*Baku State University Journal of Life Sciences & Biology **2026**, v.3 (1), p. 6-14*

journal homepage: http://bsuj.bsu.edu.az/en

https://doi.org/10.30546/300045.2026.03.1.509

STRENGTHS AND LIMITATIONS OF AI-BASED ARTICLE ANALYSIS TOOLS: A CASE STUDY ON RESEARCH AND REVIEW LITERATURE

Elnur Haciyev, Nigar Huseynova, Aysu Tinayeva, Rovshan Khalilov, Mahammad Nadirov

Baku State University, 23 Z. Khalilov st., AZ-1073, Baku, Azerbaijan

Received 23 february 2026; accepted 31 march 2026

Abstract

Artificial intelligence (AI)-based tools are increasingly used in academia to support literature reviews, preliminary research exploration, and manuscript screening. However, their ability to perform higher-order scientific reasoning and critically evaluate research remains unclear. This study critically assessed an AI program's analytical performance across two biomedical article types: a Q1 original research article and a Q3 review article. Using a qualitative, comparative case-study design, AI-generated analyses were evaluated by human experts based on interpretative accuracy, methodological awareness, quantitative reasoning, mechanistic depth, and contextual integration. The AI performed well on the review article, demonstrating strong thematic organization, conceptual summarization, and identification of broad research gaps. In contrast, its analysis of the original research article was largely descriptive, with limited engagement in quantitative data interpretation, mechanistic insight, and methodological critique, and included overgeneralizations of biological conclusions. These findings indicate that while AI tools can support literature exploration and education, they are not yet suitable as independent evaluators of high-impact research. Future development should focus on quantitative reasoning, mechanistic interpretation, and article-type–aware strategies to enhance scientific rigor.

Keywords: *artificial intelligence; literature analysis; scientific reasoning; review article; original research; biomedical research*

---

1. Introduction

1.1. Background

In academic settings, AI programs are now frequently used by undergraduate and graduate students to

* Corresponding author. Tel.: +994 70 244 15 42
E-mail address: n.huseynova@live.com

support literature reviews, by researchers to gain rapid overviews of unfamiliar fields, and by reviewers or editors as auxiliary tools for manuscript screening [8, 10, 16]. The appeal of these systems lies in their ability to quickly extract key themes, identify main conclusions, and generate structured summaries that would otherwise require substantial human effort. However, while AI excels at syntactic and semantic pattern recognition, scientific literature analysis requires more than surface-level comprehension [2]. Robust academic evaluation depends on critical appraisal, including assessment of methodological rigor, interpretation of experimental data, identification of conceptual gaps, and contextualization within the broader scientific landscape [1, 15].

Consequently, the increasing reliance on AI-generated analysis raises important questions about the depth, reliability, and critical value of such outputs. While AI tools can efficiently summarize content, it remains uncertain to what extent they can replicate higher-order scientific reasoning that is traditionally expected from human experts [3, 16].

1.2 Problem Statement

Despite the widespread adoption of AI-based tools for analyzing scientific articles, systematic evaluations of their analytical performance remain limited [5]. Most available reports focus on technical aspects such as language fluency, coherence, or summarization accuracy, rather than on the quality of scientific reasoning embedded in the analysis [12]. In particular, there is insufficient evidence regarding whether AI- generated analyses can meaningfully distinguish between different types of scientific publications and adapt their critique accordingly [13].

Original research articles and review articles represent fundamentally different forms of scientific communication [4]. Original research papers emphasize experimental design, data integrity, statistical validity, and mechanistic interpretation, whereas review articles prioritize synthesis, comparison of studies, conceptual frameworks, and identification of knowledge gaps [14]. An effective analytical tool should therefore demonstrate the capacity to adjust its evaluation strategy depending on the article type [17]. However, the extent to which current AI programs can capture these distinctions and provide article-type–specific insights remains unclear [11].

The lack of systematic, comparative assessments creates uncertainty about the appropriate role of AI in scholarly workflows and raises concerns about potential overreliance on AI-generated interpretations, particularly in educational and research training contexts[(9].

1.3 Aim and Scope

The aim of this study is to critically evaluate the performance of an AI-created program in analyzing scientific literature by examining its strengths and limitations across two distinct publication formats: one original research article and one review article. Rather than assessing correctness alone, the evaluation focuses on qualitative dimensions of scientific analysis, including accuracy of interpretation, depth of reasoning, critical appraisal, and contextual understanding.

Specifically, this work investigates how effectively the AI program identifies key findings, methodological considerations, and conceptual implications in an original experimental study, and how well it synthesizes themes, evaluates evidence, and highlights gaps in a review article. By applying a consistent evaluation framework to both article types, the study aims to reveal differences in AI performance that may be masked by general-purpose summarization metrics.

The scope of this study is intentionally qualitative and exploratory, emphasizing interpretative quality over quantitative benchmarking. The analysis is designed to reflect real-world usage scenarios in which researchers or students might rely on AI-generated critiques to support their understanding of scientific literature.

1.4 Novelty

This study presents a case-based comparative assessment of an AI-generated scientific analysis using two fundamentally different article formats—original research and review articles—within the same disciplinary context. Unlike prior evaluations that primarily assess summarization accuracy or linguistic quality, this work emphasizes qualitative scientific critique, focusing on reasoning depth, methodological awareness, and contextual integration.

By systematically comparing AI-generated analyses across article types, the study offers novel insights into the strengths and limitations of current AI tools as analytical assistants rather than mere summarization engines. The findings contribute to ongoing discussions on responsible AI integration in research and education, highlighting areas where AI can meaningfully support scholarly work and where human expertise remains indispensable.

2. Materials and Methods

2.1 Study Design

This study employed a qualitative, comparative case-study design to evaluate the performance of an AI-generated program for scientific article analysis. The evaluation was conducted using two representative scientific publications from the biomedical field: one original research article and one narrative review article. The study design aimed to assess how the AI program analyzes different article formats and whether it adapts its analytical depth, structure, and critical reasoning accordingly.

Rather than benchmarking the AI output against a predefined "correct" answer, the analysis focused on qualitative dimensions of scientific reasoning, including interpretative accuracy, depth of critique, contextual awareness, and identification of limitations. This approach reflects realistic academic use cases in which AI-generated analyses are used as supportive tools rather than authoritative evaluations.

2.2. Development of the web application

The platform's technical architecture combines the Laravel PHP framework with Livewire for reactive, server-rendered interfaces. Laravel ensures modular development, API integration, and maintainability, while Livewire enables real-time interactivity within Blade templates, minimizing reliance on client-side JavaScript frameworks.

SQLite serves as the primary database, providing lightweight storage for user-uploaded documents, metadata, chat histories, and AI responses. Its simplicity supports rapid development and small-scale deployments, with potential scalability to PostgreSQL or MySQL for larger implementations.

OpenAI's GPT models are integrated via the openai-php/laravel package, utilizing the chat/completions endpoint for streaming responses. This approach delivers feedback incrementally, enhancing the conversational experience. Laravel Reverb (WebSocket server) and Echo (JavaScript library) handle real-time communication, ensuring instant updates and smooth rendering of long-form outputs.

The work-flow includes:

1. **Document Upload and Extraction**: Users upload manuscripts in PDF or text formats, and the system extracts content using parsing libraries.
2. **Context Initialization**: Prompts are generated based on document content, user preferences, and review criteria (e.g., structure, clarity, novelty).
3. **Real-Time API Interaction**: The system queries the LLM, streaming responses to the interface in real time.
4. **Response Storage**: Feedback, interactions, and metadata are saved in SQLite for traceability and analysis.
5. **Session Management**: Users can resume sessions, review past feedback, and track manuscript progress.

2.3 Selection of Scientific Articles

The selection of a Q1 original research article and a Q3 review article was intentional to represent two distinct levels of scientific rigor and analytical demand (Huseynova et al., 2024, 2025). Q1 articles typically require in-depth methodological, quantitative, and mechanistic evaluation, whereas Q3 review articles emphasize conceptual synthesis and thematic organization. This contrast enabled a more robust assessment of whether the AI system adapts its analytical strategy according to article type and complexity. Full references of the analyzed articles are provided in the Supplementary Materials. Both articles addressed closely related scientific themes to minimize disciplinary bias while allowing evaluation of the AI platform's ability to adapt its analytical strategy to different article types and journal standards.

2.4 Evaluation Criteria

In addition to content-based evaluation, the AI-generated analyses were assessed in light of journal quartile expectations, considering whether the analytical depth, critical reasoning, and contextual interpretation were appropriate for a Q1 original research article versus a Q3 review article.

2.5 Analytical Evaluation Framework

To ensure a systematic and transparent evaluation of the AI-generated analyses, a qualitative analytical framework was applied across both article types. The framework was designed to capture higher-order scientific reasoning rather than surface-level summarization quality. Each AI-generated analysis was assessed according to the following dimensions:

Interpretative accuracy: the extent to which the AI correctly identified the main scientific claims, conclusions, and intended interpretations of the original authors without distortion or omission.

Methodological awareness: the AI's ability to recognize and comment on key aspects of experimental design or review methodology, including strengths, limitations, and appropriateness of the chosen approaches.

Quantitative reasoning: the degree to which numerical data, statistical indicators, and quantitative outcomes (e.g., IC_{50} values, fold changes, combination indices) were meaningfully interpreted rather than merely restated.

Mechanistic depth: the capacity of the AI to link observed results to underlying biological or molecular mechanisms, including signaling pathways, causal relationships, and mechanistic plausibility.

Contextual integration: how effectively the AI situated the analyzed article within the broader scientific literature, including comparison with prior findings and identification of unresolved questions.

Generalization control: the AI's tendency to avoid overgeneralization, particularly when drawing conclusions from limited datasets or compound-specific findings.

This framework was applied consistently to both the original research article and the review article to enable structured comparison of AI performance across distinct publication formats.

2.6 Comparative Analysis Approach

A side-by-side qualitative comparison was conducted to identify similarities and differences in AI performance across the two article types. Particular attention was given to whether the AI program adjusted its analytical strategy to match the structural and conceptual demands of original research versus review literature. The overall workflow of article selection, AI analysis, and human-led evaluation is illustrated in Figure 1.

The comparative analysis focused on:

- Responsiveness to experimental design and data interpretation in the original research article
- Effectiveness of literature synthesis and conceptual integration in the review article
- Balance between descriptive summarization and critical evaluation in both cases

Observations were systematically documented and synthesized to derivethe overarching strengths and weaknesses of the AI program.

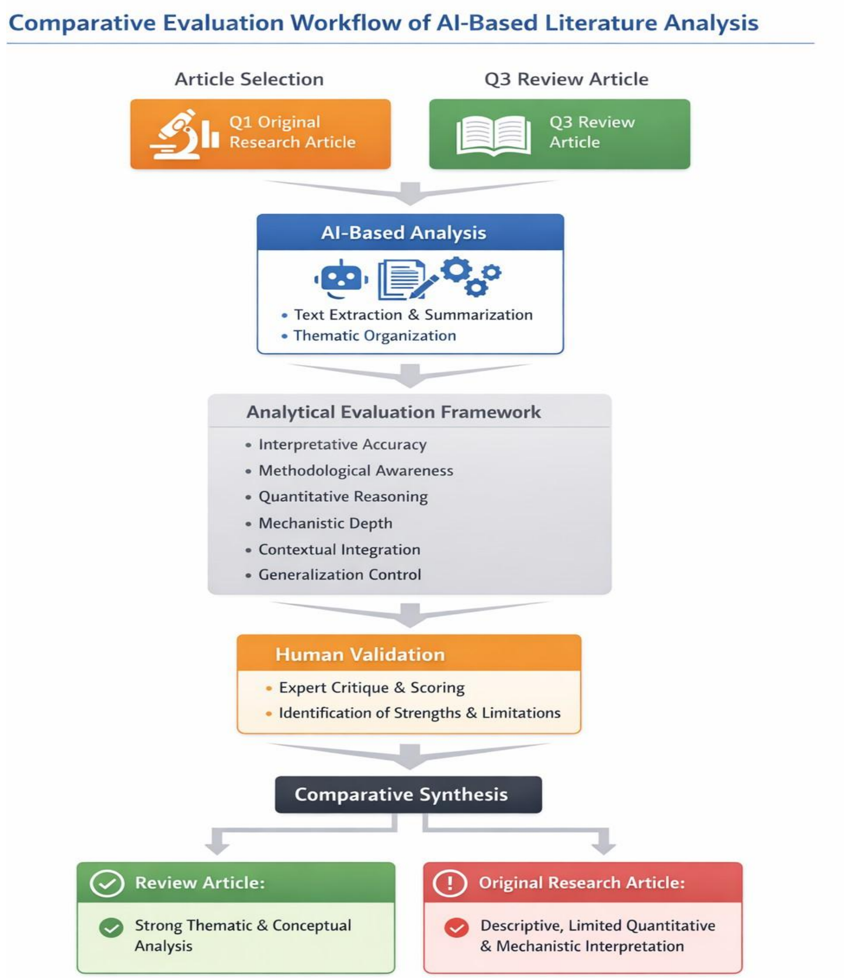

Fig. 1. Comparative evaluation workflow of AI-based literature analysis, illustrating the stepwise process from article selection to AI analysis, human-led validation, and final comparative synthesis of strengths and limitations

2.7. Fine-tuning

To ensure the platform could deliver academically rigorous feedback in Azerbaijani with sufficient detail, we undertook an initial fine-tuning process targeting the LLM's language generation capabilities. The fine-tuning was guided by specific system-level requirements: the model had to communicate exclusively in Azerbaijani (avoiding any Turkish influence), provide academic-grade analysis and summarization for scientific articles, and produce a minimum 500-word analysis per response to meet the depth expected in peer review contexts.

2.8. Ethical Considerations

This study did not involve human participants, animal subjects, or sensitive personal data. All analyzed articles were publicly available and properly cited. The AI program was used solely as an analytical tool, and all interpretations and evaluations of its outputs were performed by the authors. No automated decisions or conclusions were adopted without human critical oversight.

3. Results

3.1 Performance on Review Article (Q3)

The AI-generated analyses of the review article demonstrated strong performance in thematic organization and conceptual synthesis. The system successfully identified major research themes, including the role of flavonoids as chemosensitizers, synergistic versus antagonistic interactions, and key limitations such as lack of clinical validation and dose optimization challenges.

Interpretative accuracy and contextual integration were high, as the AI appropriately summarized the scope and conclusions of the review. Mechanistic depth remained moderate, focusing on general biological effects rather than pathway-specific explanations. Quantitative reasoning was not applicable due to the narrative nature of the article.

3.2 Performance on Original Research Article (Q1)

In contrast, the AI analysis of the original research article showed limited depth in critical evaluation. While the study objectives and experimental design were correctly summarized, the system failed to meaningfully interpret key quantitative data, including IC_{50} values, combination indices, and fold-change measurements.

Methodological awareness was limited, with no substantial critique of experimental design (e.g., single cell line use, lack of in vivo validation). Mechanistic interpretation remained superficial, with generalized references to apoptosis rather than pathway-specific insights.

AI performs better on narrative synthesis than on data interpretation. It does not adequately adapt its analytical depth to different article types. Quantitative and mechanistic reasoning remain major limitations.

4. Discussion

This study presents a human-validated evaluation of an AI-based program designed to analyze scientific articles, using two representative cases: a Q3 review article and a Q1 original research article within the biomedical field. The AI program was employed exclusively to generate the initial analyses, while all subsequent evaluation, critique, and interpretation of these analyses were carried out by human experts—specifically, the authors of the original articles. This human- led assessment ensures that the conclusions drawn reflect domain expertise, methodological awareness, and critical scientific judgment, rather than automated self-evaluation by the AI system. AI analyses were assessed according to interpretative accuracy, methodological awareness, quantitative reasoning, mechanistic depth, contextual integration, and generalization control.

A comparative synthesis of the two evaluated cases reveals a clear divergence in AI performance depending on article type. In the review article, the AI demonstrated strong performance in thematic organization, conceptual summarization, and identification of broad research gaps. Its analysis aligned well with the narrative nature of review literature, where synthesis and high-level interpretation are prioritized over detailed data interrogation. In contrast, the analysis of the original research article exposed substantial limitations. While the AI accurately summarized the study objectives and experimental setup, it did not adequately engage with the quantitative and mechanistic complexity of the work. Key experimental metrics were acknowledged but not critically interpreted, and methodological constraints were largely overlooked. This contrast indicates that the AI's analytical strategy remains largely invariant across article types, resulting in a mismatch between analytical depth and the epistemic demands of high-impact experimental research.

The findings demonstrate that the AI program performs relatively well when analyzing review literature. In the case of the Q3 review article, the AI-generated analysis was logically structured, conceptually accurate, and pedagogically useful. It successfully identified the general role of flavonoids as chemosensitizers, highlighted synergistic versus antagonistic interactions, and recognized broader research gaps such as the lack of clinical trials, dose optimization challenges, and interpatient variability. These strengths suggest that the AI program is suitable for introductory literature exploration, teaching contexts, and early-stage research orientation, where high-level synthesis is often sufficient.

However, the human evaluation revealed clear limitations in the AI program's capacity for deep scientific critique. The analysis of the Q1 original research article remained largely descriptive and did not adequately reflect the experimental rigor, quantitative depth, or mechanistic complexity of the study. Although the AI correctly summarized the study aim and methodology, it failed to meaningfully interpret key quantitative parameters such as IC_{50} values, combination indices, and fold-change metrics. Importantly, the AI did not distinguish between metabolic activity measurements derived from MTT assays and true cytotoxic effects, a distinction that is fundamental for correct biological interpretation.

From a methodological perspective, the AI analysis did not critically assess the strengths and limitations of the experimental design. Human reviewers noted the absence of discussion regarding the reliance on a single in vitro cell line, lack of in vivo validation, and limited generalizability of the findings. Additionally, the AI did not compare the employed methods with alternative or complementary approaches, nor did it evaluate their suitability in addressing the study's mechanistic hypotheses. This indicates that the program currently lacks the capacity to evaluate experimental robustness and translational relevance—core elements of expert scientific critique.

Mechanistic interpretation was another major limitation. While the AI referenced apoptosis and mitochondrial dysfunction, it did so in broad terms, without identifying specific signaling pathways, molecular checkpoints, or causal mechanisms supported by the data. Furthermore, the AI demonstrated a tendency to overgeneralize biological conclusions, particularly by treating flavonoids as a homogeneous class, despite the original article's emphasis on compound-specific effects and the possibility of antagonistic interactions. Human feedback underscored that such generalizations risk misrepresenting the scientific message and could mislead non-expert readers.

Taken together, the human-authored evaluations indicate that the AI program does not yet adapt its analytical depth to article type, journal impact level, or experimental complexity. While it is capable of organizing information and identifying general trends, it does not currently perform the integrative reasoning required to link quantitative data, experimental design, mechanistic insight, and biological context into a cohesive critical analysis. As a result, its outputs should be interpreted as supportive summaries rather than authoritative scientific evaluations. Beyond these analytical limitations, the findings of this study have important implications for how AI-generated analyses should be used across research practice, educational contexts, and the peer review process.

Implications for Research, Education, and Peer Review

The findings of this study have distinct implications for different stakeholders within the academic ecosystem. For students and early-career researchers, AI-generated analyses may serve as useful entry points for understanding unfamiliar topics and organizing foundational knowledge. However, reliance on such tools without critical supervision risks reinforcing superficial understanding, particularly in data-driven disciplines. For active researchers, AI tools can support preliminary literature exploration and hypothesis contextualization but should not be relied upon for evaluating experimental rigor, quantitative validity, or mechanistic plausibility. Human expertise remains essential for interpreting complex datasets and assessing translational relevance. For peer reviewers and editors, the results underscore that AI-based tools are currently unsuitable as independent evaluators of original research manuscripts. While they may assist with initial screening or structural assessment, they lack the depth required for authoritative scientific judgment and should only be used as supplementary aids under human oversight.

Limitations

This study has several limitations that should be acknowledged. First, the analysis is based on a small case sample comprising only two scientific articles, which limits the generalizability of the findings. Second, only a single AI-based analytical platform was evaluated, and performance may differ across models or system configurations. Third, although the evaluation was conducted by domain experts, the assessment remains qualitative in nature and may be influenced by disciplinary expectations and expert judgment. Despite these limitations, the case-based approach provides valuable insights into real-world usage scenarios and highlights systematic patterns in AI performance that warrant further investigation.

Future Directions

Future research should aim to enhance the analytical capabilities of AI-based literature review tools by integrating quantitative reasoning, mechanistic interpretation, and article-type–specific strategies. Specifically, AI models could be trained to meaningfully interpret experimental metrics such as IC_{50} values, fold changes, and combination indices, rather than simply reporting them.

Incorporating pathway-level and molecular mechanisms would allow the AI to contextualize findings within broader biological processes, improving mechanistic depth. Adaptive analytical strategies tailored to article type are also essential, enabling high-level synthesis for review articles and rigorous quantitative and mechanistic evaluation for original research papers. Furthermore, the integration of figure and table interpretation would enhance the AI's capacity to engage directly with experimental data. Employing dynamic prompting frameworks and human-in-the-loop approaches can further calibrate analytical depth according to journal standards and study complexity. Finally, cross-model validation across multiple AI platforms is recommended to systematically assess performance differences and identify areas for improvement, ultimately guiding the development of more robust and scientifically reliable AI-assisted literature analyses.

Conclusion

This study provides a human-validated assessment of an AI-based program for scientific literature analysis, comparing its performance on a Q1 original research article and a Q3 review article within the biomedical field. The results indicate that the AI performs well in synthesizing and summarizing review literature, effectively organizing concepts, identifying thematic patterns, and highlighting broad research gaps. In contrast, its analysis of original research articles exhibited substantial limitations, particularly in engaging with quantitative data, experimental design, and mechanistic complexity. Key metrics were acknowledged but not critically interpreted, methodological constraints were often overlooked, and mechanistic reasoning remained superficial. These findings underscore that while AI tools can serve as useful educational aids and support early-stage research exploration, they cannot replace expert human judgment in evaluating experimental rigor, interpreting complex datasets, or assessing translational relevance. Human oversight remains essential to ensure accuracy, contextualization, and scientific validity.

Future development of AI-assisted literature analysis should focus on integrating quantitative reasoning, pathway-level mechanistic interpretation, and article-type–aware analytical strategies. Incorporating the interpretation of figures and tables, dynamic prompting, and human-in-the-loop approaches can further enhance analytical depth. Cross-model validation is also recommended to systematically evaluate performance and identify areas for improvement. Collectively, these steps will help transform AI tools from supportive summarization engines into reliable, scientifically robust analytical assistants capable of meaningfully contributing to research, education, and scholarly review processes.

References

[1] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

[2] Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6). https://doi.org/10.1073/pnas.2218523120

[3] Else, H. (2023). Abstracts written by ChatGPT fool scientists. *Nature*, 613(7944), 423. https://doi.org/10.1038/d41586-023-00056-7

[4] Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal, 26*(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

[5] Helms Andersen, T., Marcussen, T. M., Termannsen, A. D., Lawaetz, T. W. H., & Nørgaard, O. (2025).

Using artificial intelligence tools as second reviewers for data extraction in systematic reviews: A performance comparison of two AI tools against human reviewers. *Cochrane Evidence Synthesis andMethods*, *3*(4), e70036. https://doi.org/10.1002/cesm.70036

[6] Huseynova N., Baran Z., Khalilov R., Mammadova A., & Baran Y. Chemosensitizing effect of apigenin on T-ALL cell therapy. *Frontiers in Pharmacology*, 2025, *16*. https://doi.org/10.3389/fphar.2025.1631505

[7] Huseynova N., Çetinkaya, M., Baran, Z., Khalilov, R., Mammadova, A., & Baran, Y. *Flavonoids as Chemosensitizers in Leukemias* (2024, pp. 205–234). https://doi.org/10.1007/5584_2024_828

[8] Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

[9] Liu, J. Q. J., Hui, K. T. K., Al Zoubi, F., Zhou, Z. Z. X., Samartzis, D., Yu, C. C. H., Chang, J. R., & Wong, A.Y. L. (2024). The great detectives: Humans versus AI detectors in catching large language model-generated medical writing. *International Journal for Educational Integrity*, *20*(1), 8. https://doi.org/10.1007/s40979-024-00155-6

[10] Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: How may AI and GPT impact academia and libraries? *Library Hi Tech News*, *40*(3), 26–29. https://doi.org/10.1108/LHTN-01-2023-0009

[11] Mostafapour, M., Fortier, J. H., Pacheco, K., Murray, H., & Garber, G. (2024). Evaluating literature reviews conducted by humans versus ChatGPT: Comparative study. *JMIR AI*, *3*, e56537. https://doi.org/10.2196/56537

[12] Parmar, M., Patel, N., Varshney, N., Nakamura, M., Luo, M., Mashetty, S., Mitra, A., & Baral, C. (2024). LogicBench: Towards systematic evaluation of logical reasoning ability of large language models (No. arXiv:2404.15522). *arXiv*. https://doi.org/10.48550/arXiv.2404.15522

[13] Shcherbiak, A., Habibnia, H., Böhm, R., & Fiedler, S. (2024). Evaluating science: A comparison of human and AI reviewers. *Judgment and Decision Making*, *19*, e21. https://doi.org/10.1017/jdm.2024.24

[14] Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

[15] Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science*, *379*(6630), 313. https://doi.org/10.1126/science.adg7879

[16] van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature*, *614*(7947), 224–226. https://doi.org/10.1038/d41586-023-00288-7

[17] Wells, B. J., Nguyen, H. M., McWilliams, A., Pallini, M., Bovi, A., Kuzma, A., Kramer, J., Chou, S.-H., Hetherington, T., Corn, P., Taylor, Y. J., Cuison, A., Gagen, M., Isreal, M., & FAIR-AI Consortium. (2025). A practical framework for appropriate implementation and review of artificial intelligence (FAIR-AI) in healthcare. *NPJ Digital Medicine*, *8*(1), 514. https://doi.org/10.1038/s41746-025-01900-y