Review of: "OpenAI ChatGPT Generated Content and Similarity Index: A study of selected terms from the Library & Information Science (LIS)"

Silvia Casola[1]

1 Fondazione Bruno Kessler

Potential competing interests: No potential competing interests to declare.

The paper addresses whether content generated by ChatGPT will be flagged as plagiarized by popular automatic plagiarism software. Authors generate content for ten popular terms in the LIS domain and report that the Turnitin plagiarism checker assigns a similarity index of only 13% to the generated text.

The work addresses an interesting question in a timely manner.

There are, however, some possibilities for improvement:

1. Adding more details on the tools used would be helpful. ChatGPT is currently being actively developed (with changes introduced over time), so it would be important to report the date experiments were performed. Moreover, the exact prompt the authors used should be specified. In the case of "Library and Information Science," for example, did authors only input "Library and Information Science" or a more dialogic prompt, e.g., "Can you explain to me what Library and Information Science mean?"?
2. I would be interested, as a reader, to find more details on how the Turnitin software works. While Turnitin is a commercial solution (and its exact implementation details are likely private), the authors might explain how the software computes the similarity, e.g., based on tokens; details on which kind of documents are used for comparison would also be of help. The paper could also report why Turnitin is chosen over other similarity checkers, e.g., in terms of availability or popularity.
3. In the result Section, it would be interesting to see similarity scores per topic so that the reader can easily understand whether there is variability between topics or whether the similarity index is relatively constant.
4. As the authors note, ten topics is a relatively small sample size. A larger number of topics could be considered to improve the paper. At the same time, the current small sample size would allow for a better qualitative analysis of the results. For example: are answers generally correct/factual? What are the portions of the responses flagged by the similarity checker software? Are they content-related or usually contain standard terms and expressions? Some of these considerations could be made from the included picture with the outputs, but it would be interesting to have the authors insights (considering they are domain experts).
5. Similarly, considering a larger set of software could improve the paper.

On a more presentation-related note, some improvements can be made. For example, the authors should reference

transformer models (Vaswani et al., 2017); while no paper describing ChatGPT exists to date, to my best knowledge, authors could include references to Instruction GPT (Ouyang et al., 2022), to which the model is largely inspired. Finally, the Literature Review section could mention the line of research on flagging automatically-generated text.

I am also interested in the authors' insights on the repercussions of their findings: do they believe their findings show that plagiarism checkers should be improved/updated? How will their findings impact students and Universities?

Recommended references:

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems (p./pp. 5998--6008).
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L.E., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., & Lowe, R.J. (2022). Training language models to follow instructions with human feedback. *NEURIPS*