Review of: "Can ChatGPT code the technical part of a Bachelor's Thesis in Informatics?"

Sonia Bergamaschi[1]

[1] Università degli Studi di Modena e Reggio Emilia

Potential competing interests: No potential competing interests to declare.

The presented paper explores the potential of ChatGPT in aiding code development within the context of a Bachelor's thesis focusing on "Information Systems." While the paper is certainly interesting, it lacks details, clarity, and organization, limiting its contribution to the literature.

Clarity:

- The introduction to LLMs and their training is highly limited. It would strengthen this section to delve into the specifics of deep learning techniques used in LLM training *["Obviously, the developers of ChatGPT trained their LLM using deep learning" from 1. Introduction].*
- A critique about generalization metrics has been brought up, but no mention is made of the expansive literature on testing sets or benchmarks used to assess model generalizability and performance. *["These discrepancies, termed "loss," form the foundation for the backpropagation algorithm (Zhang, Bengio, Hardt, Recht, and Vinyals, 2021). Moreover, juxtaposing accuracies from test and training datasets offers insights into the model's generalization capabilities. Nonetheless, while these metrics are definitive within the machine learning realm, they offer limited insights into the broader applicability of an LLM" from 1. Introduction]*
- The claim about extensive research on human-human collaboration needs substantiation with citations.*["While challenges in collaboration have been extensively explored in the context of human-human interactions" from 1. Introduction]*
- ChatGPT's version should be reported, as it is essential for reproducibility and understanding the model's limitations at the time of the study.
- A literature review on LLM's usage in code generation should be given in Section 1, Introduction, since it is central to the research.

Details and Organization:

- Section 2.1 Data Collection should be expanded or a new section created to incorporate examples of the listed human-AI interactions. *["While this interface can accommodate a wide range of queries, we primarily used it in our study to define problems, seek clarifications, request code, or report code-related errors" from 2.1 Data Collection]*
- Section 2.2 Code Evaluation should be expanded or a new section created to incorporate examples and further

Qeios ID: T5ESCH · https://doi.org/10.32388/T5ESCH

categories of errors found to exemplify the problems encountered. *["We adopted one of two approaches: if the error was evident or within our capacity to rectify, we addressed it directly; otherwise, we relayed the issue to ChatGPT for further guidance." from 2.2 Code Evaluation]*
- A new section should be introduced discussing the "critical dialog" as it is a cornerstone of the research conducted. We suggest going in depth into the type of interactions that the students had with the model, categorizing them, and providing examples.
- A more detailed view of the errors encountered in tables 1, 2, and 3 should be reported using the previously mentioned categorization of errors. *[Tables from 3.1 Critical Dialog]*
- In section 3.1 Critical Dialog, a fine-tuning problem has been presented but marked as solved given proper context to ChatGPT *["Once we provided ChatGPT with the correct context, the project resumed smoothly, and the results achieved consistency."]*. However, in a later section, specifically 4. Discussion, the same "technical oversight" has been marked as unresolved *["Secondly, we encountered a significant technical oversight on ChatGPT's part: it incorrectly assumed two classes instead of three while fine-tuning the RoBERTa model. Despite our considerable efforts, this error remained unresolved by ChatGPT." from 4. Discussion]*. Clarification is required.

Concerns:

- In section 5. Conclusion, it is stated that *"One could contend that exposure to novel technologies (in this case, deep learning) via ChatGPT offers students a valuable learning experience"*. This assumption needs empirical support. Pre- and post-assessments have to be conducted to verify whether the "critical dialog" resulted in a real improvement of the students' knowledge or ability in the subject.

Style:

- We suggest changing citation formatting as each reference often occupies whole lines, making the text difficult to follow. *[e.g. "a refined variant of BERT (Devlin, Chang, Lee and Toutanova, 2018; Touvron, Lavril, Izacard, Martinet, Lachaux, Lacroix, Rozière, Goyal, Hambro, Azhar et al., 2023)" from 3.1 Critical Dialog]*
- Chapter 3.1 Critical Dialog should be divided into further sections as it can be segmented into four distinct phases. *["The critical dialog with ChatGPT can be segmented into four distinct phases" from 3.1 Critical Dialog]*