

Performance of Artificial Intelligence Chatbots on Standardized Medical Examination Questions in Obstetrics & Gynecology

Angelo Cadiente^{1*}, Natalia DaFonte¹, Jonathan D. Baum^{1,2}

¹Hackensack Meridian School of Medicine, Nutley, NJ, USA

²Department of Obstetrics and Gynecology, Jersey Shore University Medical Center, Neptune, NJ, USA

Email: *angelo.cadiente@hmn.org

How to cite this paper: Cadiente, A., DaFonte, N. and Baum, J.D. (2025) Performance of Artificial Intelligence Chatbots on Standardized Medical Examination Questions in Obstetrics & Gynecology. *Open Journal of Obstetrics and Gynecology*, 15, 1-9. <https://doi.org/10.4236/ojog.2025.151001>

Received: October 3, 2024

Accepted: January 3, 2025

Published: January 6, 2025

Copyright © 2025 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Objective: This study assesses the quality of artificial intelligence chatbots in responding to standardized obstetrics and gynecology questions. **Methods:** Using ChatGPT-3.5, ChatGPT-4.0, Bard, and Claude to respond to 20 standardized multiple choice questions on October 7, 2023, responses and correctness were recorded. A logistic regression model assessed the relationship between question character count and accuracy. For each incorrect question, an independent error analysis was undertaken. **Results:** ChatGPT-4.0 scored a 100% across both obstetrics and gynecology questions. ChatGPT-3.5 scored a 95% overall, earning an 85.7% in obstetrics and a 100% in gynecology. Claude scored a 90% overall, earning a 100% in obstetrics and an 84.6% in gynecology. Bard scored a 77.8% overall, earning an 83.3% in obstetrics and a 75% in gynecology and would not respond to two questions. There was no statistical significance between character count and accuracy. **Conclusions:** ChatGPT-3.5 and ChatGPT-4.0 excelled in both obstetrics and gynecology while Claude performed well in obstetrics but possessed minor weaknesses in gynecology. Bard comparatively performed the worst and had the most limitations, leading to our support of the other artificial intelligence chatbots as preferred study tools. Our findings support the use of chatbots as a supplement, not a substitute for clinician-based learning or historically successful educational tools.

Keywords

Large Language Models, ChatGPT, Bard, Claude, Medical Education

1. Introduction

Artificial intelligence (AI) in medicine has made strides in medical imaging, diagnostic accuracy, safety, complication prediction, and drug development [1]. Moreover, students are exploring the utility of AI chatbots, such as ChatGPT, as an adjunct to traditional studies. Thus, educators need to proactively be familiarized to encourage a cohesive, safe learning environment [2] [3].

As a potential study tool, ChatGPT has been assessed across medical examinations from United States Medical Licensure Examinations, American College of Gastroenterology, Ophthalmic Knowledge Assessment Program, and Urology Self-Assessment Study Program, with mixed results [4]-[7]. Chatbot literature in obstetrics and gynecology (OB/GYN) remains behind. Levin *et al.* conducted a bibliographic analysis finding no published OB/GYN studies involving ChatGPT from its inception through February 2023 [8]. Another study showed ChatGPT scored equivalent to human candidates on a mock virtual OB/GYN objective structured clinical examination, illustrating specialty-specific fluid reasoning with factually accurate answers [9]. Grünebaum *et al.* and Wan *et al.* found ChatGPT had mixed response quality, inconsistently providing appropriate references for common pregnancy questions. ChatGPT's apparent lack of question insight and reference verification was concerning [10] [11].

Evidence on accuracy of OB/GYN information from AI chatbots is limited and largely excludes chatbots Bard and Claude. Our objective is to assess, compare, and characterize the accuracy of four AI chatbots on standardized National Board of Medical Examiners (NBME) OB/GYN questions. Our study assesses AI chatbot strengths and limitations as study adjuncts for medical students.

2. Methods

ChatGPT-3.5 (September 25 version), ChatGPT-4.0 (September 25 version), Bard (September 27 version), and Claude were the chatbots utilized for this study on October 7, 2023. Questions were outsourced from the *NBME Obstetrics & Gynecology Sample Items* [12]. Question stem and answer choice character count ("Question Character Count") was recorded. Each question was classified as either "obstetrics" or "gynecology" depending on the subject area of the question. The following standardized prompt was used for each question across the four chatbots:

"Please select the correct answer and explain why the other answer choices are incorrect:

Question Stem + Answer Choices

Each chatbot's answer choice, accuracy (number of incorrect responses versus total question amount), and full response were recorded. A qualitative review by two independent analyzers of each incorrect response was conducted to characterize the error made and cross reference with the existing literature. Resources from the American College of Obstetricians and Gynecologists (ACOG) were prioritized as the gold standard. PubMed indexed articles were used as a second-line

resource to support the correct response. A logistic regression was utilized to assess the relationship between accuracy and question character count. We further investigate accuracy related to question character count, a functional measure of question complexity, *i.e.* the degree of nuance behind a question. This study was exempt from review by the Institutional Review Board as no patient-level data was used.

3. Results

We assessed all 20 questions available from the *NBME Obstetrics & Gynecology Sample Items*. ChatGPT-4.0 scored the highest compared to the other three models at 100%; due to its perfect score, an odds ratio was not calculated in relation to character count. Likewise, ChatGPT-3.5 scored 95% with an odds ratio of 1.015 [95%CI: 0.986 - 1.045], Claude scored 90% with an odds ratio of 0.996 [95%CI: 0.989 - 1.003], and Bard scored 77.8% with an odds ratio of 0.999 [95%CI: 0.994 - 1.004]. Bard would not answer two questions; thus, they were both removed from any calculations involving this model. None of the odds ratios were statistically significant; thus, there was no association between character count and accuracy across the different models (**Table 1**).

Table 1. Accuracy and character count relationship across each chatbot.

Chatbot	Accuracy	Character Count Odds Ratio [95% Confidence Interval]	P-value
ChatGPT-3.5	95% (19/20)	1.015 [0.986 - 1.045]	0.3072
ChatGPT-4.0	100% (20/20)	N/A	N/A
Claude	90% (18/20)	0.996 [0.989 - 1.003]	0.2911
Bard	77.8% (14/18)	0.999 [0.994 - 1.004]	0.7434

In terms of obstetrics questions, there were 7 questions classified under this category. ChatGPT-4.0 and Claude both scored 100% (7/7). ChatGPT-3.5 scored 85.7% (6/7). Bard scored 83.3% (5/6), as it would not answer one obstetrics question. In terms of gynecology questions, there were 13 questions classified under this category. ChatGPT-4.0 and ChatGPT-3.5 both scored 100% (13/13). Claude scored 84.6% (11/13). Bard scored 75% (9/12), as it would not answer one gynecology question. In summary, the only incorrect answer given by ChatGPT-3.5 was in response to an obstetrics question. Both incorrect answers given by Claude were in response to gynecology questions. Of the questions Bard answered, a majority of the incorrect choices were in gynecology with one incorrect in obstetrics.

Regarding the subsequent error analysis, there were tendencies among the models to respond incorrectly to the same questions. Questions that sought a test-taker's ability to determine next steps tended to yield incorrect answers. While each model had their respective limitations, they tended to avoid stepwise protocols, prioritizing direct, possibly invasive, approaches in diagnosis. Additionally,

Claude and Bard incorrectly responded to a question related to healthcare privacy and medical decision-making in minors, recognizing the concept but misunderstanding the application thereof (Tables 2-4).

Table 2. ChatGPT-3.5 question summary and error analysis.

Scenario [12]	Error Analysis
<p>Question Presented: A 30-year-old woman, gravida 2, para 1, at 26 weeks' gestation with uterine size greater than expected and ultrasonography showing fetal hydrops.</p> <p>Question Focus: Determining the most appropriate next step in diagnosis</p>	<p>Both independent analyses concluded that ChatGPT-3.5's incorrect response was due to its prioritization of diagnostic procedures over necessary preliminary testing and its lack of consideration for the patient's medical history and potential risks associated with suggested procedures [13] [14].</p>

Table 3. Claude question summary and error analysis.

Scenario [12]	Error Analysis
<p>Question Presented: A 27-year-old nulligravid woman unable to conceive for 12 months with regular menstrual cycles, a history of pelvic inflammatory disease treated with antibiotics, and unremarkable vaginal exam and cervical cultures.</p> <p>Question Focus: Determining the most appropriate next step in diagnosis</p>	<p>Both independent analyses concluded that Claude's incorrect response stemmed from a misunderstanding or an overreliance on evolving medical data. While it acknowledges the importance of assessing for anatomical abnormalities in infertility workups, Claude's response did not align with the most appropriate and established diagnostic step [15].</p>
<p>Question Presented: A 15-year-old girl, known to be sexually active, is brought to the physician by her mother for contraception advice. The girl uses condoms consistently and is not interested in other forms of contraception.</p> <p>Question Focus: Applying healthcare privacy laws and ethical guidelines related to autonomy</p>	<p>Both independent analyses concluded that Claude's response indicates a misunderstanding of patient confidentiality in the context of adolescent health care. There appears to be a gap in Claude's interpretation of the ethical and legal aspects of patient privacy and the communication of sensitive information to parents [16].</p>

Table 4. Bard question summary and error analysis.

Scenario [12]	Error Analysis
<p>Question Presented: A 27-year-old nulligravid woman unable to conceive for 12 months with regular menstrual cycles, a history of pelvic inflammatory disease treated with antibiotics, and unremarkable vaginal exam and cervical cultures.</p> <p>Question Focus: Determining the most appropriate next step in diagnosis</p>	<p>Both independent analyses concluded Bard's response is a misunderstanding of standard infertility evaluation protocols. Bard failed to appropriately prioritize assessing potential tubal pathology in the context of the patient's presenting history. This error highlights a gap in understanding of the stepwise approach to infertility evaluation, emphasizing the importance of structural investigation [15].</p>
<p>Question Presented: A 30-year-old woman, gravida 2, para 1, at 26 weeks' gestation with uterine size greater than expected and ultrasonography showing fetal hydrops.</p> <p>Question Focus: Determining the most appropriate next step in diagnosis</p>	<p>Both independent analyses concluded Bard's response highlights oversight in its diagnostic approach. It prioritized a specific diagnostic test without first considering more basic and less invasive tests, demonstrating a lack of consideration for the clinical context and patient history. The model's response reflects a gap in integrating routine screenings and patient history into its reasoning process [13] [14].</p>

Continued

<p>Question Presented: 42-year-old woman, gravida 3, para 3, presenting with amenorrhea of 2 months and an episode of spotting 3 weeks ago. Despite no history of abnormal Pap smears and regular use of condoms, the patient exhibited a slightly enlarged uterus without palpable adnexal masses on examination. Question Focus: Determining the most appropriate next step in diagnosis</p>	<p>Both independent analyses concluded Bard's response was due to its failure to prioritize a pregnancy test in the initial evaluation of a sexually active woman of reproductive age presenting with amenorrhea. This oversight suggests a lack of proper weighing of clinical possibilities in Bard's decision-making process, particularly the consideration of pregnancy, despite the use of contraception [17] [18].</p>
<p>Question Presented: A 15-year-old girl, known to be sexually active, is brought to the physician by her mother for contraception advice. The girl uses condoms consistently and is not interested in other forms of contraception. Question Focus: Applying healthcare privacy laws and ethical guidelines related to autonomy</p>	<p>Both independent analyses concluded that Bard's response was due to a lack of understanding of patient confidentiality rights in the context of minors and sexual health. This case emphasizes the importance of context-specific ethical considerations in healthcare, where blanket policies may not be appropriate. Bard's response indicates a need for improved application of healthcare confidentiality laws of minors [16].</p>

4. Discussion

Our study supports ChatGPT-3.5, ChatGPT-4.0 and Claude as reliable and accurate study tools in obstetrics and gynecology. ChatGPT-3.5 and ChatGPT-4.0 performed strongly in obstetrics and gynecology; Claude showed proficiency in obstetrics with minor deficits in gynecology. Bard performed less robustly in both subcategories and refused to answer two questions, hindering its utility for students seeking comprehensive explanations. There was no statistically significant relationship between character count and accuracy, indicating question complexity lacks association with chatbot accuracy.

As the most popular AI chatbot, ChatGPT has been assessed in comparatively more studies.

Cadiante *et al.* describe glaring performance gaps between ChatGPT-3.5 and ChatGPT-4.0 across medical examination question sets [4]. Our study shows no such gap, with one single incorrect response by ChatGPT-3.5. In Levin *et al.*'s meta-analysis assessing 19 medical examinations, ChatGPT-3.5 averaged 61.1% with two studies showing performances comparable to ours: Nakhleh *et al.* indicating ChatGPT-3.5 scored 100% on 24 diabetes-related questions and Subramani *et al.* indicating ChatGPT-3.5 scored 85% on 20 medical physiology questions [19]-[21]. Our error analysis determined ChatGPT-3.5 misprioritized diagnostic steps, bypassing standard preliminary testing in favor of a direct diagnostic procedure. This error calls to question whether ChatGPT-3.5 may be aggressive in clinical settings where a stepwise approach is preferred. Given this was the chatbot's only inaccuracy, a definitive conclusion of diagnostic overzealousness cannot be made. ChatGPT-3.5 and ChatGPT-4.0 answer multiple choice questions accurately and provide comprehensive, clinically-based explanations, illustrating strong potential as study tools for medical students in obstetrics and gynecology.

Claude performed at a high level, particularly in obstetrics questions. Literature on Claude's performance in medical examinations is scarce, with scores of 71.6%

on the Peruvian National Medical Licensing Examination and 72.7% on institution-made human physiology questions [22] [23]. In our study, it responded incorrectly to gynecology questions. One inaccuracy follows the same folly as ChatGPT-3.5, lending itself to the diagnostic zeal conclusion, while the other results from misapplying patient confidentiality. Claude's performance shows promise as a study tool for obstetrics but caution should be undertaken for gynecology questions.

Bard performed poorly in comparison and refusing to generate answers to two questions raises further concern. These findings are consistent with a urologic study, wherein Bard did not answer patient-level questions and underperformed compared to ChatGPT and Claude [24]. In three incorrect responses from our study, Bard made the same error as ChatGPT-3.5 and Claude: failure to identify the best next diagnostic step. Similar to Claude, Bard failed to manage patient confidentiality. Bard declined to answer a gynecology diagnosis question describing a skin lesion and an obstetric clinical management question involving a pregnancy complication. Its refusal to answer may be related to censorship, as language describing a genital lesion and vaginal bleeding in pregnancy may be perceived as overtly graphic. Given Bard's less robust performance overall and in each subcategory, hesitancy to its use is warranted. This urge is further supported by its censorship, presiding as a barrier for use in OB/GYN questions where graphic descriptions are often necessary. Bard may not be a reliable adjunctive study tool for OB/GYN.

Claude and Bard's deficits highlight the significance of clinician critical reasoning skills. When facing patient confidentiality and sexual health in minors, these chatbots did not successfully navigate the complex social and ethical scenarios. While they demonstrated baseline understanding of the concepts—recognizing the question assessed patient privacy and acknowledging the importance of patient autonomy and confidentiality—their responses misapplied this concept in the context of minors. This error may underscore that AI chatbots can function as adjuncts, but cannot replace clinicians. While ChatGPT appears to perform comfortably in nuanced ethical scenarios, indicated by its correct responses in this study and modest performances in soft-skill medical examinations, there may be a deficit in other AI chatbots that has not been unveiled thus far [9] [25] [26]. Given this discrepancy in performances, it is unclear whether AI chatbots will continually underperform on this question type, or if they are capable of correctly applying ethical concepts but lacked experience prior to data collection.

The strengths of this study are that we use a primarily qualitative analysis in ascertaining the deficits in using different models for educational benefit. The use of qualitative work is hypothesis generating to the extent that it lays the groundwork for future, larger sample studies. Moreover, the use of different models allows for comparison across modalities. This study is limited by its small sample size potentially influencing the overall performance of each AI chatbot. The 20 question set may have been part of each model's training due to their variable internet accessibility and free access to this sample test. Moreover, multiple choice

questions are not the only modality to assess medical knowledge; yet, this assessment is used as a standardized measure to assess the foundation of OB/GYN knowledge of third year medical students. Other methods to assess medical knowledge are outside the scope of this study. Studies with larger sets of standardized OB/GYN questions may better reflect overall AI chatbot accuracy in this field.

Our study highlights the strengths and limitations of AI chatbots as study tools in obstetrics and gynecology. ChatGPT-3.5 and ChatGPT-4.0 stood out for their accuracy and reliability, making them valuable resources for medical students. Claude performed well in obstetrics but showed some weaknesses in gynecology, suggesting that it may be a useful but not fully reliable option. In contrast, Bard struggled in both areas and even refused to answer some questions, raising concerns about its practicality for learning in the field. While these tools show promise, particularly ChatGPT, these models are not perfect and cannot replace the critical thinking and decision-making skills of clinicians. Their occasional missteps – especially on complex or sensitive questions – reminds users that AI works best as a supplement, not a substitute for clinician-based learning or historically successful educational tools. Moving forward, larger studies with more diverse and standardized question sets will be important to better understand how AI can fit into medical education and support students throughout their undergraduate medical career.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- [1] AI in Medicine. <https://www.nejm.org/ai-in-medicine>
- [2] Nietzel, M.T. (2023) More than Half of College Students Believe Using ChatGPT to Complete Assignments Is Cheating. *Forbes*. <https://www.forbes.com/sites/michaelt Nietzel/2023/03/20/more-than-half-of-college-students-believe-using-chatgpt-to-complete-assignments-is-cheating/?sh=6c07110518f9>
- [3] Sun, L., Yin, C., Xu, Q. and Zhao, W. (2023) Artificial Intelligence for Healthcare and Medical Education: A Systematic Review. *American Journal of Translational Research*, **15**, 4820-4828.
- [4] Cadiente, A., Chen, J., Nguyen, J., Sadeghi-Nejad, H. and Billah, M. (2023) RETRACTED: Artificial Intelligence on the Exam Table: ChatGPT's Advancement in Urology Self-Assessment. *Urology Practice*, **10**, 521-523. <https://doi.org/10.1097/upj.0000000000000446>
- [5] Suchman, K., Garg, S. and Trindade, A.J. (2023) Chat Generative Pretrained Transformer Fails the Multiple-Choice American College of Gastroenterology Self-Assessment Test. *American Journal of Gastroenterology*, **118**, 2280-2282. <https://doi.org/10.14309/ajg.0000000000002320>
- [6] Mihalache, A., Popovic, M.M. and Muni, R.H. (2023) Performance of an Artificial Intelligence Chatbot in Ophthalmic Knowledge Assessment. *JAMA Ophthalmology*,

- 141, 589-597. <https://doi.org/10.1001/jamaophthalmol.2023.1144>
- [7] Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., *et al.* (2023) Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models. *PLOS Digital Health*, **2**, e0000198. <https://doi.org/10.1371/journal.pdig.0000198>
- [8] Levin, G., Brezinov, Y. and Meyer, R. (2023) Exploring the Use of ChatGPT in OB-GYN: A Bibliometric Analysis of the First ChatGPT-Related Publications. *Archives of Gynecology and Obstetrics*, **308**, 1785-1789. <https://doi.org/10.1007/s00404-023-07081-x>
- [9] Li, S.W., Kemp, M.W., Logan, S.J.S., Dimri, P.S., Singh, N., Mattar, C.N.Z., *et al.* (2023) ChatGPT Outscored Human Candidates in a Virtual Objective Structured Clinical Examination in Obstetrics and Gynecology. *American Journal of Obstetrics and Gynecology*, **229**, 172.e1-172.e12. <https://doi.org/10.1016/j.ajog.2023.04.020>
- [10] Grünebaum, A., Chervenak, J., Pollet, S.L., Katz, A. and Chervenak, F.A. (2023) The Exciting Potential for ChatGPT in Obstetrics and Gynecology. *American Journal of Obstetrics and Gynecology*, **228**, 696-705. <https://doi.org/10.1016/j.ajog.2023.03.009>
- [11] Wan, C., Cadiente, A., Khromchenko, K., Friedrichs, N., Rana, R.A. and Baum, J.D. (2023) ChatGPT: An Evaluation of AI-Generated Responses to Commonly Asked Pregnancy Questions. *Open Journal of Obstetrics and Gynecology*, **13**, 1528-1546. <https://doi.org/10.4236/ojog.2023.139129>
- [12] Obstetrics & Gynecology Subject Exam—Content Outline|NBME. <https://www.nbme.org/subject-exams/clinical-science/obstetrics-and-gynecology>
- [13] Kitchen, F.L. and Jack, B.W. (2023) Prenatal Screening. StatPearls Publishing. <http://www.ncbi.nlm.nih.gov/books/NBK470559/>
- [14] Prevention of Rh D Alloimmunization. <https://www.acog.org/clinical/clinical-guidance/practice-bulletin/articles/2017/08/prevention-of-rh-d-alloimmunization>
- [15] Infertility Workup for the Women's Health Specialist. <https://www.acog.org/clinical/clinical-guidance/committee-opinion/articles/2019/06/infertility-workup-for-the-womens-health-specialist>
- [16] Confidentiality in Adolescent Health Care. <https://www.acog.org/clinical/clinical-guidance/committee-opinion/articles/2020/04/confidentiality-in-adolescent-health-care>
- [17] Klein, D.A. and Poth, M.A. (2013) Amenorrhea: An Approach to Diagnosis and Management. *American Family Physician*, **87**, 781-788.
- [18] Diagnosis of Abnormal Uterine Bleeding in Reproductive-Aged Women. <https://www.acog.org/clinical/clinical-guidance/practice-bulletin/articles/2012/07/diagnosis-of-abnormal-uterine-bleeding-in-reproductive-aged-women>
- [19] Levin, G., Horesh, N., Brezinov, Y. and Meyer, R. (2023) Performance of ChatGPT in Medical Examinations: A Systematic Review and a Meta-Analysis. *BJOG: An International Journal of Obstetrics & Gynaecology*, **131**, 378-380. <https://doi.org/10.1111/1471-0528.17641>
- [20] Nakhleh, A., Spitzer, S. and Shehadeh, N. (2023) ChatGPT's Response to the Diabetes Knowledge Questionnaire: Implications for Diabetes Education. *Diabetes Technology & Therapeutics*, **25**, 571-573. <https://doi.org/10.1089/dia.2023.0134>
- [21] Subramani, M., Jaleel, I. and Krishna Mohan, S. (2023) Evaluating the Performance of ChatGPT in Medical Physiology University Examination of Phase I MBBS. *Advances in Physiology Education*, **47**, 270-271.

<https://doi.org/10.1152/advan.00036.2023>

- [22] Torres-Zegarra, B.C., Rios-Garcia, W., Ñaña-Cordova, A.M., Arteaga-Cisneros, K.F., Chalco, X.C.B., Ordoñez, M.A.B., *et al.* (2023) Performance of ChatGPT, Bard, Claude, and Bing on the Peruvian National Licensing Medical Examination: A Cross-Sectional Study. *Journal of Educational Evaluation for Health Professions*, **20**, 30. <https://doi.org/10.3352/jeehp.2023.20.30>
- [23] Agarwal, M., Goswami, A. and Sharma, P. (2023) Evaluating ChatGPT-3.5 and Claude-2 in Answering and Explaining Conceptual Medical Physiology Multiple-Choice Questions. *Cureus*, **15**, e46222. <https://doi.org/10.7759/cureus.46222>
- [24] Song, H., Xia, Y., Luo, Z., Liu, H., Song, Y., Zeng, X., *et al.* (2023) Evaluating the Performance of Different Large Language Models on Health Consultation and Patient Education in Urolithiasis. *Journal of Medical Systems*, **47**, Article No. 125. <https://doi.org/10.1007/s10916-023-02021-3>
- [25] Chen, J., Cadiante, A., Kasselmann, L.J. and Pilkington, B. (2023) Assessing the Performance of ChatGPT in Bioethics: A Large Language Model's Moral Compass in Medicine. *Journal of Medical Ethics*, **50**, 97-101. <https://doi.org/10.1136/jme-2023-109366>
- [26] Brin, D., Sorin, V., Vaid, A., Soroush, A., Glicksberg, B.S., Charney, A.W., *et al.* (2023) Comparing ChatGPT and GPT-4 Performance in USMLE Soft Skill Assessments. *Scientific Reports*, **13**, Article No. 16492. <https://doi.org/10.1038/s41598-023-43436-9>