

# Verbalized Uncertainty and Citation Match in LLM Question Answering: A Pilot Study on Adoption and Verification

Minjoon Sohn  
Department of Industrial  
Engineering  
Seoul National University  
Seoul, Korea  
mjsohn@snu.ac.kr

San Hong  
Department of Industrial  
Engineering  
Seoul National University  
Seoul, Korea  
saanhong@snu.ac.kr

Woojin Park  
Department of Industrial  
Engineering  
Seoul National University  
Seoul, Korea  
woojinpark@snu.ac.kr

## ABSTRACT

Large Language Model (LLM) such as ChatGPT are increasingly used as one-stop tools for everyday information seeking, but their reliance risks include over-adoption of hallucinated or unsupported answers. Prior research has shown that citations can elevate trust, sometimes excessively, while verbalized uncertainty may encourage users to interpret responses more cautiously. Yet little is known about how these cues interact when presented together in LLM-based question answering (QA). This pilot study examined how verbalized uncertainty and citation match influence user adoption and verification behavior. Six participants (Mean = 27.8 years, SD = 2.2) completed quiz-style QA tasks in a  $2 \times 2$  within-subjects design that manipulated verbalized uncertainty (with vs. without) and citation match (match vs. mismatch). Results showed that verbalized uncertainty increased citation checking and, when verification occurred, adoption aligned with citation match (higher for matched than mismatched sources). Moreover, verbalized uncertainty reduced adoption when citations were mismatched while having little effect when they were matched. These findings highlight the double-edged nature of citations and suggest that verbalized uncertainty can serve as a lightweight design cue for promoting appropriate reliance. The study offers preliminary insights into human-automation trust calibration and provides early guidance for designing LLM interfaces that balance efficiency with verification.

## KEYWORDS

Large Language Models, Question Answering, User Adoption (Reliance), Verbalized Uncertainty, Citation Relevance

## 1. Introduction

Large Language Model (LLM) chatbots (e.g., ChatGPT (OpenAI)) have rapidly entered mainstream use, and some commentators have argued that such tools could even replace

traditional search engines [1]. In fact, generative AI systems like ChatGPT are emerging as a one-stop alternative for everyday information needs, with evidence showing that ChatGPT users are as likely as Google users to reach correct answers while completing tasks, and that they do so at a faster pace [2].

Accordingly, research on LLMs such as ChatGPT has been rapidly expanding, particularly in the context of question answering (QA) and information seeking. Prior studies have compared LLMs with traditional search engines or retrieval-augmented systems to evaluate their effectiveness in both domain-specific contexts such as healthcare [3] and in everyday search tasks [2, 4]. At the same time, users are increasingly turning to LLM-based QA for everyday information needs, a trend attributed to their advantages in delivering quick summaries and single-answer responses [5].

Alongside these advantages, users also tend to place considerable trust in LLM responses. Prior research has identified two main mechanisms underlying this tendency. First, people often accept fluent and readily available answers, sometimes prioritizing convenience over scrutiny [5]. Second, conversational interfaces can evoke anthropomorphic impressions, prompting users to attribute human-like qualities such as empathy or reasoning, which in turn fosters unwarranted trust [6, 7]. Another increasingly emphasized feature in contemporary AI systems is the use of citation and reference cues, which act as provenance signals and peripheral “social proof” that lead users to treat the answer as evidence-backed. Empirical findings show that the presence of citations tends to elevate perceived trust [8]. Accordingly, providing citations in chatbot responses has been highlighted as a powerful design strategy for enhancing user trust [9].

Providing citations or references in LLM outputs is a representative strategy for enhancing explainability, and provenance-based explanations have been widely discussed in the field of Explainable AI (XAI) as an important means to improve trust and transparency [10]. The influence of citation on user adoption and trust can be distinguished in two ways.

(A) Presence of citations: Simply displaying a citation can elevate trust, even when the source is irrelevant, highlighting the strong impact of interface cues [8, 9].

(B) Relevance of citations: Trust holds only when citations genuinely support the answer. Yet about 25% of ChatGPT citations fail to do so, and users who verify such mismatched sources report sharply lower trust [8, 11].

At the same time, however, while LLM-based QA can improve efficiency in real-world decision-making contexts, it also carries the risk of overconfidence and over-adoption of incorrect answers [12]. Although LLMs generate fluent responses, they cannot guarantee factual accuracy and frequently produce plausible but incorrect statements, often referred to as hallucinations [13, 14]. The technical architecture of LLMs is designed to predict the most likely next token, without reliably encoding whether a statement is factually true [7, 15]. This explains why LLMs often produce convincing yet inaccurate information and suggests that such hallucinations are not merely temporary issues, but inherent limitations likely to persist in future systems.

LLM hallucinations are not only persistent but also difficult for users to recognize in practice. Users often find it difficult to detect hallucinations in LLM outputs. As a result, the concept of appropriate reliance—long emphasized in the tradition of human–automation interaction research [16]—provides an important theoretical guideline for the design of LLM-based QA systems. In parallel, researchers have underscored the need for practical mitigation strategies, such as strengthening AI self-verification mechanisms or providing user training, to prevent breakdowns of trust in critical domains like science and medicine [17].

One promising approach to fostering appropriate reliance is to design interfaces that actively encourage verification. To this end, a representative strategy is verbalized uncertainty. Verbalized uncertainty, commonly referred to in linguistics as hedging [18], denotes linguistic devices that reduce the degree of certainty or strength of a statement. For example, expressions such as “I cannot say for sure” or “it may be the case” explicitly signal the limits of an AI system’s answer. Prior research suggests that hedging can sometimes undermine perceived credibility, but in other contexts it may create an impression of being “objective and cautious,” thereby positively shaping user perceptions [19].

In sum, the reliability and adoption of LLM-generated responses are strongly influenced by two key cues: citations and verbalized uncertainty. In current LLM interactions, users are simultaneously exposed to citations, which can enhance trust, and verbalized uncertainty, which can promote verification. Citations, however, function as a double-edged sword: users often perceive answers as evidence-based simply because references are present [8], yet active verification is cognitively demanding and may undermine the convenience that makes LLMs attractive in the first place. Within this setting, verbalized uncertainty offers a potential design cue that can naturally prompt users to reconsider blind reliance and engage in verification when needed. However, prior

research has largely examined these cues in isolation, highlighting a gap in understanding how their interaction influences user adoption behavior.

This highlights a critical research gap. In LLM outputs, citations and verbalized uncertainty often appear together, and their interplay is central to fostering appropriate reliance. However, experimental investigations of how these cues jointly affect user adoption remain limited.

Accordingly, this study empirically examines how verbalized uncertainty, the match between citations and answers, and their interaction influence users’ adoption of LLM responses and their citation-checking behavior in a QA context. By doing so, it aims to contribute to theoretical understanding of appropriate reliance and provide practical implications for the design of LLM-based systems.

To examine these effects in detail, we formulated the following hypotheses:

H1: (Verification Effect): When verbalized uncertainty is presented, participants will be more likely to engage in citation checking.

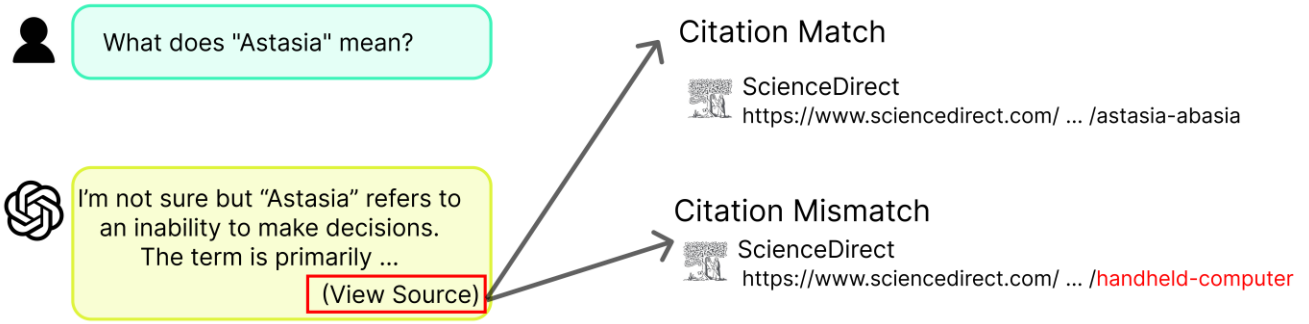
H2: (Adoption Adjustment Effect): When verification occurred, participants’ adoption rates were modulated by the citation match (matching vs. mismatching).

H3: (Interaction Effect): Verbalized uncertainty and citation authenticity will interact. Specifically, when citations are mismatched, the presence of verbalized uncertainty will reduce adoption of incorrect information.

## 2. Method

A total of six adults (5 male, 1 female) participated in this pilot study. All reported no health conditions that could interfere with task performance. Their mean age was 27.8 years ( $SD = 2.2$ ). Data collection was conducted in a controlled laboratory setting, where participants completed the experiment on laptop computers while interacting with an AI agent implemented using the custom GPT functionality of ChatGPT.

The experiment employed a  $2 \times 2$  within-subjects design with two independent variables: (1) verbalized uncertainty (with vs. without) and (2) citation match (matching vs. mismatching). In the verbalized uncertainty condition, the LLM’s answer included a brief linguistic disclaimer which was “I’m not sure...”, signaling uncertainty in the context of information seeking [19]. For citation match, answers were accompanied by either a citation that directly supported the content (matching) or one that was clearly unrelated (mismatching) (Figure 1). Mismatching citations were implemented as semantically unrelated but technically valid links (rather than broken URLs or fake DOIs) to avoid confounds from interface or system errors. Participants were randomly assigned to different condition sequences.



**Figure 1: Illustration of the experimental setting.** Participants asked a user query (“What does Astasia mean?”) followed by an LLM-generated answer with or without a verbalized uncertainty disclaimer (“I’m not sure but ...”). Each answer was accompanied by either a citation that matched the content (Citation Match) or a citation that was clearly unrelated but technically valid (Citation Mismatch). The verbalized uncertainty phrase appeared only in the designated conditions.

At the beginning of the session, participants completed four practice trials to become familiar with the interface; these trials were excluded from the analyses. In each experimental trial, participants were presented with a quiz-style question and the corresponding answer generated by the LLM chatbot (Figure 1). A “View Source” button was available, which participants could choose to click if they wished to examine the citation. External search was not permitted.

The question set was constructed to ensure both balance and control. For each item, the correctness of the LLM’s answer was balanced (50% correct, 50% incorrect), allowing assessment of whether adoption was maintained for correct answers and suppressed for incorrect ones across the combinations of verbalized uncertainty and citation match. To minimize confounds from prior knowledge, questions were drawn from non-specialist but less familiar knowledge domains, avoiding items from common trivia or current events. All items were written to be understandable in terms of language difficulty, while ensuring that background knowledge alone was insufficient to answer them correctly. In a preliminary self-check (N = 4), items were screened to ensure they could not be answered solely based on personal knowledge, and participants were instructed to report if they encountered an item they already knew. The dataset was sampled from the Open Trivia Database [20].

For each trial, participants were asked to decide whether to adopt the LLM’s answer as their own final response (coded as 1) or to reject it (coded as 0). In this study, we operationalize reliance behavior as adoption—whether users accepted the LLM’s output as their final answer. This procedure was repeated until all experimental trials were completed. The dependent variables were (a) adoption of the LLM’s answer and (b) citation-checking behavior, measured by whether the participant clicked the “View Source” button.

### 3. Results

#### Verification behavior (H1)

Participants engaged in citation checking more frequently when verbalized uncertainty was present than when it was absent. On average, the rate of citation-checking increased in the condition with verbalized uncertainty (M = 70%, SD = 4.60%) compared with the condition without verbalized uncertainty (M = 30%, SD = 4.61%) (Table 1). This pattern supports H1, indicating that verbalized uncertainty served as an effective cue to prompt users to verify the provided sources.

**Table 1: Citation-checking rates (%) by verbalized uncertainty (VU) condition (mean and standard deviation)**

Condition	Verification rate mean (%)	Verification rate SD (%)
With VU	70	4.60
Without VU	30	4.61

#### Adoption adjustment (H2)

When verbalized uncertainty prompted participants to engage in citation checking, adoption rates varied systematically depending on citation authenticity. Adoption was higher when the citation was matched (n = 56; M = 57.81%, SD = 5.04%) compared with when it was mismatched (n = 64; M = 16.14%, SD = 5.01%) (Table 2). This pattern supports H2, indicating that once participants inspected the sources, they adjusted their reliance decisions in line with the actual validity of the citation.

**Table 2: Adoption rates (%) in verified trials, by citation relevance**

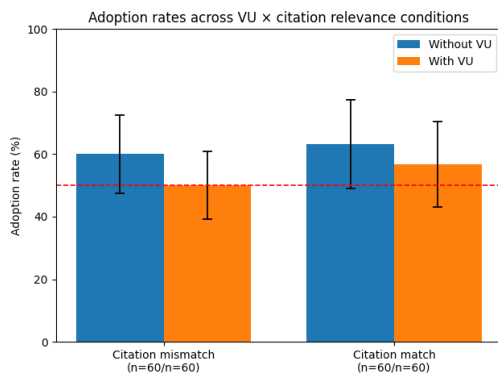
Citation match	n (verified trials)	Adoption rate mean (%)	Verification SD (%)
Match	56	57.81	5.04
Mismatch	64	16.14	5.01

**Interaction effect (H3)**

Verbalized uncertainty reduced adoption rates under both citation match and mismatch conditions, with the reduction being more pronounced when citations were mismatched than when they were matched (Table 3, Figure 2).

**Table 3: Mean adoption rates (%) and standard deviations (SD) by citation match and verbalized uncertainty (VU) conditions (N = 60 per cell)**

Citation relevance	With VU (M, SD) (%)	Without VU (M, SD) (%)
Match	56.7 (13.2)	63.3 (12.7)
Mismatch	50.0 (12.5)	60.0 (11.8)

**Figure 2: Interaction pattern between verbalized uncertainty (VU) and citation relevance on adoption rates. Error bars represent  $\pm 1$  SD.****4. Discussion**

This pilot study examined how verbalized uncertainty and the relevance of citations, which reflect common elements in LLM outputs, influence user adoption behavior in question answering. Verbalized uncertainty increased participants' likelihood of checking citations (H1), and once users inspected sources, their adoption decisions aligned with the actual validity of the citations (H2). Moreover, an interaction effect emerged (H3): Verbalized uncertainty reduced adoption in a conditional manner—lowering it when citations were mismatched, while the reduction was comparatively smaller when they were matched.

The results suggest that verbalized uncertainty can serve as an effective nudge to promote verification and counteract the double-edged nature of citations. While citations generally enhance perceived credibility, they also risk promoting over-adoption when

users accept them at face value. By prompting participants to verify the sources, verbalized uncertainty functioned as a lightweight safeguard that reduced blind adoption of incorrect information. Importantly, this effect was condition-specific: verbalized uncertainty did not substantially reduce adoption when the citation was relevant, but it noticeably constrained adoption when the citation was irrelevant. Taken together, this indicates that verbalized uncertainty, when calibrated, may help foster appropriate reliance by reducing over-trust without reducing correct adoption.

From a design perspective, these findings suggest that LLM systems could benefit from integrating light-touch verbalized uncertainty strategies to naturally stimulate user verification, especially in contexts where hallucinated or mismatched citations are likely. Rather than undermining trust wholesale, verbalized uncertainty may encourage more discerning use, aligning with human-automation interaction principles that emphasize calibrated reliance [16].

The present findings extend prior research on citation cues in LLM responses, which has shown that users often perceive answers as evidence-based simply because references are present [8], while actual verification is rare and cognitively costly [9]. Our results also resonate with linguistic and HCI research on verbalized uncertainty, which can sometimes lower credibility but, in certain contexts, foster impressions of objectivity and caution [19]. By demonstrating that verbalized uncertainty reduces blind adoption particularly when citations are irrelevant, this study connects two previously separate lines of research—citation-based trust cues and verbalized uncertainty—into a unified perspective on fostering appropriate reliance in LLM-based question answering.

Several qualifications should be noted. The study relied on general-knowledge quiz items, which may not reflect domains shaped by prior knowledge or high-stakes decision-making [21, 22]. The sample was also small (N = 6) with limited trials per condition, reducing statistical power. As a pilot study, the findings should be viewed as preliminary rather than conclusive.

As a preliminary study, these findings require validation through larger-scale experiments with more diverse participant populations. Future research should also investigate domain-specific contexts (e.g., healthcare, politics, science), where the balance between convenience, verification, and trust may operate differently. In addition, systematic manipulation of citation presentation formats (e.g., evidence granularity, verifiability nudges) could clarify how interface design shapes verification behavior. Finally, examining individual differences such as need for cognition, prior AI familiarity, and propensity to trust automation will help refine understanding of how adoption varies across users.

**ACKNOWLEDGMENTS**

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022-II2209842)

## REFERENCES

- [1] Peskoff, D., & Stewart, B. M. (2023, July). Credible without credit: Domain experts assess generative language models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 427-438).
- [2] Kaiser, C., Kaiser, J., Schallner, R., & Schneider, S. (2025, April). A New Era of Online Search? A Large-Scale Study of User Behavior and Personal Preferences during Practical Search Tasks with Generative AI versus Traditional Search Engines. In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (pp. 1-7).
- [3] Fernández-Pichel, M., Pichel, J. C., & Losada, D. E. (2025). Evaluating search engines and large language models for answering health questions. *npj Digital Medicine*, 8(1), 153.
- [4] Xu, R., Feng, Y., & Chen, H. (2023). Chatgpt vs. google: A comparative study of search performance and user experience. *arXiv preprint arXiv:2307.01135*.
- [5] Church, K. (2024). Emerging trends: When can users trust GPT, and when should they intervene?. *Natural Language Engineering*, 30(2), 417-427.
- [6] Mitchell, M. (2023). How do we know how smart AI systems are?. *Science*, 381(6654), eadj5957.
- [7] Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.
- [8] Ding, Y., Facciani, M., Joyce, E., Poudel, A., Bhattacharya, S., Veeramani, B., ... & Weninger, T. (2025, April). Citations and trust in llm generated responses. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 22, pp. 23787-23795).
- [9] Yun, H. S., & Bickmore, T. (2025, April). Framing Health Information: The Impact of Search Methods and Source Types on User Trust and Satisfaction in the Age of LLMs. In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (pp. 1-7).
- [10] Rodani, T., Osmenaj, E., Cazzaniga, A., Panighel, M., Cristina, A., & Cozzini, S. (2023). Towards the FAIRification of scanning tunneling microscopy images. *Data Intelligence*, 5(1), 27-42.
- [11] Zhang, M., & Zhao, T. (2025). Citation accuracy challenges posed by large language models. *JMIR Medical Education*, 11, e72998.
- [12] Spatharioti, S. E., Rothschild, D., Goldstein, D. G., & Hofman, J. M. (2025, April). Effects of LLM-based Search on Decision Making: Speed, Accuracy, and Overreliance. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (pp. 1-15).
- [13] McGrath, M. J., Cooper, P. S., & Duenser, A. (2024). Users do not trust recommendations from a large language model more than AI-sourced snippets. *Frontiers in Computer Science*, 6, 1456098.
- [14] Sun, Y., Sheng, D., Zhou, Z., & Wu, Y. (2024). AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanities and Social Sciences Communications*, 11(1), 1-14.
- [15] Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why Language Models Hallucinate. *arXiv preprint arXiv:2509.04664*.
- [16] Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1), 50-80.
- [17] Mittelstadt, B., Wachter, S., & Russell, C. (2023). To protect science, we must use LLMs as zero-shot translators. *Nature Human Behaviour*, 7(11), 1830-1832.
- [18] Hyland, K. (1996). Writing without conviction? Hedging in science research articles. *Applied linguistics*, 17(4), 433-454.
- [19] Kim, S. S., Liao, Q. V., Vorvoreanu, M., Ballard, S., & Vaughan, J. W. (2024, June). "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In Proceedings of the 2024 ACM conference on fairness, accountability, and transparency (pp. 822-835).
- [20] Open Trivia Database. 2025. Open Trivia DB. Retrieved September 15, 2025 from <https://opentdb.com/>
- [21] Kennedy, B., Atari, M., Davani, A. M., Hoover, J., Omrani, A., Graham, J., & Dehghani, M. (2021). Moral concerns are differentially observable in language. *Cognition*, 212, 104696.
- [22] Lenz, T., Brackey, R., & Liu, J. (2025). The Effects of Citations and Confirmation Bias on Trust in Chatbots. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (p. 10711813251357884). Sage CA: Los Angeles, CA: SAGE Publications.