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

Thomas Hanne[1]

1 University of Applied Sciences and Arts Northwestern Switzerland

Potential competing interests: No potential competing interests to declare.

The paper discusses an experimental evaluation of four large language models across 10 academic tasks based on a qualitative and quantitative evaluation. Capabilities, gaps, risks, and validation needs of using such models are discussed.

Major remarks

The contribution could be more clearly specified (such as in the abstract).

The experimental setup is not clear enough. For instance, I would expect a full description of used prompts and not just 2 examples.

Are the detailed data available somewhere?

Also details of the qualitative thematic analysis are missing.

The level of details regarding the investigation is in sharp contrast to the lengthy discussion which often does not seem to be directly related to the investigation.

Detailed remarks

p. 2 “This constrains understanding system differences based on training methodologies.” – not clear enough.

“multiple diverse generative” – skip “multiple” or “diverse”.

p. 4: “neutral prompts” – What is meant by “neutral”?

“to results communication, providing comprehensive coverage” – not clear enough

“Briefly summarize the key findings from this research summary” - So a summary of a summary is expected?

“Identifying information was anonymized for secure collection” – not clear enough

adept at” – something missing?

Bibliographic data is frequently incomplete (e.g., for the first two references).