Journal of Digital Information Management

DLINE JOURNALS

ISSN: 0972-7272

JDIM 2025: 23 (4)

Can AI Replace Human Peer Reviewers? A Comparative Analysis of AI-Generated and Human Expert Reviews

Pit Pichappan
Digital Information Research Labs
Chennai 600017. India
pichappan@dirf.org

Preethi Pichappan
London Metropolitan University
Holloway Rd, London N7 8DB. UK

ABSTRACT

This work analyses the performance of AI-generated peer reviews compared to human expert reviews across 62 manuscripts in the Real-Time Intelligent Systems track of the Springer Lecture Notes in Networks and Systems. Using four large language models ChatGPT 3.5, Perplexity AI, Qwen-3 Max, and DeepSeek the study evaluated 141 reviews using both AI and human reviews against 12 quality criteria, scored by five domain experts. Results show human reviews scored higher overall (mean = 3.98 vs. 3.15 for AI) with greater consistency and depth, particularly in methodological critique, literature contextualisation, and review confidence. AI reviews were more generic and less specific, struggled with scholarly subtlety, though they excelled at summarisation and formatting checks. While the difference approached statistical significance (p = 0.08), the effect size (Cohen's d = 0.56) indicated a moderate practical gap. The study concludes AI cannot replace human reviewers but may ethically augment the process in hybrid models under human oversight.

Subject Categories and Descriptors:
[**I.2.11 Distributed Artificial Intelligence];** *Languages and structures*: **[H.5.2 User Interfaces];** Natural language: **[I.2.7 Natural Language Processing]** Language models

General Terms: Large Language Models, Human Reviews, Scientific Paper Review, Peer Review, Human-AI Collaboration

Keywords: Peer Review, AI Tools, Large Language Models, Scholarly Evaluation, Human–AI Collaboration, Review Quality

Received: 27 July 2025, Revised 4 September 2025, Accepted 19 September 2025

Review Metrics: Review Scale: 1-6, Review Score: 4.95, Inter-reviewer consistency: 82.4%

DOI: https://doi.org/10.6025/jdim/2025/23/4/220-233

1. Introduction

The number of manuscripts submitted to journals has been increasing rapidly over the years. One quantitative analysis suggested that the volume of scientific manuscript submissions to journals doubles approximately every 15 years [1]. While we have data on the number of papers published, we do not have accurate figures for the number of submissions. If we add other publication formats, such as conferences, it becomes more challenging to obtain the data.

The peer review process has been used in the scientific publication system to handle the scientific manuscripts for three centuries. The peer review process involves cognitive alignment between the reviewers and the target papers. It enables the elimination of mediocre research and improves the quality and level of submissions by supporting review comments. While reviewers' assessments are interpretive, journals impose norms and editorial oversight to ensure consistency in evaluations. In the last decade, many open peer review models have emerged to infuse confidence and transparency in the review process, helping convince authors that the evaluation process is free of bias.

In the age of AI-generated content and the rise of preprints (such as arXiv and bioRxiv), the significance of peer review as a reliable validation process has become even more critical, particularly as automated models, though not yet fully mature, begin to replicate the subtle judgment of human experts. This highlights the importance of assessing aspects such as review confidence, specificity, and contextualization in the literature: these represent fundamental characteristics of genuine scholarly evaluation that current AI technologies struggle to replicate.

Many reviewers have been ignoring the review process because it is time consuming and offers no rewards. The absence of a comprehensive or complete review of submissions further characterises many reviews. The surge in submissions and the shortage of reviewers, coupled with time constraints, led to the resort to AI-based reviews, and the practice of using AI-dependent models is labelled as 'cheating' [2], or an unethical scientific practice. Nearly one third of the authors prefer to send their submissions based on the journal's publication speed [3]. Hence, timely review is the essence of the publication process.

2. Review Tools

The efficiency gains from automation can be substantial, as the process of reviewing alone takes up a considerable amount of expert time. It was estimated that 'more than 15 million hours' are dedicated to reviewing rejected submissions annually [4]. For instance, 1.2 million manuscripts are submitted to 2300 Elsevier journals each year, with only 30% (around 350,000) making it to publication [5]. In light of the rising volume of submissions and the increasing burden of peer review, a report by BioMed Central and Digital Science titled 'What might peer review look like in 2030?' suggested utilizing technology to support and enhance the peer review process, including finding automated methods to detect inconsistencies that may be challenging for reviewers to identify [6].

Journal reviewers either frequently or infrequently employ AI tools to support the peer review process; however, journal editors do not recognise AI for manuscript assessment and explicitly oppose its use. The International Committee of Medical Journal Editors (ICMJE) restricts the application of AI in peer review. It requires reviewers to obtain prior journal approval before using such tools in their evaluations (ICMJE) [7]. Despite this caution, existing literature and real world examples show that AI can effectively identify suitable reviewers and offer compelling evidence of its potential to assist with initial quality control of submitted manuscripts [8]. The benefits and limitations of using AI assistants to support scientific review are investigated in detail in many studies [9].

Besides the generic LLModels, a few specific review tools, such as UNSILO, *Enago, Stat Check, Stat Reviewer*, and *Review Adviser* [10, 11 12, 13, 14], are employed to assist the review process. *Review Advisor* is a toolkit for natural language processing designed to assist in selecting high-quality manuscripts for journals and to provide feedback to help authors enhance their submissions. While its effectiveness on the authors' massive dataset was limited, it provides a foundation. It may aid reviewers by proposing comments on aspects of the papers they might have missed. Another research project developed an AI tool using a neural network trained on features from manuscripts, including word frequencies, readability assessments, and formatting metrics, revealing that automated systems developed biased tendencies, such as against grammatical and formatting mistakes, which contributed to their accuracy [15] Likewise, the *pReview* software package was created to automatically generate summaries, detect contributions, analyse writing quality, and identify potentially related academic works to assist reviewers [16]. It is also evident that Natural Language Processing models could simplify and enhance the peer-review process by generating reviews for scientific papers. However, the quality of those reviews was not sufficient to replace human experts [17]. The possibility of human–AI collaboration in the decision making process for reviews has been mooted by the model *PEERRec* by Bharti et al. [18]

2.1 Human-AI Hybrid models

Modern AI has demonstrated accuracy surpassing that of domain experts across a growing number of fields. Recent studies have paved the way for future research into human AI collaboration within Hybrid Intelligence systems. These analyses underscore the transformative impact of AI technology and the emerging challenges HCI professionals encounter when implementing a human centred AI (HCAI) approach in AI system development [19-25].

3. Background

3.1 How Reviewers Use AI Tools?

We categorise AI-assisted peer reviewing into three distinct modalities, reflecting varying degrees of human involvement and reliance on artificial intelligence:

1. Fully AI-generated reviews: Reviewers submit reviews produced entirely by an AI tool without any human intervention. GPTZero LLM detector estimated that at least 15.8% of reviews were generated with AI assistance [26].

2. AI-drafted, Human edited reviews: Reviewers use AI-generated text as a draft, then revise it to remove irrelevant or superficial content, enhance clarity, and align the tone and substance with scholarly expectations.

3. Human-led reviews with AI augmentation: Reviewers compose their evaluations independently but consult AI-generated feedback to identify potential omissions, strengthen critique, or supplement specific sections (e.g., methodological checks or literature context).

While the first modality may be detectable through stylistic analysis and is typical of large language models (e.g., synthetic phrasing, hedging, or structural uniformity), the latter two are significantly more challenging to identify. In these hybrid approaches, human judgment actively reshapes or selectively integrates AI output, obscuring algorithmic fingerprints and blurring the boundary between human and machine contributions. The word level ratio check enables the detection of AI, which is investigated in [27, 28]. Checco et al [14], sought to understand the extent to which AI can assist reviewers and authors, rather than replace human decision making processes.

4. Dataset and Methodology

4.1 Study Goal

This study aims to compare peer reviews generated by human experts with those produced by selected artificial intelligence (AI) language models. The central research question is: To what extent do AI-generated reviews approximate the quality, depth, and insight of human expert reviews? How do AI-assisted reviews mimic or align with human reviews?

4.2 Dataset Composition

The dataset comprises 62 reviews of the original research manuscripts submitted to the *Real-Time Intelligent Systems* track within the *Lecture Notes in Networks and Systems* (LNNS) series. These submissions represent peer reviewed scholarly work spanning topics such as real time decision making, embedded intelligence, adaptive control, edge AI, and time sensitive data processing. The selection ensures domain coherence and topical relevance, which is critical for meaningful evaluation by both human and AI reviewers. The available human reviews are extracted from the paper submission system. The study comprises 141 human reviews from 62 papers as the first group.

4.3 AI Review Generation

Each of the 141 reviews was independently evaluated using four prominent large language models (LLMs):

- ChatGPT 3.5 (OpenAI)
- Perplexity AI
- Qwen-3 Max (Alibaba Cloud)
- DeepSeek

These models were prompted with standardised review instructions mirroring those typically provided to human peer reviewers, including expectations to:

- Summarise the manuscript's core contribution
- Evaluate methodological soundness

- Assess experimental design and data analysis
- Identify limitations
- Provide specific, actionable suggestions for improvement
- Comment on literature coverage and presentation quality
- Deliver a clear recommendation (e.g., accept, minor/major revision, reject)

The prompt query is: "*Critically evaluate the scientific merit of the attached file*," with the full text of the article uploaded. All AI-generated reviews were produced under controlled prompting conditions to minimise variability due to instruction ambiguity.

4.4 Human Review Baseline

The original human peer reviews authored by domain-specialised researchers during the actual peer review process serve as the ground truth benchmark. These reviews underwent standard content screening and reflect the expectations of research publishing. The human review reports served as the dataset one. The AI-generated reviews of all the papers from dataset 2.

4.5 Evaluation Framework

To ensure a systematic and objective comparison, we designed a multidimensional evaluation model informed by established peer review quality indicators in scholarly publishing. Five independent domain experts (Post-Ph.D.-level researchers with ≥ 5 years of experience in publishing and reviewing) were recruited as evaluators.

Each expert was presented with anonymised pairs of reviews (one human, one AI-generated) per manuscript and asked to rate both on a 6-point Likert scale (0–5) for the following 12 criteria. The evaluation experts use the Likert scale to assign a numerical score for their assessment, with 5 for an excellent review and 0 for a poor review. AI-generated reviews using the four tools for sample papers from the dataset, and the two datasets are available for comparison.

Criterion	Description
Summarise	Accuracy and conciseness in capturing the paper's main contributions
Review Depth	Level of critical engagement and analytical insight
Limitations	Ability to identify substantive weaknesses or gaps
Suggestions	Constructiveness and specificity of improvement recommendations
Review Confidence	Perceived certainty and authority in the reviewer's judgmen
Methodology	Evaluation of the soundness and appropriateness of research methods
Data/Experimentation /Inference	Assessment of empirical rigour, statistical validity, and logical inference

Literature Review	Coverage and contextualization within existing scholarship
Presentation	Clarity, structure, and writing quality
Recommendation	Justification and alignment of the final decision with the review content
Degree of Relevance	Focus on core contributions versus peripheral issues
Specificity	Use of concrete examples, equations, figures, or section references

The raters were blinded to the origin (human vs. AI) of each review to mitigate bias. Inter-rater reliability was assessed using Fleiss' Kappa to ensure consensus robustness.

4.6 Analytical Approach

Quantitative analysis includes:

- Mean score comparisons across criteria (paired *t*-tests or non-parametric equivalents)
- Effect size estimation (Cohen's *d*) to assess practical significance
- Principal component analysis (PCA) to identify latent dimensions differentiating human and AI reviews

Qualitative analysis involves thematic coding of evaluators' free text comments to capture delicate strengths and shortcomings that are not fully reflected in numerical scores.

4.7 Ethical Considerations

All human-generated reviews were used with editorial permission and anonymised to protect reviewer and author identities. AI outputs were treated as synthetic data; no proprietary model internals were reverse engineered. The study complies with ethical guidelines for research involving human subjects and computational text generation.

For the 141 reviews, we have a database of human reviews, which are then reviewed by the generic LLMs. For each review, we have both human and AI. These reviews are passed to the evaluators who observed the submitted analytics scores based on the Likert scale. Each variable is assessed, scores are generated for each review, and the final mean values are reported in Table 1.

5. Analysis

The mean evaluation scores of both human and AI reviews for the target reviews are presented. Out of a maximum score of 5, reviewers assign a score for each review, and the final mean values are presented.

Variable	**Human**	**AI**
Summarise	2.25	4.4
Review Depth	3.1	3.6

Limitations	4.2	4.3
Suggestions	3.9	4.1
Review Confidence	4.2	0.8
Methodology, data	3.6	4.1
Experimentation/Inference	3.85	3.2
Literature	4.1	2.2
Presentation	4.5	2.8
Recommendation	4.9	3.2
Degree of Relevance	4.8	2.65
Specificity	4.35	3.2

Table 1. Mean Scores of the dataset given by evaluators

The statistical significance of the variable scores is presented below. (Table 2)

	Human	**AI**
mean	3.979167	3.145833
Std deviation	0.736842	1.209707

Table 2. Statistical parameters of the review data

We produce the summary of the statistical analyses. In the Paired *t*-test, the *t*-statistic is 1.93, whereas the p-value is 0.080. → The difference between Human and AI scores is not statistically significant at α =0.05, but it approaches significance ($p < 0.10$).

The Effect Size is measured using (Cohen's *d* for paired samples), and the Cohen's *d* is found to be 0.56. This represents a medium effect size, suggesting a moderate practical difference between Human and AI scoring patterns.

5.1 Principal Component Analysis (PCA) to identify latent dimensions differentiating human vs. AI reviews

We prepared a PCA biplot by coercing the Human and AI columns to numeric and dropping empty rows. Looks clean and suitable for a 2D PCA (since there are only two features: Human and AI).

PCA biplot points are variables, red arrows are Human and AI loadings. (Figure 1)

• Principal Component 1 contrasts Human vs. AI (loadings are roughly opposite signs on PC1). That means PC1 captures the overall tendency for a review dimension to be scored higher by Humans vs. by AI, i.e., the human–AI gap axis.

• Principal Component 2 loads positively on both Human and AI (both loadings are positive), so it's a shared "overall quality/intensity" axis where both move together.

• In the biplot, dimensions on the side of the Human arrow reflect areas where humans rate higher than AI; dimensions on the side of the AI arrow reflect the reverse. Points far from the origin are the dimensions that most differentiate human and AI ratings.

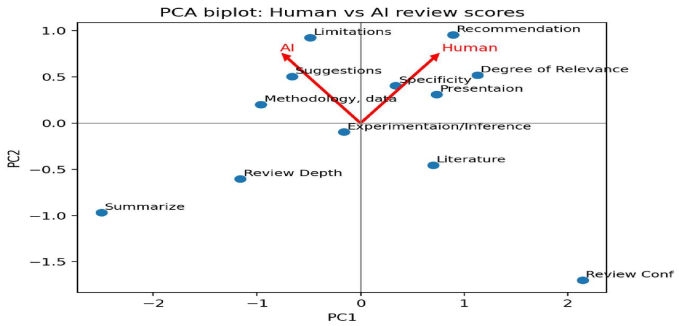

Figure 1. Principal Component Analysis biplot for human vs AI

5.2 Agreement (Bland-Altman) Plot

The plot shows the difference (Human–AI) versus the average score for each dimension. The mean bias is approximately +0.48, indicating that Human scores tend to be slightly higher than AI scores on average. Most points lie within the 95% limits of agreement (mean ± 1.96 SD), suggesting reasonable agreement, though some dimensions (e.g., Literature, Presentation, Degree of Relevance) show larger discrepancies.

Human scores average higher and are tighter (lower SD) than AI scores, and the confidence intervals don't overlap much, suggesting a notable difference in central tendency.

The differences between AI and human reviews are analysed, and the value of variations is presented in Figure 2.

To measure the Missing Inter-Rater Reliability, we computed Cohen's Kappa by first converting the continuous Human and AI scores into three ordinal categories (Low/Medium/High) using tertiles, which is essential for establishing that the five expert evaluators were scoring consistently. Then we built the confusion matrix and calculated the agreement, which is -0.199. This binary version yields a kappa of around 0.20, also suggesting worse than chance alignment at this particular threshold.

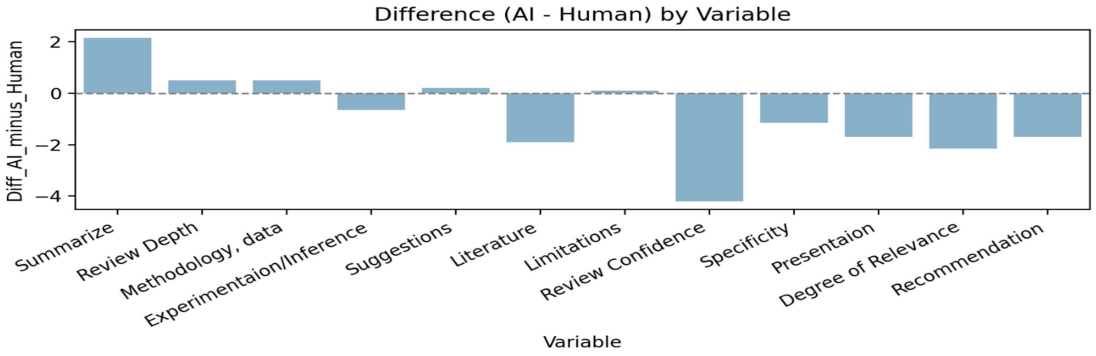

Figure 2. Variations between human and AI scores

5.3 Difference chart: AI-Human by variable

Bars above zero mean AI rated higher; bars below zero mean AI rated lower than Human.

• Humans generally rate higher than AI across variables, with a relatively consistent positive bias.

• A few variables buck this trend (negative differences), pointing to specific areas where Human is stricter or AI is more conservative.

• Agreement is moderate: dispersion around the mean difference suggests non-trivial variability but no apparent systematic drift with score level.

This grouped bar chart compares Human and AI scores side by side for each variable. One can quickly see where AI rates higher or lower than humans across items, and the magnitude of those gaps. Overall, AI tends to rate higher than humans on most variables, though a few exceptions exist in which humans rate higher. This pattern suggests AI is generally more lenient or more generous for several criteria, but not universally.

Each AI tool uses a predefined set of scales for assessment, and the four generic LLMs are now compared across several parameters. The parameters they use are outlined in Table 4. The parameters used by the AI tools for the review process are provided in this table.

Summary	Strength	Limitations	Methodology	Future	Research	Reproducability	Literature	Re Novelty	Clarity	Impact	Recommendation	Scale (L to H rating)	Presentation	Results
Perplexity	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	Yes	No	No	No
Qwen	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No
ChatGPT	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	Yes	Yes	Yes	Yes
Deepseek	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	Yes	No	Yes	No

Table 4. Comparison Matrix of the parameters used in AI tools

Note: 'Yes' denotes used and 'No' denotes not used by the AI tool

To understand and reach a conclusion about the use of parameters, we computed Pairwise *phi* correlations for the used criteria. Figure 3 shows the Pairwise criteria comparison of the four AI tools.

• With only four models, many correlations are extreme or where a column has no variance. Treat these as exploratory signals rather than definitive.

• Some criteria move together positively (e.g., clusters around literature/presentation/clarity/impact/reproducibility/novelty), while others are negatively associated with future research in this tiny sample.

• The *NaNs* (Not a number) indicate columns with no variation across the four models (all Yes), so correlations cannot be computed for those with standard Pearson.

• Figure 4 is a compact "bar strip" grid: each row is a model, each column is a criterion, green = Yes and red = No. It's excellent for quick side by side comparison.

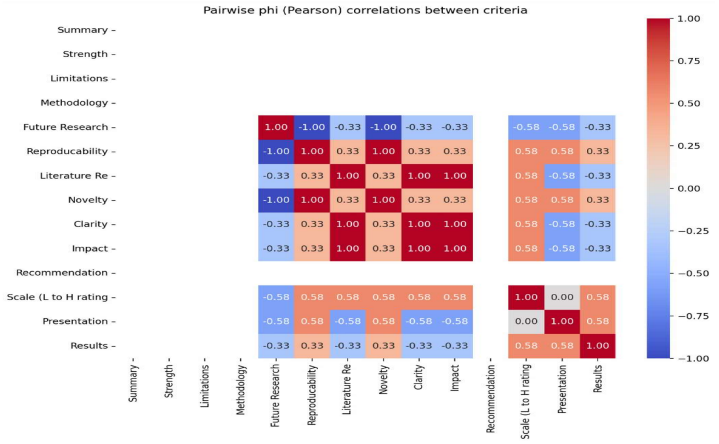

Figure 3. Pairwise criteria comparison of the four AI tools

Note: Blank columns denote Not a number (NaNs)

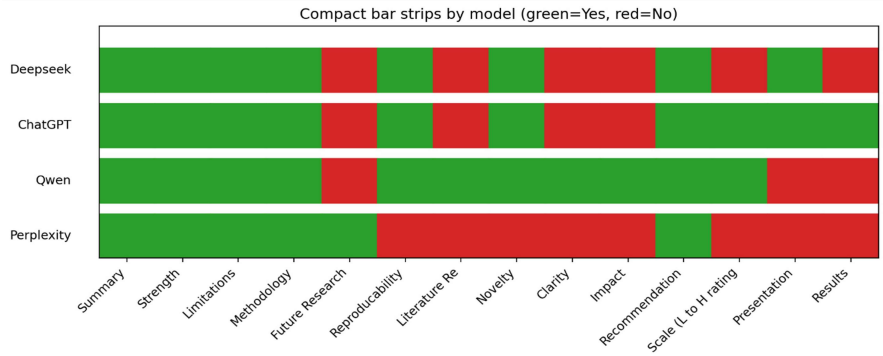

Figure 4. Consistent marking of the AI tools reviews

The strip view makes it easy to see blocks of agreement and disagreement:

• Consistent "green blocks": Several criteria are uniformly green across most models, suggesting broadly satisfied evaluation dimensions. These are the columns with contiguous vertical green bands.

• Divergence around a few criteria: Columns with mixed red/green indicate points of differentiation among models these are where the models' profiles actually separate.

• Sparse reds: Where one sees a red in an otherwise green column, that model is out of step on that specific dimension, and it likely explains its relative positioning in other visuals (e.g., radar shapes).

• Potential clusters: Rows that look visually similar (similar red/green patterns) belong to models with identical evaluation footprints; rows that differ in a few consistent columns identify "signature" strengths/weaknesses.

6. Summary and Discussions

The overall scores for the two models show that human reviews scored significantly higher than AI reviews:

the mean human score is 3.98, whereas the mean AI score is 3.15. Non-overlapping 95% confidence intervals indicate a statistically and practically significant difference. We further observed that the AI reviews showed higher variability (standard deviation = 1.21 vs. 0.74 for humans), suggesting inconsistency. AI tended to be more lenient/generous in most categories, though less critical in identifying limitations and offering specific, actionable feedback.

We infer from the data that the AI reviews often lacked depth, contextual understanding, and scholarly subtlety. While AI could summarise contributions and assess surface level clarity, it struggled with features such as critiquing methodological rigour, identifying minor imperfections and providing confident, authoritative judgments. Hybrid approaches (e.g., AI-drafted + human-edited) were noted as harder to detect but potentially valuable if ethically implemented.

While assessing the four studied LLMs, we infer their characteristic features.

- Qwen performed most comprehensively, meeting all criteria except "Future Research."

- Perplexity excelled in core areas (summary, strength, methodology) but failed in reproducibility, novelty, and impact.

- ChatGPT and Deep Seek showed mixed results strong in clarity and recommendations but inconsistent on reproducibility, novelty, and results assessment.

- AI tools themselves differ considerably in the review process when assessing scientific manuscripts.

7. Conclusion

AI cannot yet replace human reviewers, especially for high stakes scholarly evaluation. AI may augment the review process (e.g., flagging formatting issues, suggesting literature), but human oversight remains essential. Fully AI-generated reviews are ethically problematic and often constitute "unreliable" in the current scholarly ecosystem. The peer review system is under strain due to rising submissions and a shortage of reviewers, making responsible AI integration an urgent topic of discussion.

While AI tools show promise in supporting peer review through automation and augmentation, they fall short of replicating the critical insight, cognitive and contextual depth, and scholarly judgment of human experts. A similar inference was obtained in comparative studies, which suggested a hybrid approach. [29, 30] The study advocates for human–AI collaboration under clear ethical guidelines rather than full automation.

AI can assist the review process on a few issues, such as detecting inconsistencies and correlating the content with documented data.

At the same time, AI reviews are monolithic and follow a typical pattern regardless of the paper's content or discipline. AI-assisted reviews are more generic, fail to identify specific errors, and do not correlate with earlier papers when generating reviews. The use of AI tools in the review process is inevitable in the future, but should be used in conjunction with human intervention, as many recent studies have shown [31, 32]. The

direction of a new type of analysis of typical human processes, organised with the help of machine learning systems, paves the future path [14].

References

[1] Larsen, P. O., Ins, von, M. (2010). The Rate of growth in scientific publication and the decline in coverage provided by the Science Citation Index. *Scientometrics*. 84:575–603.

[2] Palmer, Kathryn. (2025). AI-Enabled Cheating Points to 'Untenable' Peer Review System, *Inside Higher Ed*, July 2025.

[3] Gaston, Thomas, E., Ounsworth, Francesca., Senders, Tessa., Ritchie, Sarah., Jones, Emma. (2020). Factors Affecting Journal Submission Numbers: Impact Factor and Peer Review Reputation. *Author manuscript*. doi: 10.1002/leap.1285

[4] American Journal Experts. (2018). Peer review: How we found 15 million hours of lost time. www.aje.com/en/arc/peer-review-process-15-million-hours-lost-time/

[5] Tedford, A. (2015). Rolling out our new editorial system: EVISE®: Find out how the new system will help reviewers streamline their workload.

[6] Burley, R., Moylan, E. (2017). What might Peer Rev look like in 2030? A report from BioMed Central and Digital Science. https://figshare.com/articles/journal_contribution/What_might_peer_review_look_like_in_2030_/4884878/1

[7] ICMJE. (2025). Issues Latest Guidelines for Medical Journal Publication with Emphasis on AI Vigilance. Accessed on Feb. 17,]. Available at: https://lifesciences.enago.com/blogs/latest-guidelines-for-medical- journal-ublication-with-emphasis-on-ai- ila nce# :~:tex t=Rest ricts%20 reviewers% 20on %20us ing%20 AI,notif y%2-othe %20guideline%20committee's%20secretariat.

[8] Kousha, Kayvan., Thelwall, Mike. (2024). Artiûcial intelligence to support publishing and peer review: A summary and review. *Learned Publishing*. 37 (1) p. 1-68

[9] Silva, Julio., C. M. C. Gouveia, Rafael., P. Zielinski, Kallil., M. C. Cristina, Maria., F. Diego, Oliveira., R. Odemir, Amancio, M. Bruno, Oliveira, Osvaldo., N. (2025). Jr. AI-Assisted Tools for Scientific Review Writing: Opportunities and Cautions*ACS Applied Materials Interfaces* 17 (34), 47795-47805

[10] Juric, Mario., Rydahl, Mads., Reckman, Hilke. (2019). Comparing UNSILO concept extractionto leading NLP cloud solutions, (PDF) White Paper Comparing UNSILO concept extraction to leading NLP cloud solutions. Available from:https://www.researchgate.net/publication/38194 839 1_Wh ite_Pa per_C ompar ing_U NSI LO_co nce pt_ext raction_to_leading_NLP_cloud_solutions [accessed Nov 14 2025].

[11] Academy, Enago. Launching 'Review Assistant: An AI-powered Tool for Peer Reviewers, https://www.ena go.com.br/academy/accelerating-peer-review-with-review-assistant/.

[12] Michèle, B., Joshua, Nuijten., R. (2020). Polanin statcheck: Automatically Detect Statistical Reporting Inco nsistencies to Increase Reproducibility of Meta Analyses, *Research Synthesis Methods* 11 (1).

[13] Aries, Systems. (2018). New Decision Support Tool, StatReviewer, Available in 15.0. https://www. ariess ys.c om/newsletter/february-/new-decision-support-tool-statreviewer-available-in-15-0/.

[14] https://github.com/neulab/ReviewAdvisor Elsevier www.elsevier.com/connect/reviewers-update/rolling-out-our-new-editorial-system-evise.

[15] Checco, A., Bracciale, L., Loreti, P., Pinfield, S., & Bianchi, G. (2021). AI-assisted peer review. *Humanities and Social Sciences Communications,* 8 (1) 1–11. https://doi.org/10.1057/s41599-020-00703-8.

[16] Roberts, J., Fisher, D. (2020). pReview: The artificially intelligent conference reviewer. *In:* 2020 19th IEE E International Conference on Machine Learning and Applications (ICMLA) *IEEE* p. 665–668.

[17] Yuan, W., Liu, P., Neubig, G. (2021). Can we automate scientific reviewing? arXiv preprintarX iv:21 02.0 0176.

[18] Bharti, P. K., Ghosal, T., Agarwal, M. et al. (2024). PEERRec: An AI-based approach to automatically gen erate recommendations and predict decisions in peer review. *Int J Digit Libr* 25, 55–72.

[19] Korteling, J. E., (Hans) Visschedijk, van de Boer, G. C. , Blankendaal, R. A. M., Boonekamp, R. C., Eikelboom, A. R. (2021). Human versus Artificial Intelligence, *Frontiers in Artificial Intelligence*. Volume 4- 2021 Article 622364.

[20] Vaccaro, M., Almaatouq, A., Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nat Hum Behav 8*, 2293–2303.

[21] Maadi, M., Khorshidi, Akbarzadeh., H., Aickelin, U. (2021). A Review on Human–AI Interaction in Mach ine Learning and Insights for Medical Applications. *Int. J. Environ. Res. Public Health*, 18, 2121.

[22] Hemmer, Patrick., Schemmer, Max., Vössing, Michael., Kühl, Niklas, (2021). Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review, *In*: PACIS 2021 Proceedings. *Paper* 472. 78.

[23] Hemmer, Patrick., Schemmer, Max., Vössing, Michael., Kühl, Niklas (2021). Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review, *In:* Twenty fifth Pacific Asia Conference on Information Systems, Dubai, UAE.

[24] Wei, Xu., Dainoff, Marvin., J. Ge, Liezhong., Gao, Zaifeng. (2021). From Human-Computer Interaction to Human-AI Interaction: New Challenges and Opportunities for Enabling Human-Centred AI. (arXiv:2105.05424).

[25] Shcherbiak, A., Habibnia, H., Böhm, R., Fiedler, S. (2024). Evaluating science: A comparison of human and AI reviewers. *Judgment and Decision Making*. 19, e21.

[26] The AI Review Lottery. (2024). Widespread AI-Assisted Peer Reviews Boost Paper Scores and Acceptance Rates Giuseppe Russo Latona, Manoel Horta Ribeiro,Tim R. Davidson,Veniamin Veselovsky, *Robert West*. (rXiv:2 405.02150v1 [cs.CY] 3 May.

[27] Liang, Weixin., Izzo, Zachary., Zhang, Yaohui., Lepp, Haley., Cao, Hancheng., Zhao, Xuandong., Chen, Lingjiao., Ye, Haotian., Liu, Sheng., Huang, Zhi., et al. (2024). Monitoring aimodified content at scale: *A case study on the impact of chatgpt on ai conference peer reviews*. (arXiv preprint arXiv:2403.07183).

[28] Liang, Weixin., Zhang, Yaohui., Wu, Zhengxuan., Lepp, Haley., Ji, Wenlong., Zhao, Xuandong., Cao, Han cheng., Liu, Sheng., He, Siyu., Huang, Zhi, et al. *Mapping the increasing use of llms in scientific papers*. (arXiv preprint arXiv:2404.01268) 2024.

[29] Shai, Farber. (2025). Comparing human and AI expertise in the academic peer review process: towards a hybrid approach (January 21,). *Higher Education Research Development,* 0[10.1080/0729436 0.202 4.2445 575], Bar Ilan University Faculty of Law Research Paper No. 5105196, *Available at SSRN:* https://ssrn.com/abstract=5105196 or http://dx.doi.org/10.1080/07294360.2024.2445575.

[30] Renata, V., Lee, J. (2025). AI Reviewers: Are Human Reviewers Still Necessary *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 0 (0).

[31] Seghier, M. L. (2025). AI-powered peer review needs human supervision. *Journal of Information, Communication and Ethics in Society*, 23 (1) p. 104–116.

[32] Kankanhalli, Atreyi. (2024). Peer Review in the Age of Generative AI, *Journal of the Association for Information Systems*, 25 (1) 76-84.