SRYAHWA PUBLICATIONS

Open Access Journal of Gynecology and Obstetrics
ISSN: 2638-5244 | Volume 7, Issue 1, 2025
https://doi.org/10.22259/2638-5244.0701003

RESEARCH ARTICLE

Comparing the Accuracy of Large Language Models (LLM) in Trending Obstetrical Topics

Amber Khemlani, BA[1], Joshua Singavarapu, BA[1], Ranjitha Vasa, MD[1,2], Huber Rodriguez-Tejada, MD[1,2], Harsh Reshamwala, BS[3], Ozgul Muneyyirci-Delale, MD[1], Mudar Dalloul, MD[1]

[1]Department of Obstetrics and Gynecology, State University of New York Downstate Health Sciences University, Brooklyn, NY, United States of America.
[2]Department of Obstetrics and Gynecology, King's County Hospital Center, Brooklyn, NY, United States of America.
[3]Cooper Union University, New York, NY, United States of America.

Received: 9 April 2025 Accepted: 26 April 2025 Published: 9 May 2025

Corresponding Author: Ranjitha Vasa, Department of Obstetrics and Gynecology, State University of New York Downstate Health Sciences University, Brooklyn, NY, United States of America.

Abstract

Generative artificial intelligence (AI) is rapidly expanding in medicine, where both patients and healthcare providers are increasingly relying on large language model (LLM) chatbots for information. In this study, we evaluated four AI chatbots—ChatGPT 4.0, Gemini 3.7, Copilot AI, and Perplexity AI —by analyzing their responses to queries related to three obstetrical pathologies: preeclampsia, placental abruption, and gestational diabetes mellitus. Queries for the top five obstetrical pathologies were obtained from U.S. Google Trends data spanning December 10, 2019, to December 10, 2024. AI-generated responses were assessed using validated evaluation tools: the Patient Education Material Assessment Tool (PEMAT) for understandability and actionability, DISCERN for information quality, and the Flesch-Kincaid formula for readability. AI-generated content was reviewed for alignment with guidelines from the American College of Obstetricians and Gynecologists (ACOG). PEMAT scores for understandability and actionability were analyzed using chi-square tests, while DISCERN and Flesch-Kincaid scores were evaluated using the Kruskal-Wallis test. ChatGPT showed promising results through PEMAT actionability, PEMAT understandability, and DISCERN scores. The Flesch-Kincaid readability scores of all the chatbots were similar, as they all were written at a high school grade level. This indicates a need for AI chatbots to formulate responses that cater to varying grade levels of knowledge. Furthermore, there is a future where AI becomes the primary source of information, and it is important to continually challenge and evaluate LLMs for potential misinformation and accurate data.

Keywords: Preeclampsia, Placental Abruption, Gestational Diabetes Mellitus, Artificial Intelligence, Obstetrics.

1. Introduction

With the release of ChatGPT on November 2022, a large language model (LLM) that has rapidly expanded its user base, generative artificial intelligence (AI) has been on the rise [1]. Since then, numerous companies have released their own chatbots akin to ChatGPT such as Gemini by Google and CoPilot by Microsoft. LLMs, also colloquially known as AI chatbots, have the potential to elevate the medical field through their presentation and exhibition of knowledge, synthesis of complex medical information, and direct user interaction. These chatbots have been noted to be expressive, interactive, and have demonstrated an increased empathy when responses were compared with real-time physicians [1]. Additionally, AI solves various medical challenges such as enhancing pathological results, reducing diagnostic costs, providing proper monitoring of patients, and improving hospital safety [2]. ChatGPT was found to pass all

Citation: Amber Khemlani, Joshua Singavarapu, Ranjitha Vasa. *et.al* Comparing the Accuracy of Large Language Models (Llm) in Trending Obstetrical Topics. Open Access Journal of Gynecology and Obstetrics 2025;7(1): 18-22.

three parts of the United States Medical Licensing Examination (USMLE) and has demonstrated high diagnostic accuracy when presented with clinical vignettes [3]. Because of their ability to synthesize and distill complex medical information, AI chatbots are being gradually incorporated clinically for both patient and physician use. However, with the fast-paced rate at which various AI chatbots are being produced, there are concerns regarding not only their accuracy, but also how content is being presented for user consumption [4].

There are a multitude of uses of AI chatbots in obstetrics, specifically when it comes to patient education regarding various pregnancy associated pathologies. AI is already being used in obstetrics using machine learning algorithms to predict preterm births and asymptomatic short cervical lengths [5]. In addition, prenatal ultrasonography is essential to detect fetal abnormalities, and AI is currently being used to assess fetal head biometry and cranial capacity [6]. While LLMs offer an accessible approach to patient communication and education, what persistently remains unclear is the accuracy of information being generated and relayed.

This study aims to evaluate four of the leading AI chatbots in terms of their ability to address questions related to national trending pathologies in obstetrics – pre-eclampsia, gestational diabetes, and placenta previa. Pre-eclampsia is a complication of pregnancy secondary to elevated blood maternal blood pressure (diagnosed with two blood pressures greater than or equal to 140/90 on two occasions, four hours apart) and proteinuria (urine protein to creatinine urine ratio of 0.3 or greater, urine dipstick 2+, or 24 hour urine protein greater than or equal to 300), diagnosed at and/or greater than 20 weeks gestational age. As of 2020, pre-eclampsia leads to about 46,000 maternal deaths and 500,000 fetal deaths annually worldwide [7]. Gestational diabetes is diagnosed in patients that are greater than or equal to 20 weeks gestational age secondary to likely elevated levels of human placental lactogen leading to increased insulin resistance [8]. Globally, about 14% of pregnancies are impacted by gestational diabetes [8]. Further, placental abruption is an obstetrical complication when the placenta separates from the uterus prior to delivery of the fetus causing excessive bleeding, fetal distress, and even premature birth affecting 0.6% to 1.2% of pregnancies as of 2023 [9].

Our project aims to assess the accuracy and comprehensiveness of each LLM and determine the degree of variability between them using the following text evaluations: patient education materials assessment tool-5 (PEMAT-5), DICERN4, and Flesch-Kincaid Reading Score.

2. Materials and Methods

This cross-sectional study uses publicly available data across four AI chatbots. Google trends was used to identify the top three national obstetrical pathologies from December 10, 2019 to December 10, 2024, which were preeclampsia, gestational diabetes, and placenta previa. Thereafter, Google trends was used to evaluate the top four search queries pertaining to each of the three obstetrical patholgies previously mentioned. The responses were inputted into four AI chatbots: ChatGPT 4.0, Gemini 3.7, Copilot AI, and Perplexity AI. The most updated and publicly available versions of the LLMs as of December 17, 2024 were used to extrapolate answers for each query. For every new search term, the previous conversation was deleted, and a new conservation was initiated to maintain anonymity and to not allow subsequent responses to be influenced by previous ones. The search terms that were input into each AI chatbot contained the same phrasing as used in the Google Trends search query.

Four scoring systems were used to evaluate the chatbot responses: PEMAT-5 understandability, PEMAT-5, DISCERN4, and Flesch-Kincaid. PEMAT evaluates to what extent readers are able to understand demonstrated content and actionable decisions when presented content. Scores for PEMAT-5 range from 0%- 100%, with higher scores indicating a higher level of understandability and actionability. DISCERN4 was used to evaluate the quality of the chatbot responses, with a scoring range of 1 (low) to 5 (high). The Flesch-Kincaid level is a subjective system which evaluates how readable a text is from 1 (easy to read) to 16 (most challenging to read). Three members of the team (A.K., J.S., H.T.) were blinded to the AI chatbot type and each others' assessment. A.K., J.S., and H.T. scored the chatbot responses independently on the PEMAT and DISCERN questionnaires. The readability score, however, was inputted into the Flesch-Kincaid scoring system code which generated the scores.

PEMAT scores were evaluated using chi-square tests based on success/failure counts. Post-hoc pairwise comparisons with Bonferroni adjustment were used to compare the understandability and actionability of each LLM. DISCERN scores were evaluated using Kruskal-Wallis test with post-hoc pairwise comparison with Bonferroni correction conducted using Mann-Whitney U test to identify differences between each LLM. Lastly, Flesch-Kincaid Grade Level scores were analyzed using the Kruskal-Wallis test.

3. Results

Mean scores and standard deviation were reported for each chatbot: ChatGPT (M = 45.65, SD = 1.13), Gemini (M = 42.16, SD = 1.03), Copilot AI (M = 41.26, SD = 1.39), and Perplexity AI (M = 39.06, SD = 3.36) [Table 1].

Table 1. *Mean and standard deviation reported for each LLM*

AI Chatbot	Mean	Standard Deviation
ChatGPT 4.0	45.65	1.13
Gemini 3.7	42.16	1.03
Copilot AI	41.26	1.39
Perplexity AI	39.06	3.36

The understandability and actionability scores of AI chatbots scores evaluated with PEMAT-5 using chi-square tests based on success/failure counts, with 1 considered a success and 0 a failure. For understandability, the overall chi-square test revealed significant differences across chatbots, $\chi^2(3) = 73.60$, $p < .001$. Post-hoc pairwise comparisons with Bonferroni adjustment ($\alpha = 0.0083$) revealed that ChatGPT (89.04% success rate) significantly outperformed Gemini (64.76%, $\chi^2 = 72.26$, $p < .001$), Copilot AI (76.89%, $\chi^2 = 22.30$, $p < .001$), and Perplexity AI (77.13%, $\chi^2 = 21.69$, $p < .001$). Similarly, Gemini demonstrated significantly lower success rates compared to Copilot AI ($\chi^2 = 14.97$, $p < .001$) and Perplexity AI ($\chi^2 = 15.81$, $p < .001$). No significant differences were observed between Copilot AI and Perplexity AI ($\chi^2 = 0.00$, $p = .996$) [Figure 1].

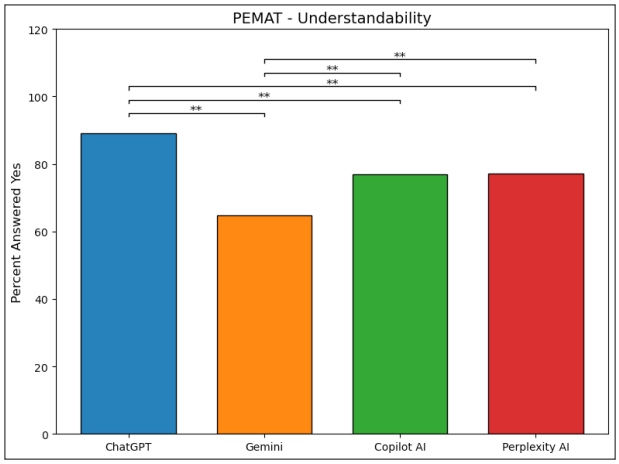

Figure 1. *PEMAT understandability scores across ChatGPT, Gemini, Copilot AI, and Perplexity AI. ** indicates highly significant (p<.001), * indicates significant p<0.0083.*

For actionability, the overall chi-square test revealed significant differences between ChatGPT and Gemini, $\chi^2(3) = 14.14$, $p = .003$. Post-hoc tests indicated that ChatGPT (48.89% success rate) significantly outperformed Gemini (29.78%, $\chi^2 = 12.91$, $p < .001$), but no significant differences were found between ChatGPT and Copilot AI ($\chi^2 = 4.81$, $p = .028$) or Perplexity AI ($\chi^2 = 3.73$, $p = .053$). Similarly, no significant differences were observed among Gemini, Copilot AI, and Perplexity AI [Figure 2].

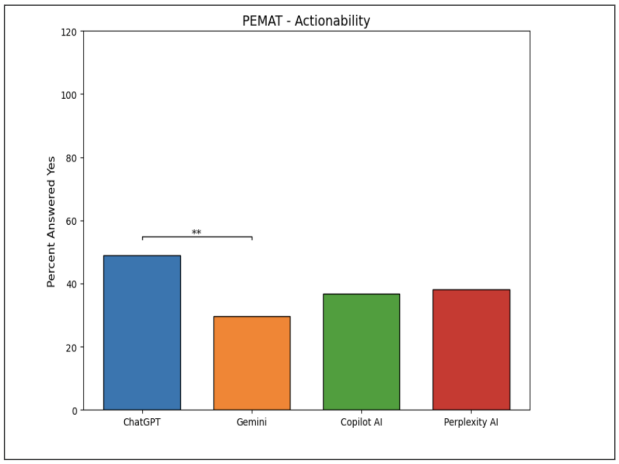

Figure 2. *PEMAT actionability scores across ChatGPT, Gemini, Copilot AI, and Perplexity AI. ** indicates highly significant (p<.001), * indicates significant p<0.0083.*

The Kruskal-Wallis test was used to assess differences across the chatbot's DISCERN scores, revealing significant variability with three degrees of freedom, H(3) = 8.41, p = .038. Post-hoc pairwise comparison with Bonferroni correction (α=0.0083) were conducted using the Mann-Whitney U test. Significant differences were found between ChatGPT and Perplexity AI (p= .008), while no significant differences were observed between ChatGPT and Gemini (p = .092), ChatGPT and Copilot AI (p = .023), Gemini and Perplexity AI (p = .632), Gemini and Perplexity AI (p = .331), or Copilot AI and Perplexity AI (p = .544) [Figure 3].

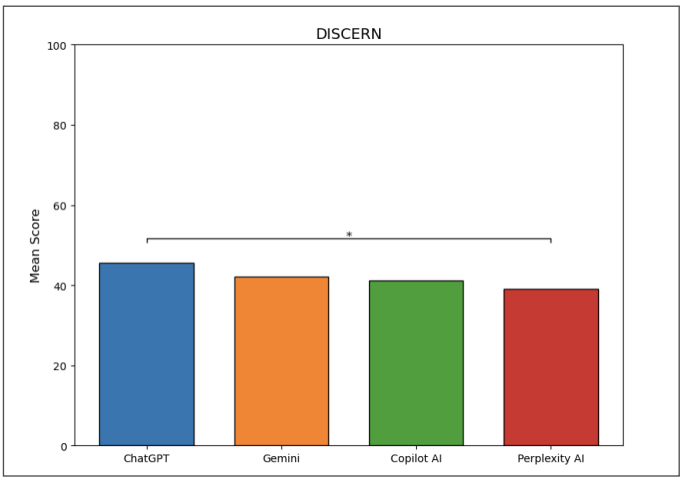

Figure 3. *DISCERN scores across ChatGPT, Gemini, Copilot AI, and Perplexity AI. ** indicates highly significant (p<.001), * indicates significant p<0.0083.*

The Flesch-Kincaid Grade Level scores were analyzed using the Kruskal-Wallis test revealing no significant variability with three degrees of freedom, H(3) = 4.83, p = .185. Mean Flesch-Kincaid Grade Level scores and their standard deviations were reported for each chatbot: ChatGPT (M = 10.55, SD = 1.13), Gemini (M = 10.01, SD = 1.03), Copilot AI (M = 11.07, SD = 1.39), and Perplexity AI (M = 11.23, SD = 3.36) [Figure 4].

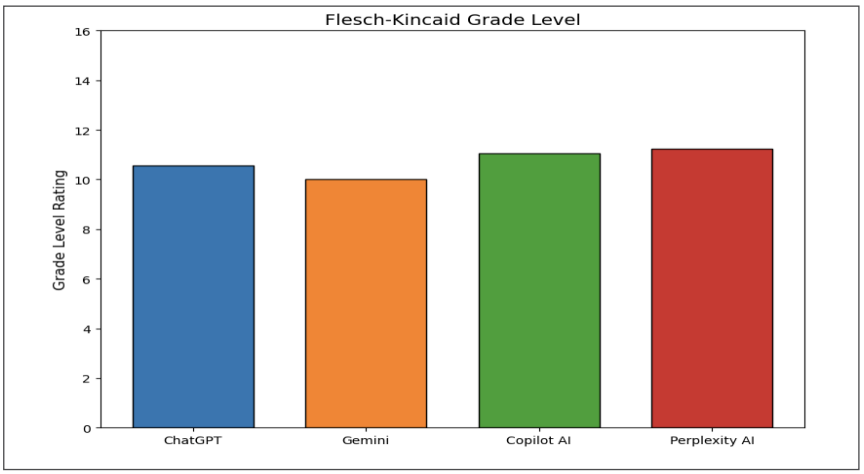

Figure 4. *Flesch-Kincaid Grade level scores across ChatGPT, Gemini, Copilot AI, and Perplexity AI. ** indicates highly significant (p<.001), * indicates significant p<0.0083.*

4. Discussion

For PEMAT understandability, ChatGPT performed significantly higher when compared to Gemini, Copilot AI, and Perplexity AI. For PEMAT actionability, ChatGPT only performed significantly higher than Gemini. When evaluating DISCERN scores, ChatGPT was only significantly higher than Perplexity. There was no significant difference between the LLM's Flesch-Kincaid reading scores. When evaluating the PEMAT understandability scores, it is important to note that ChatGPT provides easy-to-understand information regarding very complex topics in obstetrics. Additionally, ChatGPT also provides readers with clear, actionable steps after pertinent information is generated for consumption. Though ChatGPT was only significantly better than Gemini, the data shows that the average actionability scores of ChatGPT were higher than the other chatbots on average. ChatGPT also performed well on DISCERN when compared to the other chatbots, showing that the quality of ChatGPT responses were superior to other LLMs. Thus, it is important to recognize that

ChatGPT performed above the other chatbots in these three evaluative tests. The data also shows that the readability scores across the LLMs are somewhat equivalent – a refreshing result that highlights how LLMs perform well in the tasks they were built for – to condense difficult information into more simple language for easier consumption.

There were some limitations to our project. Regarding study design, our query results were accumulated over the last five years from 12/10/2019 to 12/10/2024. This Google search period coincided with COVID-19, the global pandemic that effectively altered access to healthcare for non-COVID-related hospital visits. Additionally, the construct of the study focuses on three subjective evaluations, which introduce differences on whether or not the evaluators felt a response truly answered certain questions. However, the study itself is a blind study and there were steps taken to mitigate the potential bias present, through a thorough and shared understanding of the scoring systems prior to any grading and the use of a subjective score via the Flesch-Kincaid scale. Additionally, potential bias in the chatbots themselves were ultimately eliminated with a deletion of previous conversations with each subsequent one. This allowed each chatbot response to be individualized for each query, without any confounding influences.

Though this study provides us with important insights in the use of ChatGPT to improve patient access to medical knowledge, there are several areas that need to be studied. Future research should focus on utilizing more graders to evaluate the chatbot responses to improve the validity of the study. Additionally, assessing the accuracy in the responses generated for each pathology gives us a glimpse into the most frequently searched topics, improving AI generated responses with patient needs. Further studies can also examine the readability metrics required for various subspecialties helping AI refine their content to suit different levels of literacy.

5. Conclusion

This study indicates the ability of AI to revolutionize medical care by ensuring patients receive relevant and timely information through a single reliable source, with ChatGPT showing promising results through PEMAT actionability, PEMAT understandability, and DISCERN scores. Though all the AI chatbots provided relativity similar Flesch-Kincaid readability scores, the grade levels are still high, being written at a high school level. However, the goal of AI information should be to provide easy-to-read information, since pre-eclampsia, placental abruption, and gestational diabetes mellitus patients should not be assumed to read at a high school level. Furthermore, there is a future where AI becomes the primary source of information, and it is important to continually challenge and evaluate LLMs for potential misinformation and accurate data. More studies can be done to look at how AI performs well in other areas of study, such as gynecology, even expanding to other topics in obstetrics, and how AI can respond to patient queries in real time in a clinical setting.

Sources of support: None to disclose

Author contribution: All authors listed above were involved with conception of the project, data collection, and production of the manuscript.

6. References

1. Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., ... & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*, *183*(6), 589-596.
2. Current status and applications of Artificial Intelligence (AI) in medicine. *Artificial Intelligence in Medicine*. 2019;95:102002. doi:10.1016/j.artmed.2019.07.007.
3. Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. (2023). Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS digital health*, *2*(2), e0000198.
4. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, *620*(7972), 172-180.
5. Kim HY, Cho GJ, Kwon HS. Applications of artificial intelligence in obstetrics. *Ultrasonography*. 2023;42(1):2-9. doi:10.14366/usg.22063
6. Chen Z, Liu Z, Du M, Wang Z. Artificial Intelligence in Obstetric Ultrasound: An Update and Future Applications. *Front Med (Lausanne)*. 2021;8:733468. doi:10.3389/fmed.2021.733468.
7. Magee LA, Nicolaides KH, von Dadelszen P. Preeclampsia. *N Engl J Med*. 2022;386(19):1817-1832. doi:10.1056/NEJMra2109523.
8. Eades CE, Burrows KA, Andreeva R, Stansfield DR, Evans JMM. Prevalence of gestational diabetes in the United States and Canada: a systematic review and meta-analysis. *BMC Pregnancy Childbirth*. 2024;24(1):204. doi:10.1186/s12884-024-06378-2.
9. Brandt JS, Ananth CV. Placental abruption at near-term and term gestations: pathophysiology, epidemiology, diagnosis, and management. *Am J Obstet Gynecol*. 2023;228(5 Suppl):S1313-S1329. doi:10.1016/j.ajog.2022.09.018.