Article

Not peer-reviewed version

Revolutionizing Peer Review: A Comparative Analysis of ChatGPT and Human Review Reports in Scientific Publishing

Jovan Shopovski *, Raihana Mohdali, Dejan Marolov

Posted Date: 3 February 2025

doi: 10.20944/preprints202502.0058.v1

Keywords: ChatGPT; peer review; LLM peer review; manuscript evaluation; LLM review reports

Article

Revolutionizing Peer Review: A Comparative Analysis of ChatGPT and Human Review Reports in Scientific Publishing

Jovan Shopovski [1,*], Raihana Mohdali [2] and Dejan Marolov [3]

[1] European Scientific Institute, ESI, Macedonia; Grigol Robakidze University, Georgia
[2] Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Malaysia
[3] European Scientific Institute, ESI; Goce Delchev University, Macedonia
* Correspondence: jovanpraven@yahoo.com

Abstract: ChatGPT-4 and other large language models (LLMs) and their use in academic writing have raised questions regarding their capacity to facilitate the peer review process. This research article compares AI-generated peer review reports (using ChatGPT-4 with tools) with traditional review reports generated by humans. The review reports were received for 198 manuscripts submitted to the European Scientific Journal (ESJ) between January and September 2024. Each manuscript underwent a parallel evaluation process. First by human reviewers of the journal and by ChatGPT-4 with tools paid version afterward. However, each manuscript received review reports from at least two human reviewers and only one AI-generated review report. Both review types used the ESJ's standardized evaluation form. The ChatGPT-4 was prompted to review the papers objectively and critically. Statistical analyses were conducted to compare the grades of different parts of the manuscripts, recommendation distributions, and consistency between the AI and human review reports. Kolmogorov-Smirnov test; Pearson Chi-Square test, Mann-Whitney U test; and Cohen's kappa test were used to analyze the data. Results showed that ChatGPT-4 reviewers consistently awarded higher grades and were less rigorous than the human reviewers. The ChatGPT-4 review reports mostly recommended minor revisions and have never recommended rejection of a manuscript. On the other hand, human reviewers demonstrated a more balanced distribution of recommendations, including stricter score evaluations. However, a lack of agreement between the human review reports was registered. While LLM tools can enhance the efficiency of the peer review process, their ability to uphold rigorous academic standards remains limited. Editors who use LLM tools as reviewers have to remain vigilant and not rely their decisions solely on LLM-generated reports. The existing version of ChatGPT-4 is not trained in peer review and cannot replace human expertise in peer review. However, it can be used as an assistant that under human oversight can provide useful comments and recommendations for content improvement of the manuscripts. Future research should focus on LLM tools trained for peer review in various academic fields as well as on the ethical frameworks for LLM integration in peer review.

Keywords: ChatGPT; peer review; LLM peer review; manuscript evaluation; LLM review reports

Introduction

Although the best gatekeeper of scientifically sound research thus far, the peer review process is related to many inconsistencies and flaws (Smith, 2006). Designed as a social project to benefit the scientific community, it relies on human expertise where peers in adequate academic fields critically assess the work of their colleagues (Da Silva & Daly, 2025). As such it is not only a time-consuming process but also a process accompanied by biases on various grounds such as gender, geographical origin of the authors, or confirmatory biases (Kaatz, Gutierrez, & Carnes, 2014); (Lee et al. 2013);

(Mahoney, 1997). The constant increase in the number of submissions and difficulties in securing reviewers negatively affected the efficacy of the peer-review process (Li et al., 2024); (Fox et al. 2016). The peer review is additionally burdened with the exploitative approach from the commercial publishers and the appearance of fake reviewers as part of the paper mills and review manipulation (Da Silva & Daly, 2025). Therefore, efforts to improve the process and cope with the challenges are in constant need.

The introduction of ChatGPT in November 2022 was a disruption in the area of scientific writing and publishing (Conroy, 2023). LLMs became widely used as assistants in scientific writing for preparing various parts of the manuscripts (Liang et al., 2024). With the indisputable benefits for academic writing, the LLM usage is accompanied by biases and errors, and numerous ethical concerns (Gilat & Cole, 2023).

As Generative artificial intelligence (AI) advances, many publishers and editors are interested in exploring LLM's potential to assist in the peer review process in order to overcome barriers. Analyzing the LLMs` language patterns significant increase in LLM-generated content in the review reports is already registered (Cheng et al. 2024). On the other hand, major publishers in their policies prohibit or allow limited LLM usage. Despite the ability of Generative AI to enhance the peer-review procedure concerns of biases, copyright breaches and lack of transparency are omnipresent (Li et al. 2024).

Continuous research on peer review is necessary in order to improve the process itself thus contributing to increased reliability, rigor, and relevance of the published scientific content (Squazzoni et al. 2020). Research on peer review can be crucial, especially in times of disruptions such as the LLM usage in scientific writing and publishing. It can help understand the strengths and weaknesses of the existent LLM tools and their capacity in this area. Therefore, it can facilitate the regulation process of LLM usage in peer review and accelerate the process of preparation of frameworks, policies, and best practices.

In this study, we aim to test the ability of ChatGPT-4 (part of ChatGPT Plus) in reviewing scientific articles. We will compare ChatGPT-generated reviews with those of human reviewers across manuscripts in different scientific fields. The focus is on the rigor and alignment in assessing separate parts of the manuscripts as well as the final recommendations of the reviewers. Moreover, we analyze consistency among human reviewers, providing a baseline for human-to-human alignment that can inform our understanding of LLMs' performance relative to human variability. This research not only adds to the limited body of literature on Generative AI in academic publishing but also addresses fundamental questions about the role of LLMs in supporting or enhancing the peer review process.

Literature Review

Studies confirmed the inconsistencies of human peer reviewers and the subjectivity in peer review. Bolek et al. (2022) reported that reviewers` assessments of separate manuscript sections did not fully align with their final recommendations. Cullen and Macaulay (1992) have found that only 40% of paired human reviewers for a clinical journal were in complete agreement in their review reports. Reviewers often provide divergent assessments of the same manuscript and are usually influenced by factors such as disciplinary background, personal beliefs, and individual interpretation of review criteria (Cicchetti, 1991; Bornmann et al., 2010). This kind of inconsistency imposes the questions of reliability of the peer review and confirms the necessity of additional mechanisms for improvement. Even experienced reviewers can struggle with achieving consensus, especially when evaluating manuscripts that require specialized or interdisciplinary knowledge (Lee et al., 2013; Bornmann et al., 2011).

AI tools have accelerated the peer review process by automating some routine tasks such as language and plagiarism checks. Tools like Grammarly and Turnitin are already widely used for the initial screening as part of the peer review process (Razack et al., 2021).

Large Language Models, LLMs as part of Generative Artificial Intelligence have shown promises and limitations for manuscript preparations. The free commercial usage of ChatGPT, among other things, has changed the process of scientific writing and publishing (Lozic & Stular, 2023). Empirical studies have shown that ChatGPT is the most investigated tool for this purpose and it is a great tool for preparing abstracts, titles, introductions, methods, discussions, and literature reviews of manuscripts. It also accelerates the process and it is an excellent proofreading tool (Kacena et al., 2024); (Huang & Tan, 2023). However, the efficacy of LLMs is often accompanied by errors in references, lack of critical thinking, biases, discrepancies based on academic field, potential plagiarism, and copyright uncertainties (Athaluri et al., 2023); (Altmae et al., 2023). Therefore human oversight remains crucial in the process of scientific writing (Jenko et al., 2024).

Following the assistance in manuscript preparation, recent research articles explore the potential of Large Language Models (LLMs) in the peer review process. The capabilities of ChatGPT in reviewing papers were weaker than its performance in text-writing assistance (Kadi & Aslaner, 2024). As a peer reviewer, ChatGPT has failed to recognize major inconsistencies in articles. Comparing the ChatGPT and human review reports Saad et al. (2024) found a correlation only with the ChatGPT 3.5 version but not the latest ChatGPT 4.0 version.

Hosseini & Horbach (2023) reasonably anticipated that LLMs can improve the peer review process in efficacy and can amend the reviewers` shortage with concerns about LLMs' potential to amplify biases and inequalities. They underlined the role of editors in securing transparent and responsible usage. The need for appropriate prompting strategies can also affect the performance of the LLMs in peer review (Santu et al., 2024). Comparing GPT-4's generated feedback with human peer reviewers, Liang et al. (2024) have found substantial overlap between the ChatGPT and human-generated review reports. Going further, they invited experts to evaluate both types of review reports. More than half (57.4%) of the users have found the ChatGPT-4 generated feedback helpful or very helpful. Surprisingly 82.4% found it even more beneficial than the feedback from at least some human reviewers.

While LLMs show promise in enhancing various aspects of peer review, researchers emphasize the importance of responsible use, disclosure, and further investigation into potential risks and biases associated with their implementation in academic settings. Raising public awareness about LLM usage, implications, and potential outcomes is very important and could facilitate the enactment of appropriate and timely regulation (Smith & Smith, 2023).

Many questions remain open regarding the LLMs` capacity to evaluate academic content with the same rigor as human experts. This research study will add value to the existing research on the LLMs usage in peer review and its performances. It will be one of few studies using representative sample sizes of human-generated and LLM-generated (ChatGPT Plus 4o) review reports for the same manuscripts. It will compare the evaluation scores of both AI and human reviewers and their recommendations and will conclude the consistency and the level of rigorousness of both types of reviewers. Finally, it will set a solid base for discussion and analysis of the LLMs and their future role in peer review of scientific manuscripts.

Methods and Results

Study Design and Dataset

This comparative study analyzed a dataset of 201 manuscripts in various academic fields which were submitted to the European Scientific Journal, ESJ in the period from January to September 2024. As part of the regular peer review process, each manuscript has been reviewed by both human experts and Generative AI (ChatGPT-4). Each manuscript underwent peer review by at least two human reviewers and only one AI review. The total number of human review reports was 512 and the AI reports equal to 201. The standard journal evaluation form was used by both reviewers. The ChatGPT as a reviewer was prompted to use the Evaluation Form and objectively and critically evaluate the manuscript[1]. The manuscripts were anonymized for the ChatGPT-4 in order to provide

[1] The prompt for ChatGPT-4 Plus is presented in Appendix 1.

personal data protection. The human reviewers were familiar with the names of the authors as the journal maintains a single-blind and optional open peer review.

For each manuscript reviewers were instructed to provide evaluation for specific sections (Title, Abstract, Language, Methods, Results, Conclusion, and References) using a numeric grading system (poor 1-5 excellent) and to provide a final recommendation at the end of the review report selecting one of the following options Accept, Minor Revision, Major Revision, or Reject.

Data Collection

The dataset included the following variables:

- **Reviewer Type**: Human or AI (ChatGPT-4).
- **Academic field**: Social Sciences & Humanities or Life/Natural/Medical Sciences.
- **Language of the manuscripts:** English, French, or Spanish.
- **Grades**: Numeric scores for each manuscript section and an overall total grade.
- **Final Recommendations**: Accept, Minor Revision, Major Revision, or Reject.

Three of the 201 manuscripts were excluded from the analysis due to significant missing data, leaving a usable dataset of 198 manuscripts. A Kolmogorov-Smirnov test which is generally better suited for large datasets (n=702) (Zhao, et al., 2017), was used to evaluate whether the data collected from human and AI reviews is normally distributed. This test is more robust for detecting deviations in large datasets as it compares the cumulative distribution of the data to the expected normal distribution. The findings summarized in Table 1, present strong deviations from normality ($p < .001$) across all scores for both ordinal and continuous data types. Kolmogorov-Smirnov test statistics ranged from 0.131 (total) to 0.315 (language), all with significant p-values < 0.001, implying that the data are deviant from normality. Based on these results, non-parametric statistical methods were employed in subsequent analyses to analyze the performance of human and AI review reports (McCrum-Gardner, 2008).

Table 1. Results from the Normality Test.

Review Components	Statistics	p-value
Title	0.266	<0.001
Abstract	0.244	<0.001
Language	0.315	<0.001
Methods	0.238	<0.001
Results	0.229	<0.001
Conclusions	0.250	<0.001
References	0.222	<0.001
Total	0.131	<0.001

The following analyses involved three key approaches to evaluate the differences and consistency between human and AI reviewers. First, Mann-Whitney U tests were employed to compare the scores for each component of the manuscripts provided by human versus AI reviewers (McKnight & Najab, 2010). Secondly, a Chi-square test of independence as suggested by McHugh (2013), was used to analyze nominal data and determine whether the distribution of score ratings for each manuscript component differed significantly between human and AI reviewers. Finally, inter-rater reliability was evaluated to measure the agreement between human and the AI reviewers in their final recommendations. Cohen's Kappa (for two raters – human vs AI reviewers) and Fleiss ' Kappa (for multiple raters – between human reviewers) were calculated to quantify the consistency in judgments across reviewer types (Kilic, 2015).

Results

In order to retain the variability among human reviewers while comparing their scores to the single AI reviewer's score, a Mann-Whitney U test was conducted. Table 2 presents the results of the tests, which revealed significant rank disparities between artificial intelligence and humans in the rating of the manuscripts. Overall, AI reviewers gave all components significantly higher mean ranks (the values range from 405.88 to 549.15) than the human reviewers (the values range from 273.85 to 329.5), suggesting that AI reviewers tended to make more positive reviews than the human reviewers. The results show that while AI reviewers tend to be more lenient, human reviewers possess a more discerning and diligent attitude, as reflected in their comparatively lower mean ranks.

Table 2. Comparison of Rank between Human vs AI Reviewer Scores.

Review Components	Reviewer Type	Mean Rank	Sum of Ranks
Title	Human	284.86	143,571.00
	AI	521.12	103,182.00
Abstract	Human	293.99	148,172.50
	AI	497.88	98,580.50
Language	Human	329.55	166,092.00
	AI	405.88	79,959.00
Methods	Human	275.38	138,792.50
	AI	544.46	107,258.50
Results	Human	280.92	141,584.00
	AI	531.16	105,169.00

Conclusions	Human	296.79	149,582.50
	AI	490.76	97,170.50
References	Human	280.68	141,462.00
	AI	531.77	10,5291.00
Total	Human	273.85	138,022.00
	AI	549.15	108,731.00

The results provide strong statistical evidence that human and AI reviewers differ significantly in their scoring patterns across all evaluated components as shown in Table 3. The extremely large Z-scores between -4.94 and -16.62 across all components, strongly indicate the robustness of the findings, while the consistent p-values < 0.001 confirm the statistical reliability of the observed differences. These results suggest that a hybrid review approach incorporating both human and AI perspectives may provide a more balanced and comprehensive evaluation of scientific papers.

Table 3. Mann-Whitney U Tests Result.

Review Components	**Mann-Whitney U**	**Z**	**p-value**
Title	16311.00	-14.97	<0.001
Abstract	20912.50	-12.57	<0.001
Language	38832.00	-4.94	<0.001
Methods	11532.50	-16.62	<0.001
Results	14324.00	-15.44	<0.001
Conclusions	22322.50	-11.95	<0.001
References	14202.00	-15.48	<0.001
Total	10762.00	-16.23	<0.001

A cross-tabulation of reviewer type and recommendation, shown in Table 4, was further analyzed to examine the differences in the final recommendations made by human and AI reviewers of the manuscripts. The recommendation categories were no revision, minor revision, major revision, and reject. Human reviewers most frequently recommended minor revision, with 309 out of 504 (61.3%), followed by 114 (22.6%) for major revision and 53 (10.5%) for no revision. Only 28 (5.6%) were recommended for rejection. The figures here indicate that human reviewers were commonly inclined towards providing constructive feedback requiring revision rather than outright rejecting manuscripts. On the other hand, the AI review reports primarily recommended minor revision, which was 191 of 198 (96.5%) papers, they reviewed. A small number of papers were recommended with no revision (3 papers, 1.5%), and only 4 papers (2.0%) for major revision. Interestingly, no paper was rejected by the AI reviewers, suggesting a significant difference in their evaluation approach compared to human reviewers.

Table 4. Cross-tabulation for Reviewer Type vs Recommendation.

Recommendation / **Reviewer Type**	**No Revision**	**Minor Revision**	**Major Revision**	**Reject**	**Total**
Human	53	309	114	28	504
AI	3	191	4	0	198
TOTAL	**56**	**500**	**118**	**28**	**702**

A Pearson Chi-Square test verifies the very significant differences between human and AI reviewers 'recommendation patterns, χ2(3) = 85.99, p < 0.01. This implies systematic differences between the two groups in how they assess and recommend manuscripts. These limitations of AI-based evaluation systems with the potential of assigning stricter recommendations raise the value of human oversight. Even though they are faster and probably even more objective in assessing papers, at the end of the day, such machines lack the critical judgment necessary to determine the main submissions for reworking or rejection. Human expertise may help offset some of these discrepancies by adding value to the application of technology.

A Cohen's kappa test was used to evaluate the level of agreement between the median of human raters versus AI raters across 198 manuscripts. The kappa value was 0.008, indicating an almost negligible level of agreement between the raters based on the Landis and Koch (1977) interpretation scale. On this scale, values below 0.00 indicate poor agreement, 0.00–0.20 indicate slight agreement, 0.21–0.40 indicate fair agreement and values above represent higher levels of agreement. The test statistic, T = 0.391, was not statistically significant (p = 0.696), suggesting that the observed agreement between the raters does not significantly differ from chance. The level of agreement for the final recommendation among two to five human raters on 198 manuscripts was conducted using a Fleiss' kappa analysis. The kappa value derived was 0.000, indicating an absolute lack of agreement among the raters (Landis & Koch, 1977). The test statistic calculated, z=0.000, was not statistically significant (p=1.000). In addition, the 95% confidence interval of the kappa statistic ranged from -0.215 to 0.215, thus further confirming the lack of agreement. The results show that the human reviewers disagreed consistently and their ratings were as random as chance.

Discussion

This research highlights the potential of integrating Large Language Models (LLMs) like ChatGPT-4 in the peer review process, while also underscoring its limitations. By comparing AI-generated review reports with those of human reviewers, this study offers valuable insights into the differences in evaluation rigor, recommendation tendencies, and consistency. The findings raise important questions about the future role of AI in scientific publishing and suggest potential pathways for optimization through hybrid review models.

Key Findings and Their Implications

Our findings show that ChatGPT-4 demonstrated a lenient review pattern, recommending fewer major revisions and no rejections, which contrasts with findings by Liang et al. (2024), where ChatGPT-4 feedback was more stringent in certain contexts. This difference could be attributed to variations in the design of prompts, as suggested by Santu et al. (2024), emphasizing the importance of prompt engineering in guiding AI behavior. Similarly, Saad et al. (2024) found that earlier versions of ChatGPT underperformed in recognizing critical manuscript issues, which aligns with our observation that AI reviews lack the rigor to fully replace human reviewers.

One of the most striking findings is the difference in grading rigor and recommendation distribution between human and AI reviewers. ChatGPT-4 demonstrated a clear tendency toward leniency, consistently awarding higher scores across all manuscript sections and overwhelmingly recommending minor revisions. In contrast, human reviewers offered a more balanced and critical evaluation, with a significant proportion of recommendations for major revisions or rejection.

The leniency of ChatGPT-4 poses both opportunities and risks. On the one hand, this approach may reduce the risk of unnecessary rejections and foster a more constructive review environment for authors. On the other hand, the lack of critical feedback and absence of rejections could compromise the integrity of the review process, leading to the acceptance of subpar manuscripts. Previous studies, such as those by Saad et al. (2024), have similarly noted that earlier versions of ChatGPT failed to identify major inconsistencies in manuscripts, reinforcing the necessity of human oversight.

The low inter-rater reliability among human reviewers is another important finding. The lack of agreement, as indicated by the low Fleiss' and Cohen's kappa values, reinforces existing concerns about subjectivity in peer review. This variability is consistent with previous studies, including those by Bornmann et al. (2011) and Lee et al. (2013), who highlighted that human reviewers often provide divergent assessments of the same manuscript due to personal biases and differing interpretations of review criteria. These findings suggest that human peer review alone is not a perfect standard and may benefit from supplementary AI-driven tools that can provide consistent baseline evaluations.

The Potential of Hybrid Review Models

Given the complementary strengths and weaknesses of AI and human reviewers, a hybrid review model emerges as a promising solution. AI can be employed to conduct initial assessments, focusing on objective components such as language, structure, and adherence to guidelines. This could help streamline the review process by quickly identifying manuscripts that require major revisions or immediate rejection due to fundamental issues. Human reviewers, on the other hand, would focus on evaluating the novelty, methodological soundness, and overall contribution of the manuscript.

Such a model would allow AI to act as a gatekeeper for low-risk decisions, while human reviewers provide the critical judgment needed for high-impact decisions. Importantly, the hybrid model could reduce reviewer workload and address reviewer fatigue, which has been a persistent issue in academic publishing (Fox et al., 2017). Future research should explore the development of dynamic models that assign different roles to AI and human reviewers based on manuscript complexity and disciplinary requirements.

Ethical Considerations and Bias Mitigation

The integration of AI in peer review raises ethical questions, particularly regarding bias, transparency, and accountability. Although LLMs offer the advantage of objectivity by being immune to biases related to gender, nationality, or institutional affiliation, they are not free from biases embedded in their training data. The process is ongoing an irreversible as corporate publishers already sell their content to train LLMs (Kwon, 2024). Hosseini and Horbach (2023) warned that AI models have the potential to amplify biases and inequalities, particularly if not carefully monitored.

doi:10.20944/preprints202502.0058.v1

Moreover, the AI's tendency to favor minor revisions could reflect underlying biases in the way it interprets academic writing.

To address these concerns, it is crucial to implement ethical frameworks and guidelines for the responsible use of AI in peer review. Transparency in the role of AI reviewers should be mandated, with clear disclosure of whether an AI tool was used and how its recommendations influenced editorial decisions. Additionally, ongoing monitoring and auditing of AI performance should be conducted to identify and mitigate any emerging biases.

Limitations and Future Directions

This study is limited to the evaluation of manuscripts submitted to a single journal, which may restrict the generalizability of the findings across different publication venues. Furthermore, only ChatGPT-4 was tested, leaving room for future research to explore the performance of other LLMs and AI models.

Future studies should focus on the following areas:

1. **Multi-disciplinary Analysis:** Conduct studies across multiple journals and academic fields to assess how AI performance varies by discipline.
2. **Longitudinal Studies:** Evaluate the long-term impact of integrating AI in peer review, particularly in terms of publication quality and reviewer workload.
3. **Training Specialized LLMs:** Develop LLMs specifically trained for peer review in various fields, incorporating datasets that include diverse examples of high- and low-quality submissions. The tag to be added to the way the algorithms are trained.
4. **Author Perception Studies:** Assess how authors perceive AI-generated feedback compared to human feedback, as their acceptance of and responses to reviews are critical for the success of the peer review process.

Conclusions

This research article revealed that human reviewers provide a more rigorous and balanced evaluation of academic manuscripts compared to the ChatGPT-4 version, which predominantly recommended minor revisions and did not suggest rejection. The balanced approach of human reviewers has been confirmed in the scores provided for each part of the manuscripts, the overall score, and the recommendations of the manuscripts. On the other hand, human reviewers have shown lack of agreement in their reports. This can allude to the high subjectivity of the peer review process and the necessity for improvement.

While the speed of the LLM tools is indisputable and at the same time can provide useful comments for content improvement, their leniency may compromise the peer review procedure and its integrity if used independently and without human oversight. Therefore, a hybrid model, that will combine the LLM's speed with human rigor, could optimize the peer review process. Future research should explore objective methods to train LLM tools for higher performance in peer review.

Authors` note: The authors of the paper acknowledge the use of the ChaGPT-4 version in facilitating various stages of writing and ideation for this paper. The Discussion section was fully rewritten by ChatGPT-4 for clarity. All contributions from the AI were reviewed, critically edited, and validated by the authors to ensure academic rigor and adherence to ethical standards. The authors also assessed and addressed potential biases inherent in the AI-generated content. The final content, conclusions, and assertions in this paper are the sole responsibility of the human authors.

Data accessibility statement: All data generated or analyzed during this study are included in this published article.

Ethics and Consent: The study has been approved by the Review Board of the European Scientific Institute, ESI. Written consent during the submission process was obtained from the authors for AI peer review of their manuscripts. The manuscripts were anonymized before being uploaded to ChatGPT-4 Plus for personal data protection. The principles of the Helsinki Declaration for good research practices were followed.

Funding Statement: The authors did not obtain any funding or support for this study.

Acknowledgments: Not applicable.

Competing Interests: Dr. Jovan Shopovski is a managing editor of the European Scientific Journal, ESJ.

Appendix 1: The Prompt Used for ChatGPT-4 to Review the Manuscripts

"Please review the attached research paper thoroughly, evaluating it according to the sections listed in the ESJ evaluation form. For each section, provide:

An assessment score from 1 (poor) to 5 (excellent), based on the criteria defined in the form.

An explanation for each score, with balanced feedback that highlights strengths and suggests areas for improvement.

At the end, give a final recommendation of 'accept,' 'minor revision,' 'major revision,' or 'reject,' based solely on the paper's merits. Be objective and critical—praise strengths when present, but evaluate weaknesses fairly and without hesitation. If the paper does not meet the scholarly standards expected, recommend rejection with a clear, concise summary supporting this recommendation.

Use constructive language that reflects an impartial, academically rigorous perspective, prioritizing objective critique."

References

Altmae, S., Sola-Leyva, A., & Salumets, A. (2023). Artificial intelligence in scientific writing: A friend or a foe? Reproductive BioMedicine Online, 47(1), 3-9. https://doi.org/10.1016/j.rbmo.2023.04.009

Athaluri, S. A., Manthena, S. V., Kesapragada, V. K. M., et al. (2023). Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. Cureus, 15(4). https://doi.org/10.7759/cureus.37432

Bornmann, L., Mutz, R., & Daniel, H. D. (2011). A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. PLoS ONE, 6(12), e26895.

Bolek, M., Bolek, C., Shopovski, J., & Marolov, D. (2023). The consistency of peer-reviewers: assessment of separate parts of the manuscripts vs final recommendations. Accountability in Research, 30(7), 493-515.

Charles W Fox, Arianne YK Albert, and Timothy H Vines. Recruitment of reviewers is becoming harder at some journals: a test of the influence of reviewer fatigue at six journals in ecology and evolution. Research Integrity and Peer Review, 2:1–6, 2017.

Cheng, K., Sun, Z., Liu, X., Wu, H., & Li, C. (2024). Generative artificial intelligence is infiltrating peer review process. Critical Care, 28(1), 149.

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. Behavioral and brain sciences, 14(1), 119-135.

Conroy, G. (2023). How ChatGPT and other AI tools could disrupt scientific publishing. Nature, 622(7982), 234-236.

Cullen, D. J., & Macaulay, A. (1992). Consistency between peer reviewers for a clinical specialty journal. Academic Medicine, 67(12), 856-9.

Huang, J., & Tan, M. (2023). The role of ChatGPT in scientific communication: Writing better scientific review articles. American Journal of Cancer Research, 13(4), 1148. http://www.ncbi.nlm.nih.gov/pmc/articles/pmc10164801/

da Silva, J. A. T., & Daly, T. (2025). No reward without responsibility: Focus on peer review reports. Ethics, Medicine and Public Health, 33, 101033. https://doi.org/10.1016/j.jemep.2024.101033

Hosseini, M., & Horbach, S. P. (2023). Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. Research integrity and peer review, 8(1), 4.

Jenko, N., Ariyaratne, S., Jeys, L., et al. (2024). An evaluation of AI-generated literature reviews in musculoskeletal radiology. The Surgeon. https://doi.org/10.1016/j.surge.2023.12.005

Kaatz, A., Gutierrez, B., & Carnes, M. (2014). Threats to objectivity in peer review: The case of gender. Trends in Pharmacolpgical Sciences, 35(8), 371–373. https://doi. org/10.1016/j.tips.2014.06.005.

Kadi, G., & Aslaner, M. A. (2024). Exploring ChatGPT's abilities in medical article writing and peer review. Croatian Medical Journal, 65(2), 93.

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. Journal of the American Society for Information Science and Technology, 64(1), 2-17.

Kacena, M. A., Plotkin, L. I., & Fehrenbacher, J. C. (2024). The use of artificial intelligence in writing scientific review articles. Current Osteoporosis Reports, 1-7. https://doi.org/10.1007/s11914-023-00852-0

Kilic, S. (2015). Kappa testi. Journal of mood disorders, 5(3), 142-144.

Landis, J.R. & Koch, G.G. (1977) The Measurement of Observer Agreement for Categorical Data. Biometrics, 33, 159–174.

Kwon, D. (2024). Publishers are selling papers to train AIs-and making millions of dollars. Nature, 636(8043), 529-530.

Li, Z. Q., Xu, H. L., Cao, H. J., Liu, Z. L., Fei, Y. T., & Liu, J. P. (2024). Use of Artificial Intelligence in Peer Review Among Top 100 Medical Journals. JAMA Network Open, 7(12), e2448609-e2448609.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. Cognitive Therapy and Research, 1(2), 161–175.

Lozic, E., & Stular, B. (2023). Fluent but not factual: A comparative analysis of ChatGPT and other AI chatbots ' proficiency and originality in scientific writing for humanities. Future Internet, 15(10), 336.

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., ... & Zou, J. (2024). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. NEJM AI, 1(8), AIoa2400196.

McHugh, M. L. (2013). The chi-square test of independence. Biochemia medica, 23(2), 143-149.

McCrum-Gardner, E. (2008). Which is the correct statistical test to use?. British Journal of Oral and Maxillofacial Surgery, 46(1), 38-41.

McKnight, P. E., & Najab, J. (2010). Mann-Whitney U Test. The Corsini encyclopedia of psychology, 1-1.

Razack, H. I. A., Mathew, S. T., Saad, F. F. A., et al. (2021). Artificial intelligence-assisted tools for redefining the communication landscape of the scholarly world. Science Editing, 8(2), 134-144. https://doi.org/10.6087/kcse.244

Saad, A., Jenko, N., Ariyaratne, S., Birch, N., Iyengar, K. P., Davies, A. M., ... & Botchu, R. (2024). Exploring the potential of ChatGPT in the peer review process: an observational study. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 18(2), 102946.

Santu, S. K. K., Sinha, S. K., Bansal, N., Knipper, A., Sarkar, S., Salvador, J., ... & Williams Jr, M. C. (2024). Prompting LLMs to Compose Meta-Review Drafts from Peer-Review Narratives of Scholarly Manuscripts. arXiv preprint arXiv:2402.15589.

Smith, R. (2006). Peer review: A flawed process at the heart of science and journals. Journal of the Royal Society of Medicine, 99(4), 178-182.

Squazzoni, F., Ahrweiler, P., Barros, T., Bianchi, F., Birukou, A., Blom, H. J. J., … Willis, M. (2020). Unlock ways to share data on peer review. Nature, 578, 512–514. https://doi.org/10.1038/d41586-020-00500-y.

Zhao, D., Bu, L., Alippi, C., & Wei, Q. (2017). A Kolmogorov-Smirnov test to detect changes in stationarity in big data. IFAC-PapersOnLine, 50(1), 14260-14265.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.