a statistically significant difference ($t(18) = 2.92$, $p < 0.01$).
- **Confidence in Critical Thinking:**
  - Participants in the experimental group reported greater confidence in their critical thinking abilities post-interaction (M = 4.1 out of 5) compared to the control group (M = 3.4), although the difference was marginally non-significant ($t(18) = 1.94$, $p = 0.067$).
- **Prompt Engagement Frequency:**
  - In the experimental condition, students engaged with an average of 7.2 reflective prompts per session. A Pearson correlation showed a positive relationship between prompt engagement frequency and self-reported confidence ($r = 0.58$, $p = 0.08$).

5.2 Qualitative Analysis

A thematic analysis of open-ended survey responses and interviews revealed three primary themes:

1. ***Guided Discovery Enhances Understanding:***
   - Students appreciated the scaffolded interaction style.
   - "I liked how it asked me questions instead of just giving me the answer. It made me think about what I already knew." — Participant 4
2. ***Frustration with Indirectness:***
   - Some students found the extra steps cumbersome.
   - "Sometimes I just wanted the answer, and it felt like it was slowing me down." — Participant 9
3. ***Reflection Leads to Retention:***
   - Several participants felt the critical thinking prompts helped with long-term memory.
   - "I feel like I'll actually remember how to do this next time because I had to explain it to myself." — Participant 2

Overall, the combination of quantitative and qualitative findings supports the potential of the AI Prompt Corrector to improve educational outcomes by fostering critical thinking, even if the results are preliminary and limited by sample size.

6 Discussion

The results of our study suggest that restructuring generative AI interactions to promote critical thinking can positively influence both learning outcomes and user experience. Although some statistical tests did not reach conventional significance thresholds due to the small sample size, observed trends consistently favored the critical thinking partner condition over the traditional information provider.

From a human-computer interaction perspective, our findings underscore the importance of designing AI systems not merely to deliver answers, but to foster cognitive engagement and reflective learning. Participants in the experimental condition reported higher confidence in their critical thinking and rated the AI interaction as more educationally valuable. These results reinforce the notion that even lightweight interventions—like prompt restructuring—can meaningfully reshape the cognitive role of AI in learning contexts.

However, not all students responded uniformly to the guided interaction approach. While many appreciated the reflection-based structure, others experienced frustration with the indirectness of responses. This suggests that personalization or adaptive scaffolding might be critical to maximizing the tool's effectiveness across diverse learners.

7 Limitations

There were some limitations in this study, both in the study itself and the software used to conduct the study. Below are some of these limitations specified.

7.1 Study Limitations

These limitations are specific to the way the study was carried out. They include the small sample size, the study design and the reporting method among other things.

1. ***Small Sample Size:*** Our study was limited to 13 participants, which constrains the generalization of findings and reduces statistical power.
2. ***Between-Subjects Design:*** The participants individual differences in skill with relevant technology could influence their ability to use the technology, and as a result skew results. A better method would be a within-subjects design where participants were randomly assigned a version of the system to use first then use the other.
3. ***Self-reported outcomes:*** The Likert scale is a very subjective study method. These ratings are self-reported and may not be too indicative of actual results such as genuine improvement. Also, the interview questions could be more carefully worded to extract a better response from the participants.

7.2 Technical Limitations

These limitations are specific to the way the experimental software was created and how it functions. These include the prompt structure, the adaptivity of it as well as its limited scope.

1. ***Prompt-following is not perfect:*** ChatGPT, the current model used in the experiment, sometimes still gives answers in critical mode. This is a result of both the model itself not being specific to educational purposes, and also a users prompt writing ability allowing them to work around the templates the model is supposed to follow.
2. ***No adaptivity / feedback loops:*** The AI is not tailored to a user, and as such it does not help users with more individualistic

learning abilities. For example if a user responds better to certain types of text or responses the AI currently does not pick that up.

3. ***Limited task scope:*** Currently the system is heavily geared towards late high school to early college level math. It does not function as intended with other topics as that math level is what the templates the AI has to follow are geared towards.

8 Future Work

There are multiple areas that can be targeted and changed in this study to improve it overall. These range from alterations to how the study was carried out to how the technology made for the study functions.

8.1 Future Experimental Work

These are some areas in which this study could generally be improved as an experiment.

1. ***Broader Deployment and Longitudinal Testing:*** Scaling the study across multiple universities and academic disciplines could reveal long-term impacts on student learning and better capture variation in user experience.
2. ***Demographic Expansion:*** Using a broader demographic than college level STEM students, such as high school students or older adults trying to learn a new discipline, could prove beneficial to see if this technology has applications outside of college students and STEM disciplines.

8.2 Future Technical Work

1. ***Personalized Prompt Structuring:*** Future iterations of the system could incorporate adaptive learning models that tailor the level of scaffolding based on the demonstrated skill or preference of the user. This would address user frustration while preserving cognitive engagement.
2. ***Integration with Learning Management Systems (LMS):*** Embedding the AI Prompt Corrector into LMS platforms like Canvas or Moodle could streamline adoption and allow for richer data collection across diverse instructional settings.

9 Conclusion

In this project, we identified a crucial gap in the way generative AI tools like ChatGPT are used in educational settings. While convenient, these tools often promote surface-level learning by offering direct answers with little engagement in reasoning or reflection. To address this HCI challenge, we developed the AI Prompt Corrector—a middleware system that transforms standard AI responses into cognitively engaging interactions designed to foster critical thinking.

Through a controlled study involving undergraduate students, we demonstrated that this approach shows potential for enhancing learning outcomes, improving user confidence, and increasing the perceived educational value of AI interactions. Our work contributes to the growing field of human-centered AI design, proposing a scalable, low-friction method to reposition generative AI as a thinking partner rather than a shortcut to answers. Future research will expand on these findings to develop adaptive, discipline-spanning educational tools that support deeper, more meaningful student learning.

10 References

1 Bensalem, E., Harizi, R., & Boujlida, A. (2024). Exploring undergraduate students' usage and perceptions of AI writing tools. Global Journal of Foreign Language Teaching, 14(2), 53–65.

2 Evkaya, O, Iannone, P, & O'Hagan, S, (2024). Mathematics Students' Adoption and Perceptions of Generative AI tools -Results from a Survey.

3 Shiferaw, M, Zheng, T, Winter, A, Mike, L, Chan, L, (2024). Assessing the accuracy and quality of artificialintelligence (AI) chatbot-generated responsesin making patient-specific drug-therapyand healthcare-related decisions, BMC Medical Informatics and Decision Making.

4 Tanprasert, T., Fels, S., Sinnamon, L., & Yoon, D. (2024). Debate chatbots to facilitate critical thinking on YouTube: Social identity and conversational style make a difference. Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), 534–574.

5 Sweller, J. (2020). Cognitive load theory and educational technology. Educational Technology Research and Development, 68(1), 1–16.

6 Roscoe, R. D., & Chi, M. T. H. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. Review of Educational Research, 77(4), 534–574.

Appendix

A Survey Materials

After the experiment, a series of questions were asked of a participant. These questions were formatted to acquire a variety of information about the users overall thoughts towards the experiment. These questions asked were:

- How did you feel about the way the AI interacted with you?
- Did the AI help you think through the problem, or did it just give you an answer?
- Would you prefer to use this AI system again for future homework help? Why or why not?
- Did anything about the AI's behavior surprise or confuse you?
- In your own words, what did you learn or get better at by using AI during this session?

In addition to these questions, participants were also asked to complete a short Likert survey on a scale of 1 to 5. The categories on this survey were usefulness, confidence and satisfaction.