"Is Stack Overflow Obsolete? An Empirical Study of the Characteristics of ChatGPT Answers to Stack Overflow Questions"

by Samia Kabir, David N. Udo-Imeh, Bonan Kou, and Tianyi Zhang

Reviewed By: Abdur Rafay Saleem

Paper Summary:

This paper presents an in-depth analysis of the characteristics and quality of answers generated by ChatGPT, a large language model, to programming questions on Stack Overflow. The authors aim to investigate the correctness, consistency, comprehensiveness, and conciseness of ChatGPT answers compared to human-written answers on Stack Overflow. They conduct a manual analysis of 517 ChatGPT answers, a large-scale linguistic analysis using automated methods, and a user study with 12 participants to understand programmers' perceptions of ChatGPT answers. The results show that more than half of ChatGPT answers contain incorrect information, and many answers suffer from verbosity and inconsistency. However, participants in the user study still preferred ChatGPT answers 35% of the time due to their comprehensiveness and well-articulated language style.

Critical Evaluation Criteria:

A. Soundness of Approach:

The paper employs a mixed-methods research design, combining manual analysis, linguistic analysis, and user studies, which is a robust approach for investigating the characteristics of ChatGPT answers. The manual analysis follows a standard NLP data labeling process, and the codebook is iteratively refined to ensure reliability. The linguistic analysis uses well-established tools like LIWC and sentiment analysis, providing valuable insights into the language structure and attributes of ChatGPT answers. The user study protocol is well-designed, with a within-subjects design and a semi-structured interview to capture user perceptions and preferences.

However, there are a few limitations to the approach. The manual analysis relies on subjective coding, which may introduce some bias despite efforts to ensure reliability. The user study has a relatively small sample size of 12 participants, which may limit the generalizability of the findings. Additionally, the paper does not provide detailed information about the prompts used to generate ChatGPT answers, which could impact the replicability of the study.

B. Novelty:

The paper is the first comprehensive study that investigates the characteristics and quality of ChatGPT answers to programming questions on Stack Overflow. While previous studies have examined the usability and effectiveness of ChatGPT in other domains, this work bridges the gap by focusing on the programming domain and comparing ChatGPT answers with human-written answers. The study provides novel insights into the correctness, consistency, comprehensiveness, and conciseness of ChatGPT answers and the factors influencing programmers' preferences between ChatGPT and Stack Overflow.

However, the novelty of the study could be further emphasized by highlighting the specific contributions and insights that go beyond previous work on large language models and their limitations.

C. Clarity of Relation with Related Work:

The related work section provides a clear overview of the existing literature on misinformation generated by large language models, help-seeking behavior of programmers, and human-AI collaboration in programming. The authors position their work in the context of previous studies, highlighting the lack of comprehensive analysis of ChatGPT answers to programming questions. They also discuss the limitations of existing approaches for mitigating hallucinations and improving the reasoning capabilities of language models, motivating the need for their study.

D. Quality of Evaluation & Results:

The evaluation is rigorous and technically sound. The manual analysis is conducted on a statistically significant sample of 517 ChatGPT answers, ensuring the generalizability of the findings. The linguistic analysis is performed on a large dataset of 2000 randomly sampled Stack Overflow questions, providing a comprehensive understanding of the linguistic characteristics of ChatGPT answers. The user study involves 12 participants with diverse programming backgrounds, and the thematic analysis of the interview transcripts reveals insightful patterns and themes. The results are well-presented and support the authors' claims, providing a detailed taxonomy of fine-grained issues in ChatGPT answers.

However, the evaluation has some limitations. The manual analysis focuses on a specific set of quality aspects (correctness, consistency, comprehensiveness, conciseness), but there may be other relevant factors that are not considered. The user study relies on self-reported data, which may be subject to biases and limitations in participants' ability to accurately assess the quality of answers. The paper could benefit from a more critical discussion of the limitations and potential threats to validity.

E. Ability to Replicate:

The paper provides sufficient detail to enable independent replication of the study. The authors make their data and codebooks publicly available, including the ChatGPT answers, manual analysis annotations, and interview transcripts. The methodology section describes the data collection, manual analysis, linguistic analysis, and user study procedures in detail, allowing other researchers to reproduce the study.

However, the lack of detailed information about the prompts used to generate ChatGPT answers may hinder exact replication of the study. Providing the specific prompts or a more detailed description of the prompting strategy would enhance the replicability of the study.

F. Quality of Presentation:

The paper is well-written and organized, meeting the standards of the research community. The introduction clearly motivates the research questions and highlights the significance of the study. The methodology section provides a detailed description of the data collection, manual analysis, linguistic analysis, and user study procedures. The results are presented in a logical order, with clear subheadings and informative figures and tables.

The discussion section provides valuable insights into the implications of the findings and future directions for research and practice. The limitations of the study are also acknowledged and discussed transparently.

However, there are a few areas that could be improved. Some of the figures and tables could benefit from more detailed captions and explanations to enhance readability. The discussion section could be expanded to provide a more critical analysis of the findings and their implications, considering alternative explanations and potential limitations.

Conclusion:

In conclusion, this paper makes a significant contribution to the understanding of the characteristics and quality of ChatGPT answers to programming questions. The mixed-methods approach, combining manual analysis, linguistic analysis, and user studies, provides a comprehensive and rigorous evaluation of ChatGPT answers. The findings have important implications for the design of responsible conversational chatbots and the development of strategies to mitigate misinformation in AI-assisted programming.

However, the study has some limitations, such as the subjectivity of manual analysis, the small sample size of the user study, and the lack of detailed information about prompts used to generate ChatGPT answers. The paper could benefit from a more critical discussion of these limitations and potential threats to validity.

Overall, this is a high-quality research paper that advances our understanding of the challenges and opportunities of using ChatGPT for programming tasks, while also acknowledging areas for improvement and future research.