Visible Logic: Measuring the Effects of Source-Attributing AI Transparency on Trust and Utility in Academic Settings

Rochester Institute of
Technology
Rochester, New York

ABSTRACT

The adoption of Large Language Models (LLMs) in higher education presents a critical problem. Their utility is often compromised by unverified information which undermines academic integrity and promotes blind trust. This source opacity prevents students from assessing the credibility of the information given. This proposal details a mixed-methods experimental study on an innovative transparency feature. This tool includes real-time source attribution, a dynamic confidence meter, and a visible reasoning trace. The study hypothesizes this feature will significantly increase students' perceived trustworthiness and utility of the AI output. Additionally, it believes this will lead to enhanced fact-checking and critical evaluation behaviors. The resulting findings establish empirical HCI (Human-Computer Interaction) guidelines for designing responsible AI tools that support critical learning skills, instead of replacing them.

Author Keywords

explainable AI (XAI), Large Language Models (LLMs), transparency, trust, attribution, academic integrity, Human-Computer Interaction (HCI).

INTRODUCTION

The introduction of Artificial Intelligence (AI) tools into higher education has introduced opportunities to increase productivity, but also many significant challenges to academic integrity. Large Language Models (LLMs) function as black-box systems that often hallucinate and generate unsupportive text [10]. This inability to trace information back to its source creates a critical gap in student learning, promoting either an over-reliance on the tool or wholesale rejection which may risk students falling behind [4].

The problem at hand is that the opacity and unverified attribution of LLMs fundamentally undermine their reliability and adoption as effective tools for critical learning and academic integrity in higher education. This research proposes to bridge the gap between technical capability and human perception by designing a transparency feature and embedding it into a known AI tool (e.g. Gemini, ChaptGPT, Claude). This feature would provide granular information regarding the AI's generation process in an intuitive manner. It will let the user know what sources were used to generate the text, and what analysis it went through to generate text that is not directly source-based. The research question addresses how such a feature can influence the perceived trustworthiness and utility of the information given.

This research could contribute novel findings to the field of Explainable AI (XAI) in education, specifically pertaining to transparency and accountability [3]. By focusing on source attribution and confidence scoring, this research provides a methodology to foster informed and reflective use of AI in academic settings [9]. A successful implementation of the transparency feature is expected to help AI tools assist students and actively encourage the development of their metacognitive skills and source evaluation.

RELATED WORK

The proposed study is grounded in three main areas of peer-reviewed research: the necessity of XAI in educational contexts, the technical challenge of LLM attribution, and standard HCI design for communicating factuality.

The Need for XAI in Education and Trust Calibration

The ethical concerns of AI, summarized by FATE (Fairness, Accountability, Transparency, and Ethics), are magnified in educational settings [3]. Khosravi et al. developed the XAI-ED framework, asserting that educational AI requires specialized explanations to support students' metacognition and enhance teacher confidence [3]. Without transparency, teachers show distrust, leading to the underutilization of valuable AI tools [5]. From the student perspective, a survey by Zhou et al. confirmed that while AI offers productivity benefits, the primary concerns are threats to academic integrity and the risk of over-reliance. These findings show the demand for clear usage guidelines and tools that promote critical thinking [4]. Crucially, experimental work by Wang et al. demonstrates that the presence of explanations enables users to make necessary judgments: they can "judge when [they] should trust and not trust the model" [5]. This principle of trust calibration, specifically the ability to appropriately modulate reliance is the main goal of the transparency feature.

The LLM Opacity and Attribution Gap

The challenge of blind trust is driven by two technical flaws: hallucination and opaque attribution. Gao et al. created the ALCE (Automatic LLMs' Citation Evaluation) benchmark, revealing that even state-of-the-art LLMs lacked complete citation support in 50% of cases when generating text with citations [2]. This establishes the severity of the hallucination problem as the core justification for the proposed research. To quantify this problem, Yue et al. formalized the types of attribution errors that an LLM can make into three categories: attributable, extrapolatory (lacks sufficient information), and contradictory (directly contradicts the source) [1]. This existing framework provides the necessary technical definition for the proposed confidence meter, which would quantify the risk of extrapolation or contradiction in real-time.

Technical and HCI Approaches to Transparency

The concept of granular source attribution, for instance, has been shown to be feasible. Phukan et al. demonstrates that fine-grained attribution is possible by leveraging the LLM's hidden state representations to identify and trace text segments copied verbatim from sources [6]. This technical approach directly supports the feasibility of providing real-time, click-to-source linking. Furthermore, the necessity of a reasoning trace is supported by work suggesting that to enhance reliability in complex tasks, LLMs require a built-in "chain of thought" and planning ability which could easily be visible to the user [7]. Similarly, Kommiya Mothilal et al. advocates for shifting HCI focus toward the "reasoning about reasoning" principle which supports making the underlying process steps visible for reflective use [9]. From an interface design perspective, empirical evidence exists for communicating factuality. A user study by Do et al. provides critical HCI guidance for the proposed transparency feature, as it found that participants highly preferred a design where factuality was communicated visually. It was found that specifically using color-coding of phrases within the response based on computed factuality scores was a visual component that users paid attention to and enjoyed [8]. This finding directly informs the visual elements that will be incorporated into the transparency feature.

Research Question and Hypothesis

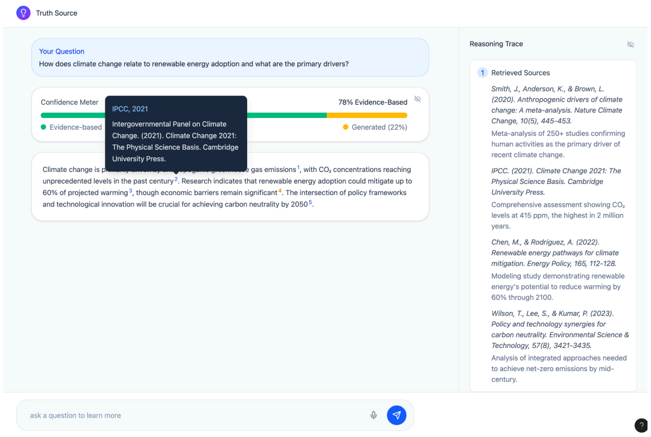
The existing literature establishes the need for transparency, defines the problem of opacity and error, and confirms the technical and design feasibility of the proposed solution. However, a comprehensive experimental study that combines all three key features (granular source attribution, confidence meter, and reasoning trace) in an academic setting and measures its effect on student trust and verification behavior is missing. Therefore, the research is guided by the following research question (RQ): How does an intuitively designed source-attributing transparency feature in an AI tool influence the perceived trustworthiness and utility of information in an academic setting? The primary hypothesis (H1) is that university students using the LLM with the transparency feature will exhibit significantly higher perceived trustworthiness and utility scores. Additionally, these students will also demonstrate a measurable increase in critical verification behavior (e.g., clicking sources and fact-checking time) compared to students using a standard LLM interface. The null hypothesis is that there will not be a significant difference between the perceived trustworthiness, utility, or critical verification behavior between group A and B.

METHODS

This study will employ a mixed-methods experimental design composed of two phases: a formative design phase and a controlled, between-subjects experiment.

Phase 1: Formative Research

The initial phase focuses on defining the trust variables and determining the visual features needed for the experiment. The goal of this formative research phase is to identify baseline user needs, trust-breakers, and the necessary level of granularity for the transparency feature. The methodology will involve administering a survey to approximately N=50 university students who have used generative AI for academic purposes before. The survey will measure baseline levels of trust/mistrust in AI output, perceived usefulness, and students' current fact-checking behaviors. It will also utilize Likert-scale questions derived from XAI trust metrics to gauge the perceived value of specific transparency cues (e.g., source links, confidence scores, reasoning steps). The final outcome will be an empirical list of design requirements. This list will help me prioritize the visual look of the transparency feature and determine color-coding for factuality based on the feedback received.



How does climate change relate to renewable energy adoption and what are the primary drivers?

IPCC, 2021

Intergovernmental Panel on Climate Change. (2021). Climate Change 2021: The Physical Science Basis. Cambridge University Press.

Climate change is p[illegible] gas emissions[1], with CO_2 concentrations reaching unprecedented levels in the past century[2]. Research indicates that renewable energy adoption could mitigate up to 60% of projected warming[3], though economic barriers remain significant[4]. The intersection of policy frameworks and technological innovation will be crucial for achieving carbon neutrality by 2050[5].

1 Retrieved Sources

Smith, J., Anderson, K., & Brown, L. (2020). Anthropogenic drivers of climate change: A meta-analysis. Nature Climate Change, 10(5), 445-453.

Meta-analysis of 250+ studies confirming human activities as the primary driver of recent climate change.

IPCC. (2021). Climate Change 2021: The Physical Science Basis. Cambridge University Press.

Comprehensive assessment showing CO_2 levels at 415 ppm, the highest in 2 million years.

Chen, M., & Rodriguez, A. (2022). Renewable energy pathways for climate mitigation. Energy Policy, 165, 112-128.

Modeling study demonstrating renewable energy's potential to reduce warming by 60% through 2100.

Wilson, T., Lee, S., & Kumar, P. (2023). Policy and technology synergies for carbon neutrality. Environmental Science & Technology, 57(8), 3421-3435.

Analysis of integrated approaches needed to achieve net-zero emissions by mid-century.

Figure 1. Initial prototype of the transparency feature

Phase 2: Controlled Between-Subjects Experiment

The main phase will be a controlled between-subjects experiment designed to test the efficiency of the transparency feature. A convenience sample of N=50 university students will be recruited and randomly assigned to one of two groups of 25 students each. For the intervention, group A (Control) will use a standard LLM interface (text input, text output), while group B (Treatment) will use the same LLM, but the output will be augmented with the transparency feature, including the visual attribution, confidence score, and reasoning trace sidebar. Figure 1 demonstrates the current interface of the transparency feature. The procedure will involve providing all participants with a complex academic task such as synthesizing competing theories on a psychological phenomenon and asking the LLM for help. This task requires external sources and challenges the AI to use complex reasoning [7]. Both groups will use their assigned LLM interface to complete the task over a set time period. Data collection will rely on system logs to capture behavioral data like time spent on task, number of clicks on sources/attribution links in Group B, and total time spent reviewing/editing the final output. After the task, all participants will complete a questionnaire that allows them to rate the trustworthiness, utility, and reliability of this transparency feature.

Data Analysis Plan

For statistical analysis, an independent samples t-test will be used to compare the means of the dependent variables: trust, utility, and verification behavior between group A and group B. Significance will be set at alpha < .05. Trust and perceived utility will be measured via a post-task survey using validated scales for perceived trustworthiness and reliability of the output. Critical verification behavior will be measured by log data that shows the time spent on attribution features and frequency of clicking on source links in group B.

Ethical Considerations

This study will require Institutional Review Board (IRB) approval. Participants will be informed that the study involves generating and evaluating AI output for a simulated academic task. Every participant will receive and sign consent forms that will detail data privacy measures, anonymous collection of log data, and the voluntary nature of participation, ensuring ethical guidelines are followed.

DISCUSSION

Expected Results and Interpretation

The study is designed to yield strong evidence supporting the benefits of human-centered Explainable AI (XAI) design in academic settings. The interpretation of results is tied directly to the success of the transparency feature in promoting the ability of a user to know when to rely on the system and when to fact-check.

If group B shows significantly higher scores in perceived trustworthiness and utility than group A, the study demonstrates that intuitive transparency features increase the subjective reliability and acceptance of AI tools. Crucially, if group B also shows a measurable increase in critical verification behavior, this would support the idea that the transparency feature promotes trust calibration and empowers students to actively evaluate sources.

Conversely, if group B reports high trust but low critical verification, it suggests the feature is seen as a black-box badge of approval rather than a tool for evaluation. This outcome would indicate a need to redesign the feature to be more cognitively demanding to encourage true fact-checking. Finally, if no significant difference is found between groups, it would suggest that students' existing mistrust of AI is too ingrained to be overcome by current XAI features, or that the design failed to intuitively communicate the necessary information. This would call for a deeper qualitative analysis, potentially through follow-up interviews to analyze the specific flaws.

Limitations

The proposed study faces several key limitations inherent in HCI research. Firstly, ecological validity is a big concern. The experiment is time-bound and conducted in a controlled lab setting, which may not perfectly reflect the long-term use and habitual behavior of students in real academic courses. Second, LLM implementation poses a substantial technical challenge. Building and maintaining a custom, source-attributing LLM for the treatment group relies on the feasibility techniques demonstrated by Phukan et al. [6] and the complex integration of Retrieval-Augmented Generation (RAG) models, which can be difficult to fully control. Third, generalizability is limited, as the sample is restricted to university students with prior AI exposure, thereby limiting the direct application of findings to other educational or professional contexts.

Despite these limitations, this research presents a necessary next step in designing AI tools that responsibly coexist with human learning. By quantifying the effect of the transparency feature on trust calibration and critical thinking, this work can provide a roadmap for developers and educators to build a foundation of accountability for the next generation of AI systems.

REFERENCES

1. Yue, X., Wang, B., Zhang, K., Chen, Z., Su, Y., & Sun, H. (2023). Automatic Evaluation of Attribution by Large Language Models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 4615–4635.
2. Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling Large Language Models to Generate Text with Citations. *arXiv preprint arXiv:2305.14627*.
3. Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable Artificial Intelligence in education. *Computers and Education: Artificial Intelligence*, *3*, 100074.
4. Zhou, X., Zhang, J., & Chan, C. (2024). Unveiling students' experiences and perceptions of Artificial Intelligence usage in higher education. *Journal of University Teaching and Learning Practice*, *21*(2).
5. Wang, D., Bian, C., & Chen, G. (2024). Using explainable AI to unravel classroom dialogue analysis: Effects of explanations on teachers' trust, technology acceptance and cognitive load. *British Journal of Educational Technology*, *55*(6), 2541–2559.
6. Phukan, A., Somasundaram, S., Saxena, A., Goswami, K., & Srinivasan, B. V. (2024). Peering into the Mind of Language Models: An Approach for Attribution in Contextual Question Answering. *Findings of the Association for Computational Linguistics: ACL 2024*, 11481–11495.
7. Leung, H., & Wang, Z. (2025). LLM SHOULD THINK AND ACTION AS A HUMAN. *arXiv preprint arXiv:2502.13475*.
8. Do, H. J., Dugan, C., Ostrand, R., Sattigeri, P., & Murugesan, K. (2024). Facilitating Human-LLM Collaboration through Factuality Scores and Source Attributions. *arXiv preprint arXiv:2405.20434*.
9. Kommiya Mothilal, R., Ahmed, S. I., Zhang, S., & Guha, S. (2025). Reasoning About Reasoning: Towards Informed and Reflective Use of LLM Reasoning in HCI. *arXiv preprint arXiv:2510.22978*.
10. Sharma, P., Liu, Y., Xia, H., Oswal, M., & Huang, Y. (2024). PersonaFlow: Designing LLM-Simulated Expert Perspectives for Enhanced Research Ideation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1-15). ACM.