DOI: 10.4274/forbes.galenos.2025.01033
Forbes J Med 2025;6(3):237-241

Original Article / Orijinal Araştırma

Artificial Intelligence Generated Questions in Medical Education: How Prompt Design in Different Chatbots Shapes Assessment in Obstetrics and Gynecology?

Tıp Eğitiminde Yapay Zeka Tarafından Oluşturulan Sorular: Farklı Chatbotlarda Prompt Tasarımı Kadın Hastalıkları ve Doğum Alanında Değerlendirmeyi Nasıl Şekillendiriyor?

Zuhal YAPICI COŞKUN[1], Yavuz Selim KIYAK[2], Özlem COŞKUN[2], Mehmet KOCA[3], İrem BUDAKOĞLU[2], Özhan ÖZDEMİR[4]

[1]Ankara Bilkent City Hospital, Clinic of Obstetrics and Gynecology, Ankara, Türkiye
[2]Gazi University Faculty of Medicine, Department of Medical Education and Informatics, Ankara, Türkiye
[3]Republic of Türkiye, Ministry of Health, Ankara, Türkiye
[4]University of Health Sciences Türkiye, Gülhane Faculty of Medicine, Department of Obstetrics and Gynecology, Ankara, Türkiye

Cite as: Yapıcı Coşkun Z, Kıyak YS, Coşkun Ö, Koca M, Budakoğlu İ, Özdemir Ö. Artificial intelligence generated questions in medical education: how prompt design in different chatbots shapes assessment in obstetrics and gynecology? Forbes J Med. 2025;6(3):237-241

ABSTRACT

Objective: The aim of this study is to assess the difficulty level of artificial intelligence (AI)-generated multiple-choice questions (MCQs) created by large language models (LLMs) using different prompts across various chatbots, compared to human-written questions.

Methods: We generated case-based MCQs on obstetrics and gynecology using two distinct prompts across four LLM-based chatbots. Expert-reviewed MCQs were administered to 97 medical students who were undergoing clerkship training in obstetrics and gynecology. Subsequently, item difficulty indices were calculated for each MCQ.

Results: The mean difficulty index of the AI-generated questions was 0.30. One prompt produced questions with a difficulty index of 0.34 (classified as difficult), while the other produced a lower difficulty index of 0.25 (classified as more difficult). In contrast, the mean difficulty index of the human-written questions was 0.63, indicating a moderate level of difficulty.

Conclusion: Our study highlights the challenges of using AI-generated MCQs in medical education. Although AI offers promising benefits for question generation, the questions produced were generally too difficult for undergraduate medical students. This underscores the need for more detailed and contextually informed prompt designs to better align AI outputs with assessment requirements. Although BDM-based chatbots enhance efficiency in question generation, expert review remains essential to ensure the appropriateness and quality of the items.

Keywords: Artificial intelligence, large language models, medical education, obstetrics and gynecology

ÖZ

Amaç: Bu çalışmada çeşitli chatbotlarda (yapay zeka robotu) farklı promptlar (istemler) kullanılarak büyük dil modelleri (BDM) ile yapay zeka tarafından üretilen çoktan seçmeli soruların (ÇSS) zorluk seviyesinin, insan tarafından yazılmış sorularla karşılaştırmalı bir şekilde değerlendirilmesi amaçlanmıştır.

Yöntem: Dört BDM tabanlı chatbotta (yapay zeka robotu) iki farklı istem kullanarak obstetrik ve jinekoloji üzerine vaka tabanlı ÇSS'lar oluşturulmuştur. Uzman grubu tarafından incelendikten sonra, ÇSS'lar kadın hastalıkları ve doğum anabilim dalında staj yapan 97 tıp öğrencisine uygulanmıştır. Daha sonra her bir ÇSS için madde (soru) güçlük indeksleri hesaplanmıştır.

Received/Geliş: 15.10.2025
Accepted/Kabul: 18.11.2025
Publication Date/ Yayınlanma Tarihi: 05.12.2025

Corresponding Author/ Sorumlu Yazar:
Zuhal YAPICI COŞKUN, MD,
Ankara Bilkent City Hospital, Clinic of Obstetrics and Gynecology, Ankara, Türkiye
✉ zuhalyapici@yahoo.com
ORCID: 0000-0002-9338-4669

Bulgular: Yapay zeka tarafından üretilen soruların ortalama zorluk endeksi 0,30'dur. İstemlerden biri 0,34 zorluk indeksine sahip (zor olarak sınıflandırılan) sorular üretirken, diğeri 0,25'lik daha düşük bir zorluk indeksi (çok zor olarak kabul edilen) ile sonuçlanmıştır. Buna karşılık, insan tarafından yazılan soruların ortalama zorluk endeksi 0,63'tür ve bu da orta düzeyde bir zorluğa işaret etmektedir.

Sonuç: Çalışmamız, tıp eğitiminde yapay zeka tarafından üretilen ÇSS'lar ile insan üretimi olan ÇSS'lar yerine kullanımında karşılaşılabilecek zorlukları vurgulamaktadır. Yapay zeka, soru üretimi açısından umut verici görünmekle birlikte, üretilen soruların genellikle tıp öğrencileri için yüksek zorlukta olduğu gözlemlenmiştir. Bu sonuçlar ölçme değerlendirme gereksinimlerini karşılayabilecek yapay zeka çıktılarına ulaşabilmek için daha detaylandırılmış ve bağlamla uyumlu istemlerin yapılması gereksinimini vurgulamaktadır. Ayrıca, BDM tabanlı chatbotlar verimlilik açısından destek sağlarken, soruların uygunluğunu ve kalitesini sağlamak için uzman incelemesi önemini korumaktadır.

Anahtar Kelimeler: Yapay zeka, büyük dil modelleri, tıp eğitimi, kadın hastalıkları ve doğum

INTRODUCTION

Case-based multiple-choice questions (MCQs) are widely used in medical education,[1] particularly in fields such as obstetrics and gynecology,[2,3] because of their efficiency in assessing knowledge and cognitive skills. However, creating high-quality MCQs is a labor-intensive and time-consuming process that demands significant expertise. To address this challenge, leveraging artificial intelligence (AI) for the development of MCQs has been proposed as a promising solution, given its potential to play a pivotal role in medical education.[4-6]

Large language models (LLMs) have demonstrated significant potential to automate the generation of MCQs for medical education.[7-12] More specifically, recent literature provides evidence supporting the validity of ChatGPT-generated MCQs.[13] However, the findings suggest that employing simple prompts, such as "write four MCQs about this topic," often leads to suboptimal outputs.[14,15] Therefore, well designed[11] and detailed[16] prompts should be used to generate higher-quality MCQs. However, it remains unclear whether these commands can generate MCQs at targeted difficulty levels—defined as the proportion of test-takers who answer an item correctly—across different LLM-based chatbots. This study aims to fill this gap by examining the difficulty levels of AI-generated MCQs in obstetrics and gynecology using various prompts across four chatbots.

This study examines how prompt design and model selection influence the difficulty of AI-generated MCQs in obstetrics and gynecology and addresses how these compare with human-written questions. Our goal was to enhance understanding of how AI can be effectively utilized for case-based MCQ generation and to provide evidence supporting the development of better MCQs that are appropriately challenging for medical students.

METHODS

Study Setting, Design and Participants

This study was conducted in the Department of Obstetrics and Gynecology at University of Health Sciences Türkiye, Gülhane Faculty of Medicine, during the 2023–2024 academic year. Undergraduate medical education at the school spans six years.

The first three years focus on basic medical sciences, offering limited clinical exposure. In the fourth year, the gynecology and obstetrics clerkship is conducted in six rotations with 35–50 students per group, and the passing grade is calculated based on an assessment consisting of a multiple-choice examination. To avoid bias, only fourth-year students who had completed the obstetrics and gynecology clerkship were included. Using convenience sampling, we recruited 97 (88.1%) volunteer participants for the study. The participants completed a test comprising 18 obstetrics and gynecology MCQs, including 14 AI-generated questions and four human-written ones. Since the participants were native Turkish speakers, the AI-generated questions in English were translated into Turkish. Following completion of the internship, the test was conducted in a supervised classroom setting after informed consent had been obtained from all participants.

Multiple-Choice Question Generation

In March 2024, we aimed to generate four MCQs for each of the four LLM-based chatbots. The chatbots used were ChatGPT-3.5, ChatGPT-4, Gemini 1.0 Pro, and Mixtral-8x7B-Instruct-v0.1. We used two prompts for each chatbot: Kıyak's[16] promptand Zuckerman et al.'s[11] prompt. Previous studies have demonstrated their effectiveness.[9,11] These prompts were also used to develop a custom GPT to generate case-based MCQs for medical education because of their proven effectiveness.[17]

To generate the questions, we selected two topics from the learning objectives of the obstetrics and gynecology clerkship: the differential diagnosis of vaginal bleeding and the management of postpartum hemorrhage. These topics were provided to the four chatbots using two distinct prompts. Ultimately, we generated 14 questions—seven each on vaginal bleeding and postpartum hemorrhage—because Gemini 1.0 Pro declined to create a question using Zuckerman et al.'s[11] prompt, likely due to its requirement for an NBME-style question. We used ChatGPT's webpage for ChatGPT-3.5, Case-based MCQ Generator[18] for ChatGPT-4, Gemini's webpage for Gemini 1.0 Pro, and Hugging Face's webpage[19] for Mixtral. We added "differential diagnosis of vaginal bleeding" and "management of postpartum hemorrhage" to the prompts and generated each question on a separate conversation page. Additionally, since Kıyak's[16] prompt requires selecting a difficulty level, we specifically requested the generation of difficult questions.

In addition to the 14 AI-generated questions, four expert-written MCQs—two on vaginal bleeding and two on postpartum hemorrhage—were provided to the participants.

The test was administered independently of the obligatory exams in the medical school and did not affect the students' grades. Informed consent was obtained from each participant. The study was approved by the University of Health Sciences Türkiye, Gülhane Scientific Research Ethics Committee (approval no: 2024/04, date: 24.04.2014).

Expert Panel

The AI-generated questions were reviewed and revised by two obstetrician–gynecologists with expertise with LLMs to ensure clinical accuracy, consistency, and suitability for the intended student level. Revisions were primarily limited to minor refinements in wording, clinical accuracy, and internal consistency, leaving the fundamental structure of the questions unchanged. None of the items necessitated major revisions.

Statistical Analysis

In line with our research question, we calculated the item (question) difficulty index for each question as the number of correct answers divided by the total number of test-takers. The difficulty index ranges from zero to one, with one representing the easiest level and zero representing the most difficult level. We classified values <0.30 as "too difficult", 0.30–0.40 as "difficult", 0.40–0.80 as "moderate", and >0.80 as "easy".[20,21]

RESULTS

Of the fourth-year medical students who completed their obstetrics and gynecology internship, 88.1% participated in an 18-item multiple-choice test. The mean difficulty index of the 14 AI-generated items was 0.30, placing the items at the threshold of the "difficult" category. In contrast, the four expert-authored items demonstrated a substantially higher mean difficulty index of 0.63, aligning with the "moderate" difficulty range (Table 1).

Notable differences were observed in the more detailed classification based on prompt type. Items generated using Kıyak's[16] prompt (n=8) were classified as "difficult", with an average difficulty index of 0.34. However, items derived from Zuckerman et al.'s[11] prompt (n=5) showed a lower average difficulty index of 0.25 and were classified as "very difficult."

Table 1. The difficulty levels of AI-generated and human-written multiple-choice questions

Focus	LLM type or human	Prompt	Difficulty index	Difficulty level
Vaginal bleeding	ChatGPT-3.5	Kıyak[16]	0.15	Too difficult
		Zuckerman et al.[11]	0.24	Too difficult
	ChatGPT-4	Kıyak[16]	0.68	Moderate
		Zuckerman et al.[11]	0.33	Difficult
	Gemini 1.0 Pro	Kıyak[16]	0.27	Too difficult
		Zuckerman et al.[11]	-	-
	Mixtral-8x7B	Kıyak[16]	0.40	Moderate
		Zuckerman et al.[11]	0.21	Too difficult
	Human-written	-	0.73	Moderate
	Human-written	-	0.60	Moderate
Postpartum hemorrhage	ChatGPT-3.5	Kıyak[16]	0.07	Too difficult
		Zuckerman et al.[11]	0.50	Moderate
	ChatGPT-4	Kıyak[16]	0.34	Difficult
		Zuckerman et al.[11]	0.09	Too difficult
	Mixtral-8x7B	Kıyak[16]	0.36	Difficult
		Zuckerman et al.[11]	0.13	Too difficult
	Gemini 1.0 Pro	Kıyak[16]	0.48	Moderate
		Zuckerman et al.[11]	-	
	Human-written	-	0.62	Moderate
	Human-written	-	0.57	Moderate

AI: Artifical intelligence, LLM: Large language model

In light of these findings, prompt design plays a critical role in shaping the cognitive accessibility of AI-generated questions.

In the two thematic areas evaluated—differential diagnosis of vaginal bleeding (n=7) and management of postpartum hemorrhage (n=7)—items generated by AI consistently exhibited lower difficulty indices compared to those written by experts. AI-generated items have a uniform level of difficulty, items written by experts provide a more balanced spectrum of difficulty. Specifically, none of the AI-generated items reached the "easy" classification (>0.80), while two of the four items written by experts fell within the "moderate-easy" range.

DISCUSSION

This study aimed to evaluate the difficulty levels of human-generated and AI-generated MCQs in obstetrics and gynecology. The findings highlighted the combined influence of prompt architecture and human judgment on the psychometric properties of MCQs in medical education. To the best of our knowledge, this is the first study to investigate the effect of different prompts and chatbots on the difficulty of MCQs.[13] Our key findings indicate that the mean difficulty index of the 14 AI-generated questions falls into the "too difficult" category when administered to a group of undergraduate medical students. Questions generated using Kıyak's[16] prompt were classified as difficult, whereas those generated using Zuckerman et al.'s[11] prompt were classified as too difficult. In comparison, the human-written questions reflected a moderate level of difficulty.

While our study found AI-generated questions difficult, previous researches reported moderate mean difficulty levels indices such as of 0.71,11 0.69,10 and 0.689 in ChatGPT-generated case-based questions. This discrepancy may be attributed to differences among the participant groups; variations in their background knowledge, experience, and familiarity with the subject matter could have influenced their performance. Our results emphasize that prompt design is not just a technical input for AI tools, but an educational intervention that can shape learning outcomes by modulating assessment difficulty. This positions prompt engineering as an emerging field of educational design.

The findings of this study highlight a critical issue in AI-based educational content creation: mismatched difficulty levels. Assessments that fail to accurately reflect the expected competency of learners may lead to suboptimal or even negative learning outcomes. When AI-generated items consistently fall outside the optimal difficulty range—particularly in formative assessments—this can compromise both learning efficiency and student motivation. Consequently, aligning the prompt structure with the curriculum level and learner profiles is not merely desirable, but a pedagogical necessity.

Our findings indicate that AI-generated questions, particularly those created using Zuckerman et al.'s[11] prompt, tend to be too difficult for undergraduates, likely because the prompt did not specify the intended difficulty. In contrast, Kıyak's[16] prompt produced questions with varying difficulty, ranging from too difficult to moderate, despite having explicitly requested difficult questions. This variability may arise from two factors: (1) limited detail provided in the prompts and (2) inherent limitations of LLMs.[22-24] The prompt templates can be improved by updating them to allow inclusion of additional context about local needs in the curriculum and target population, through prompt-engineering tactics.[25] However, the inherent limitations of LLMs will still require experts to review and revise the outputs to ensure that the MCQs are appropriate for the target population's needs. Better prompts can reduce the effort required after generation.

Study Limitations

Several limitations should be considered when interpreting the results of this study. First, the study was conducted with a specific group of fourth-year medical students at a single institution, which may limit the generalizability of the findings. Second, the study evaluated only a small number of questions generated by four AI models. Additionally, the study focused on two specific topics in obstetrics and gynecology, and the results might differ for other medical topics. Future research should expand the study to include multiple institutions and a more diverse group of participants to enhance generalizability, investigate the performance of various AI models and versions to identify those most effective in generating optimally difficult questions, and assess AI-generated question difficulty across a wider range of medical topics. Furthermore, the reviewers' expertise with LLMs could have affected their assessment of question clarity and perceived difficulty.

CONCLUSION

In conclusion, our study highlights both the challenges and potential of utilizing AI-generated MCQs in medical education, particularly in obstetrics and gynecology. Despite the promising advancements in AI, the questions generated were generally too difficult for undergraduate medical students. This underscores the necessity for more detailed and contextually informed prompt designs to better align AI outputs with assessment requirements. Additionally, while LLM-based chatbots provide valuable support in efficient question generation, expert review remains crucial to ensure compliance with ethical guidelines.

Ethics

Ethics Committee Approval: The study was approved by the University of Health Sciences Türkiye, Gülhane Scientific Research Ethics Committee (approval no: 2024/04, date: 24.04.2014).

Informed Consent: Informed consent was obtained from each participant.

Footnotes

Authorship Contributions

Surgical and Medical Practices: Z.Y.C., Y.S.K., Ö.Ö., Concept: Z.Y.C., Y.S.K., Ö.C., Ö.Ö., Design: Z.Y.C., Y.S.K., Ö.C., M.K., İ.B., Ö.Ö., Data Collection or Processing: Z.Y.C., Y.S.K., Analysis or Interpretation: Z.Y.C., Y.S.K., Ö.C., M.K., Literature Search: Z.Y.C., Writing: Z.Y.C., Y.S.K., Ö.C., M.K., İ.B.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

REFERENCES

1. Pugh D, De Champlain A, Touchie C. Plus ça change, plus c'est pareil: making a continued case for the use of MCQs in medical education. Med Teach. 2019;41:569-77.
2. Balaha MH, El-Ibiary MT, El-Dorf AA, El-Shewaikh SL, Balaha HM. Construction and writing flaws of the multiple-choice questions in the published test banks of obstetrics and gynecology: adoption, caution, or mitigation? Avicenna J Med. 2022;12:138-47.
3. Jud SM, Cupisti S, Frobenius W, et al. Introducing multiple-choice questions to promote learning for medical students: effect on exam performance in obstetrics and gynecology. Arch Gynecol Obstet. 2020;302:1401-6.
4. Çalışkan SA, Demir K, Karaca O. Artificial intelligence in medical education curriculum: An e-Delphi study for competencies. PLoS One. 2022;17:e0271872.
5. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach. 2024;46:446-70.
6. Stadler M, Horrer A, Fischer MR. Crafting medical MCQs with generative AI: a how-to guide on leveraging ChatGPT. GMS J Med Educ. 2024;41:Doc20.
7. Cheung BHH, Lau GKK, Wong GTC, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions-A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PLoS One. 2023;18:e0290691.
8. Coşkun Ö, Kıyak YS, Budakoğlu Iİ. ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: a randomized controlled experiment. Med Teach. 2025;47:268-74.
9. Kıyak YS, Coşkun Ö, Budakoğlu Iİ, Uluoğlu C. ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. Eur J Clin Pharmacol. 2024;80:729-35.
10. Laupichler MC, Rother JF, Grunwald Kadow IC, Ahmadi S, Raupach T. Large language models in medical education: comparing ChatGPT- to human-generated exam questions. Acad Med. 2024;99:508-12.
11. Zuckerman M, Flood R, Tan RJB et al. ChatGPT for assessment writing. Med Teach. 2023;45:1224-7.
12. Sathe TS, Roshal J, Naaseh A, L'Huillier JC, Navarro SM, Silvestri C. How I GPT it: development of custom artificial intelligence (AI) chatbots for surgical education. J Surg Educ. 2024;81:772-5.
13. Kıyak YS, Emekli E. ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. Postgrad Med J. 2024;100:858-65.
14. Kıyak YS. ChatGPT's ability or prompt quality: what determines the success of generating multiple-choice questions. Acad Pathol. 2024;11:100119.
15. Ngo A, Gupta S, Perrine O, Reddy R, Ershadi S, Remick D. ChatGPT 3.5 fails to write appropriate multiple choice practice exam questions. Acad Pathol. 2024;11:100099.
16. Kıyak YS. A ChatGPT prompt for writing case-based multiple-choice questions. Rev Esp Educ Med. 2023;4:98-103.
17. Kıyak YS, Kononowicz AA. Case-based MCQ generator: a custom ChatGPT based on published prompts in the literature for automatic item generation. Med Teach. 2024;46:1018-20.
18. OpenAI. Case-based MCQ generator [Internet]. United States: ChatGPT; 2024 [cited 2024 Apr 30]. Available from: https://chatgpt.com/g/g-vuyyH0jUp-case-based-mcq-generator
19. Hugging Face. Hugging Face – Chat [Internet]. United States: Hugging Face; 2024 [cited 2024 Apr 30]. Available from: https://huggingface.co/chat/
20. Franzen D, Cuddy MM, Ilgen JS. Trusting your test results: building and revising multiple-choice examinations. J Grad Med Educ. 2018;10:337-8.
21. Tavakol M, O'Brien DG, Sharpe CC, Stewart C. Twelve tips to aid interpretation of post-assessment psychometric reports. Med Teach. 2024;46:188-95.
22. Masters K. Medical Teacher's first ChatGPT's referencing hallucinations: lessons for editors, reviewers, and teachers. Med Teach. 2023;45:673-5.
23. Kirshteyn G, Golan R, Chaet M. Performance of ChatGPT vs. HuggingChat on OB-GYN topics. Cureus. 2024;16:e56187.
24. Ozgor BY, Simavi MA. Accuracy and reproducibility of ChatGPT's free version answers about endometriosis. Int J Gynaecol Obstet. 2024;165:691-5.
25. Indran IR, Paranthaman P, Gupta N, Mustafa N. Twelve tips to leverage AI for efficient and effective medical question generation: a guide for educators using Chat GPT. Med Teach. 2024;46:1021-6.