Original Article

Assessment of Chat-GPT, Gemini, and Perplexity in Principle of Research Publication: A Comparative Study

Ameer M. Salih[1,2], Jaafar Omer Ahmed[3,4], Dilan S. Hiwa[1], Abdulwahid M. Salih[1], Rawezh Q. Salih[1], Hemn A. Hassan[1,3], Yousif M. Mahmood[1], Shvan H. Mohammed[5]* ID , Bander A. Abdalla[1]

1. Scientific Affairs Department, Smart Health Tower, Madam Mitterrand Street, Sulaymaniyah, Iraq
2. Civil Engineering Department, College of Engineering, University of Sulaimani, Sulaymaniyah, Iraq
3. Kscien Organization for Scientific Research (Middle East office), Hamdi Street, Sulaymaniyah, Iraq
4. Psychology Department, Faculty of Art, Soran University, Soran, Iraq
5. Xzmat Polyclinic, Rizgari, Kalar, Sulaymaniyah, Iraq

* **Corresponding author**: shvanh80@gmail.com (Shvan H. Mohammed). Jutyaran, Building 50, Zip code 46021, Kalar, Zip code: 46001, Sulaymaniyah, Iraq

Abstract

Introduction

Many researchers utilize artificial intelligence (AI) to aid their research endeavors. This study seeks to assess and contrast the performance of three sophisticated AI systems, namely, ChatGPT, Gemini, and Perplexity when applied to an examination focused on knowledge regarding research publication.

Methods

Three AI systems (ChatGPT-3.5, Gemini, and perplexity) were evaluated using an examination of fifty multiple-choice questions covering various aspects of research, including research terminology, literature review, study design, research writing, and publication-related topics. The questions were written by a researcher with an h-index of 22, and it was later tested on two other researchers with h-indices of 9 and 10 in a double-blinded manner and revised extensively to ensure the quality of the questions before testing them on the three mentioned AI systems.

Results

In the examination, ChatGPT scored 38 (76%) correct answers, while Gemini and Perplexity each scored 36 (72%). Notably, all AI systems frequently chose correct options significantly: ChatGPT chose option (C) correctly 88.9% of the time, Gemini accurately selected option (D) 78.9% of the time, and Perplexity correctly picked option (C) 88.9% of the time. In contrast, other AI tools showed minor agreement, lacking statistical significance, while ChatGPT exhibited significant concordance (81-83%) with researchers' performance.

Conclusion

ChatGPT, Gemini, and Perplexity perform adequately overall in research-related questions, but depending on the AI in use, improvement is needed in certain research categories. The involvement of an expert in the research publication process remains a fundamental cornerstone to ensure the quality of the work.

1. Introduction

The work of John McCarthy is the foundation of modern artificial intelligence (AI) research. In 1956, at Dartmouth College, he introduced the phrase "artificial intelligence," marking the inception of formal AI research [1]. The emergence

of AI was an innovative technological frontier, promising transformative impacts across diverse sectors. Recent years have witnessed significant strides in the AI domain, particularly in the

refinement of chatbot technology. An increasingly prevalent notion suggests that AI, having surpassed human capabilities in several domains, holds promise for substantial advancements in the realm of research publications. AI stands poised to augment research writing, the accuracy of information retrieved, and referencing, thereby potentially revolutionizing the field [2].

Over the past few years, a multitude of AI tools have become readily accessible, providing a diverse array of services and functionalities. A notable instance of such an AI system is ChatGPT, an advanced language model crafted by OpenAI. It underwent training using a vast array of textual materials gathered from websites, literature, and diverse sources, engaging in language modeling tasks to enhance its capabilities. This attribute sets it apart as one of the most expansive and resilient language models ever devised, integrating an astonishing 175 billion parameters [3,4]. An additional AI system that has attracted attention is Gemini, previously identified as Google Bard, which is an AI-driven information retrieval apparatus with a sophisticated chatbot that utilizes a "native multimodal" approach to effectively process and adjust to various types of data like video, audio, and text [5,6]. Perplexity AI stands as an AI-powered research and conversational search engine, adept at responding to queries through the utilization of natural language predictive text. It synthesizes answers from web sources, accompanied by citations through embedded links within the text response [7].

Many researchers are known to utilize chatbots as aids in their research endeavors.

This study seeks to assess and contrast the performance of sophisticated AI systems—namely, ChatGPT, Gemini, and Perplexity—when applied to an examination focused on knowledge regarding research publication. It also aims to shed light on the current state of AI integration within the research publication process and identify opportunities for further development

2. Methods

In this comparative investigation, we evaluated the performance of three distinct AI systems: ChatGPT-3.5, Gemini, and Perplexity. The assessment comprised 50 multiple-choice questions, each offering four options (A-D). The questions spanned various domains including eleven research terminology queries, six literature review inquiries, twelve study design probes, twelve research writing assessments, and nine publication-related investigations.

Initially, a researcher with an h-index of 22, identified as the second author in the manuscript, composed a set of sixty multiple-choice questions. Subsequently, two other researchers with h-indices of 14 and 16, mentioned as authors seven and ten respectively, underwent the examination in a double-blinded

Table 1. The association between correct answers and AI tools

Correct	ChatGPT				
	A	B	C	D	Total
A	7 (63.6%)	0 (0.0%)	2 (18.2%)	2 (18.2%)	11 (100%)
B	0 (0.0%)	8 (72.7%)	2 (18.2%)	1 (9.1%)	11 (100%)
C	0 (0.0%)	0 (0.0%)	8 (88.9%)	1 (11.1%)	9 (100%)
D	0 (0.0%)	3 (15.8%)	1 (5.3%)	15 (78.9%)	19 (100%)
Total	7 (14%)	11 (22%)	13 (26%)	19 (38%)	50 (100%)
P-value	<0.001				

Correct	Gemini				
	A	B	C	D	Total
A	7 (63.6%)	2 (18.2%)	1 (9.1%)	1 (9.1%)	11 (100%)
B	1 (9.1%)	7 (63.6%)	2 (18.2%)	1 (9.1%)	11 (100%)
C	0 (0.0%)	0 (0.0%)	7 (77.8%)	2 (22.2%)	9 (100%)
D	2 (10.5%)	2 (10.5%)	0 (0.0%)	15 (78.9%)	19 (100%)
Total	10 (20%)	11 (22%)	10 (20%)	19 (38%)	50 (100%)
P-value	<0.001				

Correct	Perplexity				
	A	B	C	D	Total
A	8 (72.7%)	0 (0.0%)	1 (9.1%)	2 (18.2%)	11 (100%)
B	2 (18.2%)	5 (45.5%)	2 (18.2%)	2 (18.2%)	11 (100%)
C	0 (0.0%)	0 (0.0%)	8 (88.9%)	1 (11.1%)	9 (100%)
D	0 (0.0%)	3 (15.8%)	1 (5.3%)	15 (78.9%)	19 (100%)
Total	10 (20%)	8 (16%)	12 (24%)	20 (40%)	50 (100%)
P-value	<0.001				

Correct	Researcher 1				
	A	B	C	D	Total
A	10 (90.9%)	0 (0.0%)	0 (0.0%)	1 (9.1%)	11 (100%)
B	0 (0.0%)	9 (81.8%)	0 (0.0%)	2 (18.2%)	11 (100%)
C	0 (0.0%)	1 (11.1%)	8 (88.9%)	0 (0.0%)	9 (100%)
D	0 (0.0%)	2 (10.5%)	1 (5.3%)	16 (84.2%)	19 (100%)
Total	10 (20%)	12 (24%)	9 (18%)	19 (38%)	50 (100%)
P-value	<0.001				

Correct	Researcher 2				
	A	B	C	D	Total
A	10 (90.9%)	0 (0.0%)	0 (0.0%)	1 (9.1%)	11 (100%)
B	1 (9.1%)	9 (81.8%)	1 (9.1%)	0 (0.0%)	11 (100%)
C	0 (0.0%)	0 (0.0%)	9 (100%)	0 (0.0%)	9 (100%)
D	2 (10.5%)	1 (5.3%)	3 (15.8%)	13 (68.4%)	19 (100%)
Total	13 (26%)	10 (20%)	13 (26%)	14 (28%)	50 (100%)
P-value	<0.001				

fashion. Following this phase, all three researchers collaborated to review and analyze both questions and answers. Ten questions were excluded due to their lack of clarity, leaving a total of fifty questions selected for the final examination version. These selected questions were unanimously agreed upon by the

researchers as informative indicators of knowledge within the realm of research and its associated intricacies.

The questions were then uniformly inputted into each of the AI systems in March 2024, following a standardized protocol. This protocol involved initiating interactions with the AI systems by introducing a prompt starting with "Hello." Subsequently, each AI system received the same directive: "Please select the correct answer for the following multiple-choice questions." The questions were directly transcribed from a prepared Word document, and the AI-generated responses were recorded in an Excel spreadsheet. Statistical analysis was performed using Statistical Package for the Social Sciences (SPSS) version 27.0, with a significance level set at $p < 0.05$. Chi-square (Fisher's Exact Test) was employed for data analysis.

During the literature review phase of the present study, papers were selectively included from reputable journals and omitted those published in predatory journals, adhering to the criteria delineated in Kscien's list [8].

3. Results

In the examination, ChatGPT demonstrated slightly higher accuracy with a total of 38 correct answers (76%), compared to 36 correct answers (72%) by both Gemini and Perplexity. Notably, Researcher 2 excelled in terminology and literature review questions, with 15 correct answers (88.23%), surpassing ChatGPT and Gemini, with 13 correct answers (76.47%). In research writing, Perplexity, along with Researcher 1 and Researcher 2, led with 10 correct responses (83.3%). Additionally, Researcher 1 exhibited the highest accuracy in research publication, with 9 correct responses (100%), outperforming ChatGPT and Researcher 2, who achieved 7 correct responses (77.78%) (Supplementary 1).

In the examination comparing AI tools and two researchers' accuracy in identifying correct answers, researchers demonstrated superior accuracy compared to AI tools. For

example, in questions where the correct answer was C, Researcher 2 achieved a perfect 100% accuracy, outperforming ChatGPT, Perplexity, and Gemini, which scored 88.9%, and 77.8% respectively. Notably, all AI systems significantly chosen the correct options. For instance, ChatGPT correctly identified option C 88.9% of the time, Gemini correctly chose option D 78.9% of the time, and Perplexity accurately selected option C 88.9% of the time (Table 1).

In comparing AI tools and researchers' performance, significant agreement was noted with ChatGPT. For instance, out of 43 questions where researcher 1 agreed on the correct answer, ChatGPT agreed in 35 cases (81.4%) and disagreed in only 8 answers (18.6%). However, the comparison with the other two AI tools showed no significance but a slight alignment with the researchers' agreement on the correct answers (Table 2).

4. Discussion

The imitation of human intelligence functions by machines, most commonly computer systems, is referred to as AI. It involves acquiring knowledge (gaining information and understanding rules for its utilization), logical deduction (applying rules to arrive at rough or precise outcomes), and self-adjustment. In addition, AI endeavors to develop systems capable of executing tasks traditionally associated with human intelligence, including decision-making, speech recognition, language translation, and visual perception, among various others [9]. Although AI language models have been in development for years, the general population's understanding of AI's potential and use has increased dramatically recently. The academic community has already embraced language-based AI, and numerous researchers utilize chatbots as aids in their research. These bots assist in structuring ideas, offering feedback on their work, and aiding in referencing and summarizing the existing research literature [2,10,11].

Kacena et al. demonstrated that the utilization of AI, particularly ChatGPT, reduced the time invested in crafting review articles. However, it yielded the highest similarity indices, indicating a greater probability of plagiarism. In addition, they reported that ChatGPT possesses the ability to swiftly scour the internet and evaluate potential sources, potentially accelerating the literature review process. In the current study, the performance of ChatGPT regarding the principle of literature review questions

Table 2. Comparative Analysis of AI Tools' and Researchers' Performance in Research Studies

AI tools	Researcher 1		P-value*	Researcher 2		P-value*
	Agree	Disagree		Agree	Disagree	
ChatGPT 3.5						
Agree	35 (81.4%)	3 (42.9%)		34 (82.9%)	4 (44.4%)	
Disagree	8 (18.6%)	4 (57.1%)	0.048	7 (17.1%)	5 (55.6%)	0.027
Total	43 (100%)	7 (100%)		41 (100%)	9 (100%)	
Gemini						
Agree	32 (74.4%)	4 (57.1%)		30 (73.2%)	6 (72%)	
Disagree	11 (25.6%)	3 (42.9%)	0.300	11 (26.8%)	3 (33.3%)	0.697
Total	43 (100%)	7(100%)		41 (100%)	9 (100%)	
Perplexity						
Agree	33 (76.7%)	3 (42.9%)		32 (78%)	4 (44.4%)	
Disagree	10 (23.3%)	4 (57.1%)	0.085	9 (22%)	5 (55.6%)	0.094
Total	43 (100%)	7 (100%)		41 (100%)	9 (100%)	

*Fisher's Exact Test

showed a high performance, and Gemini scored just as high, further supporting the finding of the previous study [12].

Salvagno et al. reported that AI may soon be leveraged for the automated production of figures, tables, and supplementary visual components within manuscripts. This utilization could facilitate data summarization and contribute to manuscript lucidity [13]. However, the current study demonstrated that the AI systems had different scores, and their performance was influenced by the different categories they were tested on, which means that identifying the strengths and weaknesses of the currently available AIs is paramount in choosing which AI system will aid in research publications rather than hindering and jeopardizing the integrity of the research paper, For instance, Kacena et al. showcased that 70% of the references were incorrect when an AI only method was applied to writing research papers, raising controversy if these AI tools should even be used as aid in that regard [12]. The present study showed that Gemini performed poorly by only getting half of the questions wrong in the research writing principles questions. In addition, Perplexity was shown to perform poorly on principles of publication-related questions, and ChatGPT exhibited subpar performance in research terminology inquiries, further supporting the notion that leveraging AI use is dependent on recognizing their limitations in the field of research.

Concerns about biases in AI systems, stemming from their training data, are widely recognized as a significant challenge. Research indicates that AI models can perpetuate biases and exhibit skewed behavior, replicating existing discriminatory patterns. Addressing these biases is crucial and requires the implementation of effective strategies prioritizing fairness and justice during development. This is particularly important in research, where ensuring impartiality is paramount. Responsible use of advanced language models like ChatGPT, Gemini, and Perplexity is essential, given the ethical dilemmas they pose, including the potential for misinformation and emotionally persuasive content. Proactive steps are needed to mitigate these risks and promote responsible usage. Additionally, the use of AI in content generation raises concerns about unintentional plagiarism, as systems may reproduce text without proper citation. While AI tools may increase publication output, there may not be a corresponding increase in expertise or experience among researchers [3,12].

Several studies have investigated the comparison of AI and human capabilities across various domains. Long et al. noted a remarkable level of accuracy in AI, ranging from 90% to 100% when evaluating its performance against specialized doctors' diagnostic and treatment decisions for congenital cataracts [14]. Additionally, Rajpurkar et al. discovered consistency in results between AI and radiologists, particularly in diagnosing chest radiographs [15]. However, there is limited available data on the comparison of AI and human performance in research principles. In this study, the comparison between AI tools and human performance regarding predetermined correct answers on research principles revealed a significant agreement (80-85%) between ChatGPT and researchers.

One of the limitations of our study is that we evaluated only three AI systems in comparison to the vast and increasing number of AI tools becoming available in these times. In addition, a larger number of questions will lead to a more comprehensive understanding of the strengths and weaknesses of these AI systems in the field of research and their utilities in that regard.

5. Conclusion

ChatGPT, Gemini, and Perplexity perform adequately overall in research-related questions, but depending on the AI in use, improvement is needed in certain research categories. The involvement of an expert in the research publication process remains a fundamental cornerstone to ensure the quality of the work.

Declarations

References

1. Xu Y, Liu X, Cao X, Huang C, Liu E, Qian S, et al. Artificial intelligence: A powerful paradigm for scientific research. The Innovation. 2021;2(4). *doi:10.1016/j.xinn.2021.100179*
2. Altmäe S, Sola-Leyva A, Salumets A. Artificial intelligence in scientific writing: a friend or a foe? Reproductive BioMedicine Online. 2023;47(1):3-9. *doi:10.1016/j.rbmo.2023.04.009*

3. Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems. 2023. *doi:10.1016/j.iotcps.2023.04.003*

4. Abbas YN, Hassan HA, Hamad DQ, Hasan SJ, Omer DA, Kakamad SH, et al. Role of ChatGPT and Google Bard in the Diagnosis of Psychiatric Disorders: A Comparative Study. Barw Medical Journal. 2023;1(4):14-19. *doi:10.58742/4vd6h741*

5. Masalkhi M, Ong J, Waisberg E, Lee AG. Google DeepMind's gemini AI versus ChatGPT: a comparative analysis in ophthalmology. Eye. 2024:1-6. *doi:10.1038/s41433-024-02958-w*

6. Muhialdeen AS, Mohammed SA, Ahmed NH, Ahmed SF, Hassan WN, Asaad HR, et al. Artificial Intelligence in Medicine: A Comparative Study of ChatGPT and Google Bard in Clinical Diagnostics. Barw Medical Journal. 2023. *doi:10.58742/pry94q89*

7. Perplexity AI. Perplexity AI [Internet]. www.perplexity.ai. 2022. Available from: *https://www.perplexity.ai/*

8. Muhialdeen AS, Ahmed JO, Baba HO, Abdullah IY, Hassan HA, Najar KA, et al. Kscien's list; a new strategy to discourage predatory journals and publishers (second version). Barw Medical Journal. 2023;1(1):24-26. *doi:10.58742/bmj.v1i1.14*

9. Helm JM, Swiergosz AM, Haeberle HS, Karnuta JM, Schaffer JL, Krebs VE, et al. Machine learning and artificial intelligence: definitions, applications, and future directions. Current reviews in musculoskeletal medicine. 2020; 13:69-76. *doi:10.1007/s12178-020-09600-8*

10. Ahamed ZM, Dhahir HM, Mohammed MM, Ali RH, Hassan SH, Muhialdeen AS, et al. Comparative Analysis of ChatGPT and Human Decision-Making in Thyroid and Neck Swellings: A Case-Based Study. Barw Medical Journal. 2023;1(4):2-6. *doi:10.58742/bmj.v1i2.43*

11. Salih AM, Mohammed NA, Mahmood YM, Hassan SJ, Namiq HS, Ghafour AK, et al. ChatGPT Insight and Opinion Regarding the Controversies in Neurogenic Thoracic Outlet Syndrome: A Case-Based Study. Barw Medical Journal. 2023;1(3):2-5. *doi:10.58742/bmj.v1i2.48*

12. Kacena MA, Plotkin LI, Fehrenbacher JC. The use of artificial intelligence in writing scientific review articles. Current Osteoporosis Reports. 2024:1-7. *doi:10.1007/s11914-023-00852-0*

13. Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? Critical care. 2023;27(1):75. *doi:10.1186/s13054-023-04380-2*

14. Long E, Lin H, Liu Z, Wu X, Wang L, Jiang J, et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. Nature biomedical engineering. 2017;1(2):0024. *doi:10.1038/s41551-016-0024*

15. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS medicine. 2018;15(11):e1002686. *doi:10.1371/journal.pmed.1002686*