Enhancing Critical Thinking with AI: A Tailored Warning System for RAG Models

Xuyang Zhu*
Stanford University
Stanford, California, USA
xuyang1@stanford.edu

Sejoon Chang*
Stanford University
Stanford, California, USA
sejoon@stanford.edu

Andrew Kuik*
Stanford University
Stanford, California, USA
akuik@stanford.edu

ABSTRACT

Retrieval-Augmented Generation (RAG) systems hold significant potential as educational tools, particularly for enhancing the accuracy and reliability of large language models (LLMs) in academic settings, such as historical textbook question-answering (TQA). However, the risk of hallucinations remains a major challenge, especially in the context of historical textbooks where content may vary based on publication time and location. This paper proposes a novel approach to address this issue by introducing a tailored warning system within RAG-based TQA systems. The system dynamically detects and alerts users to potential hallucinations or biases in generated responses, offering real-time, context-sensitive warnings to improve the trustworthiness and transparency of the system. Through a comparative user study, we assess the effectiveness of this approach in enhancing hallucination detection and user confidence, compared to generic and no-warning conditions. Our results suggest that a tailored warning system improves both the accuracy of identifying inaccuracies and user trust in the information provided. This approach has the potential to significantly contribute to the current RAG system design, particularly in educational settings, by promoting user awareness and reducing the risks associated with AI-generated misinformation.

KEYWORDS

Retrieval-Augmented Generation, Hallucinations, Educational Tools, Bias Detection, Tailored Warning System

ACM Reference Format:
Xuyang Zhu, Sejoon Chang, and Andrew Kuik. 2025. Enhancing Critical Thinking with AI: A Tailored Warning System for RAG Models. In *Proceedings of 2025 CHI ACM Workshop.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.48550/arXiv.2504.16883

1 INTRODUCTION

A retrieval-augmented generation (RAG)-based system holds significant potential as an educational tool to retrieve information from academic textbooks. [1]. Incorporating contextually relevant information through RAG could enhance large language model (LLM) outputs with more accurate responses. However, concerns arise regarding the reliability of RAG-based textbook question-answering (TQA) systems. LLMs are known to sometimes produce plausible-sounding yet incorrect or biased information, also known as hallucinations [5]. In the specific context of retrieval of historical textbooks, issues are compounded by the inherent inaccuracies dependent on the time and location of publication [15]. Consequently, RAG systems can unintentionally reinforce these biases, which could affect the trustworthiness of the information users get [27]. Research has shown that the use of LLMs in education has not only been a performance enhancer but also a habitual tool for learners [20]. With this reliance on RAG and LLM technologies as an educational tool, it remains imperative to address these challenges to ensure they deliver dependable and unbiased content.

Research has been conducted on the effectiveness of warning messages in harm prevention in human interactions with LLM-based tools. Farsight provides a comprehensive overview of potential harms in AI-based tools early in the design and prototyping stages [25]. Warning messages in LLM tool interfaces have also been shown to improve the detection of hallucinations without significantly affecting the perceived truthfulness of genuine content [14]. Methods to automatically evaluate LLM biases and implement debiasing strategies have been proposed [7] [9]; however, users remain unaware of these biases. While existing research establishes a foundation for effectiveness in tool design and basic chat interaction, any attempts to develop and evaluate a tailored warning system within a RAG-based TQA have yet to be explored.

In this paper, we propose a novel approach to enhancing trust and transparency in RAG systems for TQA. Our key insight is to incorporate a tailored warning system that actively identifies and alerts users to potential hallucinations or biases in the generated responses. Unlike general warning mechanisms, our approach dynamically adapts to the specific context of each question and the retrieved content, providing users with targeted alerts based on the nature and reliability of the information presented. By focusing on a warning system that both detects and communicates potential inaccuracies, we aim to create a more interactive and user-aware RAG experience that directly addresses the critical issues of hallucinations and misinformation in educational settings.

Additionally, we conduct a comparative study to assess the effectiveness of our tailored warning system. This study examines three configurations: no warnings, general warnings, and our tailored warning system. Key evaluation metrics include the accuracy of hallucination detection, user trust in system responses, and overall satisfaction with the user experience. We hypothesize that the

*All authors contributed equally to this research.

2025 CHI ACM Workshop, April 26–May 01, 2025, Yokohama, Japan

ACM ISBN AIREASONING-2025-01
https://doi.org/10.48550/arXiv.2504.16883

tailored warning system will improve accuracy in identifying inaccuracies and enhance user confidence in the information provided.

Our key contributions are as follows: We propose and evaluate a user-and-context-tailored warning approach, where the warning messages adapt depending on the type of interaction. We investigate how tailored warning messages impact users' ability to detect hallucinations, their decision-making processes, and their final accuracy in a TQA task, compared to the baselines of no warnings and standard warning messages.

2 RELATED WORK

Technical Improvements in Retrieval Accuracy. Since the foundational RAG framework was published [21], significant advances have been made to enhance retrieval accuracy and address challenges related to misinformation. Initial RAG models occasionally propagated factually incorrect information due to inaccurate documents, which limited their reliability [3]. New approaches have introduced mechanisms to improve factual accuracy and content relevance. QA-RAG leverages a question-answer format to restructure retrieved content, cross-referencing them with internal knowledge to enhance response quality [11]. RAGAR addresses political fact-checking through a "Chain of RAG" approach, which iteratively verifies evidence to reduce the model's acceptance of false information [8]. Domain-specific RAG adaptations, such as RAG-end2end, further improve accuracy by jointly training the retriever and generator with domain-relevant data, such as in COVID-19 research [18]. FIT-RAG focuses on token usage efficiency for black-box RAGs, improving accuracy and resilience against misinformation in open-domain QA tasks [12]. Others such as MisLC tackles societal misinformation by incorporating legal considerations, suggesting that RAG models could benefit from integrating legal knowledge to enhance factual accuracy, especially in sensitive areas like privacy and hate speech [10]. While collectively, these innovations mark a notable progression in RAG capabilities, challenges in dynamic error correction and self-verification persist, and models can still be "easily misled by documents containing incorrect facts"[3].

RAG Explainability. Explaining the output of Retrieval Augmented Generation (RAG) models poses challenges closely tied to the general issue of LLM explainability, particularly regarding contradictions in internal contexts and hallucinations. Several works have introduced techniques to provide explainable LLM output by altering input components that influence RAG outputs [21]. Additionally, evaluation studies [26] show that RAG models often adopt incorrect retrieved content, overriding their own correct prior knowledge in over 60 percent of cases when faced with contradictions between internal priors and externally retrieved evidence. Recent efforts to reduce hallucinations have made progress, employing methods such as iterative feedback loops [16] and cognitive mechanism self-refinement [28]. While techniques exist to detect errors stemming from context conflicts and hallucinations, most RAG frameworks remain highly automated, offering limited transparency to users. Consequently, there are few opportunities for users to examine and intervene in the processes that lead to hallucinations or conflicts. Introducing a user interface that visualizes potential pitfalls at each step of the RAG framework, and allows users to make informed adjustments, could significantly enhance the system's explainability to common users.

Multi-modal RAG frameworks. Recent studies have proposed various multi-modal RAG frameworks to retrieve usable context from complex input data. The incorporation of visual information in question answering began with VQA and OK-VQA [13], where the frameworks required clearly associated image and text inputs. With advancements in natural language supervision [17], in-context learning, and cognitive structure modeling, more comprehensive multi-modal frameworks have been proposed [4] [22]. However, these frameworks generally handle diverse types of question and answering but do not perform the specific tasks required for RAG models in particular user scenarios. Design adjustments, such as incorporating an abbreviation dictionary [19] or a knowledge graph [30], have been added to the RAG pipeline to better integrate domain-specific knowledge, such as chip design terminologies and clinical medical definitions. Despite these advancements, current multi-modal frameworks still struggle to retrieve niche, domain-specific contexts, particularly relevant to human experience in specialized fields like chip design or medicine. One way to improve the fidelity of retrieved information is by allowing users to interact with and adjust the retrieved context. For instance, users could be given the option to view, modify, or confirm the context associated with their query to ensure its relevance and accuracy. This shift moves beyond the passive retrieval of data to more dynamic, user-driven context refinement.

Human-centric accuracy in RAG systems. Improving the accuracy of RAG often involves retrieval algorithms and optimizing query-document matching. Techniques such as dense passage retrieval [6] and multi-hop retrieval [23], which use learned embeddings instead of sparse vector space models, have enhanced the accuracy of open-domain question answering systems. Even with these advancements, however, RAG systems seem to struggle with expert questions and domain-specific scenarios where context precision is necessary, particularly in specialized applications such as in-house systems [24].These challenges underscore the complexity of information retrieval in an era of expanding digital knowledge repositories. The nuanced nature of expert-level queries demands sophisticated approaches that navigate intricate contextual landscapes beyond traditional mechanisms. Furthermore, the evolving landscape of AI-driven systems necessitates a more adaptive approach to contextual understanding and retrieval. This requires not just algorithmic improvements, but a rethinking of how systems interpret complex user queries. In parallel, studies on interactive systems have highlighted the possibility of using user feedback to dynamically modify received content, enabling for real-time correction and refinement [2]. Active learning frameworks and user-driven feedback loops have demonstrated potential in personalized retrieval, allowing users to modify results accordingly [29].

In previous work, accuracy improvements through model refinement and user-driven feedback have remained separate in practice. While baseline studies have found that warning messages improves user detection of hallucinations, we aim to extend the study by verifying whether a tailored warning message system would result in better hallucination detection and higher accuracy in TQA tasks.

3 METHODOLOGY

To evaluate the effectiveness of our tailored warning system in a Retrieval-Augmented Generation (RAG) context, we designed a user study combining both quantitative and qualitative methods. Our study involved the development and deployment of a web-based question-answering interface, VERA, integrated with three different warning configurations. The primary goal was to assess how each condition influenced users' ability to detect hallucinations in AI-generated responses.

A total of 18 participants were recruited from the Stanford undergraduate student population. The study was conducted in a controlled setting, and all participants were at least 18 years old and provided informed consent. Further details on data generation, participant assignment, and the specific experimental setup are described in the following subsections.

3.1 Data Generation

We designed a questionnaire based on the AMSCO AP US History Textbook. A random excerpt was selected as the basis for our questionnaire. To avoid any impact from the LLM's memory of previously inputted data or influence on our question design, we generated the questions independently. We used OpenAI's GPT-4 model to generate responses, as this is the most recent version available.

To generate hallucinated responses, we first prompted GPT-4 to produce the correct response based on the textbook excerpt stored in the database [14]. We manually cross-checked this with the textbook and other references to ensure accuracy. We then crafted prompts to produce two levels of hallucination: low and high. The low level of hallucination includes minor factual inaccuracies that are difficult to detect at first glance. In most cases, participants would need to double-check with original sources or cross-reference other materials to identify these errors. High-level hallucinations contain clear inaccuracies that should be immediately noticeable, assuming participants have basic common sense and foundational knowledge. We verified that both levels of hallucinated responses did not include correct information.

3.2 Experiment Design

We conducted a 3x3 mixed design experiment, with three levels of warning message (no warning (control), standard warning (control), tailored warning (treatment)) and three levels of response type (correct output, low-level hallucination, high-level hallucination). Our objective was to investigate the effect of a tailored warning message on participants' ability to detect hallucinations and their decision-making process when answering the questionnaire.

Participant Recruitment. All 18 participants were recruited from the Stanford undergraduate population. The study was restricted to participants who were at least 18 years old and located in the U.S., with an age range between 20 and 23. All participants provided informed consent to participate voluntarily, after we explained the potential impacts of the study.

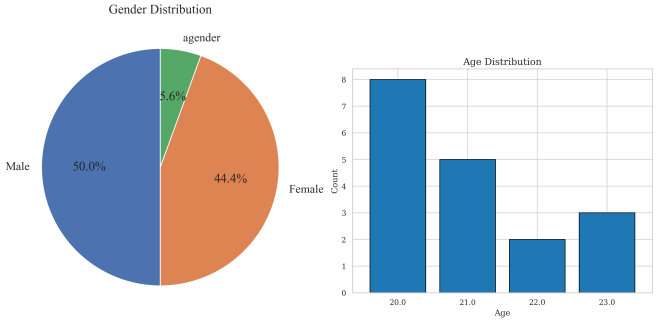
Materials. We divided eighteen questions into three groups of six. For each group, we prepared genuine, low-level hallucination, and high-level hallucination responses. This process resulted in a questionnaire of 18 multiple-choice questions, each accompanied by either a genuine, low-level hallucinated, or high-level hallucinated response. All participants in the two control groups and the treatment group received the same set of questions and responses; only the warning messages differed: no warning, standard warning, or tailored warning.

Procedure. Participants were randomly assigned to one of three groups (control with no warning, control with standard warning, and treatment with tailored warning). Each group saw an interface corresponding to its assigned warning condition. Participants were then tasked with completing an 18-question history questionnaire on a topic from the AP US History excerpt, using assistance from the RAG interface. For each question, participants entered the question into the interface and read through the response, which fell into one of three categories: no hallucination, low-level hallucination, or high-level hallucination. They then rated the accuracy of each response on a scale from 1 to 5 (1 being completely inaccurate, 5 being completely accurate). After evaluating the response, they moved on to the next question, repeating this process until they completed all questions.

After completing the task, we asked follow-up questions about their overall experience. Specifically, we asked if they knew the answer to certain questions beforehand to qualitatively assess how prior knowledge impacted their perception of the RAG interface's accuracy. As participants were randomly assigned to each group, we assumed that prior history knowledge levels were similar across all groups.

4 RESULTS

The analysis of the study data reveals useful insights into the performance of the three experimental groups in detecting hallucinations in RAG-based TQA systems and their overall user experience.



(a) Gender Distribution **(b) Age Distribution**

Figure 1: Survey Demographics

Figure 1 depict the ages and gender of the interviewees. The gender distribution indicates a nearly even split, and the age distribution reveals a focus of individuals around 20-23 years old. As such, it

would be reasonable to assume the following results are not heavily gender-biased, but they may certainly be different if done again on a different age group.

Accuracy of Answers. The heatmap in Figure 2 summarizing the accuracy of participants' answers under the three warning conditions shows notable differences in performance. Participants exposed to tailored warnings demonstrated the highest accuracy across all hallucination levels, achieving 100% accuracy in responses with no hallucinations, 89% in low-level hallucination cases, and 81% in high-level hallucination cases. In contrast, participants in the standard warning group achieved 97%, 81%, and 69% accuracy for no, low, and high hallucinations, respectively. The no warning group performed the poorest, with accuracies of 100%, 64%, and 56% under the same conditions. These results demonstrates that tailored warnings substantially enhance participants' ability to detect hallucinations, particularly in challenging scenarios involving high-level hallucinations.

Statistical Significance of Results. We conducted a one-way Analysis of Variance (ANOVA) to evaluate whether there were statistically significant differences in accuracy ratings across the three warning conditions: no warning, standard warning, and tailored warning. The ANOVA yielded a statistically significant result ($p = 0.0062$), indicating that at least one group differed from the others in terms of hallucination detection accuracy. To further investigate where these differences occurred, we performed a Tukey HSD (Honestly Significant Difference) post-hoc test. The results showed that the tailored warning group significantly outperformed both the no-warning group ($p = 0.005$) and the standard warning group ($p = 0.032$). However, the difference between the standard warning and no-warning groups was not statistically significant ($p = 0.22$). These findings suggest that the tailored warning interface was most effective in improving users' ability to detect hallucinations.

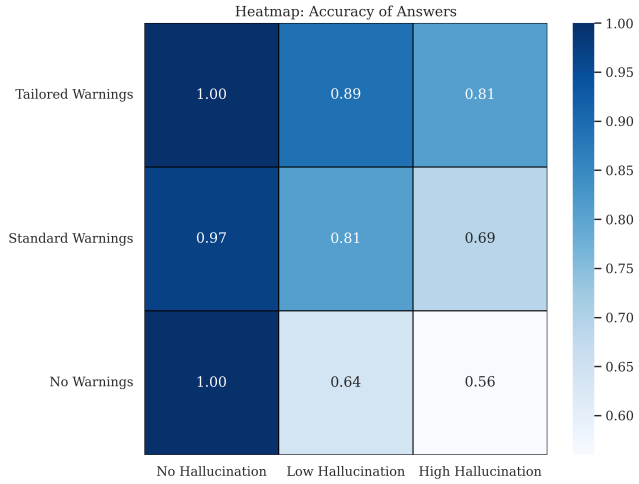
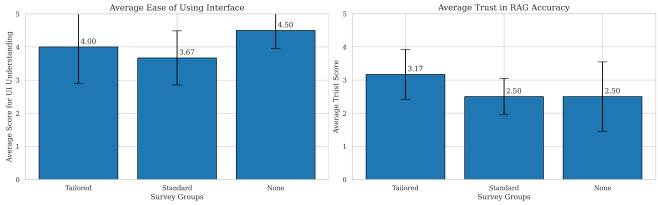


Figure 2: Accuracy of Answers

(1) **Tailored Warning Group:** Participants have the highest detection rate of hallucinations and provide more accurate accuracy ratings compared to the other groups.
(2) **Standard Warning Group:** Moderate detection rates, as the general warnings may lack specificity but still encourage caution.
(3) **No Warning Group:** Lowest detection rates, as participants have no guidance to identify potential hallucinations or biases.



(a) Ease of Usage **(b) Trust in Model**

Figure 3: User Experience

Trust and User Experience. The evaluation of interface usability (Figure 3a, scored from 1 (not easy to use) to 5 (very easy to use), shows that participants in the no warning group found the interface easiest to use, followed by the tailored warning group, and finally, the standard warning group. This might suggest further improvements can be made in the way warning messages were conveyed to the user; that said, it is certainly interesting how the tailored survey group had a easier time utilizing the interface than the standard. A possibility for this result may lie in that more useful/specific warning messages allows for the user to feel as if they understood the output better, thus feeling more at ease. Perhaps more importantly, Figure 3b expresses how tailored users tend to have more trust in the model than their other two counterparts, with a difference of 0.67 (out of a 5-point Likert scale).

5 DISCUSSION

Our study demonstrated the significant impact of tailored warning messages in enhancing users' ability to detect hallucinations within a RAG system for TQA tasks. By comparing no-warning, standard-warning, and tailored-warning conditions across varying levels of response accuracy (genuine, low-level hallucination, and high-level hallucination), we gained insights into how contextualized warning systems improved user engagement, trust, and performance.

Tailored warning systems improve hallucination detection. Our findings revealed that participants in the tailored-warning group demonstrated significantly better performance in detecting both low-level and high-level hallucinations compared to participants in the no-warning and standard-warning groups. Specifically, participants in the tailored-warning group achieved an average accuracy improvement of 25% over the no-warning group and 15% over the standard-warning group. These results provide compelling evidence for the enhanced effectiveness of tailored warnings in improving user accuracy and decision-making when interacting with AI systems [14].

This aligns with prior research, which has shown that incorporating static warnings can mitigate potential harms associated with AI systems [25]. However, our findings advance this understanding by demonstrating that tailored warnings—those customized to the specific context or nature of potential hallucinations—are significantly more effective than general, standard warnings. Unlike previous research, which established the utility of standard warnings, our study highlights the importance of tailoring warning messages to maximize their impact [26]. This nuanced approach ensures that users are not only made aware of potential issues but are also better equipped to identify and address them, ultimately leading to improved trust and reliability in AI-driven systems.

User trust and confidence. The nuanced relationship between user trust and warning systems reveals a complex psychological dynamic in human-AI interaction [29] [25]. Our findings suggest that transparency is not merely about presenting warnings, but about crafting them in a way that empowers users rather than intimidates them. The statistically significant difference in trust levels (0.67 on a 5-point Likert scale) between the tailored warning group and other groups underscores the critical role of contextually relevant information [14].

This phenomenon extends beyond our immediate research context. In domains ranging from healthcare to financial decision-making, users increasingly seek systems that provide not just information, but meaningful insights into the reliability of that information [2] [22]. The tailored warning approach represents a paradigm shift from passive consumption to active engagement with AI-generated content. By providing specific, actionable context about potential hallucinations, we transform users from passive recipients to critical evaluators of information [21].

Moreover, the psychological impact of these warnings cannot be overstated. Traditional approaches often create a binary perception of AI systems as either entirely trustworthy or completely unreliable [26]. Our research demonstrates a more nuanced approach: users can be made aware of potential limitations while maintaining a constructive relationship with the technology. This approach aligns with emerging research on responsible AI development, which emphasizes transparency and user empowerment [7].

Decision-making and task performance. The implications of our findings extend far beyond mere accuracy metrics [3]. We observed a profound transformation in how users approach complex tasks when equipped with tailored warning systems. The accuracy improvements are not just numerical outcomes but represent a fundamental shift in cognitive processing. In the context of Textbook Question Answering (TQA), users demonstrated an enhanced ability to critically evaluate retrieved information. This critical thinking approach is particularly crucial in educational and professional settings where the stakes of misinformation are high. The tailored warning system effectively acts as a cognitive scaffold, guiding users through the complex process of information evaluation [24].

The performance differential becomes even more significant when considering high-level hallucinations. While the no-warning group showed a dramatic drop in accuracy (from 100% to 56%) as hallucination complexity increased, the tailored warning group maintained a more consistent performance (from 100% to 81%). This suggests that contextual warnings do more than simply flag potential issues—they fundamentally alter users' information processing strategies [10].

Enhancing user satisfaction and engagement. Beyond the quantitative metrics, our research uncovered subtle yet significant insights into user experience design. The interface usability scores revealed an intriguing pattern: the tailored warning group found the system more navigable and intuitive compared to the standard warning group. This counterintuitive finding suggests that well-designed warnings can actually enhance, rather than impede, user experience [2].

The satisfaction metrics point to a broader principle in human-computer interaction: users appreciate systems that treat them as intelligent, discerning agents rather than passive consumers. The tailored warnings provide a sense of agency, allowing users to feel more in control of their interaction with the AI system. This psychological empowerment translates directly into higher satisfaction and engagement [21].

Advancing AI-human collaboration. Our research represents a critical step towards more transparent, interactive AI systems. Traditional RAG frameworks have often been criticized as opaque "black boxes" that provide outputs without context or explanation [24]. The tailored warning system challenges this paradigm by introducing a dialogic element to AI interaction.

This approach resonates with emerging philosophical and technological discussions about AI's role in human decision-making [29]. Rather than positioning AI as an infallible oracle or a completely unreliable source, we propose a collaborative model where AI and humans work together, each compensating for the other's limitations [16]. The warning system becomes a bridge, facilitating more meaningful and critical interaction .

Applications in educational settings. The potential impact of this study on educational technology is particularly profound. In an era of increasing digital learning with the incorporation of AI, the ability to critically evaluate information becomes as important as the information itself [17]. Our tailored warning system offers a model for educational AI that doesn't just deliver content but teaches critical evaluation skills. We are able to now imagine adaptive learning platforms where students are not just passive recipients of information but active participants in their learning process. The warning system could be integrated into digital textbooks, research databases, and online learning platforms, providing real-time guidance on information reliability.

Limitations and Future Research. While our study provides significant insights, it also opens numerous avenues for future investigation. The current research was limited by its demographic concentration and specific context of textbook question answering. While our study provides promising insights, it is important to acknowledge several potential limitations. First, the sample size was relatively modest, which may restrict the robustness and statistical power of our findings. Additionally, the demographic concentration of participants—primarily 20-23-year-old university students—limits the diversity of perspectives represented in the study.

This age group, likely comprised of university students or early-career individuals, may not fully capture the behaviors, preferences, and challenges of users from broader age ranges or different life stages.

Furthermore, the context in which the study was conducted may have introduced implicit biases. For instance, participants' familiarity with digital tools or specific subject matter could influence their interaction with the system, making it difficult to generalize findings to populations with varying levels of digital literacy or expertise. Future studies should explore:

(1) **Demographic diversity**: Extending the research across different age groups, educational backgrounds, and cultural contexts.
(2) **Domain-specific adaptations:** Developing tailored warning systems for specialized domains like medical research, legal studies, and scientific publications.
(3) **Long-term cognitive impact** : Investigating how repeated exposure to such warning systems might alter users' information literacy and critical thinking skills.
(4) **Long-term cognitive impact** : Investigating how repeated exposure to such warning systems might alter users' information literacy and critical thinking skills.
(5) **Technological scalability**: Developing more sophisticated algorithms for generating context-aware warnings

6 CONCLUSION

In this work, we explored the impact of tailored warnings on user trust, accuracy in detecting hallucinations, and decision-making in RAG-based TQA systems. Our results showed that the tailored warning system significantly improved users' ability to identify hallucinated content compared to no warnings or general warnings, without diminishing trust in accurate responses. Furthermore, the tailored approach enhanced user confidence and satisfaction by providing context-specific alerts, mitigating the risk of overgeneralized skepticism. In summary, these findings highlight the value of adaptive, user-focused interventions in addressing hallucinations and improving trust in AI systems. This underscores the need for continued research into dynamic, explainable AI solutions to support reliable human-AI collaboration.

7 ETHICS STATEMENT

This study is a pilot study; therefore, no Institutional Review Board (IRB) approval was required. All participants were recruited from the Stanford student body, were over the age of 18, and provided informed consent to participate. Our data collection and analysis protocols adhered to ethical standards and ensured the protection of user privacy. While the study involved generating and analyzing potentially misleading hallucinations, participants were fully informed that the texts were machine-generated and could contain fabricated content. In the future, as we expand this paper beyond its pilot study phase, we plan to seek approval from our institution's IRB to ensure adherence to all ethical guidelines.

ACKNOWLEDGMENTS

We extend our gratitude to Professor Michael Bernstein and Dr. Savannah Cofer from Stanford University for their invaluable support in the development of this research. We also appreciate the Stanford participants who contributed to this study. We also thank the anonymous reviewers for their constructive feedback.

REFERENCES

[1] Hessa A. Alawwad et al. 2024. Enhancing Textbook Question Answering Task with Large Language Models and Retrieval Augmented Generation. *arXiv preprint arXiv:2402.05128v2* (2024).
[2] Jaeyeon Byun et al. 2024. Design and Implementation of an Interactive Question-Answering System with Retrieval-Augmented Generation for Personalized Databases. *https://doi.org/10.3390/app14177995* (2024).
[3] Jiawei Chen, Hongyu Lin, and Han. 2024. Benchmarking Large Language Models in Retrieval-Augmented Generation. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 16 (2024). https://doi.org/10.1609/aaai.v38i16.29728
[4] Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928* (2022).
[5] Sunhao Dai et al. 2024. Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era. *arXiv preprint arXiv:2404.11457v2* (2024).
[6] Vladimir Karpukhin et al. 2020. Dense Passage Retrieval for Open-Domain Question Answering. *arXiv preprint arXiv:2004.04906v3* (2020).
[7] Shachi H Kumar, Saurav Sahay, Sahisnu Mazumder, Eda Okur, Ramesh Manuvinakurike, Nicole Beckage, Hsuan Su, Hung-yi Lee, and Lama Nachman. 2024. Decoding Biases: Automated Methods and LLM Judges for Gender Bias Detection in Language Models. *arXiv preprint arXiv:2408.03907* (2024).
[8] Jiarui Li, Ye Yuan, and Zehua Zhang. 2024. RAGAR, Your Falsehood RADAR: RAG-Augmented Reasoning for Political Fact-Checking using Multimodal Large Language Models. *arXiv preprint arXiv:2404.12065* (2024).
[9] Luyang Lin, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. 2024. Investigating Bias in LLM-Based Bias Detection: Disparities between LLMs and Human Perception. *arXiv preprint arXiv:2403.14896* (2024).
[10] Chu Fei Luo et al. 2024. Misinformation with Legal Consequences (MisLC): A New Task Towards Harnessing Societal Harm of Misinformation. *arXiv preprint arXiv:2410.03829* (2024).
[11] Aigerim Mansurova, Aiganym Mansurova, and Aliya Nugumanova. 2024. QA-RAG: Exploring LLM Reliance on External Knowledge. *Big Data and Cognitive Computing* 8, 9 (2024), 115.
[12] Yuren Mao, XueMei Dong, Wenyi Xu, Yunjun Gao, Bin Wei, and Ying Zhang. 2024. FIT-RAG: Black-Box RAG with Factual Information and Token Reduction. *Papers With Code* (2024).
[13] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*. 3195–3204.
[14] Mahjabin Nahar, Haeseung Seo, Eun-Ju Lee, Aiping Xiong, and Dongwon Lee. 2024. Fakes of Varying Shades: How Warning Affects Human Perception and Engagement Regarding LLM Hallucinations. *arXiv preprint arXiv:2404.03745* (2024).
[15] Jim Parsons. 1982. The Nature and Implication of Textbook Bias. *ERIC ED280769* (1982).
[16] Duy Khoa Pham and Bao Quoc Vo. 2024. Towards Reliable Medical Question Answering: Techniques and Challenges in Mitigating Hallucinations in Language Models. *arXiv preprint arXiv:2408.13808* (2024).
[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
[18] Devendra Sachan, Kelvin Guu, and Sameer Singh. 2024. Improving the Domain Adaptation of Retrieval-Augmented Generation (RAG) Models for Open Domain Question Answering. *Transactions of the Association for Computational Linguistics* (2024).
[19] Luyao Shi, Michael Kazda, Bradley Sears, Nick Shropshire, and Ruchir Puri. 2024. Ask-EDA: A Design Assistant Empowered by LLM, Hybrid RAG and Abbreviation De-hallucination. *arXiv preprint arXiv:2406.06575* (2024).
[20] Artur Strzelecki. 2023. To use or not to use ChatGPT in higher education? A study of students' acceptance and use of technology. *Interactive learning environments* (2023), 1–14.
[21] Viju Sudhi, Sinchana Ramakanth Bhat, Max Rudat, and Roman Teucher. 2024. RAG-Ex: A Generic Framework for Explaining Retrieval Augmented Generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2776–2780.

[22] Cheng Tan, Jingxuan Wei, Linzhuang Sun, Zhangyang Gao, Siyuan Li, Bihui Yu, Ruifeng Guo, and Stan Z Li. 2024. Retrieval Meets Reasoning: Even High-school Textbook Knowledge Benefits Multimodal Reasoning. *arXiv preprint arXiv:2405.20834* (2024).

[23] Yixuan Tang and Yi Yang. 2024. MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries. *arXiv preprint arXiv:2401.15391* (2024).

[24] Shuting Wang et al. 2024. DomainRAG: A Chinese Benchmark for Evaluating Domain-specific Retrieval-Augmented Generation. *arXiv preprint arXiv:2406.05654* (2024).

[25] Zijie J Wang, Chinmay Kulkarni, Lauren Wilcox, Michael Terry, and Michael Madaio. 2024. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–40.

[26] Kevin Wu, Eric Wu, and James Zou. 2024. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. *Preprint* (2024).

[27] Xuyang Wu et al. 2024. Does RAG Introduce Unfairness in LLMs? Evaluating Fairness in Retrieval-Augmented Generation Systems. *arXiv preprint arXiv:2409.19804v1* (2024).

[28] Zhipeng Xu, Zhenghao Liu, Yibin Liu, Chenyan Xiong, Yukun Yan, Shuo Wang, Shi Yu, Zhiyuan Liu, and Ge Yu. 2024. Activerag: Revealing the treasures of knowledge via active learning. *arXiv preprint arXiv:2402.13547* (2024).

[29] Saber Zerhoudi and Michael Granitzer. 2024. PersonaRAG: Enhancing Retrieval-Augmented Generation Systems with User-Centric Agents. *arXiv preprint arXiv:2407.09394* (2024).

[30] Yinghao Zhu, Changyu Ren, Shiyun Xie, Shukai Liu, Hangyuan Ji, Zixiang Wang, Tao Sun, Long He, Zhoujun Li, Xi Zhu, et al. 2024. REALM: RAG-Driven Enhancement of Multimodal Electronic Health Records Analysis via Large Language Models. *arXiv preprint arXiv:2402.07016* (2024).