Regeneration, Reconstruction & Restoration

Original Article

Open Access

Evaluation of Artificial Intelligence-Generated Information on Cell Culture and Laboratory Protocols in the Field of Tissue Engineering: A Comparison between GPT-3.5, Cluade-Instant, and Microsoft Copilot

Shaghayegh Najary[a], Ali Azadi[b], Sayna Shamszadeh[c], Forough Shams[d], Hanieh Nokhbatolfoghahaei[e*]

[a]Student Research Committee, School of Dentistry, Shahid Beheshti University of Medical Sciences, Tehran, Iran; [b]Dentofacial Deformities Research Center, Research Institute for Dental Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran; [c]Endodontic Research Center, Research Institute for Dental Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran; [d]Department of Tissue Engineering and Applied Cell Sciences, School of Advanced Technologies in Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran; [e]Dental Research Center, Research Institute for Dental Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

Corresponding author: Hanieh Nokhbatolfoghahaei, Dental Research Center, Research Institute for Dental Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. ***E-mail:*** hanieh.nokhbe@sbmu.ac.ir

Submitted: 25 May, 2024**; Accepted:** 8 June, 2024**; Published Online:** 21 October, 2024**; DOI:** *10.22037/rrr.v9.45423*

Abstract

Background and objectives: The current study aimed to evaluate and compare the validity and precision of the information provided by three online artificial intelligence (AI) platforms, including GPT-3.5 (Open AI), Claude-Instant (Anthropic), and Copilot (Microsoft), in the context of laboratory protocols in the field of tissue engineering.

Materials and methods: Three specialists in the field of tissue engineering research and molecular and cellular laboratory methods created a survey with 20 open-ended questions. The questions included a variety of subjects regarding cell culture processes and cell analysis techniques. The chatbots' responses to the open-ended inquiries were assessed according to the modified global quality scale. A P-value less than 0.05 was considered statistically significant.

Results: GPT-3.5, with a median score of 4 out of 5, performed significantly better than Claude-Instant and Copilot. GPT-3.5 showed superior performance compared to Claude-Instant ($P < 0.001$) and Copilot ($P < 0.007$). Both Claude-Instant and Copilot chatbots received a median score of 3 out of 5 in total.

Conclusion: In conclusion, AI language models provided satisfactory responses for laboratory procedures in the field of tissue engineering. Specifically, GPT-3.5 outperformed both Claude-Instant and Microsoft Copilot chatbots. Nevertheless, the information offered by all three chatbots was insufficient in terms of the crucial details required for the design and execution of an in vitro analysis.

Keywords: Artificial intelligence, Cell culture techniques, Dentistry, Machine learning, Tissue engineering.

Please cite this paper as: Najary Sh, Nokhbatolfoghahaei H, Shamszadeh S, Shams F, Azadi A. Evaluation of Artificial Intelligence-Generated Information on Cell Culture and Laboratory Protocols in the Field of Tissue Engineering: A Comparison between GPT-3.5, Cluade-Instant, and Microsoft Copilot. Journal of "Regeneration, Reconstruction & Restoration"(Triple R). 2024; Volume 9: e11. Doi: 10.22037/rrr.v9.45423

1. Introduction

The design of in vitro studies is regarded as the fundamental stage of any research project in the field of tissue engineering (1). To prevent the recurrence of previous errors, it is necessary to establish and follow rigorous criteria of reproducibility and reliability in the realm of in vitro biomedical research. A comprehensive set of instructions outlining the proper procedures for handling cultures and cell analysis has been written; however, it was interpreted differently by various

divisions within the field of biomedical research (2). Thus, it is crucial for researchers and lab staff to be equipped with an accurate source of information on various cell culture models and test protocols.

Natural Language Processing (NLP) systems using Artificial Intelligence (AI) are rapidly advancing in their complex operations and consistently influencing every facet of society. NLP enables users to interact with machines using a language-based interface, eliminating the need for coding (3, 4). These models are capable of engaging in text summarization, classification, and data gathering. Language models such as ChatGPT, GoogleBard, BingAI, and Claude use artificial intelligence to generate text-based responses that closely mimic the human language (5).

The Chatbot Generative Pre-Trained Transformer (ChatGPT), created by OpenAI, is an AI software developed to mimic human-like discussions with users. This chatbot operates using algorithms that are specifically designed to comprehend natural language inputs and provide suitable responses, which can be either pre-written or newly created by artificial intelligence. ChatGPT undergoes continuous enhancements using reinforcement techniques, natural language processing, and machine learning to enhance its capacity to comprehend and provide comprehensive responses to users' questions (5, 6).

Integrating NLP models in preclinical research has the potential to greatly improve the availability of information for both researchers and lab technicians. Researchers can utilize AI as an assistant to reduce their workload and obtain updated data (7). Nevertheless, the final output may lack accuracy. It has also been suggested that ChatGPT should not be relied upon as a substitute for human judgment, and its results should always be evaluated by specialists before being utilized in any important decision-making process or application (8, 9).

The objective of this study is to evaluate and compare the precision and reliability of information provided by various AI chatbots, including GPT-3.5, Claude-Instant, and Microsoft Copilot, for the design and protocol of cellular and molecular laboratory studies in the field of tissue engineering. Furthermore, through the assessment of the efficacy of these AI models, our aim is to offer a thorough comprehension of their potential applications for aiding researchers and laboratory staff in all aspects of cellular and molecular biology.

2. Materials and methods

2.1. Questionnaire design

A group of three specialists in the field of tissue engineering research and molecular and cellular laboratory procedures prepared a survey with 20 open-ended questions. The questions addressed a range of topics related to the procedures involved in cell culture, as well as various techniques used for cell analysis. These techniques included the Methyl Thiazolyl Tetrazolium (MTT) assay, Alamar Blue assay, 4′,6-diamidino-2-phenylindole (DAPI) staining, scanning electron microscopy, Real-time Polymerase Chain Reaction, Western blot, Alizarin Red staining, Alkaline phosphatase assay, Oil Red O staining, and Toluidine Blue staining. The details of the questionnaire are outlined in the supplementary section. The chatbots were asked with open-ended queries to generate unstructured text responses. All queries of open-ended questions began with ‘’What was your comprehensive answer to the following question if you were an expert in regenerative medicine research and molecular and cellular laboratory techniques:’’ to restrict avoiding chatbots’ answers and make them provide as much information as possible. The study’s questionnaire is provided below:

1. What was your comprehensive answer to the following question if you were an expert in regenerative medicine research and molecular and cellular laboratory techniques: “What are the steps of the MTT assay for experiments not comprising scaffolds?”
2. What was your comprehensive answer to the following question if you were an expert in regenerative medicine research and molecular and cellular laboratory techniques: “What are the steps of the (methyl thiazolyl tetrazolium) MTT assay for experiments comprising scaffolds?”
3. What was your comprehensive answer to the following question if you were an expert in regenerative medicine research and molecular and cellular laboratory techniques: “What are the steps of the Alamar Blue assay for experiments not comprising scaffolds?”
4. What was your comprehensive answer to the following question if you were an expert in regenerative medicine research and molecular and cellular laboratory techniques: “What are the steps of the Alamar Blue assay for experiments comprising scaffolds?”
5. What was your comprehensive answer to the following question if you were an expert in regenerative medicine research and molecular and cellular laboratory techniques: “What are the steps of the 4′,6-diamidino-2-phenylindole (DAPI) staining?”
6. What was your comprehensive answer to the following question if you were an expert in regenerative medicine research and molecular and cellular laboratory techniques: “What are the steps of preparing hydrogel scaffolds for scanning electron microscope (SEM)?”
7. What was your comprehensive answer to the following question if you were an expert in regenerative medicine research and molecular and cellular laboratory techniques: “What are the steps of preparing composite polymer-ceramic scaffolds for scanning electron microscope (SEM)?”
8. What was your comprehensive answer to the following question if you were an expert in regenerative medicine research and molecular and cellular laboratory techniques: “What are the steps of the degradation tests of a scaffold?”
9. What was your comprehensive answer to the following question if you were an expert in regenerative medicine research and molecular and cellular laboratory techniques: “What are the

steps of the cell RNA extraction in the Real-time Polymerase Chain Reaction (RT-qPCR) for experiments not comprising scaffolds?
10. What was your comprehensive answer to the following question if you were an expert in regenerative medicine research and molecular and cellular laboratory techniques: "What are the steps of the cell RNA extraction in the Real-time Polymerase Chain Reaction (RT-qPCR) for experiments comprising scaffolds?"
11. What was your comprehensive answer to the following question if you were an expert in regenerative medicine research and molecular and cellular laboratory techniques: "What are the steps of the cDNA synthesis in the Real-time Polymerase Chain Reaction (RT-qPCR) for experiments not comprising scaffolds?"
12. What was your comprehensive answer to the following question if you were an expert in regenerative medicine research and molecular and cellular laboratory techniques: "What are the steps of the Real-time Polymerase Chain Reaction (RT-qPCR) on the cultured cells?"
13. What was your comprehensive answer to the following question if you were an expert in regenerative medicine research and molecular and cellular laboratory techniques: "What are the steps of the Western blot on the cultured cells?"
14. What was your comprehensive answer to the following question if you were an expert in regenerative medicine research and molecular and cellular laboratory techniques: "What are the steps of the Alizarin Red staining on the cultured cells?"
15. What was your comprehensive answer to the following question if you were an expert in regenerative medicine research and molecular and cellular laboratory techniques: "What are the steps of the Alkaline phosphatase assay on the cultured cells?"
16. What was your comprehensive answer to the following question if you were an expert in regenerative medicine research and molecular and cellular laboratory techniques: "What are the steps of the Oil Red O staining on the cultured cells?"
17. What was your comprehensive answer to the following question if you were an expert in regenerative medicine research and molecular and cellular laboratory techniques: "What are the steps of the Toluidine Blue staining on the cultured cells?"
18. What was your comprehensive answer to the following question if you were an expert in regenerative medicine research and molecular and cellular laboratory techniques: "How many human buccal fat pad stem cells should be seeded per each well in 24- and 48-well plates for experiments?"
19. What was your comprehensive answer to the following question if you were an expert in regenerative medicine research and molecular and cellular laboratory techniques: "How many human buccal dental pulp stem cells should be seeded per each well in 24- and 48-well plates for experiments?"
20. What was your comprehensive answer to the following question if you were an expert in regenerative medicine research and molecular and cellular laboratory techniques: "What are the components and their concentrations in an osteogenic culture medium?"

2.2. Data collection

Three AI chatbots were chosen based on being widely used and freely accessible: GPT-3.5, Claude-Instant, and Microsoft Copilot. No limitations were applied to the length or content of the responses. Responses were collected for both qualitative and quantitative analysis. The selection of these chatbots was based on their advanced artificial intelligence capabilities and widespread utilization. Queries were submitted via the primary application programming interface (API) of GPT3.5 (accessed through https://chat.openai.com), Claude-Instant (accessed through https://claude.ai/), and Microsoft Copilot (accessed through https://copilot.microsoft.com). The questions for each chatbot were run all at once on March 12, 2024. Every question was presented in a new chat, and earlier discussions were erased to prevent any contamination or recall.

2.3. Assessment protocol

A group of three specialists in the field of tissue engineering or regenerative medicine assessed the responses provided by the chatbots. The assessors were unaware of the chatbots and evaluated the responses using a modified version of the Global Quality Score (GQS) [10], which is a valid and rigorous evaluation tool. The GQS is a 5-point Likert scale evaluation measure used to assess the quality of various materials. In this study, the tool has been adapted to specifically evaluate the quality of responses by referees (Table 1). A response from a chatbot was considered the final score if it was endorsed by at least 2 out of 3 assessors. However, if all assessors gave different scores, a consensus meeting was organized to allow the assessors to discuss and reach a final decision.

2.4. Statistical analysis

The data analysis was conducted using SPSS version 27 software (SPSS, Chicago, IL). The Kruskal-Wallis test was employed to compare the scores assigned by referees to each chatbot for open-ended questions. A pairwise comparison was conducted to assess the performance of different Chatbots using the Dunn Post hoc test with Bonferroni Correction. The tests were conducted using a 95% confidence interval. Significance was attributed to P-values below 0.05.

3. Results

3.2. Performance of each chatbot on the open-ended questions

Median and interquartile range (IQR) were used to describe the given scores to the chatbots' answers. The scores assigned to the

Table 1. The modified version of the global quality score (GQS).

Score	Description
1	Poor quality, poor flow of the information, most information missing, not at all useful for researchers and technicians
2	Generally poor quality and flow, some information listed but many important topics missing, very limited useful for researchers and technicians
3	Moderate quality, suboptimal flow, some important information adequately discussed but others poorly discussed, somewhat useful for researchers and technicians
4	Good quality and generally good flow. Most of the relevant information is listed but some topics are not listed. useful for researchers and technicians
5	Excellent quality and flow, very useful for researchers and technicians

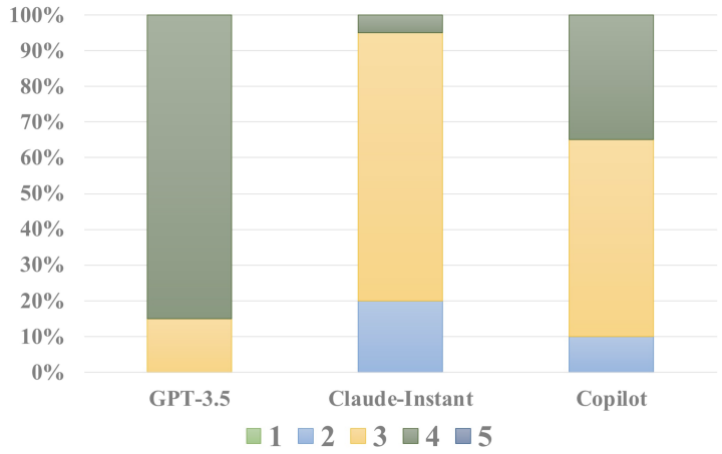

Figure 1. The frequency of each score given to each chatbot by each referee.

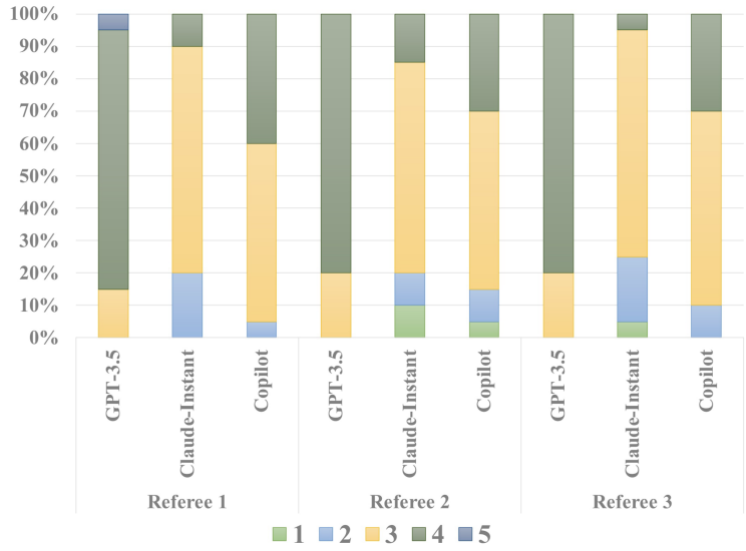

Figure 2. The frequency of each score in the performance of chatbots after taking a consensus of referees' opinions on each question.

Table 2. Description of agreement's level of different referee's assessment of different chatbots performance in responding to the questions.

	GPT-3.5			Claude-Instant			Copilot		
	Referee 1 & Referee 2	Referee 1 & Referee 3	Referee 2 & Referee 3	Referee 1 & Referee 2	Referee 1 & Referee 3	Referee 2 & Referee 3	Referee 1 & Referee 2	Referee 1 & Referee 3	Referee 2 & Referee 3
K-Score	0.72	0.72	0.69	0.73	0.79	0.71	0.75	0.65	0.69

Poor agreement: K-score < 0

Slight agreement: 0.0 <= K-score <= 0.20

Fair agreement: 0.21 <= K-score <=0.40

Moderate agreement 0.41<= K-score <=0.60

Substantial agreement: 0.61 <= K-score <= 0.80

Almost perfect agreement: 0.81 <= K-score <= 1.0

Table 3. Description of median and interquartile range (IQR) of scores given by each refree to different chatbot responses to the questions.

	GPT-3.5			Claude-Instant			Copilot		
	Referee 1	Referee 2	Referee 3	Referee 1	Referee 2	Referee 3	Referee 1	Referee 2	Referee 3
Median	4.00	4.00	4.00	3.00	3.00	3.00	3.00	3.00	3.00
IQR	0.00	0.00	0.00	0.00	0.00	0.75	1.00	1.00	1.00
	Overall			Overall			Overall		
Median	4.00			3.00			3.00		
IQR	1.00			0.00			1.00		

Table 4. Comparison between the performance of different chatbots in responding to the questions using the Kruksal-Wallis test.

	Mean Rank	Kruskal-Wallis P-value
GPT-3.5	43.95	
Claude-Instant	18.85	< 0.001 *
Copilot	28.70	

Table 5. Pair-wise comparison between performance of different chatbots using the Dunn Post Hoc with Bonferroni correction.

	Dunn Post hoc P-value
GPT-3.5 vs. Claude-Instant	< 0.001 *
GPT-3.5 vs. Copilot	0.007 *
Claude-Instant vs. Copilot	0.144

each chatbot by each judge are summarized in Figure 1. Figure 2 illustrates the frequency distribution of scores obtained from consensus of reviewers' evaluations on the performance of each chatbot. GPT-3.5 obtained an overall median (IQR) score of 4.00 (1.00), with three grades being '3' and 17 grades being '4' based on the final results. Score ‘’2’’ was given to the Claude-Instant chatbot four times, followed by ‘’3’’ and ‘’4’’ 15 and one times, respectively. Microsoft Copilot received 11 grades of ‘’3’’, seven grades of ‘’4’’, and two grades of ‘’2’’ score. An overall median (IQR) score of 3.00 (0.00) and 3.00 (1.00) was given to both Claude-Instant and Copilot chatbots, respectively.

3.3. Inter-rater agreement level

Cohen’s weighted Kappa statistic indicated that the K-score values exceeded 0.6, indicating a level of agreement classified as "substantial" in evaluating the open-ended responses of the chatbots (Table 2).

3.4. Evaluation of the performance of chatbots

Table 3 presents the ratings assigned to each chatbot by individual referees, as well as the overall performance of each chatbot based on a collective evaluation of referees' scores on each question. The results of the Kruskal-Wallis test revealed statistically significant differences ($P < 0.001$) among three chatbots in terms of their performance and overall score. GPT-3.5, with a median score of 4, demonstrated superior performance compared to both Claude-Instant ($P < 0.001$) and Microsoft Copilot ($P = 0.007$) chatbots. This was verified by the results of the Dunn Post hoc test (Tables 4 and 5).

4. Discussion

It has been shown in our study that information provided by GPT-3.5, Claude-Instant, and Microsoft Copilot chatbots in the field of regenerative medicine research and molecular and cellular laboratory techniques were mostly graded as ‘’3’’ and ‘’4’’ in terms of precision and accuracy. No ‘’2’’ score was given to any of the answers provided by GPT-3.5; however, some information given by Claude-Instant and Microsoft Copilot chatbots were graded as ‘’2’’. GPT-3.5 showed significantly better performance in terms of accuracy and

correctness of the information provided when compared to two other chatbots.

ChatGPT, utilizing the GPT-3.5 framework, is highly regarded for its remarkable ability to produce logical and human-like replies (11). GPT-3 remains a robust tool for general purposes, while GPT-3.5 excels in nuanced and specialized contexts. In a comprehensive study, GPT-4 has shown to be the most successful chatbot in all categories when compared to Claude and Bard, attaining an impressive success rate of 84.1% (12). A further study analyzed six AI chatbots specifically in the context of scholarly writing in the fields of humanities and archaeology. ChatGPT-4 has the highest level of quantitative correctness and scientific contribution, followed by ChatGPT-3.5, Bing, and Bard, based on the accuracy of the facts. However, Claude 2 and Aria achieved significantly lower scores (13). Claude benefits from a 52 billion scale of parameters and has shown desirable behavior in conversations; however, there have been doubts about the accuracy of the data provided by this chatbot (14).

Microsoft has incorporated artificial intelligence (AI) into its Edge browser and Bing search engine, using the advanced technology that OpenAI used to create ChatGPT. Bing Chat functions in a manner identical to ChatGPT, allowing users to type in any question and obtain an answer in natural human language via the large language model (LLM) (15, 16). The Edge Copilot feature significantly improves the Bing Chat experience by providing additional suggestions and adjustments. The chat tab prioritizes conversational language and provides a variety of suggestions for possible queries. This involves hyperlinks for additional information, suggested subsequent queries, and functions in a manner similar to a typical search engine (17). Like other AI-powered systems, Bing Chat may sometimes offer false or irrelevant information, which might cause uncertainty or inaccurate information (5).

A similar study assessed the accuracy of responses provided by GoogleBard, GPT-3.5, GPT-4, Claude-Instant, and Bing chatbots when answering clinical questions related to oral and maxillofacial surgery and revealed that Bard, GPT-3.5, GPT-4, Claude-Instant, and Bing answered 34%, 36%, 38%, 38%, and 26% of the questions correctly, respectively. GPT-4 achieved the highest scores in open-ended questions, with the majority of responses being rated as grades "4" or "5." Conversely, Bing mostly obtained scores classified as grades "1" or "2", indicating its lowest performance compared to other chatbots (18). In our study, GPT-3.5 had markedly superior accuracy and precision in delivering information compared to Claude-Instant and Copilot chatbots.

5. Limitations

One limitation of this study is the small number of questions, which limits the range of topics covered and thus reduces the generalizability of the results. Additionally, as chatbot databases continuously evolve, the authors recommend that future studies include more questions and assessors to thoroughly evaluate chatbot performance in providing necessary information within the cell laboratory techniques field. The study's single-time response capture also hindered the assessment of chatbot consistency. Furthermore, the limited knowledge of these AI chatbots in a highly specialized area, compared to technicians, can lead to inconsistent or inaccurate responses. Therefore, further training or specific modifications to their datasets are necessary for their effective use in this field.

6. Conclusion

In conclusion, artificial intelligence language models were able to deliver replies that were acceptable for laboratory procedures in the field of tissue engineering. In particular, GPT-3.5 performed better than both the Claude-Instant and the Microsoft Copilot chatbots. On the other hand, the information that was provided by all three chatbots was inadequate in terms of the essential details that are necessary for the design and execution of an in vitro analysis. AI-generated responses may exhibit ambiguity, occasional errors, or misinterpretations. Additionally, comprehending AI-generated answers requires a higher level of reading comprehension. Hence, it is important to use the data offered by chatbots under the guidance of professionals and regard them as assistants, not as substitutes for human expertise.

Ethics

The study protocol was approved by the Shahid Beheshti University of Medical Sciences' ethics committee (IR.SBMU.RETECH.REC.1403.180).

Acknowledgment

The authors appreciate the "Students Research Committee" and "Research and Technology Chancellor" in Shahid Beheshti University of Medical Sciences for their financial support of this study.

Authors' contribution

Shaghayegh Najary: Conceptualization, Methodology, Investigation, Validation, Writing – Original draft, Writing – Review & editing

Hanieh Nokhbatolfoghahaei: Conceptualization, Methodology, Investigation, Validation, Supervision, Writing – Original draft, Writing – Review & editing

Sayna Shamszadeh: Investigation, Validation, Writing – Original draft, Writing – Review & editing

Forough Shams: Investigation, Validation, Writing – Original draft, Writing – Review & editing

Ali Azadi: Conceptualization, Methodology, Project administration, Investigation, Formal analysis, Software, Visualization, Writing – Original draft, Writing – Review & editing

Conflict of Interest
The authors declare no conflict of interest.

Funding/support
This study is related to the project No. 1403/34459 from Students Research Committee, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

References
1. Godbey, W. and A. Atala, In vitro systems for tissue engineering. Annals of the new York Academy of Sciences, 2002. **961**(1): p. 10-26.
2. Hirsch, C. and S. Schildknecht, In vitro research reproducibility: Keeping up high standards. Frontiers in pharmacology, 2019. **10**: p. 492837.
3. Fanni, S.C., et al., Natural language processing, in Introduction to Artificial Intelligence. 2023, Springer. p. 87-99.
4. Stiennon, N., et al., Learning to summarize with human feedback. Advances in Neural Information Processing Systems, 2020. **33**: p. 3008-3021.
5. Ray, P.P., ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems, 2023.
6. Haleem, A., M. Javaid, and R.P. Singh, An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. BenchCouncil transactions on benchmarks, standards and evaluations, 2022. **2**(4): p. 100089.
7. Chubb, J., P. Cowling, and D. Reed, Speeding up to keep up: exploring the use of AI in the research process. AI & society, 2022. **37**(4): p. 1439-1457.
8. Salvagno, M., F.S. Taccone, and A.G. Gerli, Can artificial intelligence help for scientific writing? Critical Care, 2023. **27**(1): p. 75.
9. Kitamura, F.C., ChatGPT is shaping the future of medical writing but still requires human judgment. 2023, Radiological Society of North America. p. e230171.
10. Bernard, A., et al., A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. Official journal of the American College of Gastroenterology| ACG, 2007. **102**(9): p. 2070-2077.
11. Thoppilan, R., et al., Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239, 2022.
12. Borji, A. and M. Mohammadian, Battle of the Wordsmiths: Comparing ChatGPT, GPT-4, Claude, and Bard. GPT-4, Claude, and Bard (June 12, 2023), 2023.
13. Lozić, E. and B. Štular, ChatGPT v Bard v Bing v Claude 2 v Aria v human-expert. How good are AI chatbots at scientific writing?(ver. 23Q3). arXiv preprint arXiv:2309.08636, 2023.
14. Ray, P.P., ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems, 2023. **3**: p. 121-154.
15. Rudolph, J., S. Tan, and S. Tan, War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. Journal of Applied Learning and Teaching, 2023. **6**(1).
16. Alberts, I.L., et al., Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be? European journal of nuclear medicine and molecular imaging, 2023. **50**(6): p. 1549-1552.
17. Bai, H. A practical three-phase approach to fully automated programming using system decomposition and coding copilots. in Proceedings of the 2022 5th International Conference on Machine Learning and Machine Intelligence. 2022.
18. Azadi, A., et al., Evaluation of AI-generated responses by different artificial intelligence chatbots to the clinical decision-making case-based questions in oral and maxillofacial surgery. Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology, 2024.