Perceived Credibility of LLM Responses in Search Contexts

Abhinav Choudhry
ac62@illinois.edu
University of Illinois at Urbana-Champaign
Champaign, Illinois, United States

Kyrie Zhixuan Zhou
zz78@illinois.edu
University of Illinois at Urbana-Champaign
Champaign, Illinois, United States

Rachel F. Adler
radler@illinois.edu
University of Illinois at Urbana-Champaign
Champaign, Illinois, USA

ABSTRACT

People's search behaviours in Google have been extensively studied, but less so in Large Language Models (LLMs) such as ChatGPT. In this extended abstract, we propose a survey study to understand people's perceived credibility of LLM responses and their search behaviors. The hypotheses are synthesized based on a literature review of people's search behaviours on Google. With this study, we can potentially shed light on safer and more responsible search behaviours when using LLMs as either standalone conversational AI products or as a part of social robots.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **User studies**.

KEYWORDS

LLM, Search Behaviour, Credibility

ACM Reference Format:
Abhinav Choudhry, Kyrie Zhixuan Zhou, and Rachel F. Adler. 2018. Perceived Credibility of LLM Responses in Search Contexts. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

With the advent of text-based generative AI, search and information seeking are likely to become less keyword-driven and more based on natural language. Healthcare Q&A robots have been designed to address the shortage of medical personnel [4] and they can be easily enhanced by LLMs. However, LLMs do not always generate correct information and the stakes are higher when it comes to areas such as healthcare [5]

Existing research found that ChatGPT responses are more useful than Google featured snippets[1] [14]. However, issues with featured snippets might occur in people's interaction with LLMs as well, such as accuracy of LLM responses, higher stakes of misinformation in high-stake/sensitive areas like healthcare [12], and people's overconfidence about their internal knowledge [9] after seeing ChatGPT-generated responses.

Understanding people's trust in LLM responses and confidence in their internal knowledge after queries is thus timely and important to ensure safer and healthier search behaviour in the LLM era. In this extended abstract, we propose a survey study to compare people's perceived credibility of LLM responses and featured snippets, informed by the extensive literature on search engine behaviours. In the sections below, we first discuss existing research on (1) Generative AI for search and (2) Search engine's featured snippets; then we introduce study design including hypotheses and survey design.

[1]According to Google, Google's search results sometimes show listings where the snippet describing a page comes before a link to a page, not after as with their standard format. Results displayed this way are called "featured snippets."

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference acronym 'XX, June 03–05, 2018, Woodstock, NY
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

2 RELATED WORK

2.1 Generative AI for Search

The capabilities of generative AI for information-seeking have already been tested for their effectiveness. A study on translators found that direct answers provided by ChatGPT were typically more useful than top search results even with featured snippets, and a version that combines generative AI with references and top search results, called Perplexity AI performs even better [14]. LLMs also create a more intuitive way of using natural language to answer queries as compared to conventional search engines, leading to higher user satisfaction [19]. However, LLM can introduce errors [19], and ease of querying may not translate to higher accuracy in searches [23]. Search results have some advantages in terms of being cheaper to process than power-hungry LLMs, and benefit localised searches within an organisational domain since generative AI could return out-of-scope results. Nevertheless, time pressure to search for information prompts users to rely more on the chat with LLMs, even though many display some overall skepticism towards AI's accuracy [3].

We can also get a clue from how things could evolve based on previous changes made to search engine behaviour, with the biggest one being featured snippets, sometimes also called Q&A, that was introduced by Google roughly a decade ago, and since then, other search engines, such as Bing also have rolled out similar snippets. These snippets are populated from the search results and are displayed at the top of the search results in a separate section. We will elaborate on this aspect in the section below.

2.2 Search Engine Featured Snippets

2.2.1 Accuracy Issues. Research on search behaviour has found that featured snippets are overestimated in credibility by users and they may also lead to changes in opinions if they are uncertain [2]. Almost half of the featured snippets are usually taken from the

top-ranked search query [21] but this does not necessarily make them accurate. The accuracy of featured snippets varies a lot and Google's is better at answering where questions than who, what, or where questions [25]. There have been tweaks to the algorithm that have made snippets available to more queries, but these are often not from very credible sources [13].

2.2.2 Bad Information on Sensitive Areas. In certain sensitive categories, such as pharmaceutical drugs, Google snippets were rated as less than 50% accurate and complete in 5 out of 6 questions as against FDA-approved drug medication guides [12], and they have also been known to be misinterpreted or not presenting a viewpoint in many health-related snippets [10]. A featured snippet on health is most often picked from the Mayo Clinic but its answers are often too short, lacking depth and breadth, and when information is picked from a for-profit website, such as WebMD and Healthline, the risk is of sourcing from information not written by medical professionals, and with vested corporate interests [16]. A study found that of misleading Google search results about California representatives, 70% had a misleading featured snippet [11].

2.2.3 Change in User Behaviour. In fact, access to featured snippets risks making people more confident about their internal knowledge even though featured snippets are external knowledge [9]. An industry study of 2 million featured snippets found that they reduce the overall number of clicks on search results and the number of clicks that the top search result would normally get [18]. A survey found that a majority of Google searches were not resulting in a click at all with the query answered by the search itself [8].

2.2.4 Age Differences in Relying on Featured Snippets. A survey found that the youngest age group (13-18) was over twice as likely as the oldest age group (70-100) to consider their question answered by a featured snippet, or a knowledge panel [15]. While older people are more likely to make incorrect decisions due to reliance on heuristics or rules of thumb [1], they also show lesser sunk-cost fallacy in later life, i.e., they would not continue investing time, money, or effort into a situation when it is no longer beneficial [20]. The problematic search behaviour of younger people was also confirmed by Taylor's research – millennial generation Web searchers made limited attempts to evaluate the quality or validity of the information [22].

3 STUDY DESIGNS

3.1 Hypotheses

Users have shown a tendency for directly relying on features introduced by search engines, meant for heuristic help, including featured snippets [8, 9, 15, 18]. With the evolution towards LLMs, users have been found to prefer LLMs over Google in search, although this might be influenced by time pressure: time limits led to greater reliance on LLMs [14]. Examining how people evaluate the credibility of LLM responses and featured snippets is interesting yet under-investigated. Thus, we wish to test the following hypothesis,

Hypothesis 1: There is a difference in the level of trust for results of queries from LLMs, search engine results with featured snippets, and plain search results.

Previous studies showed that people tended to cross-validate LLMs' responses regarding healthcare with other information sources such as Google, exhibiting caution [24], a caution also reflected in other non-healthcare studies [14]. However, there has not been a comparison between users' credibility perception of LLMs vs search engine results when searching for information in high-stake areas such as healthcare. Thus, we wish to test the following hypothesis,

Hypothesis 2: There is a difference in the level of trust for results of queries from LLMs, search engine results with featured snippets, and plain search results, in high-stake areas for Q&A, such as healthcare.

Users have previously demonstrated more confidence in their internal knowledge when having easy access to search engine's featured snippets; in other words, their search fluency affected their knowledge confidence and this could be termed search-induced cognitive overconfidence [9]. However, experiments have proved that this knowledge self-assessment could be fallacious [7]. Recent evidence suggests that LLM-based searching risks exacerbating existing biases through more biased user search queries and hardening user confidence in their stances [17]. Unjustified confidence also prevents people from considering changes in opinion [6] We expect that LLMs too might have a similar influence on user behaviour, particularly overconfidence in internal knowledge. Thus, we wish to test the following hypothesis,

Hypothesis 3: There is a difference in the level of confidence in users' internal knowledge as results of queries from LLMs, search engine results with featured snippets, and plain search results.

Older adults are more cautious about assigning credence to the results shown in search engine's featured snippets, as compared to younger adults [15], and other research has shown them to differ in their biases from younger age groups [1, 20]. Whether this gap of cautious level between older vs. younger adults is widened in LLMs is unknown. Thus, we wish to test the following hypothesis,

Hypothesis 4.1: There is a difference in the degree of caution between older adults and younger adults in their trust of results from LLMs, featured snippets, and plain search results.

Hypothesis 4.2: The gap between the degree of caution between older adults and younger adults is wider for LLMs as compared to featured snippets, and plain search results.

3.2 Survey Design

We propose to conduct a survey with a crossover design. The survey will have some questions asked on regular topics and some high-stakes topics. The questions and full screenshots of responses from LLMs and search engines will be presented – the responses include (1) LLM responses presented where featured snippets usually feature, (2) featured snippets and a ranked list of search results, and (3) a ranked list of search results without featured snippets. Users will also have the option to click on the links presented for more information about search results in a popup window if they wish to. For each response, the participants will be asked questions that relate to the search query to understand their accuracy of understanding. They will be asked about their level of confidence in their internal knowledge, trust in the search results, and ease and satisfaction with the process. The ordering of LLMs, results with snippets, and plain results, which can be called search modes, will

be randomised between the participants. The number of questions, and types of questions, will be consistent for all participants.

Participants will be recruited from different age populations by targeted outreach to universities, community forums, religious organisations, health organisations, and others. Results will be analysed using tests of statistical significance and they will include demographic analysis.

REFERENCES

[1] Tibor Besedeš, Cary Deck, Sudipta Sarangi, and Mikhael Shor. 2012. Age Effects and Heuristics in Decision Making. *The Review of Economics and Statistics* 94, 2 (05 2012), 580–595. https://doi.org/10.1162/REST_a_00174 arXiv:https://direct.mit.edu/rest/article-pdf/94/2/580/1916910/rest_a_00174.pdf

[2] Markus Bink, Steven Zimmerman, and David Elsweiler. 2022. Featured Snippets and their Influence on Users' Credibility Judgements. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval* (Regensburg, Germany) *(CHIIR '22)*. Association for Computing Machinery, New York, NY, USA, 113–122. https://doi.org/10.1145/3498366.3505766

[3] Robert Capra and Jaime Arguello. 2023. How does AI chat change search behaviors? arXiv:2307.03826 [cs.HC]

[4] Yu-Hsuan Chang, Yi-Ting Guo, Li-Chen Fu, Ming-Jang Chiu, Han-Mo Chiu, and Hung-Ju Lin. 2023. Interactive Healthcare Robot using Attention-based Question-Answer Retrieval and Medical Entity Extraction Models. *IEEE Journal of Biomedical and Health Informatics* (2023).

[5] Giovanna Deiana, Marco Dettori, Antonella Arghittu, Antonio Azara, Giovanni Gabutti, and Paolo Castiglia. 2023. Artificial intelligence and public health: evaluating ChatGPT responses to vaccination myths and misconceptions. *Vaccines* 11, 7 (2023), 1217.

[6] Philip M Fernbach, Todd Rogers, Craig R Fox, and Steven A Sloman. 2013. Political extremism is supported by an illusion of understanding. *Psychological science* 24, 6 (2013), 939–946.

[7] Matthew Fisher, Mariel K Goddu, and Frank C Keil. 2015. Searching for explanations: How the Internet inflates estimates of internal knowledge. *Journal of experimental psychology: General* 144, 3 (2015), 674.

[8] Rand Fishkin. 2019. Less than Half of Google Searches Now Result in a Click - SparkToro — sparktoro.com. https://sparktoro.com/blog/less-than-half-of-google-searches-now-result-in-a-click/. [Accessed 18-02-2024].

[9] Kristy A Hamilton and Li Qi. 2023. Search Fluency Mistaken for Understanding: Ease of Information Retrieval from the Internet Inflates Internal Knowledge Confidence. *Social Media+ Society* 9, 3 (2023), 20563051231195547.

[10] Anat Hashavit, Tamar Stern, Hongning Wang, and Sarit Kraus. 2024. The Impact of Snippet Reliability on Misinformation in Online Health Search. arXiv:2401.15720 [cs.IR]

[11] Emma Lurie and Deirdre K. Mulligan. 2021. Searching for Representation: A sociotechnical audit of googling for members of U.S. Congress. arXiv:2109.07012 [cs.CY]

[12] Cambrey Nguyen. 2021. The accuracy and completeness of drug information in Google snippet blocks. *Journal of the Medical Library Association: JMLA* 109, 4 (2021), 613.

[13] Jack Nicas. 2017. Google Has Picked an Answer for You—Too Bad It's Often Wrong. *Wall Street Journal* 16 (2017).

[14] Ming Qian and Huaqing Wu. 2023. Translators as Information Seekers: Strategies and Novel Techniques. In *Artificial Intelligence in HCI*, Helmut Degen and Stavroula Ntoa (Eds.). Springer Nature Switzerland, Cham, 150–166.

[15] Lily Ray. 2019. We Surveyed 1,400 Searchers About Google - Here's What We Learned — moz.com. https://moz.com/blog/new-google-survey-results. [Accessed 18-02-2024].

[16] Amanda Scull. 2020. Dr. Google Will See You Now: Google's Health Information Previews and Implications for Consumer Health. *Medical Reference Services Quarterly* 39, 2 (2020), 165–173. https://doi.org/10.1080/02763869.2020.1726151 arXiv:https://doi.org/10.1080/02763869.2020.1726151 PMID: 32329674.

[17] Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effects of LLM-Powered Search Systems on Diverse Information Seeking. arXiv:2402.05880 [cs.CL]

[18] Tim Soulo. 2017. Ahrefs' Study Of 2 Million Featured Snippets: 10 Important Takeaways — ahrefs.com. https://ahrefs.com/blog/featured-snippets-study/. [Accessed 18-02-2024].

[19] Sofia Eleni Spatharioti, David M. Rothschild, Daniel G. Goldstein, and Jake M. Hofman. 2023. Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment. arXiv:2307.03744 [cs.HC]

[20] JoNell Strough, Tara E Karns, and Leo Schlosnagle. 2011. Decision-making heuristics and biases across the life span. *Annals of the New York Academy of Sciences* 1235, 1 (2011), 57–74.

[21] Artur Strzelecki and Paulina Rutecka. 2020. Featured Snippets Results in Google Web Search: An Exploratory Study. In *Marketing and Smart Technologies*, Álvaro Rocha, José Luís Reis, Marc K. Peter, and Zorica Bogdanović (Eds.). Springer Singapore, Singapore, 9–18.

[22] Arthur Taylor. 2012. A study of the information search behaviour of the millennial generation. *Information research: an international electronic journal* 17, 1 (2012), n1.

[23] Albatool Wazzan, Stephen MacNeil, and Richard Souvenir. 2024. Comparing Traditional and LLM-based Search for Image Geolocation. *arXiv preprint arXiv:2401.10184* (2024).

[24] Yunpeng Xiao, Kyrie Zhixuan Zhou, Yueqing Liang, and Kai Shu. 2024. Understanding the concerns and choices of public when using large language models for healthcare. arXiv:2401.09090 [cs.CY]

[25] Yiming Zhao, Jin Zhang, Xue Xia, and Taowen Le. 2019. Evaluation of Google question-answering quality. *Library Hi Tech* 37, 2 (2019), 308–324.