

Research Article

Evidence on Demand: OpenEvidence versus Established LLMs on the Orthopaedic In-Training Examination

Asim A. Khan, B.A.^{1a}, Shaan S. Lalvani, B.A.^{1b}, Sam Pourarbab, B.A.^{1c}, Lord J. Hyeamang, B.A.^{1d}, Ryan M. Lew, M.S.^{2e}, Zaamin B. Hussain, M.D. EdM^{2f}, Gregory P. Nicholson, M.D.^{2g}, Grant E. Garrigues, M.D.^{2h}

a Asim Khan is a Charlotte, North Carolina native and a graduate of Northwestern University who is currently in his third year at Rush Medical College. He is drawn to orthopedic surgery, with a particular interest in shoulder pathology and joint injury outcomes, helping patients return to the activities they love. Outside of the hospital, he enjoys playing squash and coaching underserved K-12 students in Chicago.

[Connect with Asim Khan on LinkedIn](#)

[Conflicts of Interest Statement for Asim Khan](#)

b Shaan Lalvani is a medical student at Rush Medical College in Chicago, where he conducts orthopaedic research at Midwest Orthopaedics at Rush. He completed his undergraduate degrees in biology and economics at Brown University, where his honors thesis examined Galectin-3 in glioblastoma stem cells. Before medical school, he worked as a clinical research coordinator in the Department of Neurology at Mount Sinai in New York, contributing to neuro-oncology and neurofibromatosis clinical studies. His current work spans study design, basic science, data extraction and analysis, and manuscript preparation, and he intends to pursue a career in orthopaedic surgery.

[Connect with Shaan Lalvani on LinkedIn](#)

[Conflicts of Interest Statement for Shaan Lalvani](#)

c Sam Pourarbab is a first year medical student at Rush Medical College in Chicago, Illinois. He earned a Bachelor of Science in Neuroscience from the University of California, Riverside, where he conducted research focused on learning and memory. His prior experiences also include community service, health care volunteering, and mentorship programs serving underserved populations. He is currently involved in orthopedic surgery and otolaryngology research with a focus on clinical outcomes, sports medicine, and musculoskeletal disease.

[Connect with Samuel Pourarbab on LinkedIn](#)

[Conflicts of Interest Statement for Samuel Pourarbab](#)

d Lord J. Hyeamang is a 4th year medical student at Rush Medical College.

[Connect with Lord Hyeamang on LinkedIn](#)

[Conflicts of Interest Statement for Lord Hyeamang](#)

e Ryan Lew is an incoming PGY1 at Creighton University Orthopaedic Surgery Residency.

[Conflicts of Interest Statement for Ryan Lew](#)

f Dr. Hussain is an orthopedic surgeon specializing in shoulder and elbow surgery.

[Connect with Dr. Zaamin Hussain on LinkedIn](#)

[Conflicts of Interest Statement for Dr. Zaamin Hussain](#)

[Visit the Open Payments Data Page for Dr. Zaamin Hussain](#)

g Dr. Gregory Nicholson specializes in shoulder and elbow surgery, employing advanced arthroscopic and open surgical techniques to address sports-related, traumatic, arthritic, and occupational conditions. He graduated from Indiana University School of Medicine and completed his internship and orthopedic residency at University Hospital of Cleveland. Dr. Nicholson furthered his expertise with a fellowship in shoulder and elbow surgery at the New York Orthopaedic Hospital at Columbia-Presbyterian Medical Center.

Dr. Nicholson is actively involved in designing cutting-edge shoulder replacement systems and consults for various orthopedic companies, contributing to the development of instruments and implants for shoulder and elbow surgery. He serves as the principal investigator for funded research on rotator cuff repair augmentation and shoulder arthroplasty.

Board certified by the American Board of Orthopaedic Surgery, Dr. Nicholson is a Fellow of the American Academy of Orthopaedic Surgeons and a member of the American Shoulder and Elbow Surgeons, the American Orthopaedic Society for Sports Medicine, and the Association of Bone and Joint Surgeons.

Since 1993, Dr. Nicholson has provided care for professional baseball players, starting with the Indianapolis Indians AAA team and, from 2004 onward, as one of the team physicians for the Chicago White Sox (MLB). He also contributes to the care of the Chicago Bulls (NBA).

Dr. Nicholson has published extensively on shoulder and elbow conditions, authoring papers, book chapters, technique papers, and videos. He has edited a DVD series on shoulder instability and is currently editing the upcoming AAOS Orthopaedic Knowledge Update-4 for Shoulder and Elbow.

[Visit Dr. Gregory Nicholson](#)

[Conflicts of Interest Statement for Dr. Gregory Nicholson](#)

¹ Rush Medical College, ² Orthopaedic Surgery, Midwest Orthopaedics at Rush

Keywords: artificial intelligence, large language models, medical education, orthopaedic in-training examination

<https://doi.org/10.60118/001c.162882>

Journal of Orthopaedic Experience & Innovation

Vol. 7, Issue 2, 2026

Background

Large language models (LLMs) are increasingly used in medical education, yet head-to-head comparisons of contemporary multimodal and clinically oriented (retrieval-grounded) models on orthopedic in-service examinations remain limited.

Methods

A cross-sectional comparative study was conducted using 265 multiple-choice questions from the publicly available 2014 Orthopaedic In-Training Examination (OITE). Questions were categorized by AAOS subspecialty and by cognitive taxonomy (T1 recall, T2 interpretation, T3 reasoning) and stratified into text-only (n=111) versus image-based items (n=164). ChatGPT-5 Plus and Gemini 2.5 Pro were evaluated on the full dataset, while OpenEvidence was evaluated on the text-only subset. Accuracy was scored against the official answer key and comparisons were done using McNemar and Cochran Q tests.

Results

Overall accuracy on the full 2014 OITE was 80.0% for Gemini 2.5 Pro and 78.1% for ChatGPT-5, with no statistically significant difference (p=0.59). Across the text-only subset, Gemini 2.5 Pro and OpenEvidence each scored 84.7% versus 81.1% for ChatGPT-5 (p=0.71), and OpenEvidence demonstrated strong T3 reasoning performance (82.7%). Subspecialty performance varied without statistically significant differences. LLMs demonstrated highest accuracy in basic science and the largest numerical discrepancy was in Oncology (Gemini 77.3% vs ChatGPT-5 54.5%).

Conclusion

A retrieval-grounded, evidence based LLM (OpenEvidence) performed comparably to multimodal LLMs on text-based orthopaedic in-training examination questions at or above senior resident PGY-5 benchmark-level performance with no significant differences in overall accuracy. These findings suggest that OpenEvidence may provide educational utility comparable to general-purpose models while offering source transparency.

INTRODUCTION

Generative artificial intelligence (AI) based large language models (LLMs) have expanded rapidly in both capability and adoption since the public release of ChatGPT in late 2022 (Mesko 2023). In medicine, these models have demonstrated strong performance on complex reasoning tasks and

standardized examinations, suggesting a potential role in clinical care and medical education (Singhal et al. 2023).

Several studies have evaluated LLM performance on high-stakes examinations, including the United States Medical Licensing Examination (USMLE), demonstrating passing-level performance without specific fine-tuning (Brin et al. 2023; Bicknell et al. 2024; Kung et al. 2023). More recently, benchmarking efforts have expanded to spe-

[Visit the Open Payments Data Page for Dr. Gregory Nicholson](#)

^h Dr. Garrigues is a Shoulder and Elbow specialist at Midwest Orthopaedics at Rush and co-team physician for the Chicago White Sox. Dr. Garrigues has been honored with multiple awards, including the 2017 ASES European traveling fellowship, 2025 Neer Award, and grants from OREF and an NIH R01. He received his undergraduate and medical degrees from Harvard, a research degree from MIT, and was the Head of Shoulder Reconstruction at Duke before joining MOR in 2018.

[Visit Dr. Grant Garrigues](#)

[Connect with Dr. Grant Garrigues on LinkedIn](#)

[Conflicts of Interest Statement for Dr. Grant Garrigues](#)

[Visit the Open Payments Data Page for Dr. Grant Garrigues](#)

cialty board and in-service examinations, including orthopaedic surgery, to compare LLM performance with residence physicians and practicing clinicians (Zhang et al. 2024). The Orthopaedic In-Training Examination (OITE), administered annually by the American Academy of Orthopaedic Surgeons (AAOS), serves as the primary standardized assessment for orthopaedic residents in the United States and a predictor of performance on the American Board of Orthopaedic Surgery (ABOS) Part 1 certifying examination (Swanson et al. 2013). Accordingly, multiple groups have used OITE questions to evaluate LLMs, reporting that contemporary general-purpose models such as GPT-4 and Google Bard or Gemini achieve scores comparable to junior or mid-level orthopaedic residents, though raising questions about how these tools compare with more senior trainees and how such performance should be interpreted in the context of orthopaedic education (Mo et al. 2025).

Despite this growing literature, important methodological limitations persist. Many existing studies focus on a single model or do not fully characterize question types, subspecialty domains, or prompting techniques. Recent commentaries have emphasized the need for more rigorous comparative evaluations with transparent reporting of model versions, prompting strategies, and question selection (Kunze et al. 2025).

Specialized, clinically oriented AI tools have emerged to address limitations of general-purpose models, particularly regarding source transparency and hallucinations. OpenEvidence (OpenEvidence, Inc.) is one such platform; it is a clinician-facing LLM that grounds responses in cited, peer-reviewed medical sources, which may improve transparency and reduce hallucinations but does not eliminate them entirely (Hurt et al. 2025). Although OpenEvidence is increasingly used by physicians and trainees for point-of-care questions, it has not previously been evaluated on standardized orthopaedic examinations such as the OITE as part of a broader comparative analysis of LLMs.

Given the rapid evolution of LLMs, prior benchmark studies can quickly become outdated as new model versions are released (Yang et al. 2025). Rather than identifying a single ‘best’ model, comparing performance across different model architectures may help orthopaedic trainees and educators decide which tools appropriate for evidence-based studying and rotation-level clinical reasoning, and identify where supervision is most needed given known limitations (Xu et al. 2025). Therefore, the objective of this study was to evaluate and compare the performance of three large language models ChatGPT-5 (OpenAI), Gemini 2.5 Pro (Google DeepMind), and OpenEvidence (OpenEvidence, Inc.) on performance in the 2014 OITE across cognitive taxonomies and orthopaedic subspecialties.

METHODS

STUDY DESIGN

We conducted a cross-sectional comparative study to evaluate the performance of three large language models (LLMs),

ChatGPT-5 (OpenAI), Gemini 2.5 Pro (Google DeepMind), and OpenEvidence (OpenEvidence, Inc.) on a standardized orthopaedic exam between October and November 2025. The study utilized the 2014 Orthopaedic In-Training Examination (OITE) as the question source. As this study involved publicly available, de-identified educational materials and no human subjects, it was exempt from Institutional Review Board (IRB) oversight.

QUESTION SOURCE AND CLASSIFICATION

The complete 2014 OITE question set was obtained from publicly available archives. Questions were manually screened for inclusion. Items requiring video playback or those with corrupted data were excluded. A total of 265 multiple-choice questions were included in the final dataset. Each question was categorized by orthopaedic subspecialty according to the American Academy of Orthopaedic Surgeons (AAOS) domain structure: Trauma, Spine, Sports Medicine, Pediatrics, Adult Reconstruction (Hip/Knee), Hand/Upper Extremity, Foot and Ankle, Shoulder and Elbow, Oncology, and Basic Science. Additionally, questions were classified by cognitive taxonomy based on established medical education frameworks defined by Buckwalter et al: Taxonomy 1 (T1): Recall and recognition of facts, Taxonomy 2 (T2): Interpretation of data, images, or diagnostic studies, and Taxonomy 3 (T3): Complex clinical reasoning, problem-solving, and management decisions (Buckwalter et al. 1981).

Classifications were done by two reviewers S.P and S.H and disagreements were resolved by a third review, A.K. Questions were further stratified into text-only (n=111) and image-based (n=154) subsets to accommodate the different input capabilities of the AI models. Image-based questions included radiographs, MRI/CT scans, or clinical photographs essential for determining the correct answer.

AI MODELS AND INPUT PROTOCOL

Three models representing different AI architectures were prompted: ChatGPT-5 (Auto), Gemini 2.5 Pro, and OpenEvidence. As OpenEvidence currently supports text-only input, it was evaluated solely on the text-only question subset (n=111). Consequently, direct three-way comparison across all models is limited to this shared subset, while ChatGPT-5 and Gemini 2.5 Pro were additionally evaluated across the full 265-question dataset including image and video-based items.

A standardized prompt structure was used for all interactions to ensure consistency and to minimize variability (Figure 1). The prompt instructed each model to act as an orthopaedic surgery resident, to select the single best answer choice, and to provide a step-by-step explanation for its reasoning. For multimodal models (ChatGPT-5 and Gemini 2.5 Pro), image files were uploaded alongside the text stem for all relevant questions. The chat context was reset after every question to prevent information leakage or memory bias between items.

AI Performance on Orthopedic Examination Script:

Prompt for ChatGPT-5/Gemini 2.5 PRO/Open-Evidence:

You are an orthopedic surgery resident taking the Orthopedic In-Training Examination (OITE).

Read the following question carefully and choose the single best answer.

Please explain your reasoning and provide references to standard orthopedic literature if appropriate.

Example:



Figure 21

Question 21

A 2-year-old child with a FGF3R gene defect has the spinal deformity shown in Figure 21. What is the most likely outcome?

1. Worsen with time
2. No change with time
3. Resolution at adulthood
4. Resolution at adolescence
5. Resolution with walking ambulation

Figure 1. Schematic of the standardized prompting workflow used for AI model evaluation on the 2014 Orthopaedic In-Training Examination (OITE)

EVALUATION AND SCORING

Model outputs were recorded and scored against the official OITE answer key. A response was graded as “correct” only if the model explicitly selected the keyed option. Primary outcomes included overall accuracy (percentage of correct answers) for each model, and secondary outcomes included performance stratified by cognitive taxonomy (T1–T3), question type (text-only vs. image-based), and subspecialty domain.

STATISTICAL ANALYSIS

Statistical Analysis was performed using Stata (StataCorp LLC, College Station, TX). Descriptive statistics were calculated to determine accuracy rates across all categories, with 95% confidence intervals computed for overall performance. Comparisons of accuracy between models were conducted using McNemar’s test for paired proportions and Cochran’s Q test for comparisons across all three models on the text-only subset. Statistical significance was defined as $p < 0.05$. All graphical visualizations were generated using GraphPad Prism (GraphPad Software, San Diego, CA).

RESULTS

A total of 265 questions from the 2014 OITE were analyzed. Both multimodal models demonstrated performance exceeding the national mean score for PGY-5 residents (74%). Gemini 2.5 Pro achieved the highest overall accuracy at 80.0% (212/265), followed by ChatGPT-5 at 78.1% (207/265). The difference in overall accuracy between the two

multimodal models was not statistically significant ($p = 0.59$) (Table 1).

When stratified by taxonomy, no statistically significant performance differences were observed between the models across any domain. For factual recall (T1) questions, ChatGPT-5 achieved slightly higher accuracy (88.1%) compared to Gemini 2.5 Pro (84.7%; $p = 0.59$) and OpenEvidence (86.4%; $p = 0.87$). In contrast, Gemini 2.5 Pro showed numerically higher accuracy than ChatGPT-5 on questions requiring interpretation of images or data (T2) (76.3% vs. 68.4%; $p = 0.44$), though this data was not statistically significant. Similarly, for complex reasoning (T3) items, Gemini 2.5 Pro demonstrated a slight increase over ChatGPT-5 (79.2% vs. 76.8%; $p = 0.60$).

On the subset of 111 text-only questions where all three models could be directly compared, Gemini 2.5 Pro and OpenEvidence achieved identical accuracy (84.7%, 94/111), outperforming ChatGPT-5 (81.1%, 90/111). These differences were not statistically significant ($p = 0.71$) (Table 2). Notably, OpenEvidence performed comparably to the multimodal models despite lacking visual input capabilities, excelling particularly in text-based reasoning (T3) questions (82.7%).

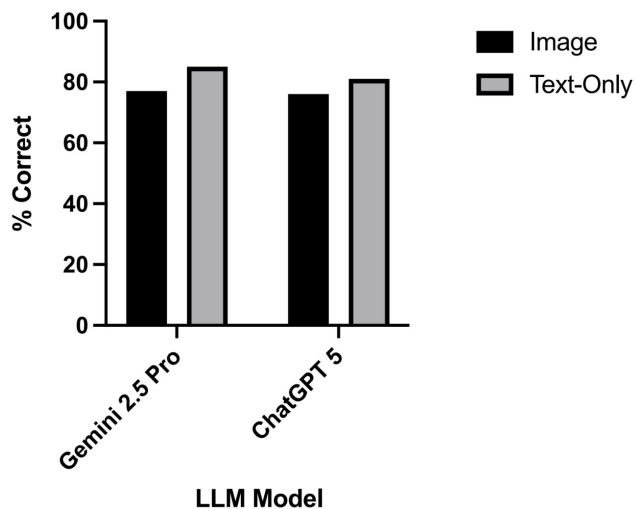
Performance varied by orthopaedic subspecialty but showed no statistically significant differences between models ($p > 0.05$ for all pairwise comparisons). Both models achieved their highest accuracy in Basic Science (>93%). The largest numerical discrepancy was observed in Oncology, where Gemini 2.5 Pro achieved higher accuracy than ChatGPT-5 by 22.7 percentage points (77.3% vs. 54.5%), though this did not reach statistical significance. Numerical differences were also observed in Foot and Ankle and Hip and Knee, but similarly should be interpreted with caution

Table 1. Overall Model Performance of LLM Models on the 2014 Orthopaedic In-Training Examination

Metric	ChatGPT-5	Gemini 2.5 Pro	OpenEvidence
Total Questions Attempted	265	265	111
Overall Accuracy, n (%)	207 (78.1%)	212 (80.0%)	94 (84.7%)
T1: Recall, n/N (%)	52/59 (88.1%)	50/59 (84.7%)	51/59 (86.4%)
T2: Interpretation, n/N (%)	26/38 (68.4%)	29/38 (76.3%)	N/A
T3: Reasoning, n/N (%)	129/168 (76.8%)	133/168 (79.2%)	43/52 (82.7%)

Table 2. Comparative Performance of AI Models on the Text-Only Question Subset

Metric	ChatGPT-5	Gemini 2.5 Pro	OpenEvidence	p-value*
Overall Accuracy, n (%)	90/111 (81.1%)	94/111 (84.7%)	94/111 (84.7%)	0.71
Taxonomy Performance				
T1: Recall, n/N (%)	52/59 (88.1%)	50/59 (84.7%)	51/59 (86.4%)	0.87
T3: Reasoning, n/N (%)	38/52 (73.1%)	44/52 (84.6%)	43/52 (82.7%)	0.29

**Figure 2. Image vs Text Only Question Performance between Large Language Models**

given the small question counts per domain and lack of statistical significance (Figure 2).

On text only questions, all three models demonstrated highest accuracy in Adult Spine, with OpenEvidence performing in the same range to the multimodal models (Figure 2). Open Evidence achieved a perfect score on the Foot and Ankle section (100%) compared to ChatGPT-5 and Gemini 2.5 Pro (82.5%).

All three models answered 72.1% (80/111) of questions correctly (Figure 4). Only 5.4% (6/111) of questions were answered incorrectly by all three models.

DISCUSSION

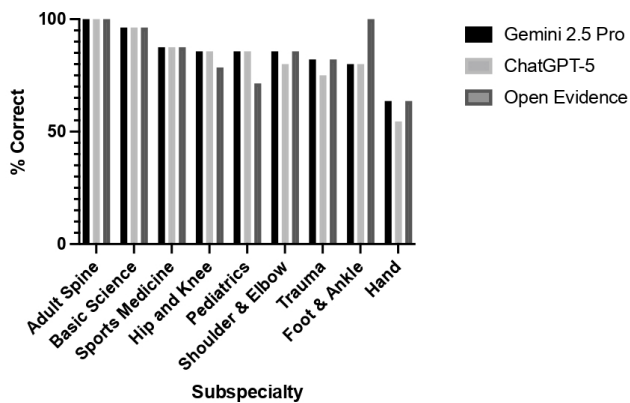
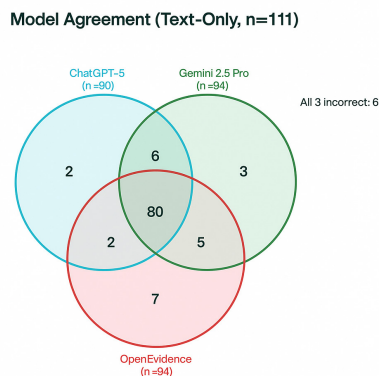
This study investigated overall accuracy between ChatGPT-5, Gemini 2.5 Pro, and OpenEvidence across a standardized orthopaedic examination, as well as variation

across cognitive taxonomies (factual recall, interpretation, clinical reasoning) and orthopaedic subspecialties. We additionally evaluated if these LLMs achieved performance at or above the PGY-5 resident benchmarks. In addressing these questions, we found that both ChatGPT-5 and Gemini 2.5 Pro exceeded contemporary PGY-5 national benchmarks from 2024 (73%) for orthopaedic residents in ACGME accredited programs, achieving overall accuracies of 78.1% and 80.0%, respectively, with no statistically significant difference between models ($p = 0.59$) (Table 1) (Orthopaedic In-Training Examination (OITE) Technical Report 2024 2024).

Because OpenEvidence does not support image interpretation, it was evaluated only on the 111 text-only questions. In this shared subset, where all three models could be directly compared, Gemini 2.5 Pro and OpenEvidence each achieved 84.7% accuracy, while ChatGPT-5 Plus slightly trailed both at 81.1%. The absolute differences were small (a 3.6 percentage point difference between ChatGPT-5 and the other two models), and they did not reach statistical significance ($p = 0.71$), indicating that no model consistently outperformed the others on purely text-based items (Table 2). Across the text-only questions, all three models agreed on the correct answer for 72.1% of items, and only 5.4% were missed by all models, demonstrating substantial overlap in their strengths (Figure 4). While this overlap likely reflects convergence on well-established clinical knowledge, it cannot be excluded that shared exposure to publicly available exam content during model training contributed to agreement on specific items, a possibility that underscores the need for contamination-resistant evaluation designs in future work. OpenEvidence also contributed the highest number of unique correct responses (7 total) (Figure 4). These results indicate that while overall accuracy was similar across models, differences in performance on specific subspecialties within orthopaedic surgery and cognitive tasks reflect variation in how each model handles distinct types of clinical knowledge.

Table 3. Comparative Performance of ChatGPT-5 and Gemini 2.5 Pro across orthopaedic subspecialties on the 2014 OITE

Subspecialty	ChatGPT-5 n/N (%)	Gemini 2.5 Pro n/N (%)
Basic Science	28/30 (93.3%)	29/30 (96.7%)
Pediatrics	31/35 (88.6%)	30/35 (85.7%)
Adult Spine	21/24 (87.5%)	20/24 (83.3%)
Sports Medicine	16/19 (84.2%)	15/19 (78.9%)
Shoulder and Elbow	19/23 (82.6%)	20/23 (87.0%)
Foot and Ankle	17/22 (77.3%)	15/22 (68.2%)
Trauma	36/49 (73.5%)	36/49 (73.5%)
Hip and Knee	18/25 (72.0%)	20/25 (80.0%)
Hand	9/16 (56.3%)	10/16 (62.5%)
Oncology	12/22 (54.5%)	17/22 (77.3%)

**Figure 3. Model Performance on Text-Only Questions across subspecialties****Figure 4. Venn Diagram illustrating model agreement on unique correct answers across models**

This study is the first to directly compare the performance of three contemporary frontier LLMs (ChatGPT 5 Plus, Gemini 2.5 Pro, and Open Evidence) on the OITE, and the first to evaluate a retrieval-augmented system (OpenEvidence) on this exam format. Prior work has examined older model generations or focused on general medical

licensing tests, but no previous study has assessed these newer architectures, head-to-head, on a specialty-specific orthopaedic assessment.

Our findings align with a quickly developing body of work showing that modern LLMs can achieve high performance on structured medical knowledge assessments. Early research demonstrated that GPT-3.5 could attain passing-level performance on licensing examinations such as the United States Medical Licensing Examination (USMLE) without specialty-specific training (Kung et al. 2023). Subsequent evaluations of GPT-4 and GPT-4 Omni documented substantially higher accuracy, frequently exceeding 88-90% correct responses, and, most notably, outperforming earlier LLMs on tasks requiring both factual recall and clinical reasoning (Bicknell et al. 2024; Chen et al. 2024).

Parallel advances have been reported for non-OpenAI systems. In specialty board examinations, Gemini 2.5 Pro has achieved ≥ 85 -90% accuracy on national pediatric surgery and obstetrics-gynecology exams, well above local passing thresholds, which were approximately 60% for each assessment (Wielochowska et al. 2025). More recent comparative evaluations of ChatGPT-5 and Gemini 2.5 on internal medicine subspecialty board examinations demonstrate higher overall accuracy for Gemini, while also showing that relative performance varies across subspecialties rather than remaining consistent across domains (Sheikhalishahi et al. 2025). This data suggests that general-purpose LLMs from different model families have converged to a similarly high performance level on structured, multiple-choice assessments, with variability concentrated more in subspecialty-specific reasoning style than in raw score.

Despite these high accuracies reported across prior studies, LLM performance remains imperfect, particularly for questions that require nuanced clinical judgment, integration of multiple pieces of information, or differentiation between closely related management options. These limitations likely reflect the probabilistic nature of current models and their reliance on pattern recognition rather than true causal clinical reasoning. Although ongoing model development may continue to improve performance, consistently perfect accuracy on complex clinical assessments is

unlikely in the near term. This highlights the role of LLMs as adjuncts rather than substitutes in educational and clinical decision making.

Retrieval-augmented systems such as OpenEvidence represent a related but distinct trajectory in LLM design. Recent work has shown that these LLMs, augmented with targeted retrievals and agentic workflows, can provide more evidence-grounded and therefore clinically actionable responses to real-world clinical questions than base models can alone (Low et al. 2025). In medical education settings, OpenEvidence and similar platforms have been reported to assist students integrate guideline-based reasoning and efficiently locate relevant literature during clinical rotations (Patel et al. 2025). OpenEvidence's strong performance in our text-only OITE questions is consistent with these findings and suggests that retrieval-based reasoning can perform as well as or better than general multimodal models on exam-style clinical questions, particularly when questions rely on standards of care (Table 2). A qualitative review of model-generated explanations for incorrect responses revealed potentially meaningful differences in error character across models. ChatGPT-5 incorrect responses most commonly reflected confident factual errors, in which the clinical reasoning framework applied was appropriate, but the specific conclusion was incorrect. In several cases, the model appeared to apply more contemporary clinical guidelines that diverge from the 2014 examination key, which may itself reflect the model's training on more recent literature. Gemini 2.5 Pro's errors more frequently reflected misinterpretation of the question stem or application of overly aggressive clinical judgement despite demonstrating sound underlying knowledge. In multiple instances, it explicitly acknowledged alternative correct approaches within its own explanation. OpenEvidence's incorrect responses showed a qualitatively different pattern: explanations were either absent or consisted of brief, citation-grounded statements that described correct clinical principles yet arrived at the wrong answer choice, suggesting errors more consistent with answer-selection limitations than reasoning or knowledge failures. This pattern is consistent with OpenEvidence contributing the highest number of unique correct responses among the three models ($n=7$, Figure 4), suggesting that when it diverged from the other models it did so on the basis of evidence-grounded reasoning rather than pattern-matched responses. While this analysis was qualitative and exploratory rather than formal, these differences in error character may have practical implications for clinical use, as the nature and transparency of model errors may be as important as aggregate accuracy when clinicians rely on these tools for point-of-care decisions.

Subspecialty analysis further highlighted meaningful differences among models. Within the text-only subset, OpenEvidence achieved 100% accuracy on Foot and Ankle questions, whereas ChatGPT-5 Plus scored 77.3% followed by Gemini 2.5 Pro scored at 68.2% here (Table 3, Figure 3). While these differences did not reach statistical significance, this pattern may warrant further investigation into whether retrieval-based systems offer advantages in sub-

specialties with well-established diagnostic and treatment pathways.

Differences also emerged across cognitive taxonomies. ChatGPT-5 Plus performed best on Taxonomy 1 (recall-oriented) items, aligning with its broad exposure to factual medical content and its strength in direct knowledge retrieval. Gemini 2.5 Pro fared relatively better on Taxonomy 3 (clinical reasoning) questions, consistent with reports that newer Gemini models often excel on complex and integrative problem solving tasks (Wielochowska et al. 2025; Sheikhalishahi et al. 2025; Boczkowski et al. 2025). OpenEvidence's Taxonomy 3 performance was comparable to that of Gemini 2.5 Pro despite its lack of multimodal input, suggesting that explicit evidence retrieval can support reasoning on clinical scenarios involving multiple steps (Table 1 and 2). Regarding image-based performance, Gemini-2.5 Pro demonstrated numerically higher accuracy than ChatGPT-5 on T2 interpretation questions, which include radiographs, MRI, CT, and clinical photograph (76.3% vs 68.4%; $p = 0.44$). While this difference did not reach statistical significance, it suggests potential variation in visual reasoning capability between model architectures. Given that image-based questions comprise the majority of OITE items, the ability of multimodal LLMs to accurately process medical imaging represents an important and underexplored dimension of their utility for orthopaedic education (Figure 2).

Our findings and those of existing work suggest that contemporary LLMs can match or exceed published PGY-5 benchmarks on structured orthopaedic assessments. However, examination accuracy reflects a narrow dimension of medical knowledge and should not be equated with the clinical reasoning or real-world competency of senior trainees. Furthermore, these results should not be interpreted as evidence that LLMs can replace traditional study methods or expert instruction, as model outputs remain imperfect and require verification against established resources (Kim et al. 2025; Bedi et al. 2025). Rather, their strong performance on standardized examinations suggests that trainees will increasingly encounter accurate automated responses to factual and structures clinical questions. From an educational and assessment perspective, these findings also have implications for exam development. As LLMs continue to perform well on lower-order knowledge and recall-based items (Schubert et al. 2023), greater emphasis on higher-order reasoning, synthesis of clinical information, and judgement-based decision making may better reflect the skills in which orthopaedic surgeons provide the greatest added value and are least replicated by automated systems.

From a practical standpoint, these findings offer several considerations for orthopaedic residents and training programs navigating the increasing availability of AI tools. The comparable performance of OpenEvidence and general-purpose models on text-based questions, combined with its qualitatively distinct error profile, suggests that retrieval-grounded platforms may offer a more transparent and verifiable starting point for evidence-based self-study – particularly for subspecialties with well-established clin-

ical guidelines where sources responses can be thoroughly cross-checked against primary literature. ChatGPT-5 and Gemini 2.5 Pro may offer broader utility across question types including image and video-based items, but their tendency toward confident errors warrants active verification of model outputs rather than passive acceptance. For program directors, the pattern of subspecialty and taxonomy-level variation observed across models suggests that LLM performance is uneven in ways that are not captured by aggregate accuracy alone, and that residents should be coached to use these tools critically rather than as authoritative sources. These findings do not support replacing structured study resources or expert instruction with LLMs, but they do suggest that understanding how different model architectures fail – not just how often – may be important as raw accuracy in guiding appropriate clinical and educational use.

Although this study showed high accuracy in these models, these results should still be interpreted in the context of limitations. The study we conducted was limited to a single OITE exam year (2014), which meaningfully limits generalizability. Orthopaedic surgery has evolved considerably since 2014 across clinical guidelines, implant technology, surgical techniques, and evidence-based management across subspecialties. Questions anchored to 2014 standards of care may no longer reflect current accepted practice. As a result, a model answering correctly by 2014 standards could theoretically provide outdated guidance by contemporary standards, and performance on this exam may not predict accuracy on more recent OITE administrations. Additionally, prior exposure of LLMs to the 2014 OITE during model training cannot be excluded. As a result, the accuracy estimates reported here may reflect, at least in part, familiarity with previously encountered content rather than generalizable clinical reasoning ability. OpenEvidence also does not allow for image interpretation, which only allows us to compare all three models using text only questions. Consequently, we could not directly compare all three platforms on the full exam, including image-based items. Subspecialty analyses were further limited by small question counts per domain, ranging from 16 questions in Hand to 49 in Trauma, which reduces statistical power and limits the reliability of domain-level comparisons. Findings from smaller subspecialties such as Oncology (n=22) and Hand (n=16) in particular should be interpreted with caution. In addition, resident benchmarks were taken from prior literature, and newer model versions such as ChatGPT 5.2 and Gemini 3.0 were not used in our study. We also used the paid versions of these models, so performance may differ for users who utilize the free versions,

which may rely on older or smaller models. These factors may influence external validity and limit conclusions made about performance on newer exams, updated model releases, or current clinical standards.

This study should allow future research to build on these findings to address our limitations and improve generalizability. Repeating this analysis across multiple OITE exam years would strengthen generalizability and allow assessment of whether model performance holds across evolving clinical content. Incorporating more recent examinations would also reduce the risk of data contamination given that older publicly available exams are more likely to have been indexed in model training datasets. Similarly, evaluating updated model releases would clarify if there are any differences in reasoning and accuracy across versions and over time. Comparing the free and paid versions of these models would provide insight on the performance and accuracy for users who may not have access to the paid versions. Future work should also compare models in a controlled setting that allows access to external sources versus when only relying on the model's internal database. To support fair comparison, this design should use the same directions across platforms and include sources used in the answers. Platforms that allow for multimodal interpretation should be used to evaluate image-based questions. This study should be conducted by either limiting comparisons to models capable of interpreting images or by reporting text only and image-based results separately. All these steps would allow for a more updated and clinically relevant depiction of the accuracy of these models and their usefulness to orthopaedic residents.

CONCLUSION

In this study, ChatGPT-5, Gemini 2.5 Pro, and OpenEvidence demonstrated high accuracy on the 2014 OITE, with performance at or above published senior resident benchmarks. While overall accuracy did not differ significantly between models, variation across subspecialties and cognitive taxonomies suggests meaningful differences in how these systems retrieve and apply orthopaedic knowledge. These findings support the use of LLMs as adjuncts in orthopaedic education and assessment, while reinforcing the continued significance of higher-order reasoning and clinical judgment that remain less reliably mastered by automated systems.

Submitted: February 03, 2026 EDT. Accepted: June 04, 2026 EDT. Published: September 26, 2026 EDT.



This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CCBY-NC-ND-4.0). View this license's legal deed at <https://creativecommons.org/licenses/by-nc-nd/4.0> and legal code at <https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode> for more information.

REFERENCES

- Bedi, S., Y. Jiang, P. Chung, S. Koyejo, and N. Shah. 2025. "Fidelity of Medical Reasoning in Large Language Models." *JAMA Netw Open* 8 (8): e2526021. <https://doi.org/10.1001/jamanetworkopen.2025.26021>.
- Bicknell, B. T., D. Butler, S. Whalen, J. Ricks, C. J. Dixon, A. B. Clark, et al. 2024. "ChatGPT-4 Omni Performance in USMLE Disciplines and Clinical Skills: Comparative Analysis." *JMIR Med Educ* 10 (November): e63430–e63430. <https://doi.org/10.2196/63430>.
- Boczkowski, D., T. Dolata, D. Radej, P. Sawina, R. Suleiman, A. Latkowska, et al. 2025. "Assessment of the Efficacy of the Google Gemini 2.5 Pro Model in Solving the Polish State Specialization Exam in Pediatric Surgery." *Cureus*, October 28. <https://doi.org/10.7759/cureus.95581>.
- Brin, D., V. Sorin, A. Vaid, A. Soroush, B. S. Glicksberg, A. W. Charney, et al. 2023. "Comparing ChatGPT and GPT-4 Performance in USMLE Soft Skill Assessments." *Sci Rep* 13 (1): 16492. <https://doi.org/10.1038/s41598-023-43436-9>.
- Buckwalter, J. A., R. Schumacher, J. P. Albright, and R. R. Cooper. 1981. "Use of an Educational Taxonomy for Evaluation of Cognitive Performance." *Academic Medicine* 56 (2): 115–21. <https://doi.org/10.1097/00001888-198102000-00006>.
- Chen, Y., X. Huang, F. Yang, H. Lin, H. Lin, Z. Zheng, et al. 2024. "Performance of ChatGPT and Bard on the Medical Licensing Examinations Varies across Different Cultures: A Comparison Study." *BMC Med Educ* 24 (1): 1372. <https://doi.org/10.1186/s12909-024-06309-x>.
- Hurt, R. T., C. R. Stephenson, E. A. Gilman, C. A. Aakre, I. T. Croghan, M. S. Mundi, et al. 2025. "The Use of an Artificial Intelligence Platform OpenEvidence to Augment Clinical Decision-Making for Primary Care Physicians." *J Prim Care Community Health* 16 (April): 21501319251332215. <https://doi.org/10.1177/21501319251332215>.
- Kim, J., A. Podlasek, K. Shidara, F. Liu, A. Alaa, and D. Bernardo. 2025. "Limitations of Large Language Models in Clinical Problem-Solving Arising from Inflexible Reasoning." *Scientific Reports* 15 (1): 39426. <https://doi.org/10.1038/s41598-025-22940-0>.
- Kung, T. H., M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, et al. 2023. "Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models." *PLOS Digit Health* 2 (2): e0000198. <https://doi.org/10.1371/journal.pdig.0000198>.
- Kunze, K. N., C. Gerhold, U. Dave, N. Abunnur, A. Mamonov, B. U. Nwachukwu, et al. 2025. "Large Language Model Use Cases in Health Care Research Are Redundant and Often Lack Appropriate Methodological Conduct: A Scoping Review and Call for Improved Practices." *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 41 (11): 4928–4945.e2. <https://doi.org/10.1016/j.arthro.2025.03.066>.
- Low, Y. S., M. L. Jackson, R. J. Hyde, R. E. Brown, N. M. Sanghavi, J. D. Baldwin, et al. 2025. "Answering Real-World Clinical Questions Using Large Language Model, Retrieval-Augmented Generation, and Agentic Systems." *DIGITAL HEALTH* 11 (May): 20552076251348850. <https://doi.org/10.1177/20552076251348850>.
- Mesko, B. 2023. "The ChatGPT (Generative Artificial Intelligence) Revolution Has Made Artificial Intelligence Approachable for Medical Professionals." *J Med Internet Res* 25 (June): e48392. <https://doi.org/10.2196/48392>.
- Mo, K., R. Lin, E. Dunn, G. Girgis, W. Fang, J. Walsh, et al. 2025. "Systematic Review on Large Language Models in Orthopaedic Surgery." *JCM* 14 (16): 5876. <https://doi.org/10.3390/jcm14165876>.
- Orthopaedic In-Training Examination (OITE) Technical Report 2024. 2024. Internet. American Academy of Orthopaedic Surgeons. <https://www.aaos.org/globalassets/education/product-pages/oite/oite-2024-technical-report-eds.pdf>.
- Patel, N., H. Grewal, V. Buddhavarapu, and G. Dhillon. 2025. "OpenEvidence: Enhancing Medical Student Clinical Rotations With AI but With Limitations." *Cureus*, January 3. <https://doi.org/10.7759/cureus.76867>.
- Schubert, M. C., W. Wick, and V. Venkataramani. 2023. "Performance of Large Language Models on a Neurology Board-Style Examination." *JAMA Netw Open* 6 (12): e2346721. <https://doi.org/10.1001/jamanetworkopen.2023.46721>.
- Sheikhalishahi, S., A. Haddadi, S. Sadeghipour, F. Rafiei, and H. Soltani. 2025. "Comparative Performance of ChatGPT-4o, ChatGPT-5, and Gemini 2.5 Flash on Persian Internal Medicine Subspecialty Board Exams." *Sci Rep*, December 3. <https://doi.org/10.1038/s41598-025-31251-3>.
- Singhal, K., S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, et al. 2023. "Large Language Models Encode Clinical Knowledge." *Nature* 620 (7972): 172–80. <https://doi.org/10.1038/s41586-023-06291-2>.
- Swanson, D., J. L. Marsh, S. Hurwitz, G. P. DeRosa, K. Holtzman, S. D. Bucak, et al. 2013. "Utility of AAOS OITE Scores in Predicting ABOS Part I Outcomes: AAOS Exhibit Selection." *The Journal of Bone and Joint Surgery-American Volume* 95 (12). <https://doi.org/10.2106/JBJS.L.00457>.
- Wielochowska, A., A. Stachowicz, M. Olender, A. Latkowska, J. Glinska, D. Boczkowski, et al. 2025. "The Effectiveness of the Multimodal Language Model, Google Gemini 2.5 Pro, in Solving the Specialization Exam in Gynecology and Obstetrics." *Cureus*, October 30. <https://doi.org/10.7759/cureus.95724>.

Xu, A. Y., M. Singh, M. Balmaceno-Criss, A. Oh, D. Leigh, M. Daher, et al. 2025. "Comparative Performance of Artificial Intelligence-Based Large Language Models on the Orthopedic in-Training Examination." *J Orthop Surg (Hong Kong)* 33 (1): 10225536241268789. <https://doi.org/10.1177/10225536241268789>.

Yang, A. J., J. J. Woo, and P. N. Ramkumar. 2025. "Editorial Commentary: Shifting From Redundancy to Rigor in Orthopaedic Large Language Model Research." *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 41 (11): 4946–49. <https://doi.org/10.1016/j.arthro.2025.06.020>.

Zhang, C., S. Liu, X. Zhou, S. Zhou, Y. Tian, S. Wang, et al. 2024. "Examining the Role of Large Language Models in Orthopedics: Systematic Review." *J Med Internet Res* 26 (November): e59607. <https://doi.org/10.2196/59607>.