Huy LD, Anh LN, Quang MH, et al. Performance of GPT-5 and Gemini 2.5 Pro on the Orthopaedic In-Training Examination. *Orthopedic Reviews*. 2026;18.
doi:10.52965/001c.160184

General

Performance of GPT-5 and Gemini 2.5 Pro on the Orthopaedic In-Training Examination

Le Duc Huy[1,2a], Luong Nhat Anh[1,2], Mai Huy Quang[1,2], Luu Huu Phuc[1,2], Nguyen Duc Trung[1,2], Vu Duc Thang[1,2], Le Hoai Nam[1,3], Tran Trung Dung, Prof. PhD. MD.[1,2]

[1] College of Health Sciences, VinUniversity, [2] Orthopaedics & Sports Medicine Center, Vinmec International Hospital, [3] Department of General Orthopedics, 108 Military Central Hospital

Keywords: ChatGPT, Gemini, Large language models, Artificial intelligence, OITE

https://doi.org/10.52965/001c.160184

Orthopedic Reviews

Vol. 18, 2026

Background

Previous studies evaluating large language models (LLMs) on the Orthopaedic In-Training Examination (OITE) have primarily focused on earlier-generation models and single-pass accuracy. These investigations did not assess newer multimodal systems such as GPT-5 and Gemini 2.5 Pro, nor did they examine the reasoning quality underlying model responses or the consistency of outputs across repeated trials. As LLMs are increasingly used as educational tools, a more comprehensive evaluation framework is needed to assess not only correctness but also reliability and explanatory validity on specialty-specific, image-rich examinations.

Methods

We conducted a controlled, parallel evaluation of GPT-5 and Gemini 2.5 Pro using 412 OITE-style questions from the 2023–2024 examination cycle obtained via an institutional AAOS ResStudy subscription. Primary outcomes included overall and subspecialty-specific accuracy. Secondary analyses evaluated explanatory quality, error-pattern classification, response consistency across repeated trials, and performance stratified by imaging burden. Paired accuracy was compared using McNemar's exact test.

Results

Gemini 2.5 Pro demonstrated higher overall accuracy than GPT-5 on the 2023–2024 OITE question set (81.1% vs 76.0), with both models exceeding published PGY-5 resident benchmarks. Accuracy declined significantly with questions containing images (74.2% vs 71.6%). Subspecialty performance varied widely, with accuracy ranging from 42.9% to 94.1% for GPT-5 and from 57.1% to 95.8% for Gemini, and both models performing poorest in Hand and Wrist questions. Among incorrect responses, faulty reasoning accounted for 52.5% of GPT-5 errors, whereas stem misinterpretation was the predominant error for Gemini (43.6%). Incorrect or partially correct explanations accompanied 45.4% of GPT-5 and 41.7% of Gemini responses. Consistency testing showed high reproducibility (fully consistent responses: 88% for GPT-5 and 84% for Gemini), with all inconsistent outputs occurring in image-containing questions.

Conclusions

GPT-5 and Gemini 2.5 Pro demonstrate strong performance on recent OITE content, exceeding prior LLM benchmarks; however, persistent limitations in multimodal reasoning, explanatory reliability, and response consistency indicate that high accuracy alone does not ensure dependable clinical reasoning, underscoring the need for cautious educational use.

a Correspondence:
Le Duc Huy [ORCID: 0000-0002-7494-8828]
24huy.ld@vinuni.edu.vn; +84982184409

BACKGROUND

Various artificial intelligence (AI) tasks, including natural language processing, image understanding, code generation, and complex reasoning, have shown considerable advancements. In the field of medicine, these capabilities are increasingly being used to retrieve knowledge, prepare examinations, and support decisions.[1-3] However, the extent to which contemporary LLMs can perform reliably on orthopedic assessments remains incompletely characterized.

The OITE is a proctored, computer-based, multiple-choice assessment administered annually to United States and several international residency programs. Since 1963, this examination has sampled knowledge across 10 blueprint categories spanning established principles, conventional procedures, and treatment modalities in orthopedic surgery. Programs use the OITE outcomes at the examinee and program levels to guide study plans, curriculum review, and quality improvement, with program year and cohort summaries reported for Accreditation Council for Graduate Medical Education (ACGME)-accredited United States programs.[4,5]

Prior OITE evaluations revealed that ChatGPT-3.5 correctly answered roughly 45–54% of questions—consistent with postgraduate year 1 (PGY1)-level performance—whereas GPT-4 achieved 63–74%, approaching or surpassing PGY5 averages depending on the question set and format.[1-3,6-9] Additionally, several studies have reported higher accuracy on text-only questions but lower accuracy on image-based or higher-order reasoning questions.[2,7-9]

However, important limitations persist in prior OITE-focused LLM studies. Most evaluations involved earlier-generation models, relied on single-pass accuracy, and did not assess response reproducibility, error mechanisms, or the quality of model explanations.

With the emergence of advanced multimodal LLMs, a more comprehensive evaluation framework is warranted. Accordingly, our study evaluates GPT-5 and Gemini 2.5 Pro using a 2023–2024 OITE question set with native image input. Beyond overall accuracy, this study incorporates structured analyses of explanatory quality, error-pattern classification, and response consistency. By leveraging the OITE as a standardized, high-fidelity benchmark, this work aims to provide an updated and more nuanced assessment of modern LLM performance on orthopaedic examination content and to address key gaps in the existing literature.

METHODS

This study was conducted as a controlled, parallel evaluation of two large language models (LLMs)—GPT-5 (OpenAI) and Gemini 2.5 Pro (Google DeepMind)—using a standardized set of orthopaedic knowledge questions. The study followed a predefined protocol including question selection, model testing, blinding procedures, and systematic error analysis. The primary objective was to compare overall and subspecialty-specific model accuracy. Secondary objectives included evaluation of explanatory quality, error-mechanism classification, consistency across repeated trials, and performance differences based on imaging content.

QUESTION SOURCE AND ELIGIBILITY

A total of 415 OITE-style questions corresponding to the 2023–2024 OITE examination were obtained through an institutional AAOS ResStudy subscription, which provides residency programs with access to archived educational questions for trainee self-assessment. These questions are educational materials and not proprietary or unreleased AAOS examination items. Three items containing embedded video media were excluded because GPT-5 did not support video processing at the time of testing, yielding a final analytical sample of 412 questions.

Questions were categorized into four media groups: text-only, single-image, two–three image, and ⩾four-image items. Images included clinical photographs, radiographs, CT/MRI sequences, histology, and diagrams.

MODEL VERSIONS AND TESTING ENVIRONMENT

GPT-5 (ChatGPT) and Gemini 2.5 Pro were tested between September 15 and September 27, 2025, using the stable production versions available during that period. No major software updates or version changes occurred during the testing window. Both models were evaluated using identical, controlled conditions to ensure fairness and reproducibility. For every question:

- A new blank session was opened to eliminate carry-over memory or contextual influence.
- A uniform input format was used, consisting of the question stem followed by answer options on separate lines.
- For image-based items, the same images were inserted into both models at their original resolution and orientation.
- No retries, additional prompts, or chain-of-thought instructions were used; each model generated a single response per question.

RANDOMIZATION AND BIAS MITIGATION

Before model evaluation, all 412 questions were randomized using a computer-generated sequence. The same random order was applied to both models. GPT-5 and Gemini 2.5 Pro were tested in parallel, meaning each model received the same question on the same day.

Grading was performed in a blinded fashion. The dataset provided to graders contained only anonymized model outputs labeled by question ID without model identity or time stamps.

PRIMARY OUTCOME: ACCURACY

Accuracy was defined as the proportion of correct answers relative to the validated answer key. Accuracy was calculated:

- overall,
- by media category (text-only vs. image), and

- by subspecialty domain (trauma, arthroplasty, sports, spine, pediatrics, hand, foot and ankle, oncology, and basic science).

SECONDARY OUTCOMES

1. EXPLANATORY QUALITY

For each item, the accompanying explanation generated by the model was evaluated qualitatively by two blinded orthopaedic surgeons across three dimensions:

- Correctness: alignment with current orthopaedic evidence and accepted guidelines.
- Completeness: inclusion of key concepts and clinically relevant details necessary to justify the answer.
- Coherence: logical structure and internal consistency of the reasoning, without contradictions or invented facts.

Explanations were then classified into three categories:

- Appropriate: factually correct, sufficiently detailed, and logically coherent.
- Partially correct: generally accurate but incomplete, oversimplified, or containing minor inaccuracies.
- Incorrect: major factual errors, flawed reasoning, or hallucinated content that could mislead learners.

The distribution of explanatory quality categories was reported descriptively for each model.

2. ERROR-PATTERN CLASSIFICATION

Incorrect responses were categorized into three predefined mechanisms:

- stem misinterpretation,
- faulty clinical reasoning,
- hallucination (fabricated or irrelevant details).

Disagreements were resolved by consensus.

3. CONSISTENCY TESTING

To assess reproducibility, 50 questions (approximately 12% of the full dataset) were selected using stratified sampling to ensure representation from all major orthopaedic subspecialties. These questions were re-administered to each model in three independent sessions spaced at least 24 hours apart, using the same standardized prompt structure and image inputs.

For each question, consistency was defined as:

- Fully consistent: identical answer in all three runs
- Partially consistent: same answer in two of three runs
- Inconsistent: different answers across runs with no majority pattern

The consistency rate was calculated as: number of fully consistent items ÷ total retested items.

STATISTICAL ANALYSIS

The paired question-level correctness between the models was compared overall and within each question-type stratum using McNemar's exact test. The ordinal association between accuracy and image burden was assessed using the Cochran–Armitage trend test.

RESULTS

Both models correctly identified the input question format as multiple-choice questions and selected the correct answers for all 412 questions. GPT-5 and Gemini 2.5 Pro achieved an accuracy of 76.9% (313/412) and 81.1% (334/412), respectively. A consensus analysis provides an actionable lens for educational deployment. When both models selected the same option, which occurred in 317 questions, the accuracy increased to 88.3% (280/317). Using exact binomial tests that treat consensus questions as Bernoulli trials with each model's overall accuracy as the null, consensus accuracy was significantly higher than GPT 5's and Gemini 2.5 Pro (P = 0.001).

TEXT-ONLY VERSUS IMAGE-CONTAINING QUESTIONS

For text-only questions (n = 218), the accuracy was 79.8% for GPT-5 and 87.2% for Gemini 2.5 Pro. When collapsing all image-containing questions (n = 194, comprising 1-image, 2–3-image, and ⩾4-image questions), the accuracy was 71.6% for GPT-5 and 74.2% for Gemini 2.5 Pro. By stratum, the accuracy of GPT versus Gemini 2.5 Pro was 67.1% versus 67.1% (n = 76) for 1-image questions, 74.4% versus 77.8% (n = 90) for 2–3-image questions, and 75.0% versus 82.1% (n = 28) for ⩾4-image questions (Fig. 2, Table 2).

ACCURACY BY SUBSPECIALTY

Of the 412 questions, 404 were categorized into 10 subject domains of the AAOS RessStudy. The performance varied significantly across these domains (Table 1).

Gemini 2.5 Pro achieved the highest accuracy in "Basic Science" (95.8%), "Shoulder and Elbow" (93.8%), and "Pediatric Orthopedics" (91.7%). The lowest accuracy was observed in "Hand and Wrist" (57.1%) and "Trauma" (69.0%).

GPT-5 attained the highest accuracy in "Sports Medicine" at 94.1%, followed by "Shoulder and Elbow" at 87.5% and "Musculoskeletal Tumors and Diseases" at 86.7%. The lowest accuracy was observed in "Hand and Wrist" at 42.9%.

ERROR PATTERN ANALYSIS

Among incorrect responses, GPT-5 most frequently demonstrated faulty reasoning, accounting for 52.5% of its errors (52/99), whereas this category represented a smaller proportion of errors for Gemini 2.5 Pro (38.5%, 30/78). In contrast, stem misinterpretation was more common in Gemini, comprising 43.6% of its errors (34/78) compared with 31.3% in GPT-5 (31/99). Errors attributable to hallucination or knowledge deficit were the least frequent for both models but still notable, representing 16.2% of GPT-5 er-

Table 1. Overall Accuracy of GPT-5 and Gemini 2.5 Pro

Model	*N* (total questions)	Correct (*n*)	Accuracy (%)
GPT	412	313	76.9
Gemini 2.5 Pro	412	334	81.1
GPT + Gemini consensus	317 (consensus)	280	88.3

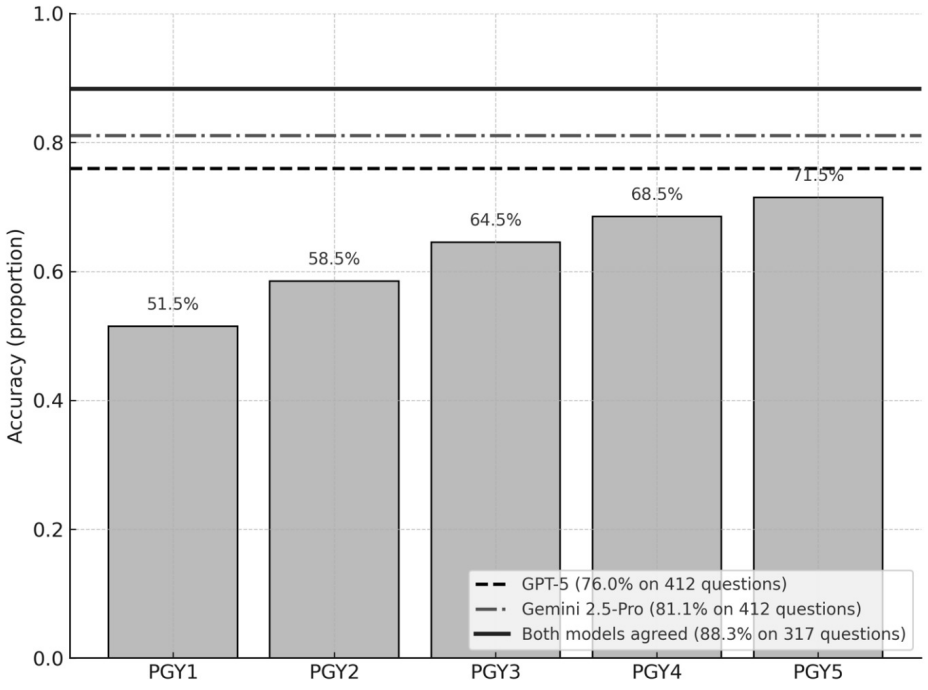

Figure 1. Overall accuracy of GPT-5, Gemini 2.5 Pro, and ACGME orthopedic residents

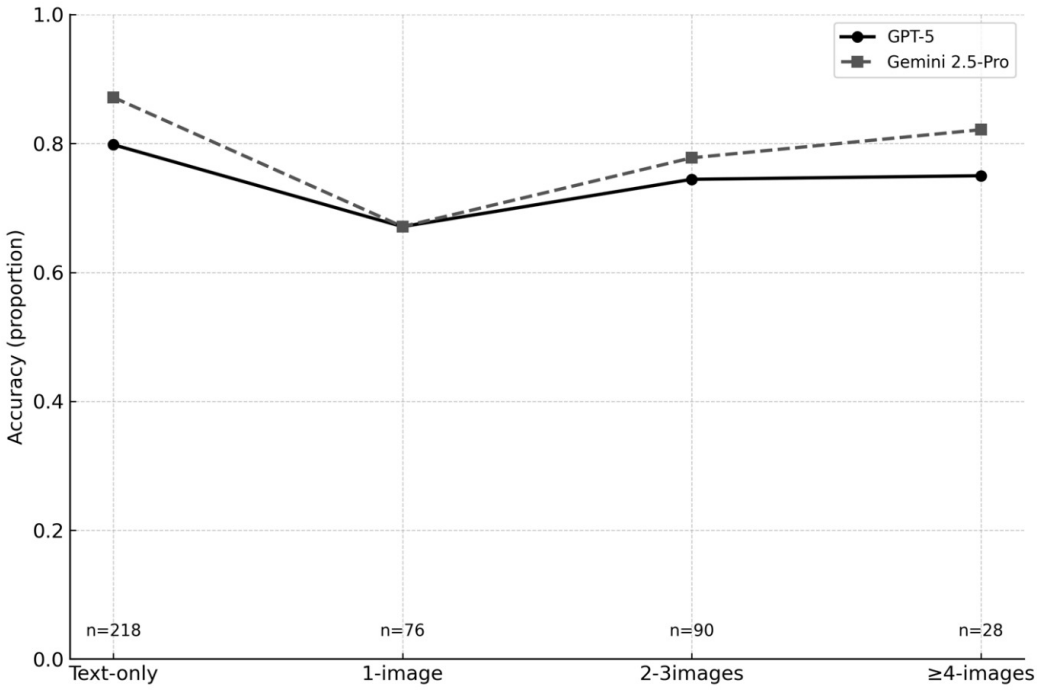

Figure 2. Accuracy by question type (text-only, 1-image, 2–3-image, and ⩾4-image questions)

Table 2. Accuracy by imaging-containing category

Category	*N* (total questions)	GPT Accuracy (n, %)	Gemini Accuracy (n, %)
Text only	218	174 (79.8)	190 (87.2)
1-image	76	51 (67.1)	51 (67.1)
2–3 images	90	67 (74.4)	70 (77.8)
≥4 images	28	21 (75.0)	23 (82.1)

Table 3. Accuracy by question's category

Category	*N* (total questions)	GPT Accuracy (n, %)	Gemini Accuracy (n, %)
Adult Reconstruction (Hip and Knee)	82	58 (70.7)	64 (78.0)
Basic Science	48	36 (75.0)	46 (95.8)
Foot and Ankle	30	24 (80.0)	22 (73.3)
Hand and Wrist	28	12 (42.9)	16 (57.1)
Musculoskeletal Tumors and Diseases	30	26 (86.7)	24 (80.0)
Pediatric Orthopedics	24	20 (83.3)	22 (91.7)
Shoulder and Elbow	32	28 (87.5)	30 (93.8)
Spine	38	28 (73.7)	32 (84.2)
Sports Medicine	34	32 (94.1)	30 (88.2)
Trauma	58	48 (82.8)	40 (69.0)

Table 4. Error-Pattern Distribution

Error Category	GPT (n, %)	Gemini 2.5 Pro (n, %)
Faulty reasoning	52 (52.5%)	30 (38.5%)
Stem misinterpretation	31 (31.3%)	34 (43.6%)
Hallucination/ Knowledge deficit	16 (16.2%)	14 (17.9%)
Total incorrect	**99 (100%)**	**78 (100%)**

Table 5. Explanatory Quality

Explanation Rating	GPT (n, %)	Gemini 2.5 Pro (n, %)
Appropriate	225 (54.6%)	240 (58.3%)
Partially correct	82 (19.9%)	87 (21.1%)
Incorrect	105 (25.5%)	85 (20.6%)
Total	**412 (100%)**	**412 (100%)**

rors (16/99) and 17.9% of Gemini errors (14/78). Overall, the models exhibited distinct failure profiles, with GPT-5 tending toward incorrect reasoning pathways and Gemini more often misinterpreting key elements of the question stem.

EXPLANATORY QUALITY OF MODEL OUTPUTS

Across all items, Gemini 2.5 Pro produced a slightly higher proportion of appropriate explanations compared with GPT-5 (58.3% vs. 54.6%) (Table 5). GPT-5 demonstrated a marginally greater frequency of partially correct explanations (19.9% vs. 21.1% for Gemini), reflecting its tendency to generate more elaborate but occasionally incomplete reasoning. Notably, GPT-5 also exhibited a higher rate of incorrect explanations, with 25.5% of its rationales categorized as incorrect compared with 20.6% for Gemini. These findings indicate that while both models frequently provided clinically sound and relevant reasoning, Gemini produced more consistently accurate explanatory text, whereas GPT-5 demonstrated greater variability with a higher proportion of flawed or misleading rationales.

Table 6. Consistency Testing (n=100 Questions)

Consistency Category	GPT (n, %)	Gemini 2.5 Pro (n, %)
Fully consistent	88 (88.0%)	84 (84.0%)
Partially consistent	8 (8.0%)	15 (15.0%)
Inconsistent	4 (4.0%)	3 (3.0%)

RESPONSE CONSISTENCY

In the 100-question reproducibility subset, GPT-5 showed greater output stability than Gemini 2.5 Pro (Table 6). GPT-5 generated fully consistent responses for 88% of items (88/100), compared with 84% for Gemini. Partial consistency occurred in 8% of GPT-5 outputs and 15% of Gemini outputs. All inconsistent responses for both models arose exclusively from image-containing questions, representing 4% of GPT-5 items and 3% of Gemini items. These findings suggest that while both models demonstrate high reproducibility overall, complex cases involving imaging introduce greater variability, with GPT-5 remaining modestly more stable than Gemini across repeated trials.

DISCUSSION

PRIOR LLM STUDIES ON THE OITE

Previous evaluations of large language models on the Orthopaedic In-Training Examination (OITE) have been limited in scope and methodological depth. The earliest structured analysis reported approximately 49% accuracy for GPT-4 on the 2019 OITE and relied on textual descriptions of imaging due to the absence of multimodal capability at that time.[10] Subsequent comparison of GPT-3.5 and GPT-4 using AAOS ResStudy question banks (2020–2022) showed that GPT-4 consistently outperformed earlier models yet continued to underperform senior residents, particularly on reasoning-intensive items.[11] A more recent multimodel comparison, including GPT-4, Claude, and Gemini, found that contemporary LLMs approached the accuracy of PGY-4 residents on selected 2022 OITE items.[12]

Across these studies, several gaps persist: limited or absent multimodal image testing, lack of structured error analysis, minimal assessment of explanatory reasoning, and no formal consistency testing across repeated runs. These limitations underscore the need for more comprehensive methodological frameworks.

Our study addresses these limitations by incorporating true multimodal image evaluation, subspecialty- and modality-stratified analysis, structured error classification, explanatory-quality assessment, and formal consistency testing.

OVERALL PERFORMANCE

Relative to earlier studies, the observed performance in our study reflects substantial generational advances in LLM capability. Nevertheless, accuracy on the OITE remains lower than reported performance on broad medical licensing examinations such as the USMLE, where ChatGPT-5 has achieved accuracy exceeding 95% [4]. This discrepancy suggests that highly specialized examinations with heavy imaging and domain-specific reasoning demands remain more challenging for LLMs than general medical assessments.

ACCURACY BY IMAGE-CONTAINING CATEGORY

The decline in performance with increasing imaging complexity is consistent with prior multimodal LLM research demonstrating persistent limitations in radiographic and visually dense tasks.[13] Gemini 2.5 Pro outperformed GPT-5 across all image categories, with the largest difference in ⩾4-image items (82.1% vs 75.0%), suggesting more effective image–text integration. Despite improvements in vision–language integration, detailed orthopaedic image interpretation remains a key constraint, particularly in questions requiring synthesis of multiple visual inputs. These findings highlight the continued gap between text-based reasoning and reliable multimodal clinical interpretation.

ACCURACY BY SUBSPECIALTY

Marked variability in performance was observed across orthopaedic subspecialties. Gemini 2.5 Pro demonstrated higher accuracy in six of ten domains, with the largest advantage in Basic Science, while GPT-5 showed relative strengths in Sports Medicine and Trauma. Both models performed poorly in Hand and Wrist, indicating a shared limitation in this subspecialty. In contrast, several domains—including Musculoskeletal Tumors, Shoulder and Elbow, and Sports Medicine—were high-performing for both models, suggesting robust representation of these topics in their training data. Such divergence across subspecialties is consistent with prior reports that LLM performance varies substantially by medical domain and knowledge structure.[14]

ERROR PATTERN AND EXPLANATORY QUALITY

The different ways GPT-5 and Gemini 2.5 Pro made mistakes match what earlier studies have shown about LLM behavior. Modern AI models usually fail not because they lack facts, but because they struggle with complex reasoning or misunderstand parts of the question.[15] GPT-5 often chose the wrong answer because its reasoning steps were incorrect, which has been described in prior work showing that LLMs can sound confident even when their logic is wrong.[16] Gemini, on the other hand, more often misunderstood the question stem.

The explanation analysis highlights the same issue. Even when the final answer was right, both models sometimes provided reasoning that was incomplete or incorrect. A detailed evaluation of GPT-4V on image-based medical exam questions revealed a high frequency of reasoning failures and flawed rationales even when the final answer was correct.[17] This gap between "choosing the right answer" and "explaining it correctly" is an important safety concern. Therefore, AI tools should support learning but should not replace expert judgment, especially for clinical reasoning in orthopaedics.

RESPONSE CONSISTENCY

Our reproducibility analysis reveals that both GPT-5 and Gemini 2.5 Pro deliver largely stable results on repeated queries, with 88% and 84% of responses fully consistent, respectively (Table 6). Such stability aligns with recent work in ophthalmology medical exams showing that LLMs can maintain substantial output reproducibility on repeated trials — albeit with variability increasing for more complex tasks.[18] Notably, in our dataset, all inconsistent responses occurred exclusively in image-containing questions, underscoring a key limitation: multimodal tasks remain inherently less reproducible than text-only reasoning.

This study had certain limitations. Expanding the category-based analysis to a larger item dataset would be valuable for confirming the stability and generalizability of the observed model strengths and weaknesses. Although error patterns and explanatory quality were assessed, these evaluations relied on qualitative judgment and may be subject to observer bias despite the use of structured criteria. In addition, this study focused exclusively on examination performance and did not assess real-world clinical decision-making, operative planning, or patient outcomes. Therefore, the findings should not be extrapolated beyond the context of standardized orthopaedic knowledge assessment.

CONCLUSIONS

In this controlled evaluation of recent OITE content, both GPT-5 and Gemini 2.5 Pro demonstrated strong performance, exceeding previously reported LLM results and approaching or surpassing senior resident benchmarks. While Gemini 2.5 Pro showed higher overall accuracy and stronger multimodal performance, GPT-5 demonstrated greater response consistency. Despite these advances, performance declined with increasing imaging questions, varied substantially across subspecialties, and revealed persistent limitations in reasoning fidelity and explanatory quality. Collectively, these findings indicate that modern LLMs have reached a high level of competence on orthopaedic examination material but remain constrained in multimodal reasoning and interpretability. LLMs may serve as useful adjuncts for orthopaedic education and exam preparation, but their outputs should be interpreted cautiously and not considered substitutes for expert clinical judgment.

ABBREVIATIONS

AAOS American Academy of Orthopaedic Surgeons
AI Artificial intelligence
LLMs Large language models
OITE Orthopaedic In-Training Examination
USMLE United States Medical Licensing Examination

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study was reviewed and determined to be exempt from ethics approval as it did not involve human participants, identifiable personal data, or clinical interventions.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The datasets analysed during the current study are available from the corresponding author.

COMPETING INTERESTS

The authors declare no competing interests.

FUNDING

This research received no specific funding.

AUTHORS' CONTRIBUTIONS

Conceptualization: L.D.H., T.T.D.
Study design and methodology: L.N.A., M.H.Q., L.H.P., N.D.T., L.D.H.
Data acquisition and analysis: M.H.Q., L.H.P., V.D.T., L.H.N.
Manuscript drafting: L.N.A., M.H.Q., L.D.H.
Supervision: L.D.H., T.T.D.

ACKNOWLEDGEMENTS

We would like to thank Editage (www.editage.com) for English language editing.

Submitted: January 22, 2026 EDT. Accepted: February 17, 2026 EDT. Published: April 16, 2026 EDT.

REFERENCES

1. Kung JE, Marshall C, Gauthier C, Gonzalez TA, Jackson JBI. Evaluating ChatGPT Performance on the Orthopaedic In-Training Examination. *JBJS Open Access*. 2023;8(3):e23.00056. doi:10.2106/JBJS.OA.23.00056

2. Guerra GA, Hofmann HL, Le JL, et al. ChatGPT, Bard, and Bing Chat Are Large Language Processing Models That Answered Orthopaedic In-Training Examination Questions With Similar Accuracy to First-Year Orthopaedic Surgery Residents. *Arthroscopy*. 2025;41(3):557-562. doi:10.1016/j.arthro.2024.08.023

3. Hofmann HL, Guerra GA, Le JL, et al. The Rapid Development of Artificial Intelligence: GPT-4's Performance on Orthopedic Surgery Board Questions. *Orthopedics*. 2024;47(2):e85-e89. doi:10.3928/01477447-20230922-05

4. *Orthopaedic In-Training Examination (OITE) Technical Report 2023*. American Academy of Orthopaedic Surgeons Accessed October 27, 2025. https://www.aaos.org/globalassets/education/product-pages/oite/oite-2023-technical-report-2024.pdf

5. *Orthopaedic In-Training Examination (OITE) Technical Report 2024*. American Academy of Orthopaedic Surgeons Accessed October 27, 2025. https://www.aaos.org/globalassets/education/product-pages/oite/oite-2024-technical-report-eds.pdf

6. Lum ZC. Can Artificial Intelligence Pass the American Board of Orthopaedic Surgery Examination? Orthopaedic Residents Versus ChatGPT. *Clinical Orthopaedics and Related Research®*. 2023;481(8):1623. doi:10.1097/CORR.0000000000002704

7. Fares MY, Parmar T, Boufadel P, et al. An Assessment of the Performance of Different Chatbots on Shoulder and Elbow Questions. *Journal of Clinical Medicine*. 2025;14(7):2289. doi:10.3390/jcm14072289

8. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT–3.5, ChatGPT-4, and Orthopaedic Resident Performance on Orthopaedic Assessment Examinations. *JAAOS - Journal of the American Academy of Orthopaedic Surgeons*. 2023;31(23):1173. doi:10.5435/JAAOS-D-23-00396

9. Ozdag Y, Hayes DS, Makar GS, et al. Comparison of Artificial Intelligence to Resident Performance on Upper-Extremity Orthopaedic In-Training Examination Questions. *Journal of Hand Surgery Global Online*. 2024;6(2):164-168. doi:10.1016/j.jhsg.2023.10.013

10. Hayes DS, Foster BK, Makar G, et al. Artificial Intelligence in Orthopaedics: Performance of ChatGPT on Text and Image Questions on a Complete AAOS Orthopaedic In-Training Examination (OITE). *Journal of Surgical Education*. 2024;81(11):1645-1649. doi:10.1016/j.jsurg.2024.08.002

11. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and Orthopaedic Resident Performance on Orthopaedic Assessment Examinations. *J Am Acad Orthop Surg*. 2023;31(23):1173-1179. doi:10.5435/JAAOS-D-23-00396

12. Nawari A, Zahir J, Kumar S, et al. Artificial Intelligence Large Language Models Are Nearly Equivalent to Fourth-Year Orthopaedic Residents on the Orthopaedic In-Training Examination: A Cause for Concern or Excitement? *J Orthopaedic Experience & Innovation*. 2025;6(1). doi:10.60118/001c.124070

13. Nam Y, Kim DY, Kyung S, et al. Multimodal Large Language Models in Medical Imaging: Current State and Future Directions. *Korean J Radiol*. 2025;26(10):900-923. doi:10.3348/kjr.2025.0599

14. Chundi G, Dawar A, Sarwar S, Prasad S, Vosbikian M, Ahmed I. Comparative evaluation of LLMs in orthopedic surgery. *Journal of Orthopaedic Reports*. Published online July 9, 2025:100728. doi:10.1016/j.jorep.2025.100728

15. Mo K, Lin R, Dunn E, et al. Systematic Review on Large Language Models in Orthopaedic Surgery. *Journal of Clinical Medicine*. 2025;14(16):5876. doi:10.3390/jcm14165876

16. Chen S, Gao M, Sasse K, et al. When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. *npj Digit Med*. 2025;8(1):605. doi:10.1038/s41746-025-02008-z

17. Yang Z, Yao Z, Tasmin M, et al. Unveiling GPT-4V's hidden challenges behind high accuracy on USMLE questions: Observational Study. *J Med Internet Res*. 2025;27:e65146. doi:10.2196/65146

18. Shvartz E, Raiyter R, Goldshtein A, Zur O, Margalit E, Bahir D. Assessing consistency of AI chatbot responses in ophthalmology medical exams. *AJO International*. 2025;2(4):100189. doi:10.1016/j.ajoint.2025.100189