Comparative Analysis of Four AI Platforms for Orthopaedic Education: Evaluation of Accuracy and Explanation Quality

Singh S[a], Kaur A[b], Singh H[c]
[a]Department of Orthopaedics, FHMS, Taylor's University, Malaysia
[b]Department of Preclinical Sciences, FMHS, UTAR, Malaysia
[c]Department of Pharmacy, Government Polytechnic College, Jalandhar, India

Corresponding Author: Assoc Prof Dr. Simerjit Singh
Department of Orthopaedics, FHMS, Taylor's University, Malaysia.
Email: simerjit.singh@taylors.edu.my

ABSTRACT

INTRODUCTION: Artificial Intelligence (AI) systems are increasingly used in medical education, which requires integrative clinical reasoning. Despite their rapid adoption, little is known about the comparative performance of different AI platforms in solving scenario-based orthopaedic multiple-choice questions (MCQs) and providing high-quality explanatory feedback. We evaluated four AI platforms; ChatGPT, Perplexity, Claude, and Gemini; on their ability to answer 45 validated orthopaedic MCQs accurately and provide clear, logical explanations. **MATERIALS AND METHODS:** Each platform received the same 45 MCQs under standardized conditions. Correctness was scored out of 45, and the explanation quality was scored on a scale of 0 to 90 using a structured rubric. Pairwise comparisons were conducted using one-way ANOVA and Tukey's post hoc tests. A composite score, comprising 70% correctness and 30% explanation weightage, further contextualized overall performance. **RESULTS:** ChatGPT and Perplexity demonstrated higher correctness scores than Claude and Gemini. Explanation quality ranged from 80% (72/90) for ChatGPT and Perplexity; to 63% (57/90) for Gemini. Both correctness and explanation quality scores were positively correlated ($r=0.84$, $p<0.01$). Composite scores paralleled these findings, placing ChatGPT and Perplexity above Claude and Gemini. **CONCLUSIONS:** The results highlight that AI platforms vary substantially in accuracy and the clarity of their explanations, thus underscoring the importance of carefully selecting a platform when integrating AI into orthopaedic education. Educators should consider that the significant inter-platform variability in correctness and explanation quality observed in this study has important implications for orthopaedic education.

KEYWORDS: Artificial intelligence, Orthopaedics, ChatGPT, Gemini, Perplexity

INTRODUCTION

Artificial intelligence (AI) is rapidly reshaping healthcare, from automated image analysis to natural language processing (NLP) chatbots that support patients and clinicians.[1,2] AI tools offer valuable opportunities to expand and deepen learning in medical education. A key advantage is their ability to provide instant feedback on practice questions, clinical scenarios, and exam preparation, helping learners identify and correct errors quickly.[3] This is particularly relevant in orthopaedics, where effective decision-making relies on integrating anatomy, biomechanics, imaging, and clinical practice knowledge.[4]

Traditional orthopaedic education relies on lectures, tutorials, and case-based learning, but advanced AI tools now enable faster and more accessible knowledge acquisition. Modern AI language models can summarize complex information and provide rapid, accurate, and relevant feedback.[5] However, little is known about how different AI platforms compare in handling specialized, complex questions typical of musculoskeletal medicine.[6]

In orthopaedics, scenario-based multiple-choice questions (MCQs) are often used for assessment and learning. These MCQs replicate real-world patient encounters, often including diagnostic images or laboratory data. This requires learners to integrate findings before selecting the best answer.[7] In orthopaedics, such questions may depend on subtle differences in fracture patterns, implant choice, or rehabilitation protocols. Thus, AI tools must provide both accurate information and clear reasoning that explains why one option is most appropriate.[8,9]

This study compared four prominent AI platforms, namely ChatGPT, Perplexity, Claude, and Gemini, to assess their effectiveness as educational tools in orthopaedics. The exact models used were 4o (ChatGPT), Sonar (Perplexity), 3.5 Sonnet (Claude), and 2.0 Flash (Gemini). ChatGPT and Perplexity are known for their natural language processing capabilities, while Claude and Gemini are recognized for their ability to handle complex medical data. The study evaluated two key metrics: i) correctness on 45 validated multiple-choice questions (MCQs) and ii) the quality of explanations. While many AI studies focus only on accuracy, assessing both correctness and explanation quality provides greater educational value.[10]

Emerging evidence suggests that medical students and residents place substantial trust in AI explanations, treating them as reliable resources for understanding complex subjects.[11] Therefore, a better understanding of how these explanations relate to correctness and whether AI platforms maintain consistent quality has significant implications for training programmes. By systematically evaluating these AI tools, we aim to provide educators with insights to inform their selection of the most reliable and pedagogically sound AI platforms. We also aim to highlight areas where AI requires further development to meet the demands of advanced clinical education in orthopaedics. Hence, the primary aims of this study were (1) to compare the correctness scores of four AI platforms across a standardized set of orthopaedic MCQs, (2) to assess the quality of the explanations provided, and (3) to explore how correctness and explanation quality might correlate, offering insight into each platform's strengths and limitations.

MATERIALS AND METHODS

Study Design and AI Platforms

The study was a cross-sectional, comparative study of four commercially available AI platforms. Each platform was evaluated using a 45-question test bank designed specifically for orthopaedic education. The question pool covered a breadth of clinical scenarios, including diagnosis (e.g., fracture patterns, musculoskeletal tumours), treatment management (e.g., surgical versus conservative approaches), and imaging interpretation (e.g., radiology, orthopaedic devices). The questions underwent multiple review rounds by two orthopaedic educators to ensure clarity and relevance. Each correct answer was awarded 1 point, with a maximum correctness score of 45 per platform. Explanation scores ranged from 0-2 per domain item. The three domains included: i) accuracy of the content (alignment with standard orthopaedic principles and evidence-based guidelines), ii) clarity of thought process (absence of ambiguity and explicit mentioning of key steps to help make a diagnosis or deciding treatment), and iii) completeness (addressing the main question and explaining the rationale of excluding distractors in the answers), resulting in a maximum of 6 points per question. This explanation score was scaled to a 0-90 range (45 questions × 2 points × 3 domains =270 raw points, normalised to 90 for simplicity). Two reviewers independently rated the explanations. Questions were presented using identical phrasing without additional hints or clarifications. Each AI platform received randomised questions to minimise systematic bias. Each platform was retested five times. Both correctness and explanation quality were recorded for each question-platform pair. Mean scores and standard deviations for correctness (out of 45) and explanation quality (out of 90) were computed. Equality of variances was tested using Levene's test (*p*=0.61), and the data were found to be normally distributed. One-way ANOVA and post hoc Tukey's test were employed to detect significant differences in scores among platforms. Pearson's correlation was used to analyse the relationship between correctness and explanation quality. A composite score was calculated for each AI platform, weighing 70% for correctness and 30% for explanation quality, using the following formula: Composite score per question =0.7 × correctness + 0.3 × explanation quality.

RESULTS

ChatGPT and Perplexity scored the highest in both correctness and explanation quality compared to Claude and Gemini (Figure 1 and Table I). However, a two-way ANOVA did not reveal significant differences in AI performance across various question subtypes (diagnosis, treatment, and image interpretation).

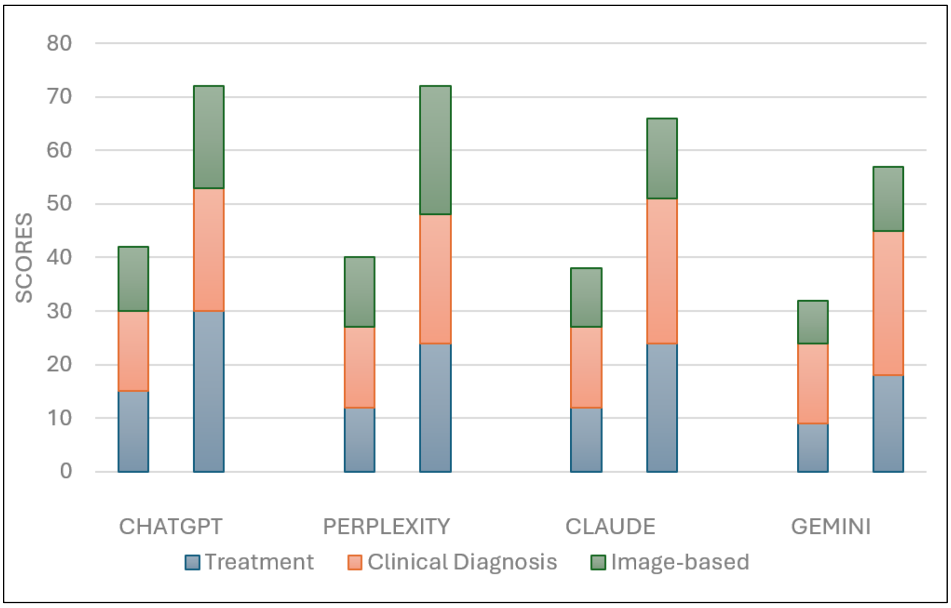

Figure 1: AI correctness and explanation quality scores among different question types

Post hoc Tukey's test analysis revealed that ChatGPT's correctness scores were significantly higher than Claude's (p<0.01) and Gemini's (p<0.01). Explanation quality scores were significantly lower for Gemini compared to ChatGPT (p<0.01), Perplexity (p<0.01), and Claude (p<0.01). A partial eta-squared value of 0.92 indicates a large effect size. All other pairwise differences were non-significant. The explanation rubric scores highlighted that ChatGPT, Perplexity, and Claude consistently produced coherent, detailed, and structured rationales.

Both the correctness and explanation quality scores were positively correlated (r=0.84, p<0.01) (Figure 2). Overall composite scoring supported the correctness-based ranking, with ChatGPT and Perplexity remaining at the top, followed by Claude and Gemini.

Table I: Association between AI platform and test scores

AI platforms	Correctness Scores Mean (SD)	Explanation Quality score Mean (SD)	def	F	*p*-value	Partial Eta Squared
ChatGPT	41.6(0.89)	72(2.82)	3,3	58.3*	<0.01	0.92
Perplexity	40.0(1.00)	72(4.69)		15.4*	<0.01	
Claude	38.0(1.58)	66(4.12)				
Gemini	31.6(1.51)	57(4.30)				

*ANOVA test was performed, level of significance at p<0.01, def. = degree of freedom

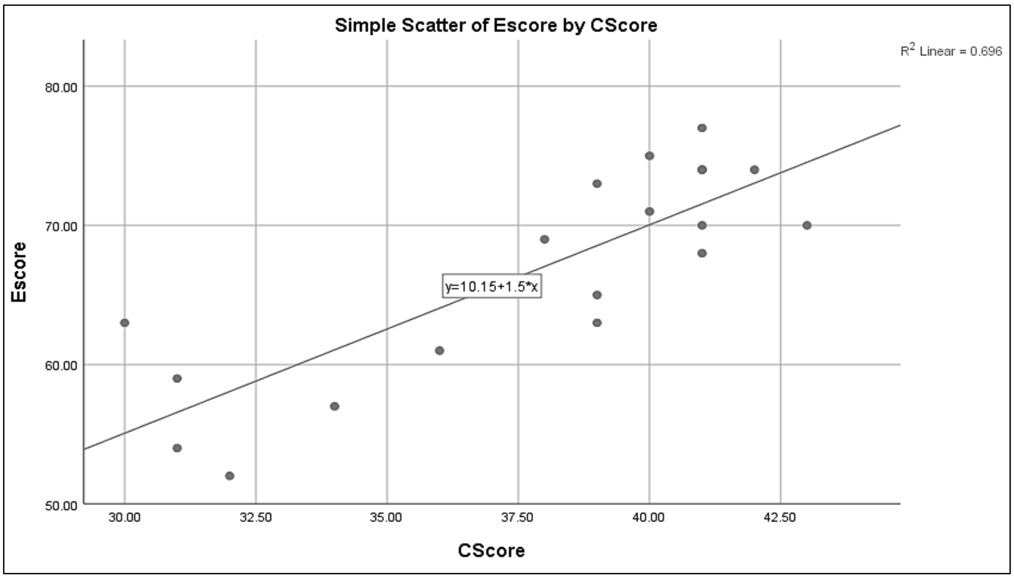

Figure 2: Correlation between the explanation and correctness scores.

DISCUSSION

The primary aim of this study was to provide evidence-based guidance for medical educators and orthopaedic trainers seeking to incorporate artificial intelligence (AI) into their teaching programmes. The study compared four platforms, namely ChatGPT, Perplexity, Claude, and Gemini; evaluating both their accuracy in answering scenario-based orthopaedic multiple-choice questions (MCQs) and the clarity and depth of their explanations. ChatGPT and Perplexity demonstrated the highest precision alongside consistently strong explanation quality. These findings indicate that well-developed language models can perform highly on MCQs, even in specialised domains such as orthopaedics.

However, accuracy alone does not guarantee meaningful learning.[12] In clinical education, a correct answer paired with an unclear or superficial explanation may offer limited value to learners, who benefit most from understanding the reasoning behind a decision. Without this, students may simply memorise correct responses without grasping the underlying principles. This will undermine the central aim of medical education, which is the development of sound clinical judgement.[13] In contrast, AI systems that provide clear, logical explanations can serve as effective instructional aids, supporting learners in understanding why certain diagnoses or treatments are more appropriate than others.[3]

Although ChatGPT and Perplexity are general-purpose natural language processing models with the capacity to address medical queries, Claude AI and Gemini are designed specifically for healthcare applications. The latter employs advanced computational methods to process complex medical datasets, supporting applications in diagnostics, personalised treatment planning, and clinical research. In this study, however, ChatGPT and Perplexity achieved higher accuracy and provided more comprehensive explanations, enabling clearer reinforcement of orthopaedic concepts for undergraduate medical learners. This advantage may be due to their transformer-based architecture, which utilises context to generate clear and connected responses.[14] ChatGPT's extensive training on large and diverse datasets likely enhances its ability to create context-sensitive answers and rationales. The comparable performance of Perplexity suggests that combining large language models with effective prompt engineering can further improve the precision and clarity of AI-generated educational content.

Claude and Gemini demonstrated comparatively lower accuracy and explanation quality than ChatGPT and Perplexity. Claude achieved 84% (38/45) correct responses and an explanation quality score of 73% (66/90), but some answers lacked the depth or completeness required for advanced clinical reasoning. Gemini's performance was lower, with 71% (32/45) correct responses; post hoc Tukey's testing confirmed that this difference was statistically significant compared with ChatGPT and Perplexity. Despite its lower overall accuracy, Gemini showed a strong positive correlation between correctness and explanation quality (r=0.84), indicating that the accompanying explanations were generally high quality when its answers were correct. Conversely, incorrect answers were typically paired with weaker explanations. This pattern may reflect inherent model architecture or training data limitations. Performance across different question subtypes namely, diagnosis, treatment, and image interpretation was broadly similar for all platforms. This likely reflects that the AI systems responded to textual descriptions of radiological or clinical findings rather than analysing actual images. As a result, "image-based" items functioned effectively as text-based questions, accounting for the consistent performance across categories.

From an educational perspective, these results have several key implications. First, platform choice is important. High-performing AI tools can shorten the feedback cycle by identifying errors and reinforcing correct reasoning. This is especially relevant in orthopaedic education, where learners must integrate imaging interpretation, surgical indications, and rehabilitation planning into clinical decisions.[15,16] By providing clear, immediate rationales, tools such as ChatGPT and Perplexity can act as accessible, self-paced learning aids outside traditional classroom or clinical settings.

Second, explanation quality is essential for developing a deeper understanding. While accuracy is important, particularly in high-stakes fields, a well-structured explanation helps students retain knowledge, think critically, and apply concepts to new situations. Educators and developers should therefore focus on AI models that provide correct answers and explain how those answers were reached.

Third, the strong correlation between correctness and explanation quality across all platforms (particularly for Gemini) suggests that sound reasoning tends to produce both accurate answers and clear, logical explanations. Conversely, weak reasoning often results in errors and poor clarity. This relationship supports using combined measures, such as a composite score, to comprehensively evaluate AI performance.

Relying only on correctness may overlook important educational gaps, while focusing solely on explanation quality may ignore the critical role of accuracy in clinical practice.[18,19]

Fourth, the findings highlight ways AI performance could be improved. While ChatGPT and Perplexity performed well, they occasionally produced incorrect answers or incomplete explanations, particularly for complex cases involving detailed imaging or advanced surgical decisions. Orthopaedic practice often hinges on subtle differences in fracture classification, comorbidities, and patient-specific factors such as bone density. Developers could strengthen domain-specific reasoning by training models on comprehensive orthopaedic resources, including textbooks, research articles, imaging datasets, operative notes, and validated question banks. Similar targeted fine-tuning could benefit Gemini and Claude in advanced medical contexts. Since medical guidelines evolve, platforms should also update their knowledge bases regularly to incorporate the latest consensus statements and clinical trial evidence.

This study has certain limitations. First, although orthopaedic specialists carefully validated our 45-question bank, the sample size remains relatively modest compared to the vast range of questions that an orthopaedic trainee might encounter in actual practice. Second, each platform was tested using its existing interface, which may be subject to updates, expansions of training data, or algorithmic improvements. AI models can evolve rapidly, so today's results may shift as new versions are deployed.[21] Third, the quality of explanations is partly subjective. Even with clear guidelines and agreement between reviewers, educators may value clarity or depth differently. To address this, the medical education community may need to agree on what constitutes a high-quality AI explanation in specialized domains.

Another area worth studying is how AI performance changes over time. As developers update models using feedback and additional training data, it will be possible to determine whether certain AI platforms consistently outperform others or whether lower-performing tools, such as Gemini, can improve more rapidly. It would also be useful to study how students use these platforms. Some may prefer ChatGPT's detailed explanations, while others may want quicker, more concise answers. It is also important to observe how beginners and advanced learners respond differently to AI reasoning, as some explanations may be too basic or too complex, depending on the learner's level.

Finally, while our combined scoring system balances accuracy and explanation quality, future research could test more detailed scoring methods. For example, in high-risk questions such as imaging interpretation or post-operative care, accuracy may need to be weighted more heavily. Conversely, some educators may prefer to emphasise explanation quality to help students better understand the process of clinical reasoning. Most AI models rely on their training data to answer questions, and they cannot be pre-programmed or fine-tuned by end users to deliver tailored responses. The answers generated also depend on the quality of the prompt. Nevertheless, these AI platforms can be used reliably for revision, tutoring, and formative assessments. Students can utilize them to clarify difficult concepts in simpler terms, ask follow-up questions, and resolve doubts in a manner similar to interacting with a tutor. Many medical schools are now developing frameworks and guidelines for adopting AI in assessments, including in high-stakes exams.

Our findings underscore that AI-based educational tools are not created equal and that thorough validation against standardised question banks is crucial before widespread adoption in any specialised medical field.

ChatGPT and Perplexity showed considerable promise as potential tutors or study companions for orthopaedic students and residents, while Claude and Gemini highlighted the need for more targeted refinement. Balancing accuracy with clear, high-quality reasoning will be essential for the effective use of AI in medical education as these technologies continue to evolve. These insights can guide the design of future AI tools to serve as reliable and valuable aids in training the next generation of orthopaedic specialists.

In conclusion, among the four AI platforms evaluated, namely ChatGPT, Perplexity, Claude, and Gemini; ChatGPT and Perplexity demonstrated the highest accuracy and explanation quality in scenario-based questions related to orthopaedics. Although Gemini scored lower overall, its correctness and explanation quality rose and fell together, indicating strong internal consistency at a lower level of performance. These findings support the potential of AI as a supplemental tool in medical education, particularly in orthopaedics while also highlighting the need for continued refinement and cautious implementation to maintain reliability and educational value.

CONFLICT OF INTEREST

None

ACKNOWLEDGEMENTS

We would like to thank Ms. Gurbani for reading the manuscript and proofreading.

DECLARATION

The authors declare that generative artificial intelligence technologies were used during the writing of this manuscript. Specifically, ChatGPT (version GPT-4o, developed by OpenAI, accessible at https://chat.openai.com) was employed to improve grammar and language clarity.

REFERENCES

1. Maleki Varnosfaderani S, Forouzanfar M. The role of AI in hospitals and clinics: Transforming healthcare in the 21st century. Bioengineering (Basel) 2024;11:337. doi: 10.3390/bioengineering11040337.
2. Al Kuwaiti A, Nazer K, Al-Reedy A, et al. A review of the role of artificial intelligence in healthcare. J Pers Med 2023;13:951. doi: 10.3390/jpm13060951.
3. Narayanan S, Ramakrishnan R, Durairaj E, Das A. Artificial intelligence revolutionizing the field of medical education. Cureus 2023;15:e49604. doi: 10.7759/cureus.49604.
4. Gan W, Ouyang J, Li H, Xue Z, et al. Integrating ChatGPT in Orthopedic Education for Medical Undergraduates: Randomized Controlled Trial. J Med Internet Res 2024;26:e57037. doi: 10.2196/57037.
5. Muasher-Kerwin C, Hughes MC, Foster ML, Al Azher I, Alhoori H. Exploring large language models for summarizing and interpreting an online brain tumor support forum. Digit Health 2025;11:20552076251337345. doi: 10.1177/20552076251337345.
6. Chen X, Wang L, You M, et al. Evaluating and Enhancing Large Language Models' Performance in Domain-Specific Medicine: Development and Usability Study With DocOA. J Med Internet Res 2024;26:e58158. doi: 10.2196/58158.

7. Al Shuriaqi S, Aal Abdulsalam A, Masters K. Generation of medical case-based multiple-choice questions. Int Med Educ 2023;3:12-22. https://doi.org/10.3390/ime3010002
8. Giorgino R, Alessandri-Bonetti M, Luca A, et al. ChatGPT in orthopedics: a narrative review exploring the potential of artificial intelligence in orthopedic practice. Front Surg 2023;10:1284015. doi: 10.3389/fsurg.2023.1284015.
9. Lum ZC, Collins DP, Dennison S, et al. Generative artificial intelligence performs at a second-year orthopedic resident level. Cureus 2024;16:e56104. doi: 10.7759/cureus.56104.
10. Owan VJ, Owan OB, Osaat SD, EO Etta, BA Bassey. Exploring the potential of artificial intelligence tools in educational measurement and assessment. Eurasia J Math Sci Tech Educ 2023;19:em2307. https://doi.org/10.29333/ejmste/13428
11. Sami A, Tanveer F, Sajwani K, et al. Medical students' attitudes toward AI in education: perception, effectiveness, and its credibility. BMC Med Educ 2025;25:82. doi: 10.1186/s12909-025-06704-y.
12. Harden RM, Laidlaw JM. Essential skills for a medical teacher: an introduction to teaching and learning in medicine. 3rd ed. Amsterdam: Elsevier; 2020.
13. Heber S, Grasl MC, Volf I. A successful intervention to improve conceptual knowledge of medical students who rely on memorization of disclosed items. Front Physiol 2023;14:1258149. doi: 10.3389/fphys.2023.1258149.
14. Gupta N, Khatri K, Malik Y, et al. Exploring prospects, hurdles, and road ahead for generative artificial intelligence in orthopedic education and training. BMC Med Educ 2024;24:1544. doi: 10.1186/s12909-024-06592-8.
15. Koo A, Almeida BA, Kerluku J, Yang B, Fufa D. Teaching in orthopaedic surgery: effective strategies for educating the modern learner in a modern surgical practice. JB JS Open Access 2022;7:e22. doi: 10.2106/JBJS.OA.22.00005.
16. Saravi B, Hassel F, Ülkümen S, et al. Artificial intelligence-driven prediction modeling and decision making in spine surgery using hybrid machine learning models. J Pers Med 2022;12:509. doi: 10.3390/jpm12040509.
17. Ruth M. Simulation and its effects on knowledge retention and critical thinking skills. Int J Nurs Health Care Res 2023;6:1426. https://doi.org/https://doi.org/10.29011/2688-9501.
18. Gordon D, Rencic JJ, Lang VJ, et al. Advancing the assessment of clinical reasoning across the health professions: definitional and methodologic recommendations. Perspect Med Educ 2022;11:108-14. doi: 10.1007/s40037-022-00701-3.
19. Jay R, Davenport C, Patel R. Clinical reasoning: the essentials for teaching medical students, trainees, and non-medical healthcare professionals. Br J Hosp Med 2024;85:1-8. doi: 10.12968/hmed.2024.0052.
20. Psidai B, Hilkert AS, Kaarre J, et al. A practical guide to the implementation of AI in orthopaedic research – part 1: opportunities in clinical application and overcoming existing challenges. J Exp Orthop 2023;10:117. doi: 10.1186/s40634-023-00683-z.
21. Rodriguez-Merchan EC. Some artificial intelligence tools may currently be useful in orthopaedic surgery and traumatology. World J Ortho 2025;16:102252. doi: 10.5312/wjo.v16.i2.102252.