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

Gilbert Lim

Potential competing interests: No potential competing interests to declare.

This paper assesses four state-of-the-art publicly-available Generative AI systems/models, on their competencies across ten scholarly core competency tasks. Neutral free-text prompts were designed for each task, and independently scored by two trained raters on a 10-point scale with substantial interrater reliability. While the study provides valuable insights into the relative strengths of the four models, there are some points that might be further considered:

1. The omission of ChatGPT from the evaluation is glaring, given its current primacy in the generative AI for text domain.

2. The number of prompts for each task might be stated.

3. In the Quantitative Benchmarking subsection, it is stated that the AI systems' prompt responses were independently scored by two trained raters. Any scoring rubric/methodology might be described.

4. It is then stated that "Human experts outscored the AI systems (8.9/10)". It might be clarified as to what task(s) this outscoring was on; if the human experts had indeed produced responses for the prompts too, their full results might be included in Table 1.

5. For the Findings section, example prompts & responses by the four AI systems (and associated rater scores) might be included, possibly in an appendix.