Human Interaction and Emerging Technologies (IHIET 2025), Vol. 197, 2025, 139–148
https://doi.org/10.54941/ahfe1006707

Exploring User Behavior and Validation Proficiency in Assessing Responses From a Conversational Agent

Jiayin Huang and Jonggi Hong

Stevens Institute of Technology, Hoboken, NJ 07307, USA

ABSTRACT

With the rapid development of large language models (LLMs) like ChatGPT, conversational agents are becoming popular alternatives to traditional search engines. However, the ability to distinguish between replies generated by conversational agents and accurate information, along with user behavior in validating these replies, remains unclear. This study examines users' behavior and their ability to detect incorrect responses from ChatGPT, both with and without Google search results for validation, through a user study with 15 participants. Participants assessed ChatGPT's answers to questions about Alzheimer's Disease, which had an accuracy rate of 93.33% (28/30) and an error rate of 6.67% (2/30). Interestingly, when Google search results were available, participants tended to view both correct and incorrect responses favorably. These findings provide insights into the strategies users employ to validate conversational agents' responses, highlighting differences in behavior with and without the assistance of search engines.

Keywords: Human-centered computing, Empirical studies in HCI, Human-computer interaction (HCI), Computing methodologies, Artificial intelligence

INTRODUCTION

The rapid advancement of large language models (LLMs) such as OpenAI's ChatGPT and Google Bard has significantly reshaped the landscape of information retrieval and human-computer interaction. These conversational agents, generating human-like text responses, are increasingly integrated across diverse domains, from customer service to healthcare, due to their unparalleled ease of use (Xu et al., 2023). However, despite their advanced capabilities, these AI systems remain susceptible to inaccuracies and misleading information, known as AI hallucination, where models like ChatGPT provide inconsistent responses (Ahmad et al., 2023). Given the potential consequences of relying on inaccurate information, it is crucial for users to validate AI-generated content against reliable sources. Although some AI systems have begun integrating external sources such as search engines to support user validation, limited research has examined how users interact with these supplementary mechanisms and their strategies during validation.

This study aims to address this gap by focusing on user behavior in evaluating conversational agent responses. Specifically, it investigates the strategies users employ to detect inaccuracies in AI-generated content.

Through a user study with 15 participants, we explore how individuals validate ChatGPT's responses independently and with supplementary tools like Google search results. By analyzing user interactions and decision-making processes, we aim to uncover insights to inform conversational agent design for critical evaluation tasks. Our results indicate that participants correctly identified incorrect ChatGPT responses in approximately 70% of cases. Notably, access to Google search results did not consistently enhance accuracy but led participants to view responses more favorably overall, highlighting how supplementary information can influence user attitudes without necessarily improving critical evaluation skills.

The contributions of this research include:

- Providing empirical results on user validation proficiency and its behavioral impact when interacting with conversational agents.
- Evaluating users' abilities to validate information from conversational agents using web search engines.
- Identifying behavioral patterns when using search engines to verify AI-generated.

RELATED WORK

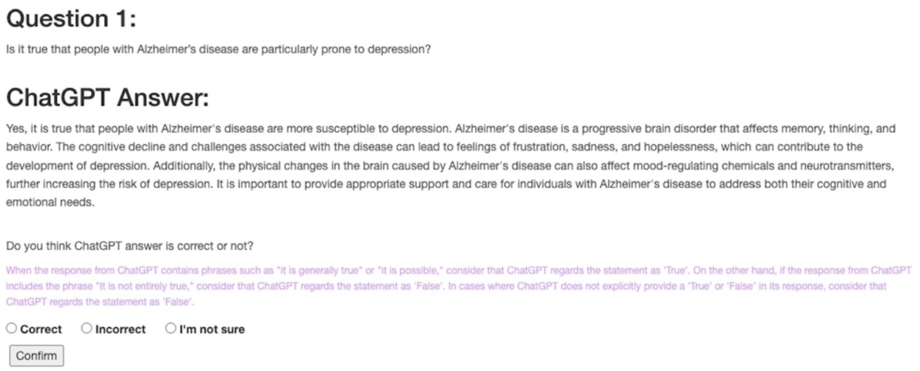
Trustworthiness of AI and Conversational Agents. In recent years, Large Language Model (LLM)-powered conversational agents, such as ChatGPT, have emerged as viable alternatives for information seeking. Studies highlight that user trust in ChatGPT's responses is often driven by fluency, directness, and human-like conversational style (Jung et al., 2024). Trust is enhanced when AI systems provide transparent explanations of their reasoning processes, and including references or interpretability cues significantly improves perceived reliability (Ehsan et al., 2021; Hassija et al., 2024). While existing research primarily investigates characteristics like consistency and transparency influencing user trust (Jang and Lukasiewicz, 2023; Wortham and Theodorou, 2017), limited attention is given to systematic methods users employ to validate AI-generated content when accuracy is uncertain.

Exploring Health Information Through Web Search. Search engines have become essential tools for accessing health-related information online due to convenience and speed, with trusted platforms like the NIH and Mayo Clinic gaining prominence (Cline and Haynes, 2001; De Choudhury, Morris and White, 2014; Kim, 2015). Trust in online health information is primarily influenced by source credibility, with reputable institutions and peer-reviewed publications viewed as more trustworthy (Abrar et al., 2023; Liu, Zhang and Kim, 2023). Users often rely on search engine rankings, assuming higher-ranked results are more reliable despite concerns about misinformation and manipulation (Cook, Ecker and Lewandowsky, 2015; Guo, 2022; Pan et al., 2007; Seckler et al., 2015; Haque, Khan and Fahim, 2023).

User Study

This study was approved by IRB (#2024-028N) as no personal-level sensitive data were used. During the study, 15 participants validated ChatGPT's answers to Alzheimer's disease-related questions by comparing them with their prior knowledge or Google search results. We developed a web-based testbed that displayed ChatGPT's responses and, in the second phase, 10 clickable Google search results alongside them.

Testbed. The testbed was designed to examine how access to Google search results influences user behavior when validating AI-generated content. Participants completed two phases: Phase 1, validating ChatGPT's responses alone; and Phase 2, validating ChatGPT's responses with Google search results displayed next to the AI's answers. Users could click any links to open the original sources in a new window. The two-phase interface as shown in Figures 1 and 2, respectively.



Question 1:

Is it true that people with Alzheimer's disease are particularly prone to depression?

ChatGPT Answer:

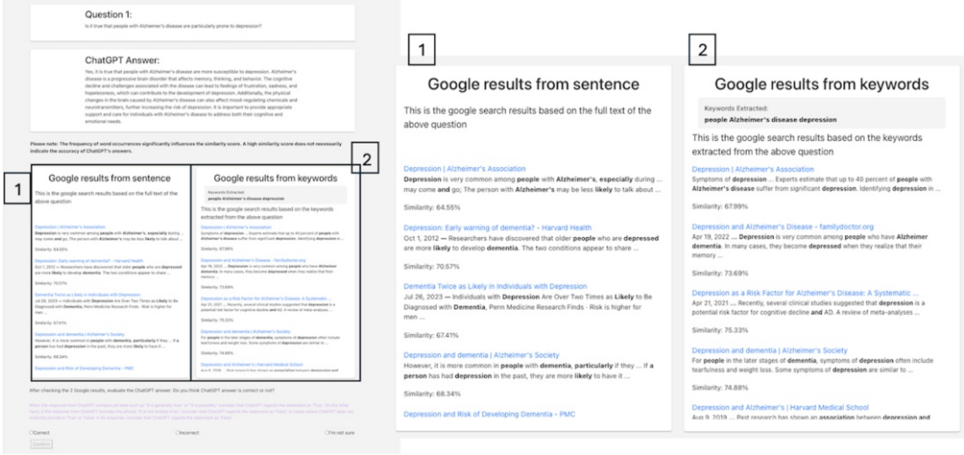
Yes, it is true that people with Alzheimer's disease are more susceptible to depression. Alzheimer's disease is a progressive brain disorder that affects memory, thinking, and behavior. The cognitive decline and challenges associated with the disease can lead to feelings of frustration, sadness, and hopelessness, which can contribute to the development of depression. Additionally, the physical changes in the brain caused by Alzheimer's disease can also affect mood-regulating chemicals and neurotransmitters, further increasing the risk of depression. It is important to provide appropriate support and care for individuals with Alzheimer's disease to address both their cognitive and emotional needs.

Do you think ChatGPT answer is correct or not?

When the response from ChatGPT contains phrases such as "it is generally true" or "it is possible," consider that ChatGPT regards the statement as 'True'. On the other hand, if the response from ChatGPT includes the phrase "It is not entirely true," consider that ChatGPT regards the statement as 'False'. In cases where ChatGPT does not explicitly provide a 'True' or 'False' in its response, consider that ChatGPT regards the statement as 'False'.

○ Correct ○ Incorrect ○ I'm not sure

Figure 1: Screenshot of the Phase 1 evaluation web application interface.



Question 1:

Is it true that people with Alzheimer's disease are particularly prone to depression?

ChatGPT Answer:

Yes, it is true that people with Alzheimer's disease are more susceptible to depression. Alzheimer's disease is a progressive brain disorder that affects memory, thinking, and behavior. The cognitive decline and challenges associated with the disease can lead to feelings of frustration, sadness, and hopelessness, which can contribute to the development of depression. Additionally, the physical changes in the brain caused by Alzheimer's disease can also affect mood-regulating chemicals and neurotransmitters, further increasing the risk of depression. It is important to provide appropriate support and care for individuals with Alzheimer's disease to address both their cognitive and emotional needs.

Please note: The frequency of word occurrences significantly influences the similarity score. A high similarity score does not necessarily indicate the accuracy of ChatGPT's answers.

Google results from sentence

This is the google search results based on the full text of the above question

Depression | Alzheimer's Association
Depression is very common among **people** with **Alzheimer's**, **especially** during ... may come **and** go; The person with **Alzheimer's** may be less **likely** to talk about ...

Similarity: 64.55%

Depression: Early warning of dementia? - Harvard Health
Oct 1, 2012 — Researchers have discovered that older **people** who are **depressed** are more **likely** to develop **dementia**. The two conditions appear to share ...

Similarity: 70.57%

Dementia Twice as Likely in Individuals with Depression
Jul 26, 2023 — Individuals with **Depression** Are Over Two Times as **Likely** to Be Diagnosed with **Dementia**, Penn Medicine Research Finds - Risk is higher for men ...

Similarity: 67.41%

Depression and dementia | Alzheimer's Society
However, it is more common in **people** with **dementia**, **particularly** if they ... If **a person** has had **depression** in the past, they are more **likely** to have it ...

Similarity: 68.34%

Depression and Risk of Developing Dementia - PMC

Google results from keywords

Keywords Extracted:
people Alzheimer's disease depression

This is the google search results based on the keywords extracted from the above question

Depression | Alzheimer's Association
Symptoms of **depression** ... Experts estimate that up to 40 percent of **people** with **Alzheimer's disease** suffer from significant **depression**. Identifying **depression** in ...

Similarity: 67.99%

Depression and Alzheimer's Disease - familydoctor.org
Apr 19, 2022 ... **Depression** is very common among **people** who have **Alzheimer dementia**. In many cases, they become **depressed** when they realize that their memory ...

Similarity: 73.69%

Depression as a Risk Factor for Alzheimer's Disease: A Systematic ...
Apr 21, 2021 ... Recently, several clinical studies suggested that **depression** is a potential risk factor for cognitive decline **and** AD. A review of meta-analyses ...

Similarity: 75.33%

Depression and dementia | Alzheimer's Society
For **people** in the later stages of **dementia**, symptoms of **depression** often include tearfulness and weight loss. Some symptoms of **depression** are similar to ...

Similarity: 74.88%

Depression and Alzheimer's | Harvard Medical School
Aug 9, 2019 ... Past research has shown an **association** between **depression and**

After checking the 2 Google results, evaluate the ChatGPT answer. Do you think ChatGPT answer is correct or not?

When the response from ChatGPT contains phrases such as "it is generally true" or "it is possible," consider that ChatGPT regards the statement as 'True'. On the other hand, if the response from ChatGPT includes the phrase "It is not entirely true," consider that ChatGPT regards the statement as 'False'. In cases where ChatGPT does not explicitly provide a 'True' or 'False' in its response, consider that ChatGPT regards the statement as 'False'.

○Correct ○Incorrect ○I'm not sure

Figure 2: Screenshot of the Phase 2 evaluation web application interface.

Dataset. We used the Alzheimer's Disease Knowledge Scale (ADKS) as the source of 30 true/false questions, covering various aspects of Alzheimer's disease. We reformatted 30 ADKS statements into yes/no questions to ensure consistent input for ChatGPT and subsequent evaluation. Each question was presented to ChatGPT three times to ensure answer consistency. Prior to the study, we collected two sets of search results for each statement, one using keywords extracted from the statement and the other using the full question to support external validation.

Participants. We recruited 15 participants aged 22–27 from our institution, with backgrounds in fields such as Computer Science, Software Engineering, and Data Science. All participants had prior experience using conversational AI systems like ChatGPT.

Procedure. The study consisted of an initial questionnaire capturing demographics and AI usage habits, two validation phases, and a follow-up questionnaire. In Phase 1, participants judged ChatGPT's responses using only their prior knowledge. In Phase 2, participants assessed the same responses with access to 10 pre-collected Google search results. Participants could choose "correct," "incorrect," or "I'm not sure" for each response. Instructions clarified that if ChatGPT's response included phrases like "it is generally true" or "it is possible," it should be treated as "True," and if it included "it is not entirely true," it should be treated as "False." Instructions were visible throughout the task.

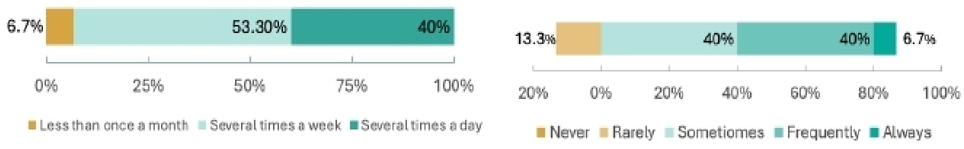
Measures and Data Analysis. We evaluated participant performance using precision, recall, and F1-score, based on classifications into true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). When "I'm not sure" was selected, the response was treated as perceived incorrect. Statistical analysis involved paired t-tests for normally distributed metrics and Wilcoxon tests otherwise.

RESULT

The findings encompass several aspects: participants' prior experience in validating responses from conversational agents, as gathered from the initial questionnaire; an analysis of their behavior and ability to validate responses during the task; and their feedback on the task itself, obtained through the follow-up questionnaire. These insights provide a comprehensive view of how users approach response validation in conversational agents and their performance throughout the study.

Participants' Experience in Validating Responses from Conversational Agents. Participants were surveyed on their use of conversational AI systems and their validation practices. Among the 15 participants, six reported using ChatGPT several times a day, eight several times a week, and one less than once a month (Figure 3a). When asked about the frequency of validating responses, six participants indicated they validated frequently, six sometimes, two rarely, and one always (Figure 3b).

Regarding validation methods, all participants reported using Google search to verify AI-generated answers. Additionally, six participants mentioned consulting other AI tools, such as Claude and Gemini. For example, P7 stated, "I use multiple AI-generated sites (Claude) along with Google," while P13 noted, "I often give different prompts and validate, and sometimes I also check with other AI tools like Claude and Gemini."



(a) (b)

Figure 3: (a) Reported frequency of ChatGPT usage, (b) Reported frequency of validating AI responses.

Accuracy of Identifying Incorrect Responses. Participants' ability to recognize correct responses improved with access to Google search results. Recall increased from 0.70 (*SD* = 0.10) in Phase 1 to 0.77 (*SD* = 0.14) in Phase 2, a statistically significant improvement ($t_{14} = -2.35$, $p = .034$, $d = -0.60$) (Figure 4). The F1 score also rose from 0.80 (*SD* = 0.07) to 0.84 (*SD* = 0.09), although this difference was not statistically significant. Analysis of participants' classification outcomes showed that true positives (TP) increased from 19.60 (*SD* = 2.90) to 21.67 (*SD* = 3.87), while false negatives (FN) decreased from 8.40 (*SD* = 2.90) to 6.33 (*SD* = 3.87) (Figure 5).

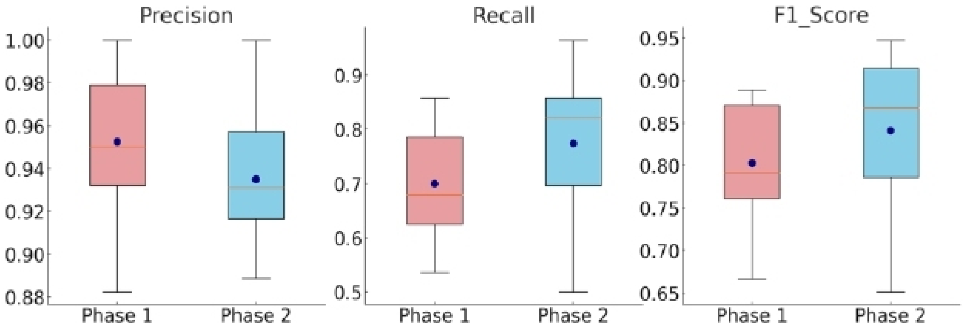


Figure 4: Box plot of each phase metrics.

However, a trade-off emerged. False positives (FP) increased from 1 (*SD* = 0.73) in Phase 1 to 1.47 (*SD* = 0.62) in Phase 2, approaching statistical significance (W = 10, p = .052, r = 2.58). True negatives (TN) decreased from 1 (*SD* = 0.73) to 0.53 (*SD* = 0.62). These findings suggest that although external search information improved participants' ability to detect correct answers, it also made them more likely to incorrectly trust false information.

Analysis of ChatGPT's Incorrect Answers. To further understand challenges in validation, we analyzed two intentionally incorrect answers from ChatGPT: Question 8 and Question 19. For Question 8 ("Is it true that in rare cases, people have recovered from Alzheimer's disease?"), ChatGPT incorrectly affirmed recovery was possible. In Phase 1, only three participants correctly identified the response as incorrect. After introducing Google search results, the number of participants incorrectly believing the answer was correct increased, suggesting that search results failed to clearly disprove the misconception.

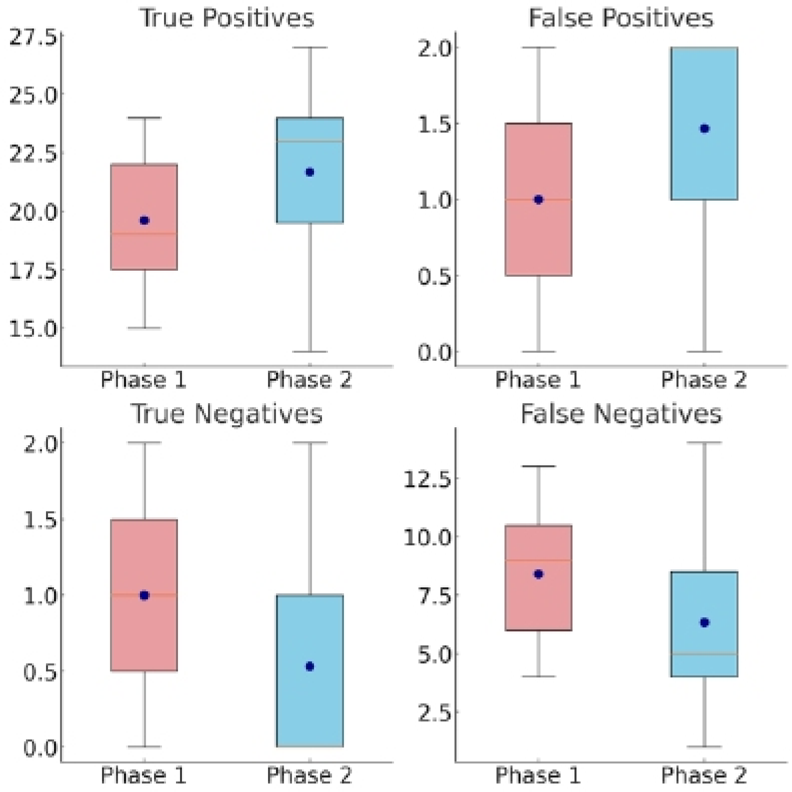


Figure 5: The number of true positives, false positives, true negatives, and false negatives in phase 1 and 2.

For Question 19 ("Is it true that tremor or shaking is common in Alzheimer's disease?"), ChatGPT again incorrectly affirmed the statement. In Phase 1, no participant judged the answer as incorrect. In Phase 2, only three participants correctly rejected it. Although search results mentioned Parkinson's disease symptoms, they did not explicitly clarify the distinction, leading to persistent misinterpretation. These examples highlight that even with external resources, ambiguous or incomplete information can reinforce misconceptions rather than correct them, emphasizing the need for clearer evidence presentation in both AI outputs and search engines.

Time Spent and Performance. Participants' average time per question significantly increased between the two phases. In Phase 1 (ChatGPT-only), the mean time was 38.71 seconds (*SD* = 14.37); in Phase 2 (with Google search), it rose to 82.00 seconds (*SD* = 23.35) (Figure 6). This increase reflects the additional cognitive effort required to cross-reference ChatGPT's responses with external sources. However, as seen in Questions 8 and 19, more time spent did not always lead to better validation accuracy. In some cases, complex or ambiguous search results prolonged decision-making without effectively improving judgment. This underscores the importance of balancing background information and clarity when designing AI systems to assist users in critical evaluation tasks.

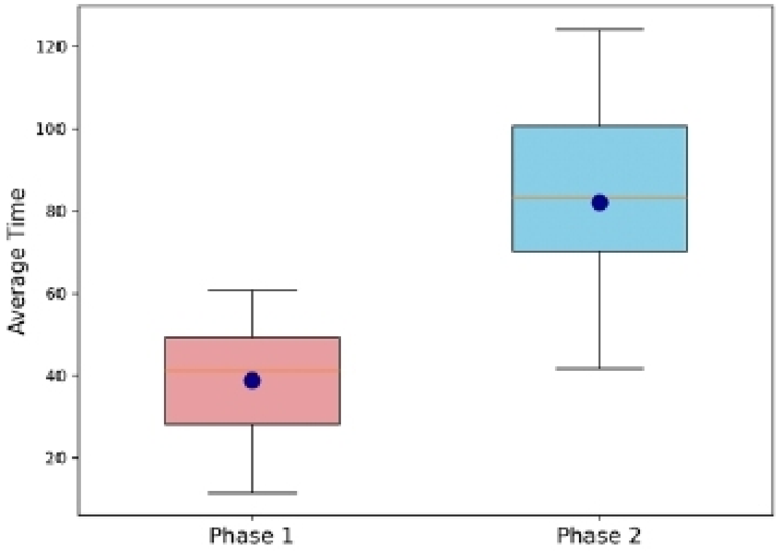


Figure 6: Average time box plot of each phase.

Correlation Between Validation Metrics and Information-Gathering Behavior. Participants clicked an average of 20 links ($SD = 12.92$) during the validation task. Correlation analyses revealed that fewer link clicks were associated with more false positives ($r = -0.63$, $p = .012$), while more clicks correlated with more true negatives ($r = 0.63$, $p = .012$) and higher precision ($r = 0.53, p = .041$). These findings suggest that active exploration of multiple information sources improves users' ability to accurately reject incorrect AI outputs. Encouraging broader information-gathering behavior may enhance validation performance when interacting with conversational agents.

Subjective Feedback. When validating ChatGPT's responses without search results, 10 participants relied on prior knowledge and experience, while six used logical reasoning. With access to Google search results, most participants ($N = 10$) compared ChatGPT responses with retrieved information, and many evaluated source reliability. As P15 noted, "I checked information from sources such as academic articles, official websites, and trusted news outlets."

As shown in Figure 7, participants preferred question-based search results over keyword-based ones. Eight participants rated question-based searches as very useful or extremely useful, citing greater relevance and reliability. In contrast, keyword-based results were generally rated as moderately useful. Participants prioritized links that closely matched the query ($N = 8$), came from trusted and authoritative sources ($N = 6$), or were from familiar websites ($N = 3$). Trusted institutions such as WHO, NIH, CDC, and Mayo Clinic were frequently mentioned as preferred sources.

These findings highlight that users' validation strategies depend heavily on source credibility, query formulation quality, and familiarity with trusted organizations, factors crucial for the design of future conversational agents and AI-generated information platforms.

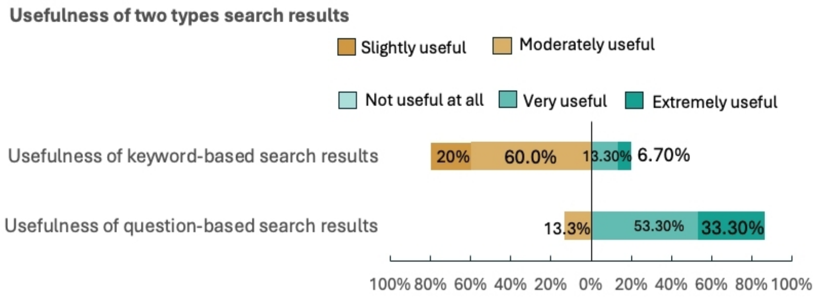


Figure 7: Reported usefulness of search results.

DISCUSSION

In this section, we will discuss the key implications of the findings, focusing on how the availability of external search results influenced participants' performance in validating ChatGPT responses. Additionally, we will address the limitations of the study, including potential factors that may have affected the results and the generalizability of the findings.

Encouraging frequent and multi-source validation. Many participants validated ChatGPT responses using Google and other AI tools. Future AI systems should promote consistent cross-validation across multiple trusted sources.

Balancing external support and critical evaluation. While access to search results improved recall, it also increased false positives. AI systems should integrate external information carefully to support recall without diminishing users' critical scrutiny.

Promoting deeper information-gathering behavior. Participants who explored more links achieved better accuracy. AI interfaces should nudge users to consult multiple sources, counteracting overconfidence from single responses.

Improving search literacy through system design. Participants favored question-style queries over keyword searches. AI tools should guide users in crafting effective queries and teach basic search strategies to enhance information retrieval.

Enhancing cognitive support in search activities. Prior research shows search is a cognitive learning process (Vakkari, 2016). Future AI should assist users with varying search abilities (Schultheiß and Lewandowski, 2023; Zhao et al., 2023), fostering better judgment rather than reinforcing blind trust in results.

LIMITATIONS

Our study involved a small, relatively tech-savvy participant pool, focused on the health domain, and used only ChatGPT-3.5 and Google search. Future work should expand to broader populations, topics, and tools to validate the generality of our findings. Nevertheless, the user behavior patterns observed in this study offer valuable insights and demonstrate potential generalizability across different domains and conversational AI systems.

CONCLUSION

This study explored how users validate responses from AI-generated systems like ChatGPT, both independently and with the support of external resources such as Google search results. Our findings reveal that while access to search results enhances recall and helps participants identify more correct responses, it also introduces a trade-off with a slight increase in false positives. These results highlight the importance of designing systems that balance the benefits of supplemental information with the need to maintain precision in user evaluations. We also observed that participants' ability to validate responses is influenced by their engagement with external resources, with more thorough information-gathering behaviors leading to better accuracy and fewer errors. This emphasizes the need for future conversational agents to integrate mechanisms that encourage users to critically evaluate AI-generated content and explore diverse information sources. These findings have broad implications for the design and deployment of conversational agents. By integrating features that promote critical evaluation, encourage the exploration of diverse resources, and support user-driven validation practices, AI systems can better foster trust and reliability. Such improvements are especially crucial in high-stakes domains, where the accuracy of AI-generated responses has significant real-world consequences. This research contributes to advancing our understanding of user behaviors and lays the groundwork for future innovations in AI-driven interactions.

REFERENCES

Abrar, M. F., Khan, M. S., Khan, I., Ali, G. and Shah, S. (2023). *Digital information credibility: Towards a set of guidelines for quality assessment of grey literature in multivocal literature review*. *Applied Sciences*, 13(7), p. 4483.

Ahmad, Z., Kaiser, W. and Rahim, S. (2023). *Hallucinations in ChatGPT: An unreliable tool for learning*. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, 15(4), p. 12.

Cline, R. J. W. and Haynes, K. M. (2001). *Consumer health information seeking on the Internet: The state of the art*. *Health Education Research*, 16(6), pp. 671–692.

Cook, J., Ecker, U. and Lewandowsky, S. (2015). *Misinformation and how to correct it*. In: *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*, pp. 1–17.

De Choudhury, M., Morris, M. R. and White, R. W. (2014). *Seeking and sharing health information online: Comparing search engines and social media*. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1365–1376.

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O. and Weisz, J. D. (2021). *Expanding explainability: Towards social transparency in AI systems*. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–19.

Guo, Y. (2022). *Digital trust and the reconstruction of trust in the digital society: An integrated model based on trust theory and expectation confirmation theory*. *Digital Government: Research and Practice*, 3(4), pp. 1–19.

Haque, E. U., Khan, M. M. H. and Fahim, M. A. A. (2023). *The nuanced nature of trust and privacy control adoption in the context of Google*. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–23.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M. and Hussain, A. (2024). *Interpreting black-box models: A review on explainable artificial intelligence*. *Cognitive Computation*, 16(1), pp. 45–74.

Jang, M. E. and Lukasiewicz, T. (2023). *Consistency analysis of ChatGPT*. *arXiv preprint*, arXiv:2303.06273.

Jung, Y., Chen, C., Jang, E. and Sundar, S. S. (2024). *Do we trust ChatGPT as much as Google Search and Wikipedia?* In: *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–9.

Kim, Y. M. (2015). *Is seeking health information online different from seeking general information online?* *Journal of Information Science*, 41(2), pp. 228–241.

Liu, J., Zhang, Y. and Kim, Y. (2023). *Consumer health information quality, credibility, and trust: An analysis of definitions, measures, and conceptual dimensions*. In: *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval (CHIIR)*, pp. 197–210.

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G. and Granka, L. (2007). *In Google we trust: Users' decisions on rank, position, and relevance*. *Journal of Computer-Mediated Communication*, 12(3), pp. 801–823.

Schultheiß, S. and Lewandowski, D. (2023). *Misplaced trust? The relationship between trust, ability to identify commercially influenced results and search engine preference*. *Journal of Information Science*, 49(3), pp. 609–623.

Seckler, M., Heinz, S., Forde, S., Tuch, A. N. and Opwis, K. (2015). *Trust and distrust on the web: User experiences and website characteristics*. *Computers in Human Behavior*, 45, pp. 39–50.

Vakkari, P. (2016). *Searching as learning: A systematization based on literature*. *Journal of Information Science*, 42(1), pp. 7–18.

Wortham, R. H. and Theodorou, A. (2017). *Robot transparency, trust and utility*. *Connection Science*, 29(3), pp. 242–248.

Xu, R., Feng, Y. and Chen, H. (2023). *ChatGPT vs. Google: A comparative study of search performance and user experience*. *arXiv preprint*, arXiv:2307.01135.

Zhao, Y., Bai, Y., Zhang, Y., Zhang, B. and Vakkari, P. (2023). *A probabilistic model toward how people search to build outcomes*. *IEEE Access*, 11, pp. 22450–22467. https://doi.org/10.1109/ACCESS.2023.3252369