Research Article

Evaluating Artificial Intelligence in Orthopaedics: A Pilot Study on Accuracy and Reliability in Medical Student Competency Tests

Domy P. Putra,[1*] Farhan A. Setyagisna,[2] Heksa Trisnawati[2]

[1]Department of Orthopaedic and Traumatology Faculty of Medicine, Universitas Brawijaya – Dr. Saiful Anwar General Hospital, Malang, Indonesia
[2]Faculty of Medicine Universitas Brawijaya, Malang, Indonesia

*Corresponding author: domy_pradana@ub.ac.id
Received 24 September 2025; Accepted 4 May 2026
https://doi.org/10.23886/ejki.14.1229.1

Abstract

Artificial intelligence (AI) is increasingly recognized as a valuable tool in medical education, yet its effectiveness across platforms remains underexplored. This study evaluated the performance of nine AI models-ChatGPT-4o, ChatGPT Mini, Gemini, Gemini Advanced, Perplexity, Perplexity Pro, Ortho Research Pro, Ortho AI, and Claude- in answering 30 expert-validated multiple-choice questions (MCQs) from the orthopaedics section of the UKMPPD. All models were evaluated concurrently between August and September 2024 using their official web interfaces. *Each model was tested five times to assess accuracy and consistency. Statistical analysis was conducted using SPSS version 30.0. Normality and homogeneity were assessed using the Shapiro-Wilk and Levene's tests. Accuracy differences were analyzed using one-way ANOVA followed by Tukey's HSD post hoc test ($p < 0.05$). Reliability was evaluated using the intraclass correlation coefficient (ICC). Gemini demonstrated the highest mean accuracy (89 ± 2.79%), while ChatGPT Mini had the lowest (66 ± 3.33%). Significant differences in accuracy were observed ($p < 0.05$), with Gemini differing only from ChatGPT Mini and Perplexity. Most models demonstrated excellent reliability (ICC > 0.90), with Ortho Res. Pro among the best (ICC 0.956; 95% CI 0.925–0.977), while the remaining models showed good reliability. Overall, AI models showed strong potential for supporting medical exam preparation, although performance varied across platforms.*
Keywords: *medical education, artificial intelligence, orthopaedics, UKMPPD, reliability.*

Evaluasi Kecerdasan Buatan dalam Ortopedi: Studi Percontohan tentang Akurasi dan Keandalan dalam Ujian Kompetensi Mahasiswa Kedokteran

Abstrak

Kecerdasan buatan (AI) semakin diakui sebagai alat yang dapat mendukung pendidikan kedokteran, namun efektivitasnya lintas platform masih belum banyak diteliti. Studi ini mengevaluasi kinerja sembilan model AI (ChatGPT-4o, ChatGPT Mini, Gemini, Gemini Advanced, Perplexity, Perplexity Pro, Ortho Research Pro, Ortho AI, dan Claude) dalam menjawab 30 pertanyaan pilihan ganda (MCQ) yang telah divalidasi oleh ahli dari bagian ortopedi UKMPPD. Seluruh model diuji secara bersamaan antara Agustus hingga September 2024 melalui antarmuka resmi masing-masing. Setiap model diuji dalam lima pengulangan untuk menilai akurasi dan konsistensi. Analisis statistik dilakukan menggunakan SPSS versi 30.0. Normalitas dan homogenitas dinilai menggunakan uji Shapiro-Wilk dan Levene. Perbedaan akurasi dianalisis menggunakan ANOVA satu arah diikuti oleh uji post hoc Tukey HSD ($p < 0,05$). Keandalan dievaluasi menggunakan koefisien korelasi intrakelas (ICC). Gemini menunjukkan akurasi rata-rata tertinggi (89 ± 2,79%), sementara ChatGPT Mini memiliki akurasi terendah (66 ± 3,33%). Perbedaan signifikan dalam akurasi diamati ($p < 0,05$), dengan Gemini berbeda hanya dari ChatGPT Mini dan Perplexity. Sebagian besar model menunjukkan keandalan yang sangat baik (ICC > 0,90), dengan Ortho Res. Pro menjadi yang terbaik (ICC 0,956; 95% CI 0,925–0,977), sementara model lainnya menunjukkan keandalan yang baik. Secara keseluruhan, model AI menunjukkan potensi kuat untuk mendukung persiapan ujian medis, meskipun kinerjanya bervariasi di berbagai platform.
Kata kunci: *pendidikan kedokteran, kecerdasan buatan, ortopedi, UKMPPD, reabilitas.*

Introduction

Artificial intelligence (AI) has evolved rapidly in recent decades and is increasingly applied across various fields, including medicine.[1] In the context of medical education, particularly in the medical profession, assessing a doctor's competence is crucial to ensuring graduates uphold the quality of healthcare, protect patient safety, and possess the necessary skills to practice ethically and professionally.[2] The term for medical student competency assessments varies by country; for example, in the United States, it is known as the United States Medical Licensing Examination (USMLE), a standardized national exam required for medical licensing. The USMLE consists of multiple-choice questions and computer-based simulations.[3] In Indonesia, this assessment is called the Medical Profession Program Student Competency Test or *Uji Kompetensi Mahasiswa Program Profesi Dokter* (UKMPPD), a competency test that medical students must pass to assess their medical knowledge and skills before becoming licensed practitioners. UKMPPD includes a computer-based test (CBT) consisting of multiple-choice questions (MCQs) and an objective structured clinical examination (OSCE) to assess clinical skills.[4]

In orthopaedics, UKMPPD tests students' ability to understand and apply clinical concepts, from taking patient histories and performing physical exams to providing appropriate management. However, medical students face significant challenges in answering UKMPPD questions. These challenges stem from the complexity of question formats, the vast amount of material to master, and the pressure to respond quickly and accurately.[5] The UKMPPD covers topics from all medical departments that students study during preclinical and clinical years. Orthopaedics, in particular, presents a challenge, as questions in this field require a multidisciplinary understanding, involving the interrelationship of various bodily systems, not just orthopaedics.[6]

To address these challenges, students need additional support to understand the material better, analyze question patterns, and identify the best answers in accordance with established guidelines. AI has emerged as a potential solution to help students prepare for exams.[7] AI excels in data analysis, understanding question patterns, and providing algorithm-based answers. Therefore, this pilot study aims to evaluate AI performance in answering UKMPPD MCQs, particularly in orthopedics, with a primary focus on assessing the accuracy and reliability of multiple AI models.

Methods

This study employed a cross-sectional design to compare the accuracy of nine AI models in answering MCQs from the orthopaedics section of the UKMPPD. All models were evaluated concurrently, with data collected between August and September 2024 through their respective official web interfaces. The AI models included ChatGPT 4-o, ChatGPT Mini, Gemini, Gemini Advanced, Perplexity, Perplexity Professional, Ortho Research Pro, Ortho AI, and Claude. These models were selected based on general accessibility, support for both Indonesian and English, ability to interpret text- and image-based questions, and availability for non-demo use. A total of 30 MCQs from the orthopaedic section of the UKMPPD, administered over the past 10 years, were selected and validated by an orthopaedic expert. Each AI model was prompted using a standardized instruction simulating a medical student taking the examination, requiring the selection of a single best answer without explanation. Each model was tested in five repetitions to assess consistency.

Accuracy was calculated based on correct responses, with correct answers assigned a score of 10 and incorrect answers a score of 0, and expressed as a percentage. All measurements were conducted under controlled conditions to minimize network interference and using the same device to ensure consistency. Statistical analysis was performed using SPSS version 30.0. Data normality was assessed using the Shapiro–Wilk test, and homogeneity of variance was evaluated using Levene’s test. Differences in accuracy between models were analyzed using one-way analysis of variance (ANOVA), followed by post-hoc comparisons with Tukey’s HSD test. A p-value of <0.05 was considered statistically significant. Reliability across repeated measurements was assessed using the intraclass correlation coefficient (ICC), with values indicating

good to excellent consistency. ICC values were interpreted as follows: values below 0.5 indicated poor reliability, values between 0.5 and 0.75 indicated moderate reliability, values between 0.75 and 0.9 indicated good reliability, and values greater than 0.9 indicated excellent reliability.[8]

Results

The mean accuracy of each AI model in answering orthopaedic UKMPPD questions is presented in Table 1.

Among the evaluated models, Gemini demonstrated the highest mean accuracy (89 ± 2.79%), whereas GPT Mini exhibited the lowest performance (66 ± 3.33%). One-way ANOVA revealed a significant difference in accuracy among the models ($p < 0.05$), indicating variability in their performance. Post hoc analysis using Tukey's HSD test showed that Gemini's accuracy differed significantly only when compared to ChatGPT Mini and Perplexity, as presented in Table 2.

Table 1. Accuracy of AI's Performance

AI	Accuracy (% ± SD)	p-value
ChatGPT 4o	81 ± 2.97	
Ortho AI	83 ± 2.23	
Ortho Res. Pro	80 ± 2.78	
Gemini	89 ± 2.79	
Gemini Adv.	81 ± 5.05	<0.001
Perplexity	75 ± 6.91	
Perplexity Prof.	82 ± 4.34	
Claude	82 ± 5.47	
ChatGPT Mini	66 ± 3.33	

Table 2. Post-Hoc Tukey Test of Gemini

Independent Variable	Dependent Variable	Mean Diff.	Std. Error	Sig.	95% CI	
					Lower Bound	Upper Bound
Gemini	ChatGPT 4o	7.998	2.694	0.106	-0.885	16.881
	ChatGPT Mini	22.666	2.694	<0.001	13.782	31.549
	Ortho Research Pro	8.666	2.694	0.061	-0.217	17.549
	Gemini Advanced	7.336	2.694	0.176	-1.547	16.219
	Perplexity	14.000	2.694	<0.001	5.116	22.883
	Perplexity Pro	6.668	2.694	0.277	-2.215	15.551
	Claude	6.668	2.694	0.277	-2.215	15.551
	Ortho AI	6.000	2.694	0.411	-2.883	14.883

Reliability analysis demonstrated that most AI models achieved good to excellent consistency across repeated trials (Table 3). Several models, including ChatGPT-4o (ICC 0.905; 95% CI 0.839–0.950), Ortho AI (ICC 0.935; 95% CI 0.890–0.966), Ortho Res. Pro (ICC 0.956; 95% CI 0.925–

0.977), Gemini (ICC 0.919; 95% CI 0.862–0.957), Gemini Advanced (ICC 0.902; 95% CI 0.833–0.948), and Perplexity (ICC 0.903; 95% CI 0.835–0.949), achieved ICC values above 0.90, indicating excellent reliability and very high consistency. In contrast, Perplexity Pro (ICC 0.845; 95% CI 0.736–0.918), Claude (ICC 0.785; 95% CI 0.633–0.887), and ChatGPT Mini (ICC 0.877; 95% CI 0.791–0.935) demonstrated good reliability. Overall, the evaluated AI models exhibited satisfactory consistency across repeated measurements when answering orthopaedic UKMPPD questions.

Table 3. Consistency of AI's Performance

AI	ICC	95% CI		ICC Interpretation
		Lower Bound	Upper Bound	
ChatGPT 4o	0.905	0.839	0.950	Excellent
Ortho	0.935	0.890	0.966	Excellent
Ortho Res. Pro	0.956	0.925	0.977	Excellent
Gemini	0.919	0.862	0.957	Excellent
Gemini Advanced	0.902	0.833	0.948	Excellent
Perplexity	0.903	0.835	0.949	Excellent
Perplexity Pro	0.845	0.736	0.918	Good
Claude	0.785	0.633	0.887	Good
ChatGPT Mini	0.877	0.791	0.935	Good

Discussion

This study evaluates the performance of nine AI models in answering orthopaedic multiple-choice questions from the UKMPPD. Their performance was assessed across five trials on accuracy and consistency. Unlike previous studies that focused on AI use in orthopaedic diagnosis or basic disease assessment, such as diagnosing osteoarthritis, this study examines the more practical potential of nine AI models: ChatGPT 4o, ChatGPT Mini, Gemini, Gemini Advanced, Ortho Research Pro, Perplexity, Perplexity Pro, Claude, and Ortho AI, to perform in a setting that simulates medical students taking the UKMPPD. The study, therefore, aims to assess the real-world accuracy and reliability of these models' responses. Overall, the findings may have broader implications for students and the academic community regarding the integration of AI into the preparation for and evaluation of national medical examinations.

The results demonstrated statistically significant differences in accuracy among AI models, indicating substantial variability in performance. Gemini achieved the highest mean accuracy (89 ± 2.79%), while ChatGPT Mini showed the lowest (66 ± 3.33%). Post hoc analysis revealed that Gemini significantly outperformed ChatGPT Mini and Perplexity, suggesting a relative advantage in handling MCQ-based clinical reasoning tasks. These findings are consistent with previous studies showing that advanced large language models outperform smaller models on medical examination-style tasks due to superior contextual understanding and reasoning.[10-12]

A deeper error analysis provides insight into the observed performance gaps. Lower-performing models, particularly ChatGPT Mini, likely exhibit limitations in multi-step clinical reasoning, which is essential for answering UKMPPD-style questions. Many MCQs require integration of patient history, physical examination findings, and diagnostic interpretation, followed by the ability to distinguish between closely related distractors. Models with reduced capacity or simplified architectures may rely more on surface-level pattern recognition rather than structured reasoning, increasing the likelihood of selecting plausible but incorrect answers.[13,14] Additionally,

these models may have difficulty differentiating subtle variations among answer options, a known limitation associated with reasoning depth and contextual processing.

Language factors may have also contributed to performance differences. UKMPPD questions are presented in Indonesian and often include localized clinical terminology. Most AI models are predominantly trained on English-language datasets, which may reduce comprehension accuracy when processing non-English inputs. This limitation may lead to misinterpretation of clinical context and contribute to incorrect responses, particularly in models not optimized for multilingual understanding.[15]

Interestingly, despite its lower accuracy, ChatGPT Mini demonstrated relatively high consistency across repeated trials, as reflected by good reliability metrics. This suggests that the model produces stable but not necessarily correct responses, potentially due to deterministic output patterns or constrained reasoning pathways. In contrast, higher-performing models such as Gemini, Ortho AI, and Ortho Research Pro achieved excellent reliability (ICC > 0.90), indicating both high accuracy and strong consistency in reasoning. This aligns with findings that larger and more advanced AI architectures tend to produce more robust and reproducible outputs.[16] Overall, these findings reinforce existing evidence that large language models show strong potential in medical question-answering tasks but remain limited by reasoning errors, contextual misinterpretation, and variability across models.[10,11,12] Therefore, careful validation is necessary before integrating AI into high-stakes educational or assessment settings.

This study has several limitations. First, only nine AI models were evaluated, while newer and updated versions continue to emerge. Second, network stability could not be fully controlled, which may have influenced response generation. Third, the number of MCQs was limited to 30 from the past 10 years, which may limit generalizability. Additionally, this study did not include image-based or multimedia questions, which are highly relevant in orthopaedic assessments and may further challenge AI performance. The absence of image-based questions may underestimate the challenges AI would face in real orthopaedic examinations. Future studies should incorporate larger, more diverse question sets, including image- and case-based formats, and perform more detailed qualitative error analysis. Evaluating AI performance in multilingual settings and optimizing models for local medical contexts will also be crucial for improving their applicability in national examinations such as the UKMPPD. These findings are particularly relevant for medical education, where AI may serve as a supplementary tool for self-assessment and practice in standardized examinations.

Conclusion

This study demonstrates that AI models show variable performance in answering orthopaedic UKMPPD multiple-choice questions, with Gemini achieving the highest accuracy among the evaluated models. Most AI models exhibited good to excellent reliability, indicating consistent performance across repeated trials. These findings suggest that AI has potential as a supportive tool for medical students in exam preparation, particularly in enhancing understanding of MCQ-based clinical reasoning. However, variability in model accuracy highlights the need for careful evaluation before integrating AI into high-stakes assessments. Future research should focus on expanding question diversity, including image-based formats, and improving AI adaptation to local language and clinical context.

Acknowledgement

There are no additional acknowledgments other than the authors included in this study.

Authors' Contributions

DPP: conceptualization, methodology, supervision, validation, writing-review, and editing. FAS: conceptualization, formal analysis, visualization, writing-original draft. HT: formal analysis, software, visualization, writing-original draft. All authors have read and approved the final manuscript.

Funding

This study did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data and Materials Availability

The datasets used in this study, including UKMPPD multiple-choice questions, are not publicly available due to institutional and copyright restrictions. However, the processed data and results generated during the study are available from the corresponding author upon reasonable request.

Consent to Participate

Informed consent was not required as the study did not involve any human participants or personal data.

Ethics Statement

Ethical approval was not required as this study did not involve human or animal participants. All procedures were conducted in accordance with relevant research guidelines.

Consent for Publication

Consent for publication was not required as the study did not include any human subjects, personal data, or identifiable images.

Conflict of Interest

The author declares that there is no conflict of interest regarding the publication of this article.

AI Usage Declaration

Artificial intelligence (AI) tools were used in this study both as the primary objects of evaluation and for drafting assistance during manuscript preparation. AI tools were not involved in data analysis or interpretation of the study results. All AI-generated content was carefully reviewed, verified, and edited by the authors. The authors take full responsibility for the accuracy, integrity, and originality of the manuscript.

References

1. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J. 2021;8:e188-e194. doi: 10.7861/fhj.2021-0095
2. Utomo PS, Randita ABT, Riskiyana R, Kurniawan F, Aras I, Abrori C, et al. Predicting medical graduates' clinical performance using national competency examination results in Indonesia. BMC Med Educ. 2022;22:254. doi: 10.1186/s12909-022-03321-x
3. Williams M, Kim EJ, Pappas K, Uwemedimo O, Marrast L, Pakmezaris R, et al. The impact of United States Medical Licensing Exam (USMLE) step 1 cutoff scores on recruitment of underrepresented minorities in medicine: A retrospective cross-sectional study. Health Sci Rep. 2020;3:e2161. doi: 10.1002/hsr2.161
4. Labobar MK, Irab SP, Anggai M, Tingginehe RM, Togodly A, Rantetoding S. Implementation pattern of CBT UKMPPD public health tutoring at the Faculty of Medicine, Cenderawasih University. Formosa J Sci Technol. 2024;3:213-222. doi: 10.55927/fjst.v3i2.8088
5. Putra AM, Syahruddin FI, Iskandar D. Factors that influence UKMPPD failure. Jurnal Eduhealth. 2024;15:530-541. doi: 10.54209/eduhealth.v15i04
6. Istadi Y, Sukestiyarno, Raharjo TJ, Azam M, Mulyono E. A scoping review of determinants of the graduation of professional competencies for medical education students in Indonesia. Adv Soc Sci Educ Humanit Res. 2021;574:681-688. doi: 10.2991/assehr.k.211125.129
7. Vieriu AM, Petrea G. The impact of artificial intelligence (AI) on students' academic development. Educ Sci. 2025;15:343. doi: 10.3390/educsci15030343
8. Portney, L.G. and Watkins, M.P. Foundations of Clinical Research: Applications to Practice. 3rd Ed. New Jersey: Pearson Education, Inc. 2009.
9. Cao M, Wang Q, Zhang X, Liang Z, Qiu J, Yung PSH, et al. Large language models' performances regarding common patient questions about osteoarthritis: A comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Perplexity. J Sport Health Sci. 2025;14:101016. doi: 10.1016/j.jshs.2024.101016
10. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2:e0000198. doi: 10.1371/journal.pdig.0000198
11. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:e45312. doi: 10.2196/45312
12. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. Nat Med. 2025;31:943-950. doi: 10.1038/s41591-024-03423-7
13. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? Patters. 2024;5:100943. doi: 10.1016/j.patter.2024.100943
14. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv. 2023. doi: 10.48550/arXiv.2303.12712
15. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res. 2024;26:e60807. doi: 10.2196/6080
16. Anil R, Borgeaud S, Alayrac JB, Yu J, Soricut R, Schalkwyk J, et al. Gemini: a family of highly capable multimodal models. arXiv. 2025. doi: 10.48550/arXiv.2312.11805