May Generative AI Be a Reviewer on an Academic Paper?*

Haichen Zhou[1,*], Xiaorong Huang[1], Hongjun Pu[1],and Zhang Qi[2]

[1] National Science Library (Chengdu), Chinese Academy of Sciences, Qunxian South Street 289, 610041 Chengdu, China
[2] Nanjing University, Xianlin Avenue 163, 210023 Nanjing, China

Abstract

The application of artificial intelligence (AI) to academic evaluation is one of the important topics within the academic community. The widespread adoption of technologies such as Generative AI (GenAI) and Large Language Models appears to have introduced new opportunities for academic evaluation. The question of whether GenAI has the capability to perform academic evaluations, and what differences exist between its abilities and those of human experts, becomes the primary issue that needs to be addressed first. In this study, we have developed a set of evaluation criteria and processes to investigate on 853 post peer-reviewed papers in the field of cell biology, aiming to observe the differences in scoring and comment styles between GenAI and human experts. We found that the scores given by GenAI tend to be higher than those given by experts, and the evaluation texts lack substantive content. The results indicate that GenAI is currently unable to provide the depth of understanding and subtle analysis provided by human experts.

Keywords

academic evaluation, Generative AI, large language models, Copilot, ChatGPT

1. Introduction

How to use AI for more objective, accurate, and efficient academic evaluation has become an important research topic[1][2]. Generative AI (GenAI) is a novel technology that uses artificial intelligence to generate content in various forms[3][4]. In the context of academic evaluation, GenAI provides a new possibility for automating academic evaluations by generating academic evaluation content[5]. Comparing the evaluations of human experts and those generated by GenAI is a very intuitive way to better understand the effectiveness and reliability of GenAI. However, there is still a lack of research on the quality of the content generated by GenAI and whether there are differences between it and the content generated by human experts. Figuring out these issues provides a basis for us to answer whether GenAI can match the depth of understanding and subtle analysis provided by human experts, what areas GenAI excels in, and where it may need further improvement.

Hence, we focus on analyzing the difference between human expert evaluations and GenAI evaluations. We aim to answer the following research questions:

RQ1: Can GenAI conduct academic evaluations?
RQ2: What differences exist between the scoring results of GenAI and human experts?
RQ3: What differences exist between the evaluation text features of GenAI and human experts?

2. Method

Our research methodology includes the following steps:

1. Select papers from H1 Connect (connect.h1.co) as research cases and establish selection criteria.
2. Generate a list of papers to be collected.

Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2024) and the 4th AI + Informetrics (AII2024)
* Corresponding author.
mailto:zhouhc@clas.ac.cn (H. Zhou); huangxiaorong@clas.ac.cn (X. Huang); puhj@clas.ac.cn (H. Pu) zhang@smail.nju.edu.cn (Q. Zhang)
0000-0002-3366-1951 (H. Zhou); 0000-0002-9164-0585 (X. Huang); 0000-0003-4787-519X (H. Pu) 0000-0001-5401-2275 (Q. Zhang)

3. Obtain data such as Paper Title, DOI, Expert Score, Review Text, etc., to form the original dataset.
4. Design evaluation dimensions and scoring system for the research field of our dataset.
5. Generate the Copilot question template (Prompt).
6. Use the template to ask questions and collected Copilot scores and evaluation text data.
7. Compare the differences in scores and texts between Copilot and experts.

2.1. Data Preparation

To minimize the influence of various factors on the evaluation results, such as the differences in evaluation standards for papers in different fields, newly published papers not yet receiving sufficient attention, and differences in evaluation preferences among different experts, we have limited the research field to Cell Biology. We focused on papers from cell biology published in 2020 that received one evaluation. We collected data on 853 papers (as of May 2022) from H1 Connect. H1 Connect is a leading platform for researchers and clinicians seeking expert opinions and insights on the latest life sciences and medical research. We collected key information about the papers, including paper title, authors, journal, DOI, PMID, recommended score etc.

2.2. Question Template Design

We designed an evaluation system specifically for the field of Cell Biology to enhance the relevance and reliability of the content generated by Copilot. By summarizing the review principles of top journals in this field, such as "Nature Reviews Molecular Cell Biology", "Trends in Cell Biology", "The Journal of Cell Biology", "Nature Cell Biology", and "Journal of Molecular Cell Biology", we extracted the following evaluation dimensions. Copilot will be required to evaluate each paper on these dimensions, provide a recommendation score, and finally give a comprehensive evaluation.

After several rounds of testing, the final template (prompt) for querying Copilot has been established as follows, with the inclusion of PubMed ID to assist Copilot in accurately targeting information on the internet:

I have summarized a set of criteria for evaluating academic papers:

*•**Originality:** The paper must report novel, innovative and influential research that does not repeat or plagiarize existing work.*

*•**Accuracy:** The paper must follow high standards of experimental design, data analysis and result presentation, without errors, biases or misleading.*

*•**Conceptual advance:** The paper must provide a deep understanding and mechanistic explanation of an important problem or area, not just superficial or incremental improvements.*

*•**Timeliness:** The paper must reflect the current hot topics in the scientific community.*

*•**Significance:** The paper must have immediate or long-term impact and implications.*

I also have a recommended scoring system: 1 star (Good), 2 stars (Very Good) , 3 stars (Exceptional) .You're acting as a scientist. I'll give you a PubMed ID for the paper. First, please display the title of the paper and search the web site. No abstract is required. Second, please according to my criteria and scoring system for evaluation and scoring the paper; Third, please according to my scoring system for the overall evaluation and scoring of the paper. Pubmed ID:XXXXXXXX

2.3. Collection of Evaluation Results

The process of collecting Copilot evaluation results is shown in Figure 1:

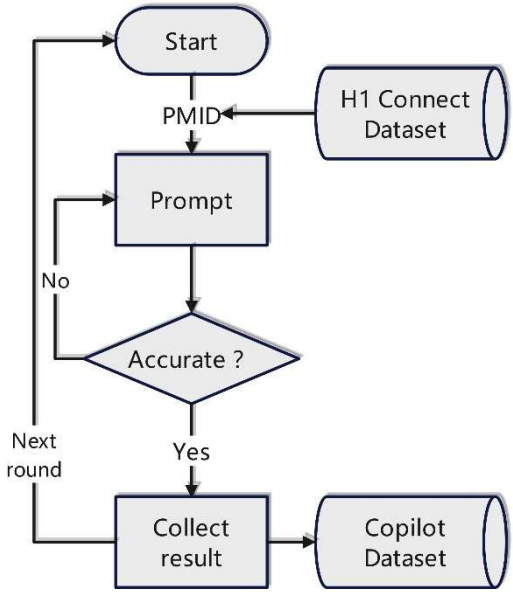

Figure 1: Flow chart of collecting Copilot evaluation results.

3. Result and Discussion

RQ1: Can GenAI conduct academic evaluations?
GenAI can conduct academic evaluations and produce readable results in form.

RQ2: What differences exist between the scoring results of GenAI and human experts?
Observing the score distribution ratio (Figure 2), papers scored 3 stars by experts only account for 15%, while those scored 2 stars and 1 star are both around 40%. However, Copilot's 3 stars evaluations account for more

than 60%, 2 stars account for 32.72%, and 1 star is less than 1%.

Figure 2: Scoring ratio between Copilot and expert.

Comparing the scoring results of Copilot and expert (Figure 3), we observe that: Copilot's scoring results are higher, indicating that it tends to give higher scores to most papers. The average score given by Copilot is 2.68 stars, while the average score given by experts is 1.76 stars. This is consistent with the experimental results of Mike Thelwall on 51 papers[2]. The fact that over 60% of papers are scored 3 stars by Copilot suggests that it may not yet possess the core ability to accurately distinguish high-value academic papers.

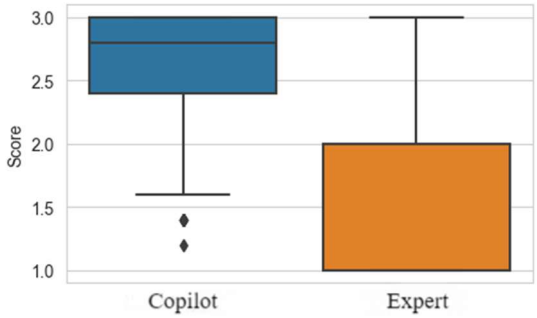

Figure 3: Scoring box plot between Copilot and expert.

Observing the source of score differences (Figure 4), we find that most papers scored 1 star by experts were scored 3 stars by Copilot, with a small portion scored 2 stars. Nearly half of the papers scored 2 stars by experts were scored 3 stars by Copilot, and the other half were scored 2 stars. Most papers scored 3 stars by experts were also scored 3 stars by Copilot. In summary, the main disagreements occur with papers scored 1 star by experts, while disagreements are lower for papers scored 3 stars by experts.

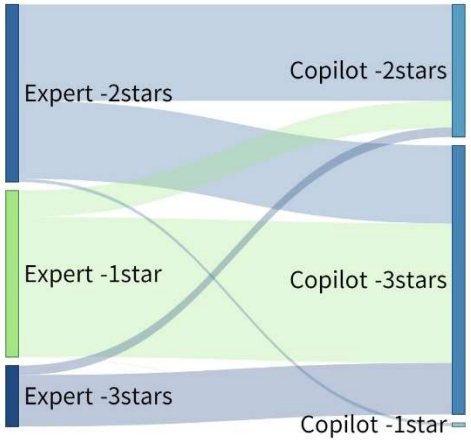

Figure 4: The source of score differences between Copilot and expert.

RQ3: What differences exist between the evaluation text features of GenAI and human experts?

From the perspective of sentences, both the number of sentences and the average sentence length in the Copilot text are less than/shorter than those in the expert text, but the difference is not significant (Figure 5).

From a lexical perspective, the overall proportion of word types between the two is not significantly different, with Copilot tending to use more adjectives (Figure 6). The high-frequency words used by experts better reflect professionalism and specificity, such as "cell", "protein", and "cancer". In contrast, the high-frequency words used by Copilot are more general, such as "significant" (Table 1).

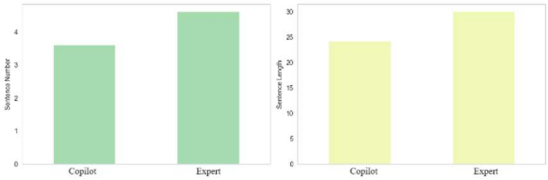

Figure 5: The number of sentences and the average sentence length between Copilot and expert.

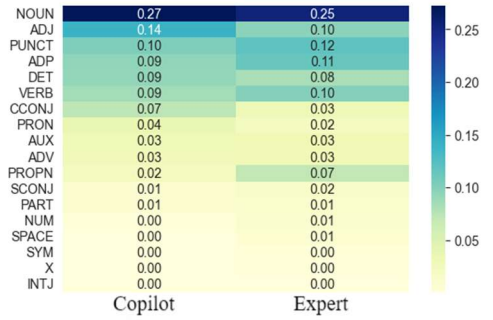

Figure 6: The proportion of word types between Copilot and expert(the meaning of abbreviations is in Appendix).

Table 1
Word counts between Copilot and expert

Word (Copilot)	Counts	Word (Expert)	Counts
paper	2539	cell	1091
research	749	protein	697
novel	677	study	481
report	560	author	436
provide	461	cancer	309
high	444	bind	295
field	421	gene	285
implication	394	expression	244
cell	354	increase	243
significant	335	role	240

4. Conclusion

In this study, we collected post peer-review scores and text data for 853 papers in the field of Cell Biology from the H1 Connect website. We also obtained the scores and evaluation text data for each of these papers from Copilot, based on our designed evaluation criteria and process. By comparing the evaluation results of Copilot and experts through quantitative analysis and text mining methods, we found that Copilot can score and evaluate under specific prompts. From the scoring perspective, there is a significant difference in the scoring patterns between Copilot and experts: the former tends to give higher star, but the high proportion of 3-star reveals that it does not have enough ability to judge the actual value of the paper. From the text perspective, Copilot's shorter sentences and generic wording indicate that its evaluation is only at the stage of imitating the features of the evaluation text, and it cannot yet carry out substantive evaluations of the originality, accuracy, and other core elements of the paper.

Overall, GenAI, represented by Copilot, is currently unable to provide the depth of understanding and subtle analysis provided by human experts. It should still not be used for academic evaluation at this stage, as its over-evaluative nature may lead to the proliferation of low-quality academic results[6]. This study presents several limitations: Firstly, a disparity exists between experts and Copilot, with the latter unable to access complete paper texts, unlike experts. Secondly, the author did not perform iterative testing nor utilized the mean outcomes of various expert assessments for analysis. Third, the evaluation criteria of Copilot and experts are not consistent. The limitations bear potential inaccuracies for the research outcomes. Consequently, we aim to address these deficiencies in the subsequent phase of analysis.

Acknowledgements

This work was supported by the key project of innovation fund from National Science Library (Chengdu), the Chinese Academy of Sciences (E3Z0000902). We sincerely appreciate the insightful comments and constructive suggestions provided by the reviewers, which have significantly contributed to the improvement of our manuscript.

References

[1] W. Liang, Y. Zhang, H. Cao, et al, Can large language models provide useful feedback on research papers? A large-scale empirical analysis, 2023. URL: http://arxiv.org/abs/2310.01783.

[2] M. Thelwall, Can ChatGPT evaluate research quality?, 2024. URL: http://arxiv.org/abs/2402.05519.

[3] Bloomberg, Generative AI to Become a $1.3 Trillion Market by 2032, Research Finds, 2023. RUL: https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/

[4] Gartner, Understand and Exploit GenAI with Gartner's New Impact Radar, 2024. URL: https://www.gartner.com/en/articles/understand-and-exploit-gen-ai-with-gartner-s-new-impact-radar.

[5] J. de Winter, Can ChatGPT be used to predict citation counts, readership, and social media interaction? An exploration among 2222 scientific abstracts, J. Scientometrics. (2024).

[6] M. B. Garcia, Using AI tools in writing peer review reports: should academic journals embrace the use of ChatGPT ?, Annals of biomedical engineering 52, 2024: 139-140.

A. Abbreviations and Examples of SpaCy Parts-of-Speech

- ADJ--adjective *big, old, green*
- ADP--adposition *in, to, during*
- ADV--adverb *very, tomorrow, where *
- AUX--auxiliary *is, has (done), will (do) *
- CCONJ--coordinating conjunction *and, or, but*
- DET--determiner *a, an, the*
- NOUN--noun *girl, cat, tree, air, beauty*
- PUNCT--punctuation *., (,), ?*
- VERB--verb *run, runs, running, eat, ate, eating*