Review of: "Towards Responsible AI-Assisted Scholarship: Comparative Assessment of Generative Models and Adoption Recommendations"

Javier Bueno[1]

[1] Universidad de Alcalá de Henares

Potential competing interests: No potential competing interests to declare.

The selected topic in the paper is very interesting given the rise in the use of generative language models in various educational contexts. The use of various artificial intelligence-based tools, such as ChatGPT, has gained popularity over the past year, and it is necessary to evaluate their strengths and weaknesses in order to make responsible use of these technologies.

Positive aspects of the work include the choice of four current models, less popular than ChatGPT, on which to base the conducted study and analyze their strengths and weaknesses in academic use.

Aspects that would need improvement to achieve greater robustness in the study include the following:

- In the Methodology section:
  - Better explain the scope of the questions in the tests conducted, as well as provide some examples of the questions asked in each of the categories in an annex.
  - Provide a more detailed explanation of how the external researchers who participated in evaluating the quality of the responses from the language models under study were selected, as well as the experts who answered the same questions as the AIs. The paper does not make it clear whether these are the same individuals, which could introduce a bias in the responses and potentially invalidate part of the study.
- In the Quantitative Benchmarking section, specify the number of questions asked in each of the evaluated categories and, in addition to the average, indicate the standard deviation in the scores of the responses.
- In the Qualitative Thematic Benchmarking section, explain how risks are identified and the disciplines evaluated in the sub-section on limitations.
- Provide a clearer explanation of the relationship between the Findings section and the described experiment, allowing for more concise conclusions to be drawn.

Finally, some inconsistencies have been detected between the cited bibliography and the accompanying text of the citation, such as the article by Ghahramani (2015) and the ten academic research categories used for the experiment, which are not mentioned in that paper. Therefore, it would be advisable to clarify their source.