Citations and Trust in LLM Generated Responses

Yifan Ding[1], Matthew Facciani[1], Ellen Joyce[1], Amrit Poudel[1], Sanmitra Bhattacharya[2], Balaji Veeramani[2], Sal Aguinaga[2], Tim Weninger[1]

[1]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556
[2]AI Center for Excellence, Deloitte & Touche LLP, New York City, NY, 10112
{yding4, mfaccian, apoudel, ejoyce3, tweninge}@nd.edu, {saguinaga, bveeramani, sanmbhattachary}@deloitte.com

Abstract

Question answering systems are rapidly advancing, but their opaque nature may impact user trust. We explored trust through an anti-monitoring framework, where trust is predicted to be correlated with presence of citations and inversely related to checking citations. We tested this hypothesis with a live question-answering experiment that presented text responses generated using a commercial Chatbot along with varying citations (zero, one, or five), both relevant and random, and recorded if participants checked the citations and their self-reported trust in the generated responses. We found a significant increase in trust when citations were present, a result that held true even when the citations were random; we also found a significant decrease in trust when participants checked the citations. These results highlight the importance of citations in enhancing trust in AI-generated content.

Code — https://github.com/yifding/TrustCitationLLM
Datasets — https://https://osf.io/yqm8z/

Introduction

Large language models (LLMs) (Achiam et al. 2023; Touvron et al. 2023) stand at the forefront of contemporary artificial intelligence (AI), wielding immense potential to reshape the way humans interact with technology and information. However, a critical question persists: Are these LLMs and their by-products trusted by users? The answer holds profound implications for the future of AI in society. Trust (Baier 1986), a cornerstone of human relationships and societal functioning, plays an indispensable role in the sharing and acceptance of information.

Human trust dynamics are complex, often deeply rooted in social norms and interpersonal relationships. Humans navigate complex social structures by building trust through shared experiences, reputations, and accountability mechanisms (Muir 1987). Leading social theories, like the Principle of Social Proof, suggest that social conventions might predispose individuals to favor human sources over algorithmic sources (Cialdini 2009). This predicts that responses from an AI system might be more trusted when a human source corroborates its response. Conversely, responses from an AI system might be more trusted when the human touch is absent because AI systems lack anthropomorphic traits and social nuances (Miller 2019). In other words, AI systems, which are largely devoid of social motives, may increase trust among users who seek impartial and objective information (Sambrook 2012; Rosen 1999).

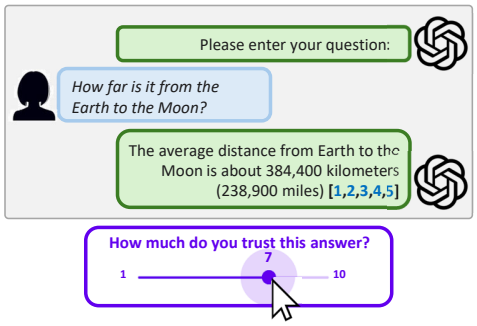

Figure 1: AI Chatbot system answering a user's question with five hyperlink citations. The presence of citations significantly increases the user's trust of the response.

Human trust in AI is a complicated subject that depends on several factors, with *accuracy* being the most important contributor (Lucassen and Schraagen 2011) and *explainability* (Rawal et al. 2021) following not far behind. Accuracy refers to the AI's ability to provide correct information consistently. Explainability entails the articulation of the model's decision-making pathways, rendering them transparent and comprehensible to users. Explainability is crucial for building a trust-based relationship between humans and AI (Stephanidis et al. 2019), significantly influencing users' willingness to engage with AI systems (Hoff and Bashir 2015).

Existing research on explainability primarily focuses on enabling AI engineers to understand model behavior. However, empirical investigations of explainability's role in user trust are limited, and the evidence is mixed regarding whether explainability indeed increases user trust (Scharowski et al. 2023; Nothdurft, Heinroth, and Minker 2013). Some studies report positive effects (Ehsan et al. 2019), while others find no significant impact (Poursabzi-Sangdeh et al. 2021; Zhang, Liao, and Bellamy 2020; Cheng et al. 2019). These mixed results

may stem from differences in how user trust is defined and measured: some studies rely on simple questionnaires, like in Fig. 1, while others assess user trust through behavior-related metrics (Poursabzi-Sangdeh et al. 2021). This distinction is important because, in studies, an increase in user trust does not necessarily translate to increased adoption or reliance on the AI (Papenmeier et al. 2022; Miller et al. 2016).

Because recent LLMs are exceedingly complex, the focus of explainability in AI has shifted from model interpretability—designing models so that their decision-making processes can be visualized or observed (Azaria and Mitchell 2023)—to *post-hoc explanations*, which denote how a specific decision was reached after it has been made (Lipton 2018). These post-hoc explanations typically take the form of feature importance and counterfactual explanations (Wachter, Mittelstadt, and Russell 2017). Feature importance explanations identify which variables most influenced a particular outcome, helping users understand the model's reasoning process (Zhao et al. 2024). Counterfactual explanations, on the other hand, describe how the model's output would change if certain inputs were different, providing insights into the decision-making logic (Mothilal, Sharma, and Tan 2020).

In LLMs systems, explainability is often conveyed via citations, which represent tangible evidence of the source of a system's knowledge and lends credibility to the response. By properly citing sources, these systems acknowledge the origins of their ideas and support their arguments with evidence. This practice not only respects the intellectual property of others but also enhances the overall quality and credibility of the response (Thornley et al. 2015).

The widespread adoption of LLMs has led to the development of Retrieval Augmented Generation (RAG) systems. These systems generate explanations by incorporating external information retrieved from various sources (Lewis et al. 2020). By extracting and integrating relevant external data, RAG systems oftentimes provide explicit citations in order to improve the transparency and reliability of their output and to encourage users to further explore the topic (Gao et al. 2023; Srinivas and Friedman 2024).

However, the role that citations play in shaping the trust relationship between AI and their users is not well understood. In the present study, we describe the results of a set of experiments that asks:

1. Do citations increase self-reported user trust in LLM-generated responses? If so, does the number of citations matter? and does the relevance of the citations matter?
2. Does the act of checking the citation decrease self-reported user trust?

Our study applies the social theories of *trust as anti-monitoring* and *the principle of social proof* to contextualize factors influencing user trust in LLMs. We hypothesized that (1) providing users with the sources behind LLM responses via citations (*i.e.*, social proof) would enhance trust in these otherwise opaque systems and that (2) providing users with the ability to check the citation would increase user trust; however, the act of actually checking citation (*i.e.*, monitoring the LLM) would be an indication of a lower level of user trust in the LLM's response.

To test these hypotheses, we deployed a bespoke QA Web site and invited participants to submit open-ended questions. LLM-generated responses were returned to the user with varying numbers of citations (zero, one, or five), which were either relevant to the answer or randomly selected from previous queries. Our analysis revealed a statistically significant increase in self-reported user trust when citations were included, a result that held true even when the citations were random. We also found a statistically significant decrease in self-reported user trust when participants checked the citations.

These findings underscore the critical role of citations in bolstering user trust in LLM-generated content. Moreover, they illuminate the nuanced interplay between citation relevance, question context, and user perceptions of trust.

Trust and Large Language Models

Trust in AI is a flourishing research area with diverse perspectives, including Trust as Anti-Monitoring and Social Proof Theory integrated into RAG systems and Chatbots. Scholars across these domains have investigated the dynamics of user trust with AI, offering insights into the mechanisms shaping human-machine interactions.

The first step when investigating users' trust in LLMs is to define trust. Trust is context-dependent, for example, a person may trust a mechanic to repair their car but not to prepare their taxes. Additionally, trust is dynamic and is built over a series of human interactions or events; each time the mechanic successfully repairs the car, the individual's trust increases, whereas a failure to fix it properly would decrease their trust.

In scenarios such as judicial decisions and wikis, citations are crucial in building trust, as they add credibility and transparency to content, regardless of the category of content. This raises an intriguing question: do individuals inherently trust LLM-generated responses more when they are accompanied by citations?

Trust as Anti-Monitoring How best to measure trust is a complicated and hotly debated topic (Baier 1986; Ferrario and Loi 2022). Trust, as Annette Baier's work suggests, can be construed as "anti-monitoring," where an indication of trusting an entity is a reduction in monitoring their behavior (Baier 1986; Archard et al. 2013). Monitoring, in the example above, refers to the intuition that if a customer trusts a mechanic, the customer is willing to allow the mechanic to repair the car without supervision. This implies that the level of monitoring someone performs is inversely related to the level of trust they have in the person or thing being monitored. The concept of trust as anti-monitoring offers a measurable framework for understanding trust. Citations serve as a monitoring mechanism, allowing people to check the LLM's response and determine if it aligns with the user's expectations and reasoning. This ties into the social-psychological Principle of Social Proof, where individuals look to external cues and validations to form trust.

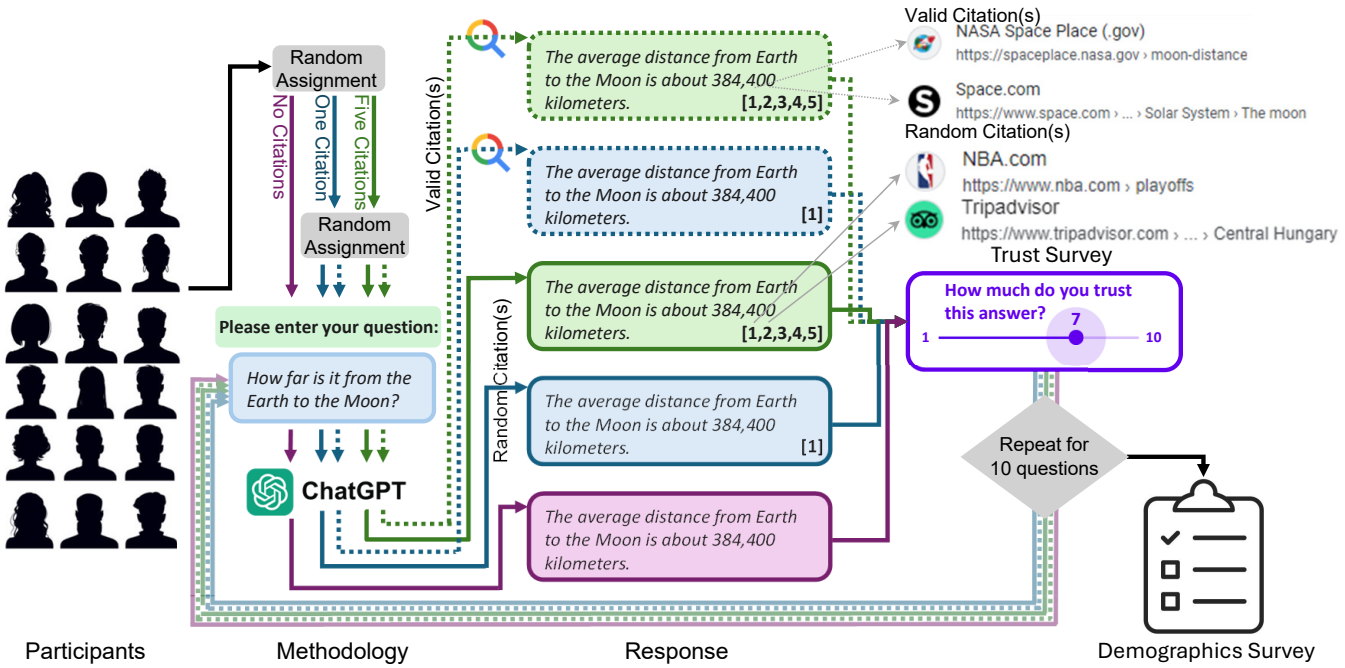

Figure 2: Methodology of the Citation Trust Experiment. Participants are assigned to zero (purple), one (blue), or five (green) citations, which can be either valid (dotted-line) or random (solid-line). A participant may ask any question, and then rates the response on a scale of 1 to 10. This is repeated for ten total questions and a demographics survey is asked at the end.

Principle of Social Proof The Principle of Social Proof is particularly useful for understanding interactions with Chatbots. This framework suggests that people are more likely to adopt a behavior if they see the social proof of others doing the same (Cialdini 2009). Social proof can, therefore, act as a proxy for trustworthiness, as individuals are more likely to use, and in turn trust, a product when they observe others using it (Lins and Sunyaev 2023; Kim, Choi, and Fotso 2024; Venkatesh and Davis 2000).

In Chatbot interactions, citations within the outputs serve as strong indicators of social proof because they signal to users that the output is endorsed by some source. Consequently, the presence of citations theoretically enhances trust, because users perceive the content as more credible and reliable.

The Principle of Social Proof aligns with the anti-monitoring framework of trust, where having the ability to monitor or check a response positively contributes to the trustworthiness of a response, regardless of whether or not the user chooses to check the source. However, it remains uncertain whether high-quality citations significantly impact trust, or if *any citation*, regardless of quality, is sufficient.

RAG and Chatbots RAG is an AI framework that incorporates information retrieved from external sources into the generation process (Asai et al. 2023, 2024; Ding, Zeng, and Weninger 2024; Ding et al. 2024). This approach tends to reduce inaccurate responses(Lewis et al. 2020) and includes citations within its generated responses, providing users with a clear trail of the sources that support the presented information.

User trust dynamics in LLMs is beginning to receive some much-needed attention and research has uncovered factors that influence users' perceptions and behaviors (Sun et al. 2022). For example, users are more likely to engage with chatbots they perceive as trustworthy (Choudhury and Shamszare 2023). However, concerns about government use of chatbots can lead to distrust (Aoki 2020). Despite occasional inaccuracy and unreliability in current chatbot versions, users often express intentions to continue using them, indicating a resilient trust in these systems (Amaro et al. 2023). Research also suggests that users tend to trust chatbots with more human-like characteristics (Kaplan et al. 2023), highlighting the interplay between trust, utility, perceived reliability and accuracy, and the humanization of AI in shaping user engagement.

Citations and Trust Experiment

We developed a bespoke Web site for data collection. On this Web site, users were introduced to the task with an animation that showed the example question: "How far is it from the earth to the moon?". If the participants agreed to participate, they were provided a simple query box, stylized to look like a standard input Web form. On this form users were prompted by instruction-text to: "Ask any question".

The participant's responses were stored on a Web server owned and managed by research team. Each question was then fed directly to ChatGPT4 and the responses were collected. Responses were truncated if they were longer than

three sentences.

The experiment was a Randomized Controlled Trial (RCT) with a between-subjects 3 by 2 factorial design (see Fig. 2). The first factor corresponded to the number of citations: zero, one, or five; the second factor corresponded to the nature of the citations: valid or random.

For the first factor: in the no citation condition, the response was taken directly from the output of ChatGPT, truncated to three sentences if necessary, and provided to the participant. In the non-zero citation conditions, we redirected the truncated response from ChatGPT to a Web search API[1], which queried the Google search engine for Web sites relevant to the response; the first five search engine results were recorded. This was invisible to the participant; however, there was a small response delay in this condition. In the one-citation condition, the top citation was provided to participant as a numeral (*e.g.*, [1]). In the five-citation condition, all five citations were provided to the participant as a list of numerals (*e.g.*, [1,2,3,4,5]). Each citation was programmed to show the URL of the citation if the participant hovered their mouse over the numeral (see Supplement C Figures S1 and S2).

For the second factor: in the valid citation condition, the search engine result(s) were provided directly to the user. In the random citation condition, the actual citations were recorded, but the citation URL(s) shown to the participant were randomly selected from citations of previous participant's questions.

Participants

We used Prolific[2] to recruit participants for this study (Palan and Schitter 2018). Data collection occurred on March 13, 2024. Participants were paid two US dollars and took a median of 17 minutes to complete. The study had 303 total participants who were randomly assigned to the experimental groups (*i.e.*, between-subjects design). Participants either saw zero (N=108), one (N=96), or five (N=101) citations. Of the two groups who saw citations (N=197), a random split of the participants received a random citation (N=87) or valid citation (N=110).

Participants were asked to enter ten questions and rate each response; finally an exit interview was conducted with a battery of demographic questions (see Supplement A and Table S1 for demographic battery and response codes).

Data, Materials, and Software Availability

All participant questions, their responses and, their ratings are available in an Excel file. This file is publicly available online at https://osf.io/yqm8z/ We used Stata Software for data and statistical analysis.

Results

Do Citations Increase User Trust?

We expect that the presence of citations in an AI chatbot's output should enhance response transparency and should improve perceived trustworthiness. The Principle of Social Proof suggests that observing evidence of others endorsing a behavior increases the likelihood of adopting that behavior ourselves. We also predict that the quality of the citations matters. Citations that accurately support the chatbot's answer will be evaluated as more trustworthy than random citations.

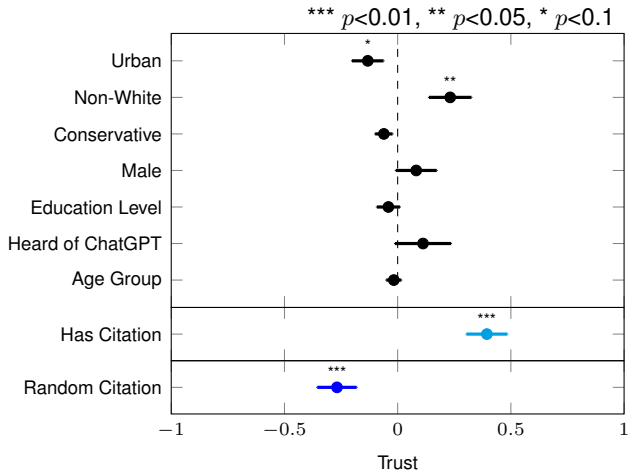

Figure 3: Citations increase perceived trustworthiness, but random citations decrease perceived trustworthiness. Regression coefficients β and their standard errors are plot on the x-axis.

In out initial analysis, the Dependent Variables (DVs) were citation and no citation coded 1 or 0 respectively. Random and Accurate citation were coded 1 or 0 respectively. Controlling for various demographic factors, results of linear regression analyses indicate a statistically significant increase in perceived trustworthiness for AI chatbot responses with citations compared to those without (See Fig. 3 and Supplement B Table S2).

Does the Quality of Citation Matter?

Additional analysis examined the influence of random versus valid citations on trustworthiness. Again, controlling for demographic factors, results revealed that answers containing random citations were significantly less trustworthy (See Fig. 3 and Supplement B Table S2).

Do the Number of Citations Matter?

Using an analysis of variance (ANOVA), we examined whether perceived trustworthiness varied among the zero citation, one citation, and five citation conditions. The analysis revealed significant differences between groups ($F(2, 3037) = 10.23, p < .001$). Post-hoc Bonferroni (Bonferroni 1936) tests indicated that both the one citation and five citation conditions were rated significantly higher on trustworthiness compared to the zero citation condition (see Supplement B Table S3).

Notably, there was no significant difference between the one citation and five citation conditions ($p > .05$). In other words, five citations in an answer are not perceived as more

[1]https://scaleserp.com

[2]prolific.com

trustworthy than an answer with one citation. This negative result is contrary to our initial hypothesis. One plausible reason for this result might be due to the principle of diminishing returns, wherein participants may perceive that a single, well-chosen citation is sufficient to confirm the AI's response.

Social Demographics and Trust

In addition to the directional hypotheses, we also explored how various social demographics predict the perceived trustworthiness of AI chatbot answers. While there is limited research specifically on how different groups respond to AI chatbots, some studies investigate how different demographics react to new technologies such as AI (Tyson and Kikuchi 2023; Rainie et al. 2022). For instance, individuals with more conservative values tend to be more skeptical toward AI and new technologies (Castelo and Ward 2021). However, both individuals with liberal and conservative views are more receptive to technology when it is framed in a way that aligns with their political values (Claudy, Parkinson, and Aquino 2024). Furthermore, higher levels of education and familiarity with chatbots may increase perceived trustworthiness; on the other hand, greater awareness of AI limitations, which often accompany higher education and familiarity, could decrease trustworthiness.

We did not find significant differences in trustworthiness ratings among most demographic categories, the slight inclination of nonwhite participants to trust the answers more suggests avenues for exploratory research. For instance, future studies could look further into the underlying factors driving this trend and explore potential cultural or societal influences on trust perceptions in AI-generated content.

Does Checking Citations Indicate a Reduction in User Trust?

The theory of trust as anti-monitoring predicts that individuals who are skeptical of an answer will be more likely to check the source of the citation. This predicts that checking a citation indicates reduced trust, as users are no longer relinquishing control and trusting the other party to be accurate. We tracked the frequency of participants that manually checked citations with their mouse while reviewing the AI chatbot's answers. There were 1,976 answers in our dataset that had at least one reference. Of these, only 193 (9.77%) were manually checked. 83 participants out of 197 participants (42.1%) in the citation groups checked at least one citation while completing the experiment. Interestingly, participants in the five citation group were significantly more likely to do a citation check (χ^2= 21.19; p<0.001).

We coded Check Citation as 0 or 1 and controlled for all demographic variables, as well as the presence of citations and random citations. Trust was included as our main independent variable (IV). Linear regression, controlling for demographics as well as the presence of citations and random citations, was performed. The analysis (see Fig. 4 and Supplement C Table S4) revealed a significant correlation between increased citation checks and lower trust ratings for the answers. These findings lend support to the trust as anti-

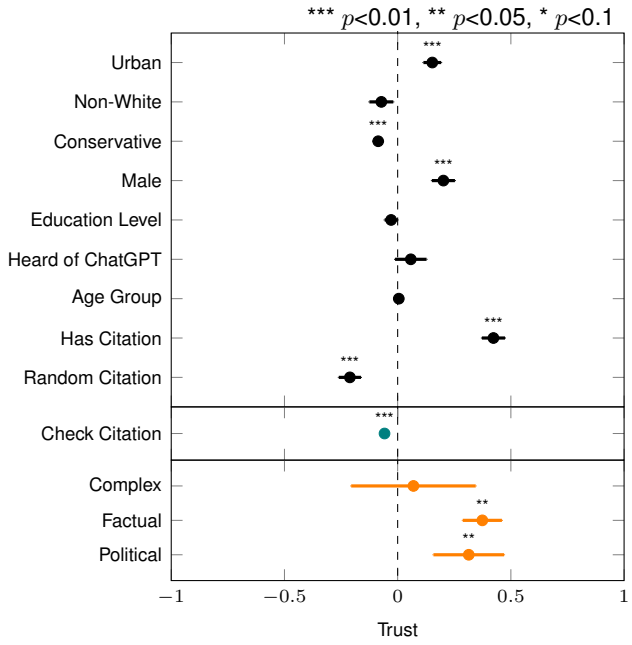

Figure 4: Checking citations decrease perceived trust. Political and factual questions have a higher perceived trust. Regression coefficients β and their standard errors are plot on the x-axis.

monitoring theory: a higher frequency of citation checking predicts lower trustworthiness.

These results raise the broader question of whether *any* citations, even if they were random, would yield higher trust ratings than zero citations. We compared the no citation trust ratings with the trust ratings of random citations for questions that were checked. We did not find a significant difference between answers that had zero citations and answers that had random citations that were checked (T=-0.877; p=0.38). There were only 193 checked-questions to analyze, but we did see a lower mean trust rating in the checked random citation answers (M=7.55; SD=2.54) compared to the no citation answers (M=7.73; SD=2.46). In other words, answers with random citations that were manually checked were no more trusted than answers with no citations at all.

We also found significant differences based on gender, political orientation, and residence. Specifically, males, those indicating liberal political orientation, and individuals residing in urban areas were significantly more likely to check the citations ($p < .001$).

Illustrating Question Semantics

The data we collected includes hundreds of interesting and unique real-world, human-generated questions, along with trust ratings for their answers.

This data permitted an exploration of the types of questions asked and if they affected trustworthiness. Political information is particularly susceptible to bias (Ditto et al. 2019; Poudel and Weninger 2024), and we believe that questions of a political nature may vary in their perceived trust-

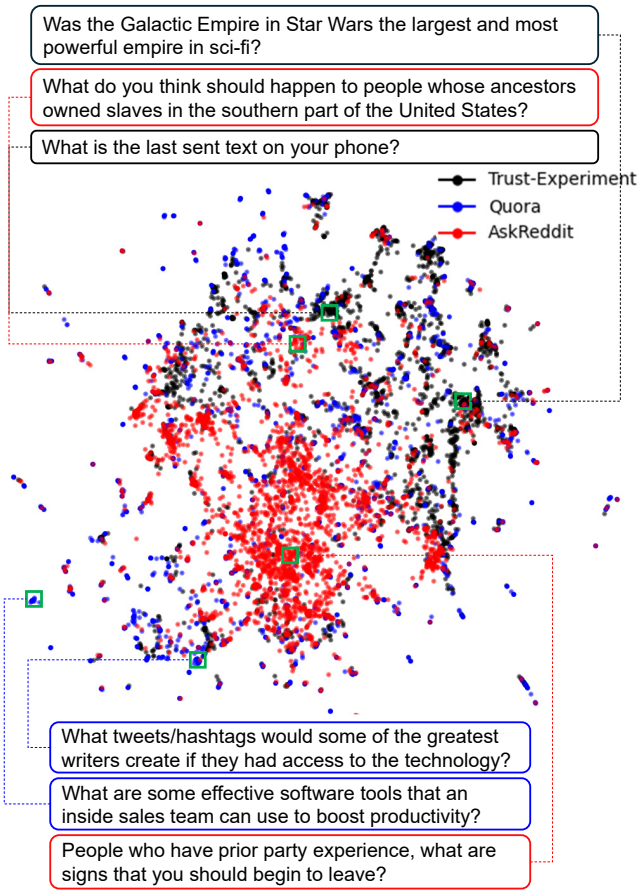

Figure 5: Visualization of question-topics asked by participants. Although we observe a substantial overlap in the kinds of questions asked by our participants (black) compared to AskReddit (red) and Quora (blue), we also identify several topical gaps. Some representative samples of these topical gaps are illustrated in on right. An interactive visualization of this figure is included in the supplementary material.

worthiness. Individuals might be more likely to ask questions that confirm their pre-existing political biases, which could increase the perceived trustworthiness of the answer if it aligns with their views. Conversely, if the chatbot's answer contradicts their beliefs, the perceived trustworthiness of the response may decrease.

We analyzed the semantics of the questions asked by participants using a comparative approach by juxtaposing them with questions from established question-answering datasets such as AskReddit and Quora. Using sentence transformers, we embedded each question into a high-dimensional vector space, capturing their semantic representations (Reimers and Gurevych 2019). Subsequently, we used UMAP, a dimensionality reduction technique, to project these question embeddings onto a 2-dimensional plot (McInnes, Healy, and Melville 2018), as depicted in Fig. 5. In this plot, participant's questions are depicted in black. An interactive plot can be found in the Supplement.

Does the Type of Question Impact User Trust?

Using linear regression, and accounting for demographics, citations, and random citations, we examined the impact of question type on user trust ratings. We then used ChatGPT4 to label each question as being Political, Factual, and Complex questions as 0 or 1; these labels need not be disjoint (*i.e.*, a question can be both factual and political).

Our results indicate that political questions received significantly higher self-reported user trust ratings compared to non-political questions , even after adjusting for demographic factors. Manual analysis of the questions and answers suggests that this effect is not due to the chatbot's awareness of the participant's political stance, but rather because political questions are often framed in a way that aligns with users' preconceived notions, thereby eliciting more favorable responses. These findings support the theory of social proof, where the chatbot's responses act as a form of social endorsement, enhancing perceived trustworthiness.

We found that fact-based questions were significantly more trusted than non-fact-based questions. This aligns with the previous result that political questions (which may also be factual) are also trusted, but it highlights different aspects of trustworthiness in AI chatbot responses. As previously discussed, political questions tend to be framed in a way that aligns with the participant's pre-existing beliefs and in-group biases, leading to higher trust ratings. Furthermore, fact-based questions are typically grounded in objective, verifiable information. Participants can cross-check these facts against known data or their existing knowledge, leading to higher trust ratings. These findings suggest that trust in AI chatbot responses can be driven by both the objective accuracy of the information (fact-based questions) and the social alignment with the participant's beliefs (political questions). While fact-based questions benefit from their verifiability, political questions benefit from framing and ingroup validation.

Complexity and Trust

Previous studies suggest that LLMs provide more accurate responses when prompted with more familiar language (Gonen et al. 2022). For example, asking "Who was the first president of the United States?" is clear and straightforward, leading to an accurate response. In contrast, a more complex question like "How have the economic policies of U.S. presidents influenced income inequality in the United States?" is less direct, causing the AI to infer more, which may reduce accuracy. This suggests that simpler, familiar prompts yield more accurate and trusted responses from AI systems.

Language perplexity was used as a proxy for the model's familiarity with the question, where perplexity was measured as $\text{PPL}(x) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log_2 P(w_i)\right)$ where N is the total number of words in the response, w_i is the i^{th} word in the response, and $P(w_i)$ is the is the probability of the i^{th} word given in ChatGPT4. In this context, lower perplexity indicates a more familiar prompt, leading to more accurate (perhaps more trusted) results. Guided by these

prior findings, we compared user-reported trust as a function of the perplexity of the prompt. We found that higher perplexity is slightly (negatively) correlated with trust (Pearson R=-0.06, p=0.002), *i.e.*, answers to complex questions are (slightly) less trusted.

In a similar exploratory analysis, we also found that the length of a prompt (number of tokens) is (slightly) negatively correlated with trust (Pearson R=-0.04, p=0.041). In other words, although we do not find statistical differences in trust between simple and complex questions, we did find that responses to simpler questions (in terms of perplexity and length) were more trusted. This suggests that users may have a higher level of trust in chatbots when the prompts are simpler, potentially indicating a preference for straightforward and concise queries that yield more understandable answers.

Discussion

The present study investigated how variations in citations influenced the perceived trustworthiness of answers provided by an AI chatbot. Drawing upon the anti-monitoring framework, we conceptualized trust and extended this framework to incorporate the Principle of Social Proof. We hypothesized that participants would trust AI chatbot responses with citations more than those without citations, as citations provided the opportunity for verification (or monitoring) of the output. Moreover, these citations, linked to supporting organizations, served as a form of social proof, enhancing trustworthiness.

Our findings supported this hypothesis, revealing that AI chatbot answers with citations were perceived as more trustworthy compared to those without citations. Furthermore, we investigated the significance of the number of citations, finding no significant difference in perceived trustworthiness between responses with one or five citations.

Beyond the number of citations, we also investigated those participants who manually inspected the citations. We found that responses containing random citations were rated lower in trustworthiness compared to those with accurate citations.

Furthermore, we explored whether a higher frequency of citation-checks, indicated by mouse hovers, correlated with lower perceived trustworthiness. Consistently, participants who checked citations tended to rate the answers as less trustworthy. This finding aligns with the theory of trust as anti-monitoring, as skeptical participants sought to verify (*i.e.*, monitor) the source of the information provided by the chatbot. Next, we investigated whether the type of question asked was associated with the perceived trustworthiness of the answers. We categorized questions into three groups: political, factual, and complex.

Our analysis revealed that political questions were rated significantly more trustworthy than non-political questions. This trend may be attributed to participants posing political questions that already aligned with their political biases, leading them to be more inclined to trust the answers. We also found that fact-based questions were significantly more trusted than others. This difference could be due to the concrete nature of factual questions, instilling confidence in participants that they already know the correct answer, whereas non-factual questions were more subjective.

Regarding complex questions, we hypothesized that they might elicit greater trust as they could potentially demonstrate the chatbot's capability to handle challenging inquiries. However, we did not observe a significant difference in trust ratings between complex problem-solving questions compared to those categorized as more straightforward. Given the limited number of questions coded as complex, we cannot assert the absence of an effect with confidence.

Along the way, we evaluated if demographic variables predicted trustworthiness. We only found that nonwhite participants were slightly more likely to trust the answers. We also found that males, individuals with liberal views, and people in urban areas were significantly more likely to check the citations (*i.e.*, mouse-hover over the references) given in the chatbot answer. Given our small sample size, we are hesitant to read too much into these results, but it may be an area for future research.

Limitations. Our study is not without several important limitations. First, our sample was comprised of participants entirely from the online data collection platform Prolific. These participants may be more technologically savvy than the typical individual who does not sign up for online research surveys. Additionally, since our sample was 65% white, we did not have sufficient statistical power to evaluate different racial groups and combined them into a simple nonwhite category. While this nonwhite category trusted their chatbot answers more than the white category, future research will have to investigate what could have caused this difference or if it was an artifact. Our study also did not control for what questions were asked so it could be that people of different demographic groups may ask the chatbot different questions, which could alter their trustworthiness. Finally, our measure of trustworthiness was a simple, one-item variable. It's possible that different forms of trustworthiness may yield interesting results. Future research can incorporate more robust measures of trustworthiness to assess how different questions and outputs influence different elements of trust.

Ethics Statement

This study received approval from the University of Notre Dame Institutional Review Board (protocol no. 23-06-7934). Participants were fully briefed on the study's purpose, ensuring informed consent and voluntary participation. Aside from broad demographic questions, personal identifiable information was not collected.

Our study aims to understand and measure trust and inform best practices in the incorporation of citations into AI systems. However, potential risks include misinterpretation of results leading to over-reliance on citations and privacy concerns related to participant data. We are committed to addressing these issues by adhering to ethical standards, ensuring transparency, and carefully evaluating the broader impact of our work.

Acknowledgements

This project was funded in part by DARPA under contract HR001121C0168 and HR00112290106. We would like to thank the anonymous reviewers for their valuable comments.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Amaro, I.; Della Greca, A.; Francese, R.; Tortora, G.; and Tucci, C. 2023. AI unreliable answers: A case study on ChatGPT. In *International Conference on Human-Computer Interaction*, 23–40. Springer.

Aoki, N. 2020. An experimental study of public trust in AI chatbots in the public sector. *Government information quarterly*, 37(4): 101490.

Archard, D.; Deveaux, M.; Manson, N. A.; and Weinstock, D. M. 2013. *Reading Onora O'Neill*. Routledge London/New York, NY.

Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Asai, A.; Zhong, Z.; Chen, D.; Koh, P. W.; Zettlemoyer, L.; Hajishirzi, H.; and Yih, W.-t. 2024. Reliable, adaptable, and attributable language models with retrieval. *arXiv preprint arXiv:2403.03187*.

Azaria, A.; and Mitchell, T. 2023. The internal state of an llm knows when its lying. *arXiv preprint arXiv:2304.13734*.

Baier, A. 1986. Trust and Antitrust. *Ethics*, 96(2): 231–260.

Bonferroni, C. 1936. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R istituto superiore di scienze economiche e commericiali di firenze*, 8: 3–62.

Castelo, N.; and Ward, A. F. 2021. Conservatism predicts aversion to consequential Artificial Intelligence. *Plos one*, 16(12): e0261467.

Cheng, H.-F.; Wang, R.; Zhang, Z.; O'connell, F.; Gray, T.; Harper, F. M.; and Zhu, H. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*, 1–12.

Choudhury, A.; and Shamszare, H. 2023. Investigating the impact of user trust on the adoption and use of ChatGPT: Survey analysis. *Journal of Medical Internet Research*, 25: e47184.

Cialdini, R. 2009. Social proof: Truths are us. *Influence: Science and practice*, 97–140.

Claudy, M. C.; Parkinson, M.; and Aquino, K. 2024. Why should innovators care about morality? Political ideology, moral foundations, and the acceptance of technological innovations. *Technological Forecasting and Social Change*, 203: 123384.

Ding, Y.; Poudel, A.; Zeng, Q.; Weninger, T.; Veeramani, B.; and Bhattacharya, S. 2024. EntGPT: Linking Generative Large Language Models with Knowledge Bases. *arXiv preprint arXiv:2402.06738*.

Ding, Y.; Zeng, Q.; and Weninger, T. 2024. ChatEL: Entity Linking with Chatbots. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 3086–3097. Torino, Italia: ELRA and ICCL.

Ditto, P. H.; Liu, B. S.; Clark, C. J.; Wojcik, S. P.; Chen, E. E.; Grady, R. H.; Celniker, J. B.; and Zinger, J. F. 2019. At least bias is bipartisan: A meta-analytic comparison of partisan bias in liberals and conservatives. *Perspectives on Psychological Science*, 14(2): 273–291.

Ehsan, U.; Tambwekar, P.; Chan, L.; Harrison, B.; and Riedl, M. O. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th international conference on intelligent user interfaces*, 263–274.

Ferrario, A.; and Loi, M. 2022. How Explainability Contributes to Trust in AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1457–1466. Seoul Republic of Korea: ACM. ISBN 978-1-4503-9352-2.

Gao, T.; Yen, H.; Yu, J.; and Chen, D. 2023. Enabling Large Language Models to Generate Text with Citations. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6465–6488. Singapore: Association for Computational Linguistics.

Gonen, H.; Iyer, S.; Blevins, T.; Smith, N. A.; and Zettlemoyer, L. 2022. Demystifying prompts in language models via perplexity estimation. *arXiv preprint arXiv:2212.04037*.

Hoff, K. A.; and Bashir, M. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3): 407–434.

Kaplan, A. D.; Kessler, T. T.; Brill, J. C.; and Hancock, P. A. 2023. Trust in artificial intelligence: Meta-analytic findings. *Human factors*, 65(2): 337–359.

Kim, Y. J.; Choi, J. H.; and Fotso, G. M. N. 2024. Medical professionals' adoption of AI-based medical devices: UTAUT model with trust mediation. *Journal of Open Innovation: Technology, Market, and Complexity*, 10(1): 100220.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Lins, S.; and Sunyaev, A. 2023. Advancing the presentation of IS certifications: theory-driven guidelines for designing peripheral cues to increase users' trust perceptions. *Behaviour & Information Technology*, 42(13): 2255–2278.

Lipton, Z. C. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3): 31–57.

Lucassen, T.; and Schraagen, J. M. 2011. Factual accuracy and trust in information: The role of expertise. *Journal of the American Society for Information Science and Technology*, 62(7): 1232–1242.

McInnes, L.; Healy, J.; and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Miller, D.; Johns, M.; Mok, B.; Gowda, N.; Sirkin, D.; Lee, K.; and Ju, W. 2016. Behavioral measurement of trust in automation: the trust fall. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 60, 1849–1853. SAGE Publications Sage CA: Los Angeles, CA.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267: 1–38.

Mothilal, R. K.; Sharma, A.; and Tan, C. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 607–617.

Muir, B. M. 1987. Trust between humans and machines, and the design of decision aids. *International journal of man-machine studies*, 27(5-6): 527–539.

Nothdurft, F.; Heinroth, T.; and Minker, W. 2013. The impact of explanation dialogues on human-computer trust. In *Human-Computer Interaction. Users and Contexts of Use: 15th International Conference, HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part III 15*, 59–67. Springer.

Palan, S.; and Schitter, C. 2018. Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17: 22–27.

Papenmeier, A.; Kern, D.; Englebienne, G.; and Seifert, C. 2022. It's complicated: The relationship between user trust, model accuracy and explanations in ai. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 29(4): 1–33.

Poudel, A.; and Weninger, T. 2024. Navigating the Post-API Dilemma. In *Proceedings of the ACM on Web Conference 2024*, 2476–2484.

Poursabzi-Sangdeh, F.; Goldstein, D. G.; Hofman, J. M.; Wortman Vaughan, J. W.; and Wallach, H. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–52.

Rainie, L.; Funk, C.; Anderson, M.; and Tyson, A. 2022. How Americans think about artificial intelligence. *Pew Research Center: Washington, DC, USA*.

Rawal, A.; McCoy, J.; Rawat, D. B.; Sadler, B. M.; and Amant, R. S. 2021. Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives. *IEEE Transactions on Artificial Intelligence*, 3(6): 852–866.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rosen, J. 1999. *What are journalists for?* Yale University Press.

Sambrook, R. 2012. *Delivering trust: Impartiality and objectivity in the digital age*. Reuters Institute for the Study of Journalism.

Scharowski, N.; Perrig, S. A.; Svab, M.; Opwis, K.; and Brühlmann, F. 2023. Exploring the effects of human-centered AI explanations on trust and reliance. *Frontiers in Computer Science*, 5: 1151150.

Srinivas, A.; and Friedman, L. 2024. Aravind Srinivas: Perplexity CEO on Future of AI, Search & the Internet Lex Fridman Podcast #434. https://www.youtube.com/watch?v=e-gwvmhyU7A.

Stephanidis, C.; Salvendy, G.; Antona, M.; Chen, J. Y.; Dong, J.; Duffy, V. G.; Fang, X.; Fidopiastis, C.; Fragomeni, G.; Fu, L. P.; et al. 2019. Seven HCI grand challenges. *International Journal of Human–Computer Interaction*, 35(14): 1229–1269.

Sun, Z.; Wang, X.; Tay, Y.; Yang, Y.; and Zhou, D. 2022. Recitation-augmented language models. *arXiv preprint arXiv:2210.01296*.

Thornley, C.; Watkinson, A.; Nicholas, D.; Volentine, R.; Jamali, H. R.; Herman, E.; Allard, S.; Levine, K.; and Tenopir, C. 2015. The role of trust and authority in the citation behaviour of researchers. *Information research*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tyson, A.; and Kikuchi, E. 2023. Growing public concern about the role of artificial intelligence in daily life. Technical report.

Venkatesh, V.; and Davis, F. D. 2000. A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management science*, 46(2): 186–204.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31: 841.

Zhang, Y.; Liao, Q. V.; and Bellamy, R. K. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 295–305.

Zhao, H.; Chen, H.; Yang, F.; Liu, N.; Deng, H.; Cai, H.; Wang, S.; Yin, D.; and Du, M. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2): 1–38.