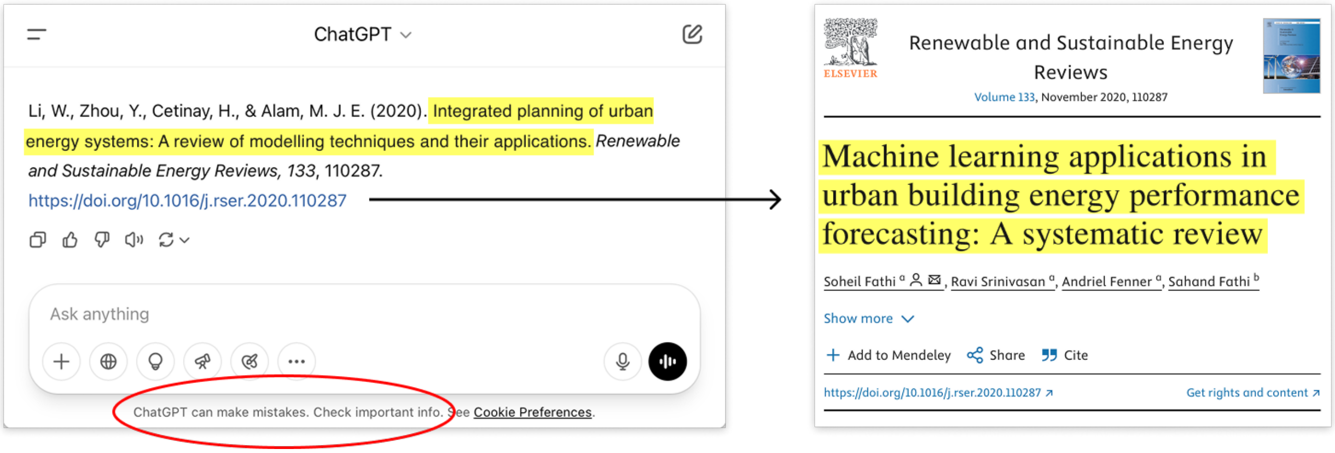
The Impact of Visual Warnings on User Trust in LLM-Generated References

Siiri Emilia Hytönen
Aalto University
Espoo, Finland
siiri.hytonen@aalto.fi

Viivi Elina Uhari
Aalto University
Espoo, Finland
viivi.uhari@aalto.fi



Li, W., Zhou, Y., Cetinay, H., & Alam, M. J. E. (2020). Integrated planning of urban energy systems: A review of modelling techniques and their applications. *Renewable and Sustainable Energy Reviews, 133*, 110287. https://doi.org/10.1016/j.rser.2020.110287

ChatGPT can make mistakes. Check important info. See Cookie Preferences.

Renewable and Sustainable Energy Reviews
Volume 133, November 2020, 110287

Machine learning applications in urban building energy performance forecasting: A systematic review

Soheil Fathi [a], Ravi Srinivasan [a], Andriel Fenner [a], Sahand Fathi [b]

https://doi.org/10.1016/j.rser.2020.110287

Figure 1: ChatGPT may produce inaccurate references but the possibility of such mistakes is communicated only through a faint disclaimer circled in red. On the right, ChatGPT has generated a reference in response to a user's request, but following the link leads the user to a different source.

Abstract

Generative large language models (LLMs) such as ChatGPT are known to confidently produce false or misleading textual information, a phenomenon often referred to as “hallucinations”. Although there are advances in detecting and mitigating these hallucinations, and subtle disclaimers such as “ChatGPT can make mistakes” exist, there is limited research on how to effectively communicate these risks to users through visual warnings. There is also limited understanding of how such warnings impact trust in human-AI collaboration. This gap is particularly relevant in common tasks requiring factual accuracy, such as generating reference lists. This study investigated how two types of visual warnings, a text warning and a modal warning, affected users' trust in AI during a reference list creation task. Participants completed the task with the help of an LLM tool, after which their trust in AI was measured using a questionnaire. During the task, participants were exposed to one of the two warnings or no warning at all. Results showed that the presence of warnings increased trust in AI, contrary to our hypothesis that a more prominent warning would reduce trust. Furthermore, the presence of warnings was associated with less verification behavior regarding the legitimacy of the generated references. Participants also provided positive feedback about the modal warning, while the text warning was perceived as unnoticeable, although some participants disregarded both. This study offers insights into how visual warnings can paradoxically increase trust in AI and how even more alerting options would be needed. Additionally, it provides a foundation for designing AI warnings and interfaces that balance trust and transparency. Further research with a larger sample size is needed to better understand the impact of warnings on trust in AI.

ACM Reference Format:
Siiri Emilia Hytönen and Viivi Elina Uhari. 2026. The Impact of Visual Warnings on User Trust in LLM-Generated References. In . ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Generative Large Language Models (LLMs) such as ChatGPT have become increasingly integrated into everyday digital interactions, assisting users with tasks such as content generation, summarization, programming, and reference list creation [1, 4, 11]. However, a persistent challenge with LLMs is their tendency to produce misleading or incorrect information, sometimes referred to as "hallucinations." These hallucinations often present fabricated information as fact, such as generating nonexistent scientific references [11]. For example, ChatGPT may generate references that look convincing

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

but contain inconsistent data. In Figure 1, the reference link generated by ChatGPT takes the user to an existing paper, but the title of the original source is different from the one provided by ChatGPT. While research has explored methods to detect and mitigate hallucinations, most efforts focus on internal model adjustments rather than user-facing interventions.

Trust and coherence are central factors in the effective adoption and use of LLMs [13], and user interface (UI) design plays a key role in shaping this trust. Overtrust can lead users to accept AI-generated outputs without scrutiny, while distrust can cause them to dismiss potentially useful responses unnecessarily. The way LLM-generated predictions are presented within an interface directly influences these trust dynamics [7]. For instance, poorly designed transparency features can distort mental models and lead to inappropriate reliance on LLM-generated content [7].

The rise of LLM tools makes it increasingly important for designers to support the formation of accurate mental models about LLM capabilities [13]. These trust-related challenges are particularly pronounced among students and non-expert users, who often lack the technical expertise to critically evaluate AI outputs [1, 4], emphasizing the importance of understanding trust calibration from their perspective. It is especially important to moderate students' trust in AI appropriately to support their effective use of these tools [1]. LLM tools have also become increasingly popular among university students, for whom creating reference lists is an inevitable academic task.

Some prior work has investigated approaches to influence users' trust in AI, with mixed success. Behavior descriptions, meaning textual explanations detailing when an AI is likely to be accurate or unreliable, have been shown to enhance human-AI collaboration by guiding users to override systematic AI failures [3]. However, they have not been shown to significantly impact trust in AI, suggesting that alternative methods deserve further exploration. Other UI features, such as transparency disclaimers, were designed with LLM-tool users to mitigate the effects of hallucinations on users [11], but their effectiveness was not empirically tested. Transparency disclosures, on the other hand, have been shown to enhance AI UI credibility, which in turn has been associated with increased trust in AI [12]. However, specific designs were not explored. These studies reveal a gap in current research lacking empirical evidence on how more prominent warnings could affect trust in AI. To address this gap, the current study aimed to answer the following research question:

RQ: How do visual warnings moderate trust in AI-generated references?

We investigated how trust in AI was influenced specifically during a reference list creation task, a common task among students that requires factual accuracy, to address the practical issues faced by this user group. The problem of prior UI-based interventions having limited impact on trust was addressed by introducing more noticeable warnings, while the challenge of moderating trust was approached by implementing one warning intended to increase trust and another intended to decrease it.

The study was executed as a between-subjects experiment. Fifteen participants completed the study from start to finish and were equally assigned to see either a text warning, a modal warning, or no warning. The warning was displayed while they completed a reference list creation task with AI assistance, after which their trust levels, task experience, and perceptions of the warnings were measured using a questionnaire. Drawing on previous research, we also formulated two hypotheses. We hypothesized that a text warning, similar to disclaimers but more noticeable, would increase trust in AI by enhancing the UI's credibility, while a more intrusive modal warning would reduce trust by encouraging skepticism toward AI. The hypotheses were as follows:

H1: The presence of a text warning will positively affect trust in AI-generated references.

H2: The presence of a modal warning will negatively affect trust in AI-generated references.

2 Related Work

Previous research has recognized the importance of moderating students' trust in AI, as well as the significant influence of AI interface design on user perceptions. Some studies have also explored how design interventions can help mitigate the effects of hallucinations and guide users' trust in AI outputs. The following sections review key findings from human-computer interaction (HCI) research on trust in human-AI collaboration with a focus on UI design.

2.1 Students' Trust in AI

Amoozadeh et al. [1] explored the role of trust in shaping university students' adoption of generative AI tools, particularly in programming tasks, as part of their potential integration into computer science (CS) education. Insights were gathered through a survey of 253 CS and non-CS students from two large universities in the US and India, featuring both multiple-choice and open-ended questions. The survey explored students' experiences with generative AI, perceptions of its utility and limitations, and whether their trust in these tools is appropriately calibrated. The results revealed varying levels of trust in AI, with 47% of participants expressing trust, 36% remaining neutral, and 16% expressing distrust. Trust was positively correlated with improvements in motivation and confidence, but skepticism about generative AI's accuracy and the need for human oversight was also common. Overall, a quarter of students found AI helpful. However, many of them, especially those with more programming experience, expressed distrust in its capabilities despite recognizing its utility. Amoozadeh et al. also highlight the importance of moderating trust, avoiding both overreliance and excessive skepticism, to ensure effective use of AI. The authors express that identifying the appropriate level of trust can help educators guide students in calibrating their trust to maximize AI's benefits. Hence, the research calls for actions to understand the factors that impact trust, which would be valuable for educators and designers working on AI models and interface design. Moderating trust in students is also important since around half of the participants perceived generative AI as transparent, even though these models lack opacity.

2.2 The Influence of UI Design in Trust in AI

Ibrahim et al. [7] examine the critical role that AI interface designs play in shaping user behavior and perceptions of AI capabilities and risks. They argue that anthropomorphic, deceptive, and immersive design features in AI interfaces can influence human-AI interactions in ways that are often overlooked in evaluations of AI systems' harms and risks. The paper introduces the Design-Enhanced Control of AI Systems (DECAI) model, grounded in control systems theory, to assess the impact of interface features. DECAI focuses on identifying affordances of design features that influence the presentation of AI output and the nature of user input in the system. As a case study, Ibrahim et al. focus on how design choices of LLM interfaces contribute to flawed perceptions of LLMs. The authors identify anthropomorphic cues and the lack of transparency disclosures focusing on the former and assessing its impact on the user state. By applying DECAI to conversational AI models like ChatGPT, the authors are able to imagine relevant hypotheses and varied user states, such as everyday users or domain experts. The research does not produce definitive answers about the impact of deceptive design pattern's, such as the absence of disclaimers, but it calls for greater scrutiny of interface designs in AI applications, emphasizing their impact on user autonomy, trust, and long-term behavior. This study aligns with our efforts to understand the effects of visual warnings and interface designs in fostering critical evaluation of AI outputs, helping users make more informed decisions when interacting with generative AI systems.

2.3 Different UI Designs' Effects on Trust in AI

Sun et al. [12] explored how different user interfaces (text-based, speech-based, and embodied) affect users' trust in health information provided by LLMs. Their mixed-methods study, involving 30 participants, found that trust in text-based interfaces was the highest followed by speech-based interfaces, while embodied interfaces received significantly lower trust ratings. The results showed a strong correlation ($r = 0.72$, $p < 0.01$) between trust in the interface and trust in the information itself, suggesting that interfaces plays a key role in shaping user perceptions of AI credibility. Notably, trust levels did not vary significantly across different types of health questions, suggesting that trust is more dependent on the perceived credibility of the interfaces rather than the nature of the health questions themselves. However, their embodied interface was not fully developed into a realistic or functional form, potentially limiting the validity of findings regarding trust. Participants may have judged the interface based on its unfinished appearance rather than its potential effectiveness in a real-world scenario. While this research focuses on trust in LLMs in healthcare, it provides insights relevant to our investigation of AI-generated references. The findings of Sun et al. highlight the importance of UI elements in shaping trust perceptions, reinforcing the need to explore how visual indicators can inform users about AI limitations and influence their critical assessment of AI-generated outputs.

Leiser et al. [11] recognize the efforts made to identify hallucinations of LLMs but draw attention to the research gap in how to effectively inform users about these deficiencies. Hence, Leiser et al. investigated strategies to mitigate the effects of LLM hallucinations on users through a participatory design approach. By conducting a workshop with eleven everyday users of LLMs, fourteen possible features were identified to address this issue. These features were then evaluated in one-on-one interviews with six machine learning (ML) experts, all holding relevant master's degrees and professional experience in related fields. They identified three additional features, resulting in seventeen features in total. Most features focused on evaluating the models' responses including the incorporation of confidence scores, source attribution, and mechanisms to categorize responses as factual or fictional. In addition, the features included adjustments to confident phrasing, the use of color-coded visual indicators, and disclaimers to inform users of LLM limitations, monetary or political interests, and potential ethical or legal concerns. Custom interfaces were also highlighted to accommodate diverse users and domains, allowing more inquisitive users to access greater detail. The experts generally found the proposed features valuable, though challenging to implement. However, some, such as static disclaimers, similar to our visual warnings, were evaluated as straightforward.

Cabrera et al. [3] investigated the role of behavior descriptions (BDs) in improving human-AI collaboration by updating people's mental models of AI behavior. Their study involved 225 participants in three distinct domains: fake review detection, satellite image classification, and bird classification. Participants labeled instances under three conditions: without AI assistance, with AI assistance labeled as 90% accurate, and with AI assistance paired with BD providing insight into AI reliability, for example, "our system often describes mountain climbing as skiing". Their findings show that BD improved the accuracy of human AI tasks by 4.2 to 9. 8 % points in all domains, helping users overcome systematic AI failures and selectively trust the system when no failures were evident. The research also found that participants improved more quickly when using BDs, learning to effectively complement the AI. Notably, participants in the BD condition corrected AI failures 32% more often than those without BD support, indicating that users were more proactive in verifying AI-generated outputs. The study also examined subjective measures of trust, helpfulness, and satisfaction through post-task questionnaires, but found no significant differences in user trust between conditions, highlighting the task-specific nature of BDs in this study. The research suggests future experiments, such as testing the types of BDs that are the most effective in improving people's performance. Although the research by Cabrera et al. focused on textual descriptions to update the mental models of users, our study expands this concept by exploring the use of more visually distinct warnings, independent of the task, to communicate AI limitations and reduce overreliance.

While previous research has addressed the calibrated trust in AI systems [1] and the influence of interface design on user perceptions [7, 12], there remains a significant gap in understanding how visual warnings specifically impact user trust in AI-generated content. Although transparency features like disclaimers and confidence scores have been proposed, most solutions require complex modifications to LLMs and lack empirical validation with real users [11]. In addition, while current studies suggest that interface designs can be strategically manipulated to affect user trust [7], actual user testing of such features has not been conducted. Cabrera et al. [3] found that behavior-based descriptions have little impact on trust levels, but they did not explore the effects of static visual warnings with varying levels of visual emphasis. Our study extends these results by empirically investigating how a text and modal warning shape trust in AI-assisted citation tasks. We focus on everyday users, specifically students, rather than experts, which distinguishes our work from most prior studies [1]. Our aim was to provide actionable insights into how transparent design can foster informed AI interactions while avoiding deceptive design patterns. This research not only addresses a critical gap in literature but also offers implications for policymakers and designers concerned with ethical and usable AI deployment.

3 Methodology

This study employed a between-subjects experimental design combined with quantitative and qualitative survey methods to examine how two different visual warnings influenced user trust in LLM-generated references. Participants engaged in an AI-assisted reference list creation task under one of three experimental conditions, two featuring a distinct visual warning design and one serving as the control condition with no warning.

The between-subjects approach ensured that participants were exposed to only one version of the warnings, preventing learning effects and fatigue that could occur if they repeatedly completed the same task under different conditions [10]. This method also enabled faster data collection and maintained ecological validity by simulating real-world scenarios where users engage with a single AI interface. However, individual differences in AI literacy and experiences in handling references may have introduced variability, which we mitigated through random assignment of participants to conditions to ensure internal validity.

By using an experimental approach, we were able to systematically manipulate interface-level interventions and assess their effects on user trust without modifying the underlying LLM model [11]. This aligns with best practices in HCI research for establishing causal relationships between interface design choices and user behavior [10].

The participants answered a post-task questionnaire that examined usage experience with AI tools, trust in AI, experiences during the task, and the effectiveness of the visual warnings. The questionnaire included closed-ended questions in addition to open-ended prompts that captured qualitative insights, following Bruhlmann et al. [2] survey design-practises. This mixed-methods approach ensured that we captured both quantifiable trust levels and contextual user perceptions. We used previous studies' Likert-scales about human-AI trust to ensure the quality of post-task questions [5, 6]

An overview of the study design is shown in Figure 2. Before finalizing the procedure, we conducted a pre-study to refine the reference list creation task and questionnaire content based on user feedback. In addition, the data analysis is explained in section 3.6.

3.1 Experimental Setup

To evaluate the impact of visual warnings on user trust in LLM-generated references, we conducted an experimental study with 15 participants, assigning five participants to each of the three warning conditions. To mitigate potential biases and individual differences in AI literacy or experience with AI-assisted reference list creation [10], participants were randomly assigned to one of the conditions.

Our study followed a between-subjects experimental design based on [10], where the independent, manipulated variable was the type of visual warning and the dependent variable was the level of user trust in AI-generated references. This approach enabled causal inference, allowing us to determine whether specific warnings significantly impacted user trust, overreliance, or skepticism toward AI-generated content.

Our target user group was university students, as they represent a key demographic of everyday LLM users who frequently engage with AI tools for academic work. Participants were recruited through various social media platforms aimed at university students to ensure representation across different majors. In addition, posters located in university buildings were used to reach our target group. Ultimately, the participant pool consisted mainly of students from Aalto University and the University of Helsinki.

Given our focus on real-world applicability, the study followed a non-controlled approach, allowing us to document natural user behavior during an LLM-assisted reference list creation task without imposing artificial constraints. This approach also helped encourage participation, as it allowed participants the freedom to complete the study at their convenience. Participants were recruited through a separate form and were given a unique ID to use during both the reference list creation task and the questionnaire. This ensured that personal information and response data remained separate, eventually enabling anonymization of responses.

3.2 Visual Warnings

We evaluated a single warning message formulation, allowing us to focus more on the warnings' visual aspects. Drawing on the findings of Cabrera et al. [3] on the effectiveness of AI behavior descriptions, we designed an actionable message that would be easy to interpret. We complemented this with insights from Leiser et al. [11], whose participants suggested that disclaimers should include explanations of LLMs' functionality and information about their inherent uncertainty. This led to the creation of an explanatory warning message that was longer than the one currently used by ChatGPT. Testing a more informative warning with novel phrasing allowed us to explore whether users value detailed explanations over the shorter messages found in current systems. Jafari and Vassileva [8] found that while most users prefer succinct, visually prominent warnings, a subset also favors more detailed content, supporting our approach. Our message highlighted the uncertainty of LLMs and included an explanation of the underlying causes of hallucinations. The warning was finalized as follows:

AI responses may be inaccurate or misleading. AI generates answers based on text patterns, and mistakes can happen due to incomplete or biased training data, or not fully understanding questions.

We implemented two distinct visual designs for the warnings. According to Johnson [9], textual warnings are a common approach, which we included as a baseline. Textual disclaimers are also used by current LLM-based tools, and by incorporating a similar warning into our design, we were able to test the effects of current designs. The warning was presented in black, using the standard body font size of the current browser, and warning icons were incorporated to capture the user's attention [8, 9]. The second warning we implemented was a modal warning dialog box that appeared at the center of the screen. While modal warnings are typically reserved for critical scenarios, they remain a common method for conveying important alerts [9]. In our study, a modal warning was chosen to serve as a more alerting and thought-provoking option.

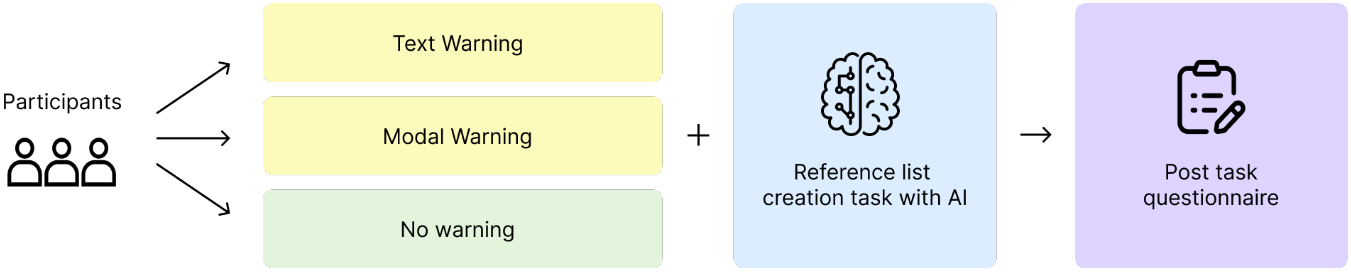


Figure 2: The study design. Participants were randomly assigned to one of three visual warning conditions. The warning was shown during an AI-assisted reference list creation task, followed by a post-task questionnaire.

The text warning remained present throughout the task, while the modal warning was shown only at the beginning. However, a button was provided to reopen the modal if the user wished to review its information again. Displaying disclaimers only at the start of a conversation with an AI was suggested by participants in Leiser et al. [11], though they also recommended showing disclaimers at the beginning of each response. However, critics interviewed by Leiser et al. argued that this would negatively impact usability, which we aligned with and therefore did not test continuously recurring warnings. Cabrera et al. [3] proposed showing descriptions only in common or significant cases, but this approach was not suitable for our study, as we aimed to design static warnings. Finally, the control group completed the collaborative task without any visual warning, since some current LLM-tools lack disclaimers altogether. Figures 3 and 4 show examples of the final UI.

3.3 Reference List Creation Task with AI

As previously outlined, we recruited participants to complete a collaborative task with an AI system designed to generate a reference list. In the following sections, we describe the task and explain its key design choices.

3.3.1 Structured Task. The reference list creation task consisted of instructions, input fields, a dialog view with AI, a view of the generated references, and a possible warning (see Figures 3 and 4). In the instructions, participants were told to compile a high-quality reference list on a given topic, and they were informed of the task's steps:

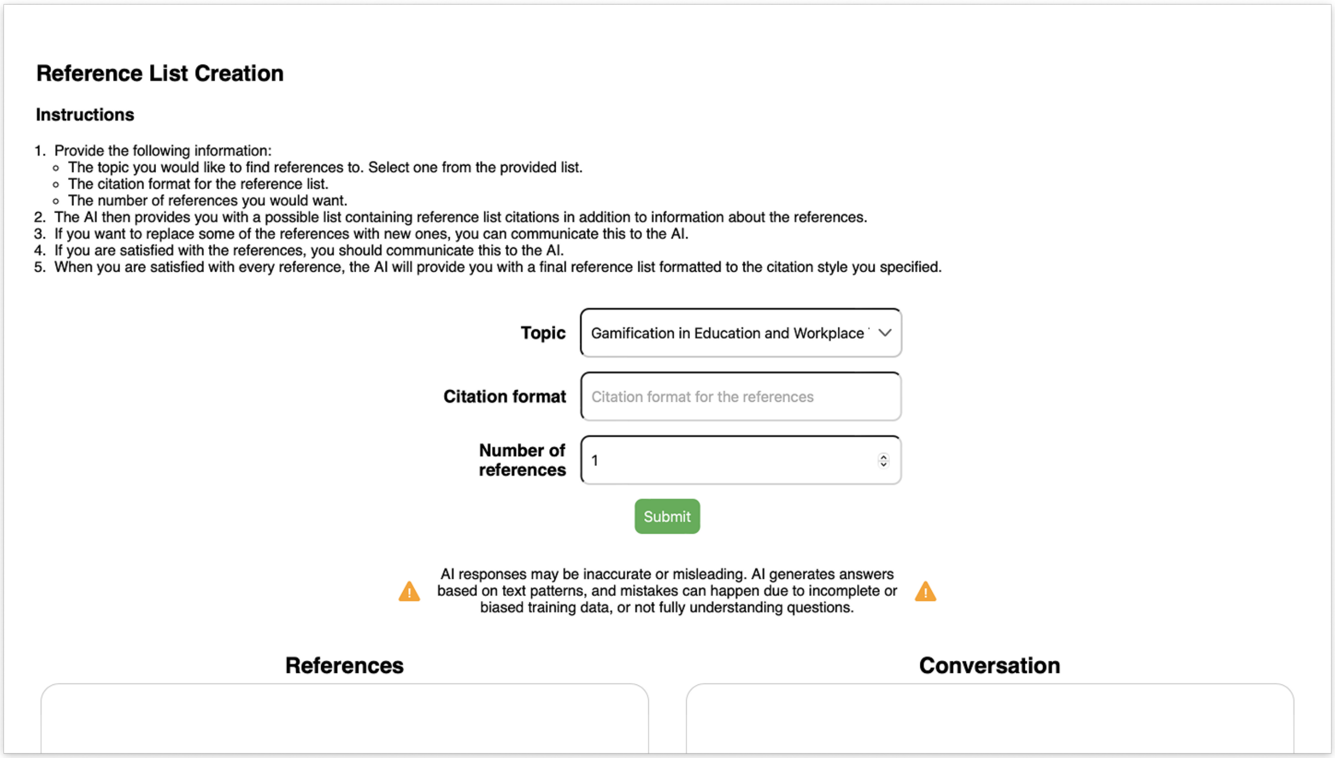
(1) The user begins the collaborative task and reads the instructions.
(2) The user selects a topic for the references.
(3) The user provides the preferred citation format for the reference list.
(4) The user provides the number of references they require.
(5) The LLM generates the requested number of references, providing each with a bibliographic citation (in the requested format), a summary of its content, where it has been published, and a link to the original work.
(6) The LLM asks if any references should be replaced with alternatives.
(7) The user indicates which references, if any, should be replaced.
(8) The LLM generates alternative references and returns an updated list, following the same structure as in step 5.
(9) Steps 6-8 are repeated until the user is satisfied with all references.
(10) The user confirms that the references are satisfactory.
(11) The LLM compiles a finalized reference list, formatted according to the specified citation style.

These steps are visualized in Figure 5, and a hierarchical task analysis (HTA) of the task is shown in Figure 6, which also helped identify informed requirements for the task. In the figure, "unread" references refer to those that the user has not yet reviewed, rather than suggesting that they must be read in full.

3.3.2 Flexible Task Completion. The reference list creation task was deployed as an online website that participants could access from their own devices at their convenience. This design choice aligns with our chosen methodology in Section 3.1, ensuring convenience for participants, ecological validity, and realistic user behavior. In real-world use cases of LLMs, students typically interact with AI tools independently and without external guidance. Reference generation and verification (HTA step 3) also require time, suggesting that participants would benefit from the ability to complete the collaborative task at their own pace and without supervision. Additionally, interpreting references provided by the LLM (HTA step 3) may require accessing external sources beyond the LLM tool, which does not support enclosing the tool in a controlled environment. It was important that participants could leave the task to access other internet resources and return back.

3.3.3 Streamlined Input Process. A structured input mechanism was necessary to enhance user efficiency and reduce cognitive load. Rather than requiring users to provide citation parameters sequentially, we designed the system to collect all relevant information, including the topic, citation format, and number of references, in a single step (HTA step 2). This approach minimizes unnecessary interaction cycles and improves the user's workflow.

To reduce variability in results, we defined a predefined collection of topics for the participants to choose from. The selected topics were relevant in contemporary discourse and moderately explored in academic research. By including options from various fields, we aimed to ensure that each participant had at least one familiar topic, regardless of their major, for which they could evaluate references critically. An alternative approach would have been to allow participants to select a topic aligned with their field of study, reflecting



Reference List Creation

Instructions

1. Provide the following information:
   - The topic you would like to find references to. Select one from the provided list.
   - The citation format for the reference list.
   - The number of references you would want.
2. The AI then provides you with a possible list containing reference list citations in addition to information about the references.
3. If you want to replace some of the references with new ones, you can communicate this to the AI.
4. If you are satisfied with the references, you should communicate this to the AI.
5. When you are satisfied with every reference, the AI will provide you with a final reference list formatted to the citation style you specified.

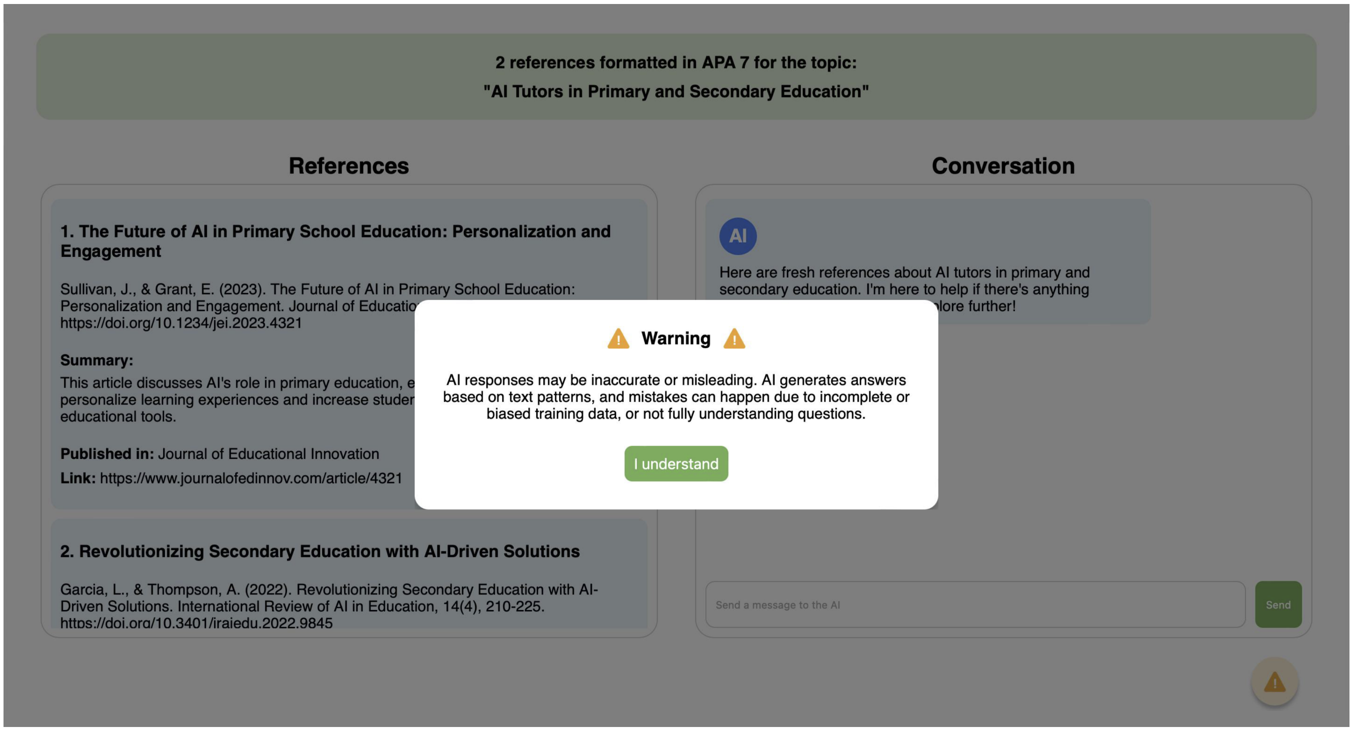
Topic: Gamification in Education and Workplace

Citation format: Citation format for the references

Number of references: 1

AI responses may be inaccurate or misleading. AI generates answers based on text patterns, and mistakes can happen due to incomplete or biased training data, or not fully understanding questions.

Figure 3: Initial UI view at the start of the reference list creation task. In this version, the user is shown a text warning.



2 references formatted in APA 7 for the topic:
"AI Tutors in Primary and Secondary Education"

1. The Future of AI in Primary School Education: Personalization and Engagement

Sullivan, J., & Grant, E. (2023). The Future of AI in Primary School Education: Personalization and Engagement. Journal of Educatio
https://doi.org/10.1234/jei.2023.4321

Summary:
This article discusses AI's role in primary education, e
personalize learning experiences and increase studer
educational tools.

Published in: Journal of Educational Innovation

Link: https://www.journalofedinnov.com/article/4321

2. Revolutionizing Secondary Education with AI-Driven Solutions

Garcia, L., & Thompson, A. (2022). Revolutionizing Secondary Education with AI-Driven Solutions. International Review of AI in Education, 14(4), 210-225. https://doi.org/10.3401/iraiedu.2022.9845

AI: Here are fresh references about AI tutors in primary and secondary education. I'm here to help if there's anything
lore further!

Warning

AI responses may be inaccurate or misleading. AI generates answers based on text patterns, and mistakes can happen due to incomplete or biased training data, or not fully understanding questions.

Figure 4: UI view after the user has entered the topic, format, and number of references, and scrolled down from the instructions. In this version, the user is shown a modal warning.

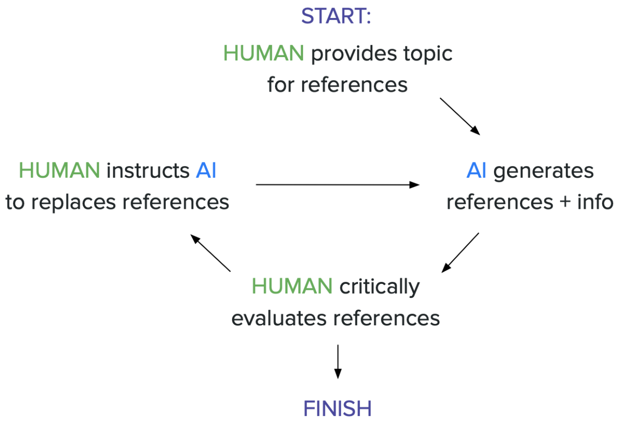

START:
HUMAN provides topic for references
AI generates references + info
HUMAN critically evaluates references
HUMAN instructs AI to replaces references
FINISH

Figure 5: Steps of the reference list creation task with AI

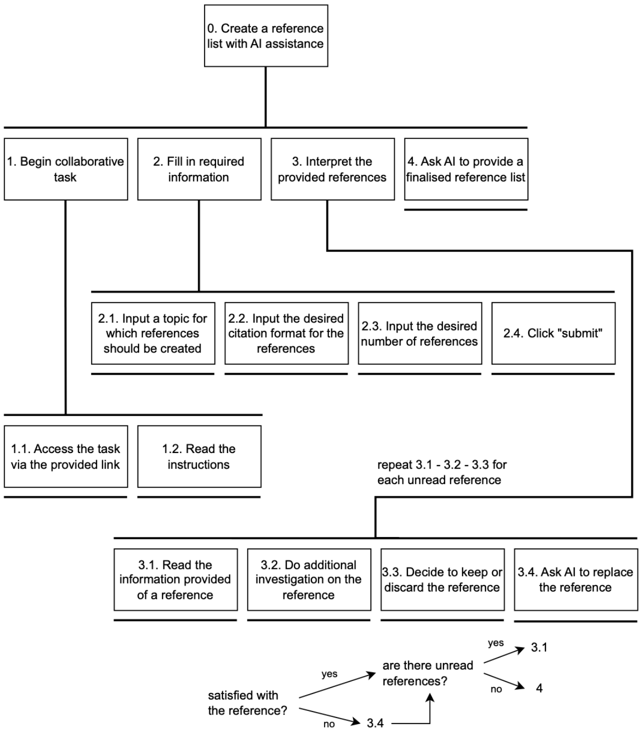

0. Create a reference list with AI assistance
1. Begin collaborative task
2. Fill in required information
3. Interpret the provided references
4. Ask AI to provide a finalised reference list
2.1. Input a topic for which references should be created
2.2. Input the desired citation format for the references
2.3. Input the desired number of references
2.4. Click "submit"
1.1. Access the task via the provided link
1.2. Read the instructions
repeat 3.1 - 3.2 - 3.3 for each unread reference
3.1. Read the information provided of a reference
3.2. Do additional investigation on the reference
3.3. Decide to keep or discard the reference
3.4. Ask AI to replace the reference
satisfied with the reference?
yes
no
3.4
are there unread references?
yes
3.1
no
4

Figure 6: Hierarchical task analysis of the reference list creation task with AI

more realistic scenarios in which students use AI for academic work. However, this would have introduced too much variance, as some topics might contain more well-documented references while others could be more obscure.

3.3.4 Support for Reference Evaluation. Given the uncertainty of the existence and relevance of LLM-generated references, our system supported the evaluation of references. Rather than presenting participants with a finalized reference list immediately, the task provided additional information about each reference, enabling participants to critically evaluate the relevance and credibility of the sources (HTA step 3). This was also supported by the non-confined environment described in Section 3.3.2. By employing an interactive, iterative process, the task fostered critical thinking and encouraged collaboration with the AI. Ultimately, users had control over whether to accept the LLM-generated references blindly or to assess them using the provided information or external resources such as online reference databases.

Providing these mechanisms also allowed us to investigate whether participants take advantage of them for reference verification, which provides insight into their trust in the AI. For example, if participants continue to perceive the references as reliable without recognizing inconsistencies in the provided information, or fail to use the information to verify references, this may indicate overreliance on AI or a lack of critical engagement. Understanding reference verification behaviors offers valuable insights into user interaction with LLM-generated references and the role of visual warnings in shaping trust.

3.4 Questionnaire

To assess the impact of visual warnings, participants completed a post-task questionnaire after finishing the reference list creation task. The questionnaire was designed to operationalize key variables related to user trust, task experience, and usability perceptions of warnings. Drawing on previous HCI research, we used questions that have been proven to be effective in measuring human-AI trust.

At the beginning of the questionnaire, we asked about participants' backgrounds, such as their study major and previous experience with LLM tools. In addition, we asked whether participants were aware of AI hallucinations prior to the study, although this question was placed at the end of the questionnaire to avoid biasing earlier responses. These background questions helped contextualize participants' answers and assess whether prior knowledge of AI influenced their responses. Some of the questions were adapted by those used by Amoozadeh et al.

Warning-specific usability questions assessed understandability, helpfulness, noticeability, and effectiveness of the warnings. The questions were presented as statements that participants rated using 5-point Likert scales (1 – Strongly disagree, 5 – Strongly agree), designed to minimize bias. The questions included: *"The visual warning was hard to understand"*, *"The visual warning was noticeable"*, and *"The visual warning was helpful"*. Additionally, two open-ended questions were included to collect qualitative feedback from the participants on the warnings [2].

By adapting human-AI trust scales from Gulati et al. [5] and Hoffman et al. [6] to our context, we measured participants' trust in AI, their confidence in AI-generated references and their perceptions AI's reliability. Trust-related questions included: *"I trust the information provided by the AI"* [5], *"If I used AI for reference list creation, I think I would be able to depend on it completely"* [5], *"The AI provided sufficient information to create a reliable reference list"* [6] and *"I was confident in detecting errors in AI-generated references"* [6] and *"I would use AI for similar tasks in the future"*. In addition, we assessed the warnings' impact on trust directly with the question

"The visual warning increased my trust towards the AI". As before, participants responded using a 5-point Likert scale.

Finally, we asked participants if and how they verified references or detected errors in them. We also provided space for open-ended feedback, allowing participants to share additional thoughts or clarify responses, ensuring a comprehensive understanding of user perspectives.

3.5 Pre-Study

Since existing literature does not specifically address how students create reference lists using AI, we conducted a pre-study to gain insights into this process. We aimed to explore whether students create reference lists in their studies, their general processes for searching and selecting references, and whether they use LLM tools for reference list creation. Additionally, we sought to determine students' awareness of AI hallucinations and to test a few questionnaire items to further validate the human-AI trust scales. Given our need for qualitative insights rather than large-scale statistical validation, we conducted semi-structured interviews instead of surveys. This provided flexibility in questioning and enabled an in-depth exploration of key topics. We also ruled out contextual inquiry and observation since it would have presented a practical challenge of requiring students to complete full reference list creation tasks within a limited timeframe.

We conducted four semi-structured interviews, each of us interviewing two participants. Participants were recruited through our personal networks. The interviews were conducted both remotely and in person and lasted about 20 minutes on average. A preliminary set of questions was used for the interviews, but deviations were made when additional questions were relevant. Three of the interviewees were students from Aalto University, majoring in Computer Science, Information Networks, and Real Estate Economics, while the fourth was a student from the University of Helsinki studying Art History. All participants were familiar with AI and had used LLM-based tools such as ChatGPT. Their use cases varied from generating references to explaining mathematical algorithms and creating programming scripts.

The processes for finding and selecting suitable references differed significantly between participants. One student primarily relied on references provided by instructors, while another mainly used Google's search engine. When evaluating references, participants primarily assessed the relevance of the content to their text. This informed our decision to make the AI include summaries of the references in our collaborative task. In addition, the interviewees tended to prefer sources that were recent, frequently cited, or appeared "academic" and reliable. Given these results, we ensured that the AI also provided the publishing years of the references and where they were published. Unfortunately, our AI model was not capable to provide information on the citation counts of the references. Regarding AI-assisted reference generation, one interviewee reported using ChatGPT to generate references for text samples, while others avoided AI tools due to their poor performance, such as broken links. This highlighted the importance of including such links to the original sources in our task as well.

To assess trust in AI, we included two differently phrased questions about human-AI trust from existing literature [5, 6], which most participants found understandable. However, responses varied depending on the wording of the trust-related question, highlighting the sensitivity of the question framing. Based on the mixed results, we included multiple trust scales in our final questionnaire. Finally, while most were aware of LLMs' tendencies to generate false information, only one interviewee was familiar with the term "hallucinations" reminding us to provide an explanation of it in our questionnaire.

3.6 Data Analysis

To analyze the results of the questionnaire, we mainly relied on descriptive statistics to identify trends in the responses. We examined the distribution of responses by calculating the means across key variables, such as trust, verification behavior, and perceived usability of visual warnings. Due to the small sample size, we did not conduct statistical significance testing. Instead, we focused on comparing group-wise averages to explore directional differences between conditions. In addition, we calculated the Pearson correlation coefficient to investigate the relationship between participants' trust in AI and their tendency to verify references in the task.

A thematic analysis was conducted for the open-ended responses, identifying emergent themes through an inductive coding process. Answers were grouped based on conceptual similarity, and responses could belong to multiple themes if relevant. To maintain traceability and support within-subject interpretation, we color-coded responses by participant. Themes were visually organized using a clustering approach similar to affinity diagramming. Both researchers collaborated in this process, allowing for a more comprehensive analysis. Using these techniques also allowed us to analyze the responses with our limited expertise and were manageable time-wise due to the small number of responses resulting from our limited number of participants.

4 Results

The data set comprised 15 participants, evenly distributed in three groups: modal warning, text warning, and control (no warning). All participants were university-level students, 60 % studying at Aalto University, 14 % at the University of Helsinki, 14 % at the LUT University and 13% at the University of Jyväskylä. Between each group, the distribution was similar. Each participant had used an AI tool before, but 26 % (N=4) were not familiar with the concept of AI hallucinations, two (N=2) being from text warning group and one (N=1) for each control and modal group. If a participant was unfamiliar with hallucinations, the term was explained later in the questionnaire. Participants reported using LLM tools 'frequently': 66 % 'often' and 20: % 'all the time'.

4.1 Quantitative Results

Quantitative results were measured from the Likert scale post-task questionnaire, where human-AI trust scales were adapted from previous studies by Gulati et al. [5] and Hoffman et al. [6].

4.1.1 Perceived Clarity and Helpfulness of Warnings. Most participants found the warnings easy to understand and visually noticeable, as seen in Table 1. Among both warning conditions, 60% of the participants (N=6) strongly agreed that the warning was noticeable, with a combined mean score of 4.3. The modal warning was perceived as more noticeable (M = 4.6) than the text warning (M = 4.0). Likewise, the warnings were considered easy to understand, with 90% (N=9) strongly disagreed with the statement "The visual warning was hard to understand" (M = 1.1 overall; modal M = 1.0; text M = 1.2).

4.1.2 Perceived Helpfulness of Warning and Effect on Trust. Perceptions of helpfulness were mixed (M = 3.4 for both groups), with 50% of the participants rating the warning as helpful or strongly helpful. The perceived impact of the warnings on trust was more varied. The Modal warning was seen lightly more trust-increasing (M=3.0), while the text warning was seen slightly less trustable (M=2.4).

4.1.3 Trust in AI and Reference Use. Trust toward AI was moderate in each group, as seen in Table 2. The responses of the participants to "I trust the information provided by the AI" varied, with the text warning participants showing slightly higher trust (M = 3.8) than those in the modal (M = 3.4) and control (M = 2.6) groups. However, most of the participants expressed reluctance to completely depend on AI for creating reference lists (M = 2.4 overall).

4.1.4 Behavioral Outcomes. Verification behavior varied. Participants who received a visual warning were less likely to verify AI-generated references (M = 2.40) than those in the control group (M = 3.20). Participants who noticed errors (N=5) reported various types of faults, including broken links or non-existent references. These participants had a slightly higher confidence in detecting errors (M = 3.6) compared to the full sample (M = 2.9).

4.1.5 Correlational Findings. We identified positive correlations between trust and perceived AI helpfulness (r = 0.73), as well as between trust and future willingness to use AI (r = 0.61) as seen in Figure 7. Trust was negatively correlated with confidence in error detection (r = -0.36), and participants who trusted AI less, verified more (r = -0.62) , implying that participants who trusted AI more were less likely to detect its inaccuracies. Participants who were confident in error detection were slightly more inclined to verify references (r = 0.27). No correlation was found between hallucination awareness and trust or verification behaviors.

4.1.6 Trust Scale. For further exploratory analysis, we calculated an aggregated 'Trust score' based on five trust-related items (Q20. – Q25.) and calculated the mean and median for each experimental group, as seen in Table 2. Although this measure is not formally validated, it serves as a supporting guide for both quantitative and qualitative results. The results indicate that trust was highest in the text warning condition, followed by the modal and control groups, which provides weak support for H1. The trust in the control group was the lowest, and while they were the most willing group to completely depend on AI, the groups several technical issues could be the consequence of the low score.

4.2 Qualitative Results

Qualitative responses were collected from open-ended questions regarding whether participants verified references or noticed any errors in them. Additionally, we gathered feedback concerning the warnings and the overall study.

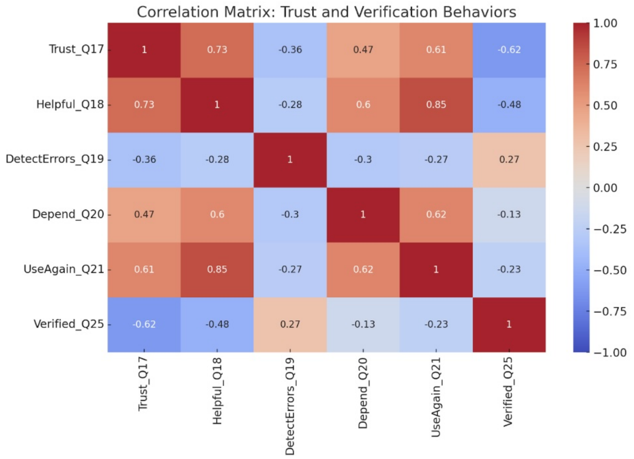


Figure 7: Correlation matrix of trust-related attitudes and verification behavior across six survey items. The matrix visualizes Pearson correlation coefficients between participants' self-reported trust in the AI (Q17), perceived helpfulness of AI-generated references (Q18), confidence in error detection (Q19), willingness to depend on AI (Q20), intention to reuse AI for similar tasks (Q21), and whether they verified the references (Q25).

4.2.1 Reference Verification Behavior. Two participants from each warning group verified the existence of references, whereas four participants from the control group engaged in reference verification. In addition, two participants from the text warning group and one from the control group reported checking the formatting of the references. Of the participants who verified references, three, one from each group, did not identify any errors, despite the fact that all final reference lists contained inaccuracies, particularly regarding the existence of sources. The most common methods used to verify references included accessing the provided links (47%, N=7) and searching for the references through external sources (33%, N=5), such as Google Scholar. Additionally, two participants mentioned that other LLM tools could be used to verify the provided references.

4.2.2 General Feedback on the Warnings. Regarding the warning, participants expressed a generally positive attitude toward the idea of visual warnings designed to inform about the limitations of AI. Of the ten participants who were shown warnings, three (30%, N=3) thought it was good to have warnings, while another two (20%, N=2) perceived them as even necessary. Additionally, 30% (N=3) perceived the warnings as informative, 30% (N=3) thought the warnings served as reminders of AI's weaknesses, and 30% (N=3) also expressed that the warnings encouraged critical thinking. The phrasing was also positively received. It was described as clear, informative, and containing enough technical detail. However, one participant expressed a desire for a more explicit explanation of whether LLMs can make up sources.

4.2.3 Feedback on the Modal Warning. With respect to the specific warnings, the modal warning received only positive feedback, while the text warning attracted more critical responses. Two participants

Table 1: Perceived usability of visual warnings (Mean ratings, 1 = Strongly Disagree, 5 = Strongly Agree)

Item	Modal	Text	Average
The visual warning was noticeable	4.6	4.0	4.3
The visual warning was hard to understand	1.0	1.2	1.1
The visual warning was helpful	3.4	3.4	3.4
The visual warning increased my trust towards the AI	3.0	2.4	2.7

Table 2: Trust-related items across conditions 1 = Strongly Disagree, 5 = Strongly Agree Mean / Median)

Item	Modal	Text	No warning
	M / Mdn	M / Mdn	M / Mdn
(Q17) I trust the information provided by the AI [5]	3.4 / 3.0	3.8 / 4.0	2.6 / 3.0
(Q18) The AI provided sufficient info to create a reliable reference list [6]	3.4 / 4.0	3.4 / 3.0	2.8 / 3.0
(Q19) I was confident in detecting errors in AI-generated references [6]	2.0 / 2.0	3.8 / 4.0	3.0 / 3.0
(Q20) I could depend completely on AI [5]	2.0 / 2.0	2.4 / 2.0	2.8 / 2.0
(Q21) I would use AI for similar tasks in the future.	3.8 / 4.0	4.0 / 5.0	3.4 / 4.0
Trust Score (Average of 6 questions)	**2.92**	**3.48**	**2.92**

from the modal warning group complimented it, and one considered it necessary. One participant from the text warning group also complimented its design, as all participants were eventually shown all warnings during the questionnaire to determine which warning they recalled seeing. However, the modal warning was also ignored by two participants, while only one participant ignored the text warning. Surprisingly, no participants reported the modal warning as annoying in open feedback, but the fact that it was ignored may indicate users' inclination to quickly bypass it.

4.2.4 Feedback on the Text Warning. Despite only one participant reporting having ignored the text warning, it was primarily criticized for being unnoticeable. Two participants suggested that the text warning could have benefited from a background color and stronger visual hierarchy. A third participant noted that it could have been more visible and that it would have benefited from an "I agree" button, similar to the one in the modal warning. Nevertheless, one participant reported that the text warning successfully captured their attention and was not disturbing.

4.2.5 Participant Reflections on the Study. Concerning open feedback about the study itself, some participants unfortunately reported technical difficulties with the collaborative task platform. However, they were still able to respond to the questionnaire items. A few participants also reported that the study prompted them to reflect on and reconsider how they use AI tools.

5 Discussion

5.1 Summary of Findings

By integrating quantitative trends and qualitative responses, we highlight four core contributions gained from this study and outline directions for future research on designing effective visual warnings and trust-calibration mechanisms in human–AI collaborative tasks

Visual warnings did not significantly affect trust or verification behavior. Although not statistically significant, average trust scores were highest in the text warning condition (M = 3.8), followed by the modal warning (M=3.4) and lowest in the no-warning control group (M = 2.6). This weakly supports H1 (the presence of a text warning will positively affect trust in AI-generated references), but does not support H2 (the presence of a modal warning will negatively affect trust in AI-generated references). Users who reported lower trust towards the AI tended to verify references more frequently (r = −0.62), while those exposed to visual warnings verified references slightly less often (M = 2.4 vs. M = 3.2 in the control). However, the results are not statistically significant and since the sample size was small, these trends require further exploration in larger samples.

Participants perceived the visual warnings as usable, yet their behavioral impact was limited. The modal warning was described as informative and clear, while the text warning was often seen as insufficient and not noticeable enough. Some participants viewed warnings as necessary features to foster critical awareness: *"I think they are very necessary... these visual warnings always make you consider how much you can actually trust the end product."* However, others admitted to disregarding them entirely: *"To be honest I kinda ignored the warning... like when a cigarette has a warning, smokers do not read it."*. In particular, even two participants ignored the modal warning, raising questions about whether these cues were truly alerting or easily overlooked, suggesting more studies on the behavioral impact of warnings on collaborative human-AI tasks and trust.

Qualitative feedback revealed preliminary design guidelines. Participants emphasized the importance of visual prominence, clarity, and integration into the user flow. The ineffectiveness of the text warning suggests that faint disclaimers, such as those currently used by tools like ChatGPT, may be insufficient. These insights offer preliminary design implications: to be effective, visual warnings must not only be visible, but also meaningfully disrupt automatic user behavior. However, trust remains complex. Some participants reported that transparency increased their trust, while others felt that the warnings raised doubt. One participant noted that "When

AI gives links to its references, it feels a lot trustworthier..." suggesting that actionable information may be more effective than warnings alone. These findings align with previous work [12] showing that transparency mechanisms improve task participation and trust, and highlight the important role of designing warnings that strike the balance of being alerting enough to raise awareness while maintaining high-level user experience.

Human–AI trust is a complex and context-dependent construct. During the study, another question was also raised: Does the confidence in detecting errors explain trust toward AI more than our visual warnings and Likert scale trust measures? Whether the goal is to increase, decrease, or precisely calibrate trust remains an open design question.

Our findings suggest that current warning mechanisms may be limited in their ability to steer user behavior. We recommend future work to explore significantly different visual or interactive warning types, or investigate complementary mechanisms, such as verifiable links, to inform users about the risks of AI hallucinations.

5.2 Limitations

Some limitations should be considered when interpreting the findings of this study. First, the sample size was small, with only a limited number of participants in each condition. This constraint reduced the statistical validity of the analyses. Replication with larger and more diverse populations is necessary to validate the observed trends.

Second, technical issues and AI hallucinations were prominent within the AI-Arena environment, which occasionally interfered with task completion. Despite prior testing, issues such as AI being unresponsive, slow, or unable to complete the task correctly due to bugs were prominent. While several participants detected errors in the references, four (N=4) participants out of fifteen reported technical bugs. Three (N=3) of them were from the control group, which might have significantly influenced their trust and verification behavior. This highlights the need for robust AI-Arenas when using AI in any experimental or evaluative setting to ensure reliability and valid user responses.

Third, the scope of warning types was limited. In addition to the visual warnings we tested, it would have been valuable to include a condition using a faint disclaimer, such as the subtle warning employed by ChatGPT. This would have provided information on the effectiveness of the current measures in place to inform users about AI hallucinations. This might even have served as a possible control condition, but we ultimately opted for a no-warning baseline, since not all LLM tools include even a minimal disclaimer. In addition, the warnings tested were designed by the research team without the involvement of professional designers or usability experts, which may have constrained their impact and clarity.

Fourth, the study was conducted in a non-controlled environment. Although this approach allowed participants flexibility, it also introduced variability in how tasks were understood and performed. This variation could have influenced trust and verification behavior in ways difficult to capture or control.

Finally, the experiment design was exploratory by nature. As novice researchers, our approach followed an iterative, trial-and-error process. While this is common in early-stage human–AI interaction research, future work would benefit from more rigorous co-design and pilot testing of warning interventions.

6 Conclusions

This study examined how different types of visual warnings influence users' trust in AI-generated references within an academic context. Using a between-subjects design, we compared a subtle inline text warning, a more prominent modal warning, and a no-warning control condition. Surprisingly, participants exposed to either warning reported slightly higher overall trust in the AI system than those in the control group. The text warning, however, led to a greater increase in trust than the modal warning, possibly due to its perceived enhancement of system transparency. According to participant feedback, the modal warning was generally preferred, while the text warning was often seen as insufficiently noticeable. Nonetheless, both warnings were overlooked by some users. Combined with the small sample size (N=15), these factors limit the strength of any conclusions about the effectiveness of the warnings.

Despite these limitations, the findings highlight a user preference for clear, informative warnings and point to the shortcomings of existing disclaimers. Moreover, the warning designs explored in this study serve as a foundation for better communicating AI limitation. However, future research should build on our preliminary findings by exploring a wider range of warning designs and other mechanisms that could affect trust in AI. Our results offer early insights and practical guidelines for developing AI systems that responsibly disclose their limitations.

References

[1] Matin Amoozadeh, David Daniels, Daye Nam, Aayush Kumar, Stella Chen, Michael Hilton, Sruti Srinivasa Ragavan, and Mohammad Amin Alipour. 2024. Trust in Generative AI among Students: An exploratory study. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1* (Portland, OR, USA) *(SIGCSE 2024)*. Association for Computing Machinery, New York, NY, USA, 67–73. doi:10.1145/3626252.3630842

[2] Florian Brühlmann and Elisa D Mekler. 2018. Surveys in games user research. *Games user research* (2018), 141–162.

[3] Ángel Alexander Cabrera, Adam Perer, and Jason I. Hong. 2023. Improving Human-AI Collaboration With Descriptions of AI Behavior. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 136 (April 2023), 21 pages. doi:10.1145/3579612

[4] Hyyryläinen Eetu. 2024. *Recognising Erroneous AI Generated References.* Master's thesis. Aalto University.

[5] Siddharth Gulati, Sonia Sousa, and David Lamas. 2019. Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology* 38, 10 (2019), 1004–1015.

[6] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* 5 (2023), 1096257.

[7] Lujain Ibrahim, Luc Rocher, and Ana Valdivia. 2024. Evaluating Harms from Design Patterns in AI Interfaces. *1st HEAL Workshop at CHI Conference on Human Factors in Computing Systems* (2024).

[8] Elaheh Jafari and Julita Vassileva. 2024. Designing Effective Warnings for Manipulative Designs in Mobile Applications. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (Cagliari, Italy) *(UMAP '24)*. Association for Computing Machinery, New York, NY, USA, 12–17. doi:10.1145/3627043.3659550

[9] Jeff Johnson. 2007. *GUI bloopers 2.0: common user interface design don'ts and dos.* Elsevier.

[10] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research methods in human-computer interaction.* Morgan Kaufmann.

[11] Florian Leiser, Sven Eckhardt, Merlin Knaeble, Alexander Maedche, Gerhard Schwabe, and Ali Sunyaev. 2023. From ChatGPT to FactGPT: A Participatory Design Study to Mitigate the Effects of Large Language Model Hallucinations on Users. In *Proceedings of Mensch Und Computer 2023* (Rapperswil, Switzerland) *(MuC '23).* Association for Computing Machinery, New York, NY, USA, 81–90. doi:10.1145/3603555.3603565

[12] Xin Sun, Yunjie Liu, Jan De Wit, Jos A. Bosch, and Zhuying Li. 2024. Trust by Interface: How Different User Interfaces Shape Human Trust in Health Information from Large Language Models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24).* Association for Computing Machinery, New York, NY, USA, Article 344, 7 pages. doi:10.1145/3613905.3650837

[13] Antonia Tolzin and Andreas Janson. 2025. Uncovering the mechanisms of common ground in human–agent interaction: review and future directions for conversational agent research. *Internet Research* (2025).