ORIGINAL ARTICLE

Comparative Evaluation of Manuscript Review Accuracy: Human Reviewers vs. AI Chatbots

Renjith George[1], Noorliza Mastura Ismail[2], Meena Anand Kukkamalla[3], Adinegara Lutfi Abas[4], Htoo Htoo Kyaw Soe[4], Preethy Mary Donald[5], Abdul Rashid Hj Ismail[2]

[1] Department of Oral Pathology, Faculty of Dentistry, Manipal University College Malaysia, Jalan Batu Hampar, Bukit Baru, Melaka. 75150 Malaysia.
[2] Department of Community Dentistry, Faculty of Dentistry, Manipal University College Malaysia; Jalan Batu Hampar, Bukit Baru, Melaka. 75150 Malaysia.
[3] Department of Periodontics, F acuity of Dentistry, Mani pal University College Malaysia, Ja/an Batu Hampar, Bukit Baru, Melaka. 75150 Malaysia.
[4] Department of Community Medicine, Faculty of Medicine, Manipal University College Malaysia, Jalan Batu Hampar, Bukit Baru, Melaka. 75150 Malaysia.
[5] Department of Oral Medicine and Oral Radiology, Faculty of Dentistry, Manipal University College Malaysia, Jalan Batu Hampar, Bukit Baru, Melaka. 75150 Malaysia.

ABSTRACT

Introduction: The advent of large language models like ChatGPT has sparked discussions about their potential in manuscript reviewing. Few studies have rigorously compared AI accuracy to human reviewers in manuscript assessment. This study aims to empirically analyse and compare manuscript review accuracy between human reviewers and AI chatbots. **Methods:** A comparative study analysed performance of two human reviewers and two AI chatbots (OpenAI ChatGPT 4o and Anthropic Claude Sonnet 3.5) in manuscript evaluation. Sixty observational studies published between 2018-2023 from Scimago Journal Ranking (first quartile [Q1] to fourth quartile [Q4]) were selected. The STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist was used for evaluation. Each item was scored as fully reported, partially reported, or not reported, generating an Overall Completeness Score (OCS). Statistical analysis included descriptive statistics, correlation coefficients, and repeated measures ANOVA tests. **Results:** AI2 demonstrated the highest mean OCS of 75.99 (SD = 6.52), significantly higher than human reviewers and AI1, which clustered around mid-40s to high-40s. Strong positive correlation existed between human reviewers ($r = 0.614$, $p < 0.001$). AI1 showed strong correlations with both human reviewers, while AI2's correlations were weaker. AI2 maintained consistent high scores across all journal tiers, whereas human reviewers and AI1 exhibited declining scores for lower-tier journals. **Conclusion:** AI chatbots, particularly AI2, show potential in manuscript evaluation while highlighting continued value of human expertise. Findings suggest AI could serve as a powerful tool to support human reviewers rather than replace them.
Malaysian Journal of Medicine and Health Sciences (2025) 21(SUPP13):26-32. doi:10.47836/mjmhs.21.s13.5

Keywords: Artificial intelligence, ChatGPT, manuscript review, STROBE checklist, peer review, academic publishing

Corresponding Author:
Renjith George, MDS
Email: renjith.george@manipal.edu.my
Tel: +606-289 6662

INTRODUCTION

The advent of large language models (LLMs) like ChatGPT has sparked discussions about their potential to perform tasks traditionally done by humans, including manuscript reviewing in academic publishing (1). As healthcare professionals face the challenge of keeping up with rapidly evolving medical research while managing clinical duties, tools that can assist in evaluating research quality become increasingly valuable (2).

Previous studies have explored ChatGPT's capabilities in generating scientific abstracts and articles (3,4,5). However, few have rigorously compared its accuracy to human reviewers in manuscript assessment (6). This gap in knowledge prompted our investigation into the proficiency of Artificial intelligence (AI) chatbots in scoring manuscripts compared to human evaluation as the standard benchmark.

The rise of AI in academic tasks has led to debates about its ability to replace or augment human efforts (7). While AI chatbots have shown promise across the healthcare sector including diagnosing diseases and creating various academic content, their efficacy in reviewing detailed manuscripts remains uncertain, especially in the healthcare sector (8). This study aims to fill this gap by empirically analysing and comparing the accuracy of manuscript review conducted by human reviewers and

AI chatbots (Anthropic AI and OpenAI).

MATERIALS AND METHODS

Study Design

Comparative study analysing the performance of human reviewers and AI chatbots in manuscript evaluation.

Sample Selection

Two human reviewers were selected through expert sampling. Human Reviewer 1 (H1) was a clinician-researcher with 20 years of experience in epidemiological research and systematic reviews, holding a master's degree in public health with expertise in observational study design. Human Reviewer 2 (H2) was also a clinician-researcher with 15 years of experience in manuscript reviewing and publication, specializing in clinical research and evaluation. Both reviewers had prior experience with the STROBE checklist and had published peer-reviewed articles in the field of observational studies.

OpenAI (ChatGPT 4o) and Anthropic AI (Claude Sonnet 3.5) were used as the AI chatbots. We selected 60 observational studies (cross-sectional, case-control, and cohort) published between 2018-2023, available on PubMed Central and ranked in Scimago Journal Ranking (Q1 to Q4). Q1 is occupied by the top 25% of journals in the list; Q2 is occupied by journals in the 25 to 50% group; Q3 is occupied by journals in the 50 to 75% group and Q4 is occupied by journals in the 75 to 100% group. The most prestigious journals within a subject area are those which occupy the first quartile, Q1. Case series, case reports, and other study designs were excluded.

Evaluation Tool

The STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist (combined) was used for manuscript evaluation (9). The combined checklist consisting of 22 items, was expanded to accommodate each item for scoring purpose. The expanded checklist was employed to assess the completeness and quality of reporting in our study both by human reviewers and AI chatbots. The checklist guided the systematic review of key elements including the title, abstract, introduction, methods, results, and discussion sections. By applying the STROBE criteria, we ensured comprehensive coverage of critical aspects such as study design, participant selection, variable definitions, data sources, bias considerations, and statistical methods.

Review Process and AI Prompting

Human reviewers and AI chatbots independently evaluated the selected articles. Human reviewers performed the review in a setting without internet access to ensure consistency with AI evaluation conditions.

AI chatbots were prompted with the following standardized instructions:

AI Prompt Used: "You are an expert manuscript reviewer tasked with evaluating observational studies using the STROBE checklist. Please assess the provided manuscript for each of the 22 STROBE items. For each item, determine if it is:

- Fully reported (assign score of 2)
- Partially reported (assign score of 1)
- Not reported (assign score of 0)

Please provide your assessment for each STROBE item with justification. Focus on:

1. Title and Abstract completeness
2. Introduction rationale and objectives
3. Methods clarity (study design, setting, participants, variables, data sources, bias, study size, quantitative variables, statistical methods)
4. Results completeness (participants, descriptive data, outcome data, main results, other analyses)
5. Discussion interpretation, limitations, generalisability, and funding information

Provide an overall completeness score as percentage based on your item-by-item evaluation."

The prompt was developed through iterative testing using a separate computer and internet connection different from the one used for data collection procedure to avoid any possible machine learning influence on the main study results.

Scoring Method

Each STROBE checklist item was scored as fully reported (2 points), partially reported (1 point), or not reported (0 points). An Overall Completeness Score (OCS) was generated for each review, calculated as the percentage of total possible points achieved.

Ethical Approval

Ethical approval for the study was obtained from the Research Ethics Committee, Manipal University College Malaysia (MUCM) (MUCM/Research Ethics Committee – 001/2024). The names of the chatbots were anonymised after the methodology for confidentiality purposes.

Statistical Analysis

The analysis focused on the Overall Completeness Score (OCS), using descriptive statistics to summarize the data and provide an overview of the scores. Correlation coefficient was calculated to examine the relationship between the evaluations provided by the human reviewers and the chatbots. Additionally, one-way and two-way repeated measures ANOVA tests were conducted to further analyze the performance of the reviewers and interactions between the reviewers and the Q status being evaluated. The one-way ANOVA allowed for comparing performance across different reviewers, while the two-way repeated measures

ANOVA examined the interaction between the reviewers' performance and the Q status. P-value below the level of 0.05 was considered significant, indicating that the findings are unlikely to have occurred due to random chance.

RESULTS

Descriptive Statistics of Overall Completeness Scores

Table I shows the descriptive statistics of Overall Completeness Scores (OCS) for STROBE Checklist Evaluation. The scores reflect the completeness of reporting in the assessed studies, with higher scores indicating more complete reporting based on the STROBE criteria.

AI2 demonstrated the highest mean score of 75.99 (SD = 6.52), significantly higher than both human reviewers and AI1. AI1 and the human reviewers showed similar performance levels, with mean scores clustered around the mid-40s to high-40s. Specifically, AI1 achieved a mean score of 46.66 (SD = 8.03), slightly higher than Human Reviewer 1 (H1) with a mean of 46.28 (SD = 15.74), but lower than Human Reviewer 2 (H2) with a mean of 48.65 (SD = 8.63).

H1 showed the highest variability in scoring, with a standard deviation of 15.74, suggesting less consistency in their evaluations compared to the other reviewers. H2 and both AI systems demonstrated more consistent scoring patterns, as evidenced by their lower standard deviations.

Correlation Analysis

Table II shows results of Pearson Correlation Coefficients for Evaluations by Human Reviewers (H1 and H2) and AI Chatbots (AI1 and AI2). There is a strong positive correlation between H1 and H2 ($r = 0.614$, $p < 0.001$), indicating substantial agreement between the human reviewers. Similarly, AI1 shows strong positive correlations with both H1 ($r = 0.634$, $p < 0.001$) and H2 ($r = 0.568$, $p < 0.001$), suggesting that its evaluations align closely with those of human reviewers. Interestingly, AI2's correlations with other reviewers are weaker, with only a weak positive correlation with H1 ($r = 0.273$, $p = 0.035$) reaching statistical significance. The correlations between AI2 and H2 ($r = 0.151$, $p = 0.250$) and AI1 ($r = 0.117$, $p = 0.372$) are not statistically significant.

Table I: Descriptive Statistics of Overall Completeness Scores (OCS) for STROBE Checklist Evaluation by Human Reviewers and AI Chatbots

	N	Mean	Std. Deviation
H1	60	46.28	15.74
H2	60	48.65	8.63
AI1	60	46.66	8.03
AI2	60	75.99	6.52

H1 = Human Reviewer 1; H2 = Human Reviewer 2; AI1 = Anthropic Claude Sonnet 3.5; AI2 = OpenAI ChatGPT 4o; N = sample size; Std. Deviation = standard deviation. All reviewers evaluated 60 observational studies using the STROBE checklist.

Table II: Pearson Correlation Coefficients Between Overall Completeness Scores (OCS) by Human Reviewers and AI Chatbots

		H1	H2	AI1	AI2
H1	Pearson Correlation	1	0.614**	0.634**	0.273*
	Sig. (2-tailed)		0.000	0.000	0.035
	N	60	60	60	60
H2	Pearson Correlation	0.614**	1	0.568**	0.151
	Sig. (2-tailed)	0.000		0.000	0.250
	N	60	60	60	60
AI1	Pearson Correlation	0.634**	0.568**	1	0.117
	Sig. (2-tailed)	0.000	0.000		0.372
	N	60	60	60	60
AI2	Pearson Correlation	0.273*	0.151	0.117	1
	Sig. (2-tailed)	0.035	0.250	0.372	
	N	60	60	60	60

*H1 = Human Reviewer 1; H2 = Human Reviewer 2; AI1 = Anthropic Claude Sonnet 3.5; AI2 = OpenAI ChatGPT 4o; N = 60 observational studies

**. Correlation is significant at the 0.01 level (2-tailed). Correlation is significant at the 0.05 level (2-tailed).

Within-Subjects Contrast Analysis

Table III presents the results of within-subjects contrast tests analysing the patterns of Overall Completeness Scores (OCS) across four reviewers (H1, H2, AI1 and AI2). The results are depicted in the profile plot (Figure 1) which shows significant linear, quadratic, and cubic trends in the data, indicating complex differences in scoring patterns among the reviewers. The high F-values

Table III: Within-Subjects Contrast Tests for Overall Completeness Scores (OCS) Across Four Reviewers

Source	Reviewer	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Reviewer	Linear	22789.464	1	22789.464	209.615	.000	.780
	Quadratic-Inverted V shape	10896.828	1	10896.828	209.650	.000	.780
	Cubic-Z shape	3821.650	1	3821.650	99.114	.000	.627
Error (Reviewer)	Linear	6414.510	59	108.721			
	Quadratic-Inverted V shape	3066.598	59	51.976			
	Cubic-Z shape	2274.924	59	38.558			

Analysis of linear, quadratic, and cubic trends in OCS patterns across H1 (Human Reviewer 1), H2 (Human Reviewer 2), AI 1(AI reviewer 1)and AI2 (AI reviewer 2). df = degrees of freedom; F = F-statistic; Sig. = significance level; Partial Eta Squared (η^2) = effect size measure. All contrasts significant at $p < .001$, indicating substantial differences in scoring patterns among reviewers.

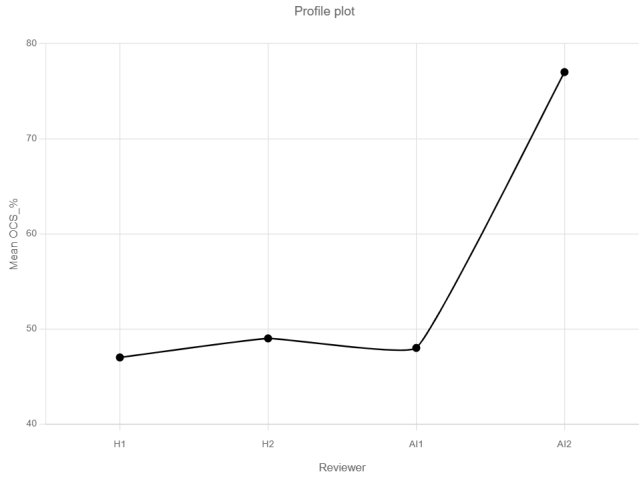

Figure 1: Profile plot of Overall Completeness Scores (OCS) across four reviewers (H1, H2, AI1, AI2), illustrating linear, quadratic, and cubic trends in scoring patterns and demonstrating significant differences among reviewers.

and low p-values (<.001) for all contrasts suggest strong, statistically significant differences, with large effect sizes (partial $\eta^2 > .6$) indicating substantial variance explained by these trends.

H1, H2, and AI1 have relatively similar mean OCS scores, clustered around the mid-range of the scale. There is a slight increase from H1 to H2, followed by a small decrease to AI1. AI2 shows a dramatic increase in mean OCS, standing out significantly from the other reviewers. The combination of trends (linear, quadratic, and cubic) captures the complex pattern of differences among the reviewers. The large effect sizes (partial η^2 values) for all three trends indicate that these patterns explain a substantial portion of the variance in the data.

Reviewer and Journal Quality Interaction

Table IV shows Results of Tests of Within-Subjects Effects for Reviewer and Q Status Interaction. The result shows highly significant effects ($p < .001$) for both the reviewer factor and its interaction with Q status, with large effect sizes (partial $\eta^2 = 0.846$ for Reviewer and 0.419 for the interaction). These results indicate substantial differences among reviewers and that these differences vary significantly across Q status levels. The results are depicted in Figure 2.

The graph reveals diverse performance patterns across Q status levels for different reviewers. AI2 consistently achieves high scores irrespective of Q status. In contrast, other reviewers exhibit a downward trend in scores from Q1 to Q4, with H1 demonstrating the most significant

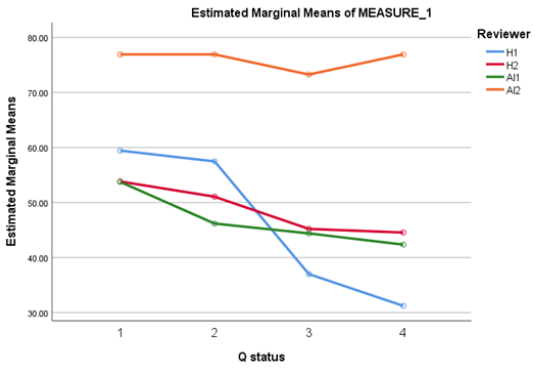

Figure 2: Interaction plot showing reviewer differences across Q status levels, illustrating how evaluation patterns vary significantly among reviewers (H1, H2, AI1,AI2) depending on Q status, with highly significant main and interaction effects.

decline. Notably, H2 and AI1 display strikingly similar scoring patterns. Both show a gradual decline in mean OCS as journal quality decreases, but this decline is less severe compared to H1.

DISCUSSION

The study compared the performance of two human reviewers (H1 and H2) and two AI systems (AI1 and AI2) in evaluating 60 observational studies using the STROBE checklist. The results show notable differences in the mean Overall Completeness Scores (OCS) across the four reviewers (Table I). The results reveal intriguing patterns that both corroborate and diverge from existing literature on AI applications in academic publishing and medical decision-making.

Performance Comparison

The results showed the AI2 chatbot consistently higher Overall Completeness Scores (OCS) compared to human reviewers and AI1, regardless of the tier of the journal. This aligns with recent research demonstrating that large language models (LLMs) can produce high-quality and safe medical responses (3,10). However, our study extends these findings by directly comparing AI performance to human reviewers in the specific context of manuscript evaluation.

Notably, the evaluation results from AI1 showed a stronger alignment with human reviewers, especially H2. This closer match in scoring trends implies that certain AI models might be able to emulate human-style reasoning when assessing manuscripts. This observation brings a new perspective to the current discussions about

Table IV: Tests of Within-Subjects Effects for Reviewer Performance and Reviewer Ч Journal Quality (Q Status) Interaction

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Reviewer	Sphericity Assumed	37507.94	3	12502.64	307.57	0.000	0.846
Reviewer * Q status	Sphericity Assumed	4926.94	9	547.43	13.46	0.000	0.419
Error (Reviewer)	Sphericity Assumed	6829.08	168	40.64			

Analysis examining differences among four reviewers (H1, H2, AI1, AI2) and how these differences vary across journal quality tiers (Q1-Q4). df = degrees of freedom; F = F-statistic; Sig. = significance level; Partial Eta Squared (η^2) = effect size measure. Large effect sizes indicate substantial reviewer differences and significant interaction effects across journal quality levels. N = 60 observational studies per reviewer.

AI chatbot's potential in academic evaluation processes. Recent studies note that while AI could assist in tasks like information retrieval, caution should be exercised against over-reliance, as this could lead to oversimplification of complex concepts and a lack of critical thinking (7,11). Concerns are also raised about the trustworthiness of AI-generated information in healthcare. The authors emphasize the need for a balanced approach, where AI is viewed as a complementary tool, not a replacement for human expertise.

Recent studies have explored the current applications and future potential of ChatGPT in medical specialties, highlighting its capability to generate high-quality content across various tasks (6,12). This aligns with our findings on AI2's high performance, suggesting that the capabilities of AI in specialized medical fields may extend to the broader context of manuscript evaluation.

Impact of Human Reviewer Expertise

The inclusion of detailed information about human reviewer backgrounds provides important context for interpreting the results. H1's clinical-research background with 20 years of experience and H2's biostatistical expertise with 15 years of experience represent different perspectives in manuscript evaluation. The higher variability in H1's scoring (SD = 15.74) compared to H2's more consistent pattern (SD = 8.63) may reflect differences in evaluation approaches between clinical and statistical perspectives. This variability underscores the inherent subjectivity in human review processes and supports the potential value of AI systems in providing more consistent evaluations.

Consistency Across Journal Tiers

Our analysis revealed that AI2 maintained consistently high OCS across all journal quality tiers (Q1-Q4), while human reviewers and AI1 exhibited declining scores for lower-tier journals. This consistency in AI performance across journal qualities is a novel finding not previously reported in the literature we reviewed. It raises important questions about the potential for AI to standardize the review process across different journal tiers, potentially addressing concerns about review quality in lower-ranked journals.

The divergence in scoring patterns for lower-tier journals (Q3 and Q4) across all reviewers highlights the increased variability in assessing articles in less prestigious publications (Figure 2). This divergence might reflect the greater challenges in evaluating research that may not have benefited from the rigorous selection processes of top-tier journals.

Methodological Considerations and Standardization

The deliberate decision not to standardize the review process between human and AI evaluators warrants discussion. While this approach aimed to maintain the natural evaluation methods of each reviewer type, it introduces potential validity concerns when making direct comparisons. Human reviewers brought their professional experience and intuitive judgment to the evaluation, while AI systems followed programmed instructions based on the provided prompts. This fundamental difference in evaluation approach may explain some of the observed variations in scoring patterns.

The lack of process standardization reflects real-world conditions where human reviewers rely on experience and judgment, while AI systems operate according to specific instructions. However, this methodological choice limits the ability to attribute performance differences solely to reviewer type versus evaluation methodology. Future studies could benefit from exploring both standardized and naturalistic evaluation conditions to better understand the sources of performance variations.

Implications for Peer Review

The high performance of AI chatbots in manuscript evaluation suggests a potential for these tools to augment or even partially automate certain aspects of the peer review process. This aligns with ongoing discussions in the field about the potential of AI to address challenges in academic publishing, such as the increasing volume of submissions and the time constraints faced by human reviewers (4).

The contrast between AI2's consistently high scores and the more varied assessments from human reviewers and AI1 highlights the importance of careful and measured deployment of AI in evaluation processes. This discrepancy emphasizes that not all AI systems may be equally suited for complex academic review tasks, underscoring the need for thorough validation and thoughtful implementation strategies.

Ethical Considerations

The robust performance of AI in our study underscores the need for careful consideration of ethical implications, as highlighted in recent guidance on AI ethics and governance for large multi-modal models (8,13,14). As our study demonstrates AI's potential in academic review processes, it is crucial to establish clear guidelines for its use to ensure transparency, accountability, and the maintenance of academic integrity.

Recent studies have examined the ethical implications of AI in healthcare, emphasizing the need for responsible AI development and deployment (15,16). Their findings highlight the importance of maintaining human oversight and ensuring that AI systems complement rather than replace human expertise, which aligns with our observations in the manuscript review process.

Limitations and Future Directions

Several limitations should be acknowledged in this study.

First, we focused exclusively on observational studies and utilized only the STROBE checklist as the evaluation tool, which limits the generalizability of findings to other study types and evaluation frameworks. Second, our sample included only two human reviewers and two AI chatbots, which may not fully capture the variability in reviewer performance and could be influenced by individual reviewer characteristics or AI model-specific attributes. Third, the lack of reviewer blinding meant that participants knew the evaluation purpose, which might have introduced evaluation bias.

Additionally, while we provided detailed information about human reviewer expertise, the AI chatbots' specific training data and model updates were not controlled for, as these systems may have been trained on different datasets and updated at different times during the study period. The prompt development process, while systematic, was based on trial-and-error methodology and may not represent the optimal prompting strategy for each AI system.

The geographical and network constraints of our AI evaluations (same IP address and region) may have influenced AI performance, though the extent of such influence remains unknown. Furthermore, our study does not address the potential for AI to learn and improve over time, a capability that could significantly impact its future role in academic publishing.

Future research should address these limitations by incorporating a larger and more diverse pool of human reviewers with varying levels of expertise, expanding evaluations to include additional study designs such as randomized controlled trials and systematic reviews using established appraisal tools including CONSORT and PRISMA checklists, and implementing blinded assessment procedures where feasible to reduce bias. Further investigations should also examine the impact of different prompting strategies and versions of AI models on performance, conduct multi-center studies across diverse geographical regions to enhance generalizability, and assess the temporal stability of AI outputs over time. In addition, future work should explore the integration of advanced search algorithms with AI-powered manuscript review systems to improve the accuracy, consistency, and efficiency of scholarly evaluation processes.

CONCLUSION

Our study demonstrates the potential of AI chatbots, in manuscript evaluation while also highlighting the continued value of human expertise. The findings suggest a future where AI could serve as a powerful tool to support and enhance human reviewers, rather than replace them. The closer alignment of AI1with human reviewer patterns, combined with the consistent but potentially over corrective performance of AI2, suggests that different AI systems may serve different roles in the peer review process.

As we move forward, it will be crucial to develop frameworks that effectively integrate AI capabilities with human insight in the peer review process, ensuring the continued quality and integrity of academic publishing in the medical field. The observed differences in AI performance patterns underscore the need for careful validation and selection of AI tools for specific aspects of manuscript evaluation.

REFERENCES

1. Sarkar M, Găman MA, Puyana JC, Bonilla-Escobar FJ. Artificial intelligence in medicine and medical education: current applications, challenges, and future directions. Int J Med Students. 2024;12(1):9-13. DOI: 10.5195/ijms.2024.2626
2. MacIntyre MR, Cockerill RG, Mirza OF, Appel JM. Ethical considerations for the use of artificial intelligence in medical decision-making capacity assessments. Psychiatry Res. 2023;328:115466. DOI: 10.1016/j.psychres.2023.115466
3. Tang L, Sun Z, Idnay B, et al. Evaluating large language models on medical evidence summarization. NPJ Digit Med. 2023;6:158. DOI: 10.1038/s41746-023-00896-7
4. Enago Academy. ChatGPT and AI tools in academic publishing: boon or bane? Enago Academy. Published August 18, 2023. Accessed January 15, 2024. https://www.enago.com/academy/chatgpt-and-ai-tools-in-academic-publishing/
5. Mitrović S, Andreoletti D, Ayoub O. ChatGPT or human? detect and explain. explaining decisions of machine learning model for detecting short ChatGPT-generated text. arXiv preprint. 2023. arXiv:2301.13852. DOI: 10.48550/arXiv.2301.13852
6. Temperley HC, Mac Curtain BM, Corr A, Meaney JF, Kelly ME, Brennan I. Current applications and future potential of ChatGPT in radiology: a systematic review. J Med Imaging Radiat Oncol. 2024;68(3):257-264. DOI: 10.1111/1754-9485.13621
7. Roberts RH, Sharma S, Halawa A. ChatGPT and medical education: friend or foe? BMJ Health Care Inform. 2023;30(1):e100830.
8. World Health Organization. WHO releases AI ethics and governance guidance for large multi-modal models. WHO. Published January 18, 2024. Accessed January 20, 2024. https://www.who.int/news/item/18-01-2024-who-releases-ai-ethics-and-governance-guidance-for-large-multi-modal-models
9. STROBE Statement. Checklists. STROBE Statement. Updated August 30, 2024. Accessed September 5, 2024. https://www.strobe-statement.org/checklists/
10. Stokel-Walker C. ChatGPT listed as author on

research papers: many scientists disapprove. Nature. 2023;614(7947):214-216. DOI: 10.1038/d41586-023-00107-z
11. Else H. ChatGPT: five priorities for research. Nature. 2023;612(7938):19-20.
12. Gusenbauer M, Haddaway NR. What everyresearcher should know about searching: clarified concepts, search advice, and an agenda to improve finding in academia. Res Synth Methods. 2023;14(2):248-261. DOI: 10.1002/jrsm.1457
13. Tilala MH, Chenchala PK, Choppadandi A, et al. Ethical considerations in the use of artificial intelligence and machine learning in health care: a comprehensive review. Cureus. 2024;16(6):e62443. DOI: 10.7759/cureus.62443
14. Gao CA, Howard FM, Markov NT, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med. 2024;7:75. DOI: 10.1038/s41746-023-00819-6
15. OpenAI. GPT-4 technical report. arXiv preprint arXiv:2303.08774. 2024. DOI: 10.48550/arXiv.2303.08774
16. Anthropic. Claude 3.5 Sonnet: advancing AI safety and capability. Anthropic. Published June 20, 2024. Accessed September 1, 2024.