TYPE Brief Research Report
PUBLISHED 08 October 2024
DOI 10.3389/fcomp.2024.1456098

Users do not trust recommendations from a large language model more than AI-sourced snippets

Melanie J. McGrath[1]*, Patrick S. Cooper[1] and Andreas Duenser[2]

[1]Commonwealth Scientific and Industrial Research Organisation (CSIRO), Clayton, VIC, Australia,
[2]Commonwealth Scientific and Industrial Research Organisation (CSIRO), Sandy Bay, TAS, Australia

OPEN ACCESS

EDITED BY
Ilaria Tiddi,
VU Amsterdam, Netherlands

REVIEWED BY
Marco A. Palomino,
University of Aberdeen, United Kingdom
Sergio Rojas-Galeano,
Universidad Distrital Francisco José de Caldas,
Colombia

*CORRESPONDENCE
Melanie J. McGrath
✉ melanie.mcgrath@data61.csiro.au

RECEIVED 28 June 2024
ACCEPTED 20 September 2024
PUBLISHED 08 October 2024

CITATION
McGrath MJ, Cooper PS and Duenser A (2024) Users do not trust recommendations from a large language model more than AI-sourced snippets. *Front. Comput. Sci.* 6:1456098. doi: 10.3389/fcomp.2024.1456098

Background: The ability of large language models to generate general purpose natural language represents a significant step forward in creating systems able to augment a range of human endeavors. However, concerns have been raised about the potential for misplaced trust in the potentially hallucinatory outputs of these models.

Objectives: The study reported in this paper is a preliminary exploration of whether trust in the content of output generated by an LLM may be inflated in relation to other forms of ecologically valid, AI-sourced information.

Method: Participants were presented with a series of general knowledge questions and a recommended answer from an AI-assistant that had either been generated by an ChatGPT-3 or sourced by Google's AI-powered featured snippets function. We also systematically varied whether the AI-assistant's advice was accurate or inaccurate.

Results: Trust and reliance in LLM-generated recommendations were not significantly higher than that of recommendations from a non-LLM source. While accuracy of the recommendations resulted in a significant reduction in trust, this did not differ significantly by AI-application.

Conclusion: Using three predefined general knowledge tasks and fixed recommendation sets from the AI-assistant, we did not find evidence that trust in LLM-generated output is artificially inflated, or that people are more likely to miscalibrate their trust in this novel technology than another commonly drawn on form of AI-sourced information.

KEYWORDS

trust, artificial intelligence, large language models, trust calibration, HCI, generative AI, ChatGPT-3, hallucination

1 Introduction

Large language models (LLMs) are deep learning models trained on vast corpora of text that give them the ability to generate general purpose natural language (Akata et al., 2023; Brown et al., 2020). The emergence of models like ChatGPT-3 was heralded by many as revolutionary, citing their scalability and adaptability, and transformative potential for industry (Grant, 2023) and even the arts (Heaven, 2022). Among the voices extolling their strengths, there are, however, those expressing reservations about the use and performance of these LLMs. These reservations include the possible use of AI-generated text to amplify the creation and distribution of misinformation (Kreps et al., 2022). Questions have also been raised about the potential for erosion of critical thinking skills as an outcome of unreflective reliance on AI output (Huschens et al., 2023). Of significant concern is a tendency for hallucination, that is language output that is nonsensical, irrelevant, or not an accurate reflection of the input data (Ji et al., 2023). The hallucinations of artificial intelligence (AI) can become a serious issue when humans rely on confidently delivered, but inaccurate, recommendations or decisions. One widely reported example is that of a lawyer who was fined after using an LLM to prepare a court filing that cited cases fabricated by the AI system (Bohannon, 2023).

Reliance on AI outputs is influenced by a number of factors, but key among them is trust. The relationship between trust and reliance on automation and AI is well established (Lee and See, 2004). In general, higher levels of trust lead to higher levels of reliance on a given AI system (Hoff and Bashir, 2015). However, trust can, of course, be misplaced. Scams and cons among humans occur when trust is given to an untrustworthy actor. Similarly, when trust is inappropriately calibrated to the capabilities of an AI system, it can result in negative outcomes. When trust exceeds the objective capability of an AI system, human complacency may lead to misuse, with impacts that range from the inconvenient to the catastrophic (Robinette et al., 2016). Conversely, lack of trust in a trustworthy system risks disuse, leading to reduced productivity and lost resources (Parasuraman and Riley, 1997).

Amidst anecdotal evidence and news stories depicting apparent naivete in reliance on LLMs (e.g., Gupta, 2023; Herbert, 2023), we were drawn to wonder whether the particular capabilities of these novel AI systems have the capacity to influence their perceived trustworthiness independently of their objective performance. That is, is there something about interacting with an LLM that artificially inflates trust in its outputs? If this is the case, it could have significant implications for organizations seeking to incorporate LLMs into their workflows.

There are several ways in which the operation of LLMs could result in patterns of trust development that diverge from those of more traditional forms of AI. These include the content of the AI output itself, i.e., what information is included or excluded. The form of language used to deliver the content may also have implications for trust, as can the form of the interactional process (Morrissey and Kirakowski, 2013). For example, researchers have previously examined how length of response influences a user's satisfaction with AI output (Huang et al., 2024). The platform itself may also be a factor. The medium of delivery has long been recognized as a factor influencing perceptions of the credibility of information, and this extends to the user interface of media such as LLMs (Huschens et al., 2023).

In the present study, we aim to establish whether trust in LLMs is inflated in relation to other forms of AI, with a particular focus on the content and forms of natural language used. We further investigate whether any such inflated trust has an impact on reliance on the outputs of these systems. We report here on the findings of an experimental study in which participants were presented with a series of general knowledge questions and a recommended answer from an AI-assistant. Participants were given the option of whether to rely on the AI-assistant's advice and offered an incentive for correct answers. We systematically varied whether the advice to participants was generated by an LLM or sourced from Google's AI-powered featured snippets function (Strzelecki and Rutecka, 2020). We further manipulated whether this advice was accurate or inaccurate.

Using this experimental paradigm we sought to address three research questions.

Research Question 1: Do participants report more trust in recommendations generated by the LLM than recommendations sourced by an alternative non-LLM AI application?

Research Question 2a: Is reported trust influenced by feedback about the accuracy or previous recommendations from the AI-assistant?

Research Question 2b: If so, are the reported changes in trust resulting from accuracy feedback greater for the LLM than the non-LLM AI-assistant? That is, does trust calibration differ significantly between the two applications?

Research Question 3: Do participants rely more on recommendations from the LLM AI-assistant than the non-LLM AI-assistant, and if so, is this influenced by reported trust?

This research complied with the Australian National Statement on Ethical Conduct in Human Research 2007 (updated 2018) and was approved by the CSIRO Social Science Human Research Ethics Committee (#113/23).

2 Methods

2.1 Participants

Participants in this study were 199 residents of the United States recruited from the online research platform Prolific (Palan and Schitter, 2018). Participants were compensated at a base rate of USD2.10 for their participation, with the opportunity to earn up to 35c in bonus payments according to the study procedure outlined below. The sample comprised 98 (49.25%) women, 99 (49.75%) men and two participants of undisclosed gender, with a mean age of 37.54 years (*SD* = 13.73).

2.2 Procedure

After providing informed consent, all participants saw the following instructions:

> *In this task you will have a series of questions to answer. You will be provided with a response by an AI-assistant and will have to decide whether you use this response.*
>
> *You will start with a 20c bonus payment.*
>
> *For each correct response you will receive an additional 5c bonus.*
>
> *For each incorrect response you will lose 5c of your bonus payment.*
>
> *You may choose not to answer in which case you will neither gain nor lose any payment.*
>
> *Please do not exit this screen or use any other sources of information to answer these questions.*

The bonus payments were used to create conditions of risk. Our participants faced a material loss or gain from trusting the recommendations of the AI-assistant. Although there are many definitions of trust in the scientific literature, the most established and widely used tend to converge on a set of key features; (a) an expectation or belief that (b) a specific subject will (c) perform future actions with the intention of producing (d) positive outcomes for the trustor in (e) situations characterized by risk and vulnerability (Castaldo et al., 2010). Risk is consequently a necessary precondition for the exercise of trust and is one of the factors that distinguishes it from frequently conflated concepts like confidence.

All participants saw the same set of three questions (see section 2.3 Materials and Measures), however, we used a 2 (AI application) × 2 (accuracy) × 3 (trial) mixed design to implement our experimental manipulation. Application was a between-participant condition referring to whether participants received recommendations generated by an LLM or non-LLM. Participants randomly allocated into each application condition saw recommendations from the same application across all three trials.

Within each application condition, participants were further randomly allocated to an accurate or inaccurate condition. In the accurate condition the AI-assistant recommendation was accurate on trials 1 and 2 and inaccurate on trial 3. In the inaccurate condition the AI-assistant recommendation was accurate on trial 1 and inaccurate on trials 2 and 3.

After being presented with a question and the AI-assistant's recommendation, participants were asked to select an answer that agreed with the AI-assistant's recommendation (e.g., *Lise Meitner's most notable achievement was contributing to the discovery of nuclear fission*), an answer that was not consistent with the AI-assistant's recommendation (e.g., *Lise Meitner's most notable achievement was NOT contribution to the discovery of nuclear fission*) or opt not to respond (*I choose not to answer*). After selecting an answer participants rated their trust in the AI-assistant before receiving feedback about the accuracy of their response and consequent implication for their bonus payment.

This process was completed consecutively for each of the three task questions. After receiving feedback on the answer for the last task question, participants were asked to give a final report on their trust in the AI-assistant. To evaluate the validity of our data, participants were also asked to report honestly, and with no impact on their payment, if they had prior knowledge of the answers or had used external sources to answer the questions.

The study was completed on the Qualtrics survey platform in January 2024, and all AI recommendations were pre-generated, that is, participants did not interact directly with an LLM.

2.3 Materials and measures

2.3.1 Task questions

All participants saw the same three task questions. To maximize the likelihood that the participant would be required to trust or rely on the AI-assistant's recommendations, questions were intended to be very difficult and outside the general knowledge of a majority of participants. To facilitate this, the following prompts were provided to ChatGPT-3:

> *Who was an obscure female scientist?*
>
> *What is a book that very few people have heard of?*
>
> *What country exports the most quartz?*

Using the responses to these prompts as a starting point, the authors identified the following task questions:

1. What was Lise Meitner's most notable achievement?
2. Who wrote Codex Seraphinanus?
3. Which country exports the most quartz?

A pilot test of 51 participants recruited from Prolific validated the difficulty of these questions with more than 80% of respondents providing incorrect answers or indicating they did not know the answer to the questions.

2.3.2 AI-assistant recommendations

To generate LLM recommendations the following prompts were provided to ChatGPT-3:

> *What was Lise Meitner's most notable achievement?*
>
> *Who wrote the Codex Seraphinianus?*
>
> *Give a convincing but incorrect response to the question "who wrote the Codex Seraphinianus."*
>
> *Give a convincing but incorrect response to the question "which country exports the most quartz."*

Non-LLM recommendations were derived from the AI-powered "featured snippet" of a Google search. The featured snippets produced by Google are excerpts extracted from websites. That is, Google Search uses AI to identify the most relevant 'snippet' from the most relevant website and the snippet text itself is not generated by Google (Strzelecki and Rutecka, 2020). Accurate recommendations were presented to participants as shown in the featured snippet. For inaccurate recommendations, the incorrect detail from the LLM-generated response (e.g., the name of the author of the Codex Seraphinianus) was substituted in the Google featured snippet. An example of recommendations of each type is given in Table 1, and the full set of AI-assistant recommendations is provided in Supplementary material.

2.3.3 Trust

After selecting a response, participants reported their level of trust in the AI-assistant using a short three-item version of the Trust in Automation Scale (Jian et al., 2000). Participants responded to the following three items on a scale from 1 = *Not at all* to 7 = *Extremely*; "I am confident in the assistant," "The assistant is reliable," and "I can trust the assistant." Cronbach's alpha for the Short Trust in Automation Scale (S-TIAS) was equal to or greater than 0.90 across all three task questions and the final trust rating.

2.3.4 Reliance

Reliance was measured behaviorally. Participants were considered to have relied on the AI-assistant if they selected an answer consistent with the assistant's recommendation, and not to have relied on the system if they chose an inconsistent response or opted not to answer.

3 Results

Means and standard deviations of trust scores across application conditions and trials are given in Table 2.

Our validation questions provided information on whether participants knew the answers to the questions prior to receiving the

TABLE 1 AI-assistant recommendations provided in response to Task Question 2.

	LLM	Non-LLM
Accurate	"Codex Seraphinianus" was written and illustrated by Luigi Serafini. Luigi Serafini is an Italian artist and designer who created this unique and surreal encyclopedia, publishing it in 1981. "Codex Seraphinianus" showcases Serafini's extraordinary imagination and attention to detail through its captivating illustrations and an invented, cryptic script. The book has garnered significant attention for its enigmatic nature, as the meaning behind the text and imagery remains largely elusive.	The Codex Seraphinianus, originally published in 1981, is an illustrated encyclopedia of an imaginary world, created by Italian artist, architect and industrial designer Luigi Serafini between 1976 and 1978. It is approximately 360 pages (depending on edition) and written in an imaginary language. Originally published in Italy, it has been released in several countries.
Inaccurate	Renowned Italian artist and writer, Aldo Cataldi, is widely believed to be the author of "Codex Seraphinianus." Cataldi was known for his avant-garde and experimental works, and "Codex Seraphinianus" was considered one of his most enigmatic and imaginative creations. Cataldi's intricate illustrations and the mysterious, invented script in the book captured the attention of readers and scholars alike, sparking endless speculation about the meaning behind its cryptic content.	The Codex Seraphinianus, originally published in 1981, is an illustrated encyclopedia of an imaginary world, created by Italian artist, architect and industrial designer Aldo Cataldi between 1976 and 1978. It is approximately 360 pages (depending on edition) and written in an imaginary language. Originally published in Italy, it has been released in several countries.

AI-assistant recommendations or, against instructions, sought external information when answering the questions. Thirteen percent of participants reported prior knowledge of the answers to the questions, while 23% reported seeking external information. Chi-square tests of independence indicated that participants with external or pre-existing knowledge did not systematically differ in their distribution across experimental conditions (AI-application or accuracy). Nor did they receive significantly higher bonus payments than other participants, suggesting that this knowledge did not systematically influence the correctness of their responses or their trust calibration.

TABLE 2 Means and standard deviations of trust scores.

Time	LLM condition	Non-LLM condition	Total
Trial 1	5.04 (0.98)	4.99 (1.03)	5.01 (1.00)
Trial 2	4.69 (1.29)	4.97 (1.16)	4.83 (1.23)
Trial 3	4.35 (1.43)	4.51 (1.31)	4.43 (1.37)
Final	3.39 (1.43)	3.53 (1.40)	3.46 (1.41)
N	99	100	199

3.1 Research question 1

Our first research question asked whether participants report more trust in recommendations generated by the LLM than recommendations sourced by a non-LLM AI application. To answer this, we conducted a factorial mixed model ANOVA with time as a within-subjects factor and AI application as a between-subjects factor. Analyses indicated a main effect of trial on trust, $F(3, 591) = 112.68$, $p < 0.001$, such that reported trust dropped significantly after each trial. There was no main effect of application on trust, $F(1,197) = 0.90$, $p = 0.34$, indicating that reported trust did not differ based on whether participants saw a recommendation from an LLM or non-LLM form of AI. However, data did show a significant interaction between trial and application, $F(3,591) = 1.01$, $p < 0.001$, with trust dropping between Trial 1 and Trial 2 to a greater degree for participants receiving LLM recommendations than non-LLM recommendations (see Figure 1). It is worth noting that at the time of reporting Trial 2 trust all participants, regardless of AI application, had received feedback that the AI-recommendation was accurate in Trial 1 and had not yet received accuracy feedback for Trial 2. There was no significant difference in reported trust between applications at subsequent time points.

3.2 Research question 2

Our second research question asked whether reported trust is influenced by feedback about accuracy of previous recommendations from the AI-assistant. The accurate and inaccurate conditions differed on the accuracy of the AI-recommendation in Trial 2. Consequently, we conducted a two-way (AI application × accuracy) ANOVA predicting the effect of accuracy feedback after Trial 2 on reported trust at Trial 3. Results indicated a significant main effect of accuracy condition, $F(1,195) = 5.78$, $p = 0.02$, such that participants in the accurate condition ($M = 4.66$, $SD = 1.33$) reported a significantly greater degree of trust in the AI recommendations than those in the inaccurate condition ($M = 4.20$, $SD = 1.38$).

The second part of this research question asked whether the effect of accuracy on trust differed between the LLM and non-LLM, which may indicate differences in trust calibration between the applications. The interaction between AI application and accuracy in the two-way ANOVA was nonsignificant, $F(1,195) = 0.9$, $p = 0.77$, indicating that accuracy of the AI recommendation had a similar effect on trust for both types of AI application.

3.3 Research question 3

Our final research question asked whether participants rely to a greater extent on recommendations from the LLM AI-assistant than the non-LLM assistant, and if so, whether this difference in reliance is influenced by reported trust. A chi-square test of independence was performed to examine the relation between AI application and reliance at each trial. The relation between these variances was

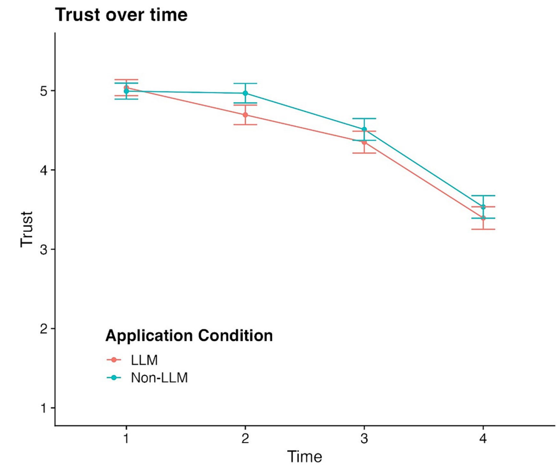

FIGURE 1
Reported trust in LLM and non-LLM AI across trials.

nonsignificant for each of Trial 1, x^2 (1, $N = 199$) = 2.13, $p = 0.14$, Trial 2, x^2 (1, $N = 199$) = 1.39, $p = 0.24$, and Trial 3, x^2 (1, $N = 199$) = 1.16, $p = 0.28$. This indicates no significant difference in reliance as an outcome of the type of AI recommendation, and consequently we did not test the influence of reported trust.

4 Discussion

In an experimental design where participants were given AI-generated answers to general knowledge questions we found no difference in reported trust between recommendations generated by an LLM and those provided by a non-LLM form of AI. When informed that the AI-assistant recommendation was inaccurate, participants subsequently reported less trust in the assistant, but the extent of the reduction in trust was not systematically different across the two types of AI application. We also did not find evidence of greater or lesser reliance on LLM-generated recommendations.

Taken together, these findings suggest that trust in LLM output is not artificially inflated, and people are no more likely to miscalibrate their trust in this novel technology than other forms of AI. This would appear to be good news for the individuals and organizations already incorporating LLMs into their knowledge workflows, nonetheless we advise caution. The scope of this study was deliberately limited to the content and form of language presented in the recommendations, yet there are a number of other elements of LLM operation that may have a bearing on trust calibration. This includes the interactivity inherent in providing queries in one's own words and seeing a response appear as if from an oracle. Further research could examine this possibility in a design delivering the same content in a static or interactive form. There may also be reputational effects on trust when engaging with LLMs (Buchanan and Hickman, 2023). Our participants were not advised of the type of AI-assistant providing the recommendations. Holding the content constant and varying information about the source of the recommendation would provide insight into this possibility. It may also be worth exploring the temporal aspects of any such effect as the reputational stocks of various LLMs fluctuate (Centre for Data Ethics and Innovation, and Department for Science, Innovation & Technology, 2024).

This study also examined responses to LLM output in a limited set of general knowledge tasks, but it is worth noting that LLMs have been incorporated into an array of tasks including creative work (Heaven, 2022) and coding (Sun et al., 2022). Factual errors in LLM output may have different implications for different forms of work, leading to important differences in trust dynamics. Future research identifying the dynamics of trust across varying LLM task-types will be necessary to appropriately evaluate the generalizability of the findings reported here.

Effective calibration of human trust levels to the objective capabilities of an AI system is critical to the efficient and safe use of these tools. However, trust is influenced by a wide range of factors beyond system performance, and the presence of these factors has the potential to artificially inflate trust and disrupt trust calibration. In this paper we investigated whether this potential for disruption of trust calibration was evident in specific aspects of use of LLMs, a highly impactful new form of AI technology. We did not find evidence of different patterns of trust calibration in response to the content generated by an LLM as compared to content sourced by a non-LLM AI system, but recommend further research into the specific characteristics that distinguish these forms of AI.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://data.csiro.au/collection/csiro:63027.

Ethics statement

The studies involving humans were approved by CSIRO Social Science Human Research Ethics Committee. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

MM: Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Writing – original draft, Writing – review & editing. PC: Methodology, Visualization, Writing – review & editing. AD: Conceptualization, Methodology, Resources, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

ChatGPT-3 and Google Search's AI assistant were used as part of this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2024.1456098/full#supplementary-material

References

Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., and Schulz, E. (2023). Playing repeated games with large language models. arXiv [Preprint]. doi: 10.48550/arXiv.2305.16867

Bohannon, M. (2023). Lawyer used ChatGPT in court—And cited fake cases. A judge is considering sanctions. Forbes. Available online at: https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/ (Accessed June 18, 2024).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). "Language models are few-shot learners" in *34th Conference on Neural Information Processing Systems*, Vancouver, Canada.

Buchanan, J., and Hickman, W. (2023). Do people trust humans more than ChatGPT? *SSRN Electron J*. doi: 10.2139/ssrn.4635674

Castaldo, S., Premazzi, K., and Zerbini, F. (2010). The meaning(s) of trust. A content analysis on the diverse conceptualizations of trust in scholarly research on business relationships. *J Bussiness Ethi.* 9, 104–117. doi: 10.1007/s10551-010-0491-4

Centre for Data Ethics and Innovation, and Department for Science, Innovation & Technology (2024). Public attitudes to data and AI: Tracker survey (wave 3). Government of the United Kingdom. Available online at: https://www.gov.uk/government/publications/public-attitudes-to-data-and-ai-tracker-survey-wave-3/public-attitudes-to-data-and-ai-tracker-survey-wave-3 (Accessed June 21, 2024).

Grant, D. (2023). Harnessing AI and ChatGPT technology: the next industrial revolution. Forbes. Available online at: https://www.forbes.com/sites/forbestechcouncil/2023/09/12/harnessing-ai-and-chatgpt-technology-the-next-industrial-revolution/ (Accessed June 21, 2024).

Gupta, A. (2023). Professor fails his entire class because ChatGPT said this | today news. Mint. Available online at https://www.livemint.com/news/world/professor-consults-with-chatgpt-fails-his-entire-class-because-ai-claimed-this-11684380038694.html (Accessed June 21, 2024).

Heaven, W. D. (2022). Generative AI is changing everything. But what's left when the hype is gone? MIT Technology Review. Available online at: https://www.technologyreview.com/2022/12/16/1065005/generative-ai-revolution-art/ (Accessed June 21, 2024).

Herbert, T. (2023). Academics apologise for AI blunder implicating big four | AccountingWEB. Accoutingweb. Available online at: https://www.accountingweb.co.uk/tech/tech-pulse/academics-apologise-for-ai-blunder-implicating-big-four (Accessed June 21, 2024).

Hoff, K. A., and Bashir, M. (2015). Trust in automation: integrating empirical evidence on factors that influence trust. *Hum Factors* 57, 407–434. doi: 10.1177/0018720814547570

Huang, S.-H., Lin, Y.-F., He, Z., Huang, C.-Y., and Huang, T.-H. K. (2024). "How does conversation length impact User's satisfaction? A case study of length-controlled conversations with LLM-powered Chatbots" in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–13.

Huschens, M., Briesch, M., Sobania, D., and Rothlauf, F. (2023). Do you trust ChatGPT?—perceived credibility of human and AI-generated content. arXiv [Preprint]. doi: 10.48550/arXiv.2309.02524

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., et al. (2023). Survey of hallucination in natural language generation. *ACM Comput Surv* 55, 1–38. doi: 10.1145/3571730

Jian, J.-Y., Bisantz, A. M., and Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *Int J Cogn Ergon* 4, 53–71. doi: 10.1207/S15327566IJCE0401_04

Kreps, S., McCain, R. M., and Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *J Exp Polit Sci* 9, 104–117. doi: 10.1017/XPS.2020.37

Lee, J. D., and See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Hum Factors* 46, 50–80. doi: 10.1518/hfes.46.1.50.30392

Morrissey, K., and Kirakowski, J. (2013). ""Realness" in chatbots: establishing quantifiable criteria" in Human-Computer Interaction. Interaction Modalities and Techniques. ed. M. Kurosu, vol. *8007* (Berlin Heidelberg: Springer), 87–96.

Palan, S., and Schitter, C. (2018). Prolific.ac—a subject pool for online experiments. *J Behav Exp Financ* 17, 22–27. doi: 10.1016/j.jbef.2017.12.004

Parasuraman, R., and Riley, V. (1997). Humans and automation: use, misuse, disuse, abuse. *Hum Factors* 39, 230–253. doi: 10.1518/001872097778543886

Robinette, P., Li, W., Allen, R., Howard, A. M., and Wagner, A. R. (2016). "Overtrust of robots in emergency evacuation scenarios" in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 101–108.

Strzelecki, A., and Rutecka, P. (2020). Direct answers in Google search results. IEEE Access 8, 103642–103654. doi: 10.1109/ACCESS.2020.2999160

Sun, J., Liao, Q. V., Muller, M., Agarwal, M., Houde, S., Talamadupula, K., et al. (2022). "Investigating explainability of generative AI for code through scenario-based design" in *27th International Conference on Intelligent User Interfaces*, 212–228.