Journal of Orthopaedic Experience & Innovation
Clinical, Economic, and Innovative Orthopaedic Experience

Nawari, Ashraf, Jamal Zahir, Sonal Kumar, Lovingly Ocampo, Olivia Opara, Hassan Ahmad, Benjamin Crawford, and Brian Feeley. 2025. "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?" *Journal of Orthopaedic Experience & Innovation* 6 (1). https://doi.org/10.60118/001c.124070.

Research Article

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?

Ashraf Nawari[1a], Jamal Zahir[2b], Sonal Kumar[2c], Lovingly Ocampo, DO[3d], Olivia Opara[2e], Hassan Ahmad[2f], Benjamin Crawford[4g], Brian Feeley, MD[1h]

---

a Ashraf's interests in orthopaedics are in global health, health equity, sports, and technology.

Conflicts of Interest Statement for Ashraf Nawari

b Jamal Zahir is a 4th-year medical student and orthopedic surgery research fellow based in the United States. Originally from Denver, Colorado, he earned his Bachelor's Degree with Honors from the University of Colorado - Denver. He is currently enrolled at Ross University School of Medicine and holds membership in the esteemed Rho Mu Delta Medical Honor Society. Jamal's interests extend beyond medicine, as he enjoys spending quality time with family and friends, practicing martial arts, exploring videography, trying out new restaurants, and engaging in clinical research. Additionally, he is the founder and CEO of Mile High MD, a platform designed to provide coaching for medical professionals and facilitate clinical research. Under his leadership, the Mile High MD team includes more than a dozen medical students, residents, and attending physicians from around the world.

Visit Jamal Zahir

Conflicts of Interest Statement for Jamal Zahir

c Sonal Kumar is a fourth year medical student.

Conflicts of Interest Statement for Sonal Kumar

d Lovingly Ocampo is a dedicated fourth-year medical student at Touro University of Osteopathic Medicine with a passion for educational research and global health, particularly within orthopedic surgery. A strong advocate for the advancement of women, minorities, and osteopaths in competitive surgical specialties, Ocampo firmly believes that mentorship and increased access to the field are vital steps toward greater inclusivity and diversity in the profession.

Connect with Dr. Lovingly Ocampo on LinkedIn

Conflicts of Interest Statement for Dr. Lovingly Ocampo

e Olivia Opara is a medical student with interest in orthopedic surgery and has three years of experience in research.

Conflicts of Interest Statement for Olivia Opara

f Hassan Ahmad is a 4th Year Medical Student.

Conflicts of Interest Statement for Hassan Ahmad

g Benjamin Crawford is from the small town of Piqua, Ohio, and is training as an orthopaedic surgery resident at University of California San Francisco St. Mary's Hospital. He is currently applying for a fellowship in orthopaedic trauma. In his free time, he enjoys spending time with his German Shepherd, Mya, and listening to music.

Visit Benjamin Crawford

Connect with Benjamin Crawford on LinkedIn

Conflicts of Interest Statement for Benjamin Crawford

h Dr. Brian Feeley is an orthopedic surgeon who specializes in using arthroscopic (minimally invasive) procedures to care for patients with athletic injuries of the shoulder or knee. In the shoulder, he treats rotator cuff tears or impingement, labral tears and other causes of instability, clavicle fractures and shoulder arthritis, and he performs shoulder replacement surgery. In the knee, he treats anterior cruciate and other ligament injuries, meniscus tears, cartilage injuries and early arthritis. He is section chief of the UCSF Division of Sports Medicine and Shoulder Surgery and director of the Muscle Stem Cell Lab. Feeley's research focuses on common knee and shoulder problems, including causes of muscle atrophy after rotator cuff tears. He also investigates how stem cells within muscle change the muscle tissue and how to stimulate these cells to improve function after injury. For the knee, he studies how to use low-cost motion analysis to decrease injury or re-injury after initial damage and surgical procedures. After earning a bachelor's degree in biology and his medical degree at Stanford University, Feeley completed a residency in orthopedic surgery at the Ronald Reagan UCLA Medical Center. He then completed a fellowship in sports medicine and shoulder surgery at the Hospital for Special Surgery, where he served as an assistant team physician to the New York Giants football team. He has been at UCSF since 2008. Feeley has published more than 230 peer-reviewed articles, review studies and book chapters, as well as a book on rotator cuff injuries. He is associate director of UCSF's residency program in orthopedic

Visit Dr. Brian Feeley

Connect with Dr. Brian Feeley on LinkedIn

Conflicts of Interest Statement for Dr. Brian Feeley

Visit the Open Payments Data Page for Dr. Brian Feeley

[1] Department of Orthopaedic Surgery, University of California, San Francisco, [2] Department of Surgery, Ross University School of Medicine, [3] Department of Surgery, Touro College, [4] Department of Orthopaedic Surgery, St Mary's Medical Center San Francisco Orthopaedic Residency Program

Keywords: natural language processing, artificial intelligence, chatgpt, orthopaedic in-training exam, orthopaedic surgery, large language models

https://doi.org/10.60118/001c.124070

Journal of Orthopaedic Experience & Innovation

Vol. 6, Issue 1, 2025

Background

The rapid improvement of generative artificial intelligence (AI) models in medical domains including answering board-style questions warrants further investigation regarding their utility and accuracy in answering orthopaedic surgery written board questions. Previous studies have analyzed the performance of ChatGPT alone on board exams, but a head-to-head analysis of multiple current AI models has yet to be performed. Hence, the objective of this study was to compare the utility and accuracy of various large language models (LLMs) in answering Orthopaedic Surgery In-Training Exam (OITE) written board questions to each other as well as orthopaedic surgery residents.

Methods

A complete set of questions from the OITE 2022 exam was inputted into various LLMs and results were calculated and compared against orthopaedic surgery residents nationally. Results were analyzed by overall performance and question type. Type A questions related to knowledge and recall of facts, Type B questions involved diagnosis and analysis of information, and Type C questions focused on the evaluation and management of diseases, requiring knowledge and reasoning to develop treatment plans.

Results

Google Gemini was the most accurate tool answering 69.9% of questions correctly. Google Gemini also performed superiorly to ChatGPT and Claude on Type A (76.9%) and Type C questions (67.4%), with Claude performing superiorly on Type B questions (70.7%). Questions without images were answered with greater accuracy compared to those with images (65.9% vs. 34.1%). All LLMs performed above the average of a first-year orthopaedic surgery intern, with Google Gemini and Claude performance approaching that of fourth- and fifth-year orthopaedic surgery residents.

Conclusion

The study assessed LLMs like Google Gemini, ChatGPT, and Claude against orthopaedic surgery residents on the OITE. Results showed that these LLMs perform on par with orthopaedic surgery residents, with Google Gemini achieving best performance overall and in Type A and C questions while Claude performed best in Type B questions. LLMs have the potential to be used to generate formative feedback and interactive case studies for orthopaedic trainees.

INTRODUCTION

Large language models (LLMs) such as ChatGPT, Google Gemini, and Claude are learning models designed to process and produce natural language in response to user input (Shen, Heacock, Elias, et al. 2023). LLMs are built on the transformer architecture, a neural network-based architecture that uses a self-attention mechanism to extract relevant context and produce compelling responses to human inputs, even in tasks for which it was not specifically trained (Li et al., n.d.). While neural networks have been discussed for many decades and the transformer was invented in 2017 (Li et al., n.d.), the LLM reached public prominence with the release of ChatGPT in 2022 (Grudin 2023). In recent years, there has been an increase in the development of LLMs which has led to a surge in the investigation of their utility in various fields. Of note, the potential role of LLMs in medicine is an avenue of active exploration (Park, Pillai, Deng, et al. 2024; Shah, Entwistle, and Pfeffer 2023; Clusmann, Kolbinger, Muti, et al. 2023; De Angelis, Baglivo, Arzilli, et al. 2023).

The most extensively researched LLM in medicine is Chat Generative Pre-trained Transformer, widely known as ChatGPT. The standard ChatGPT version utilizes GPT 3.5 which contains over 550 gigabytes of information from articles, websites, and journals to generate responses to human input (Shen, Heacock, Elias, et al. 2023). ChatGPT 4 is

an updated version, released in March 2023, and has been trained to allow visual and audio user inputs as well as text. While studies of ChatGPT 4's performance are ongoing, ChatGPT 3.5 has recently been shown to pass Step 1 of the United States Medical Licensing Exam and has been studied for performance in ophthalmology, dermatology, and radiology board-style examination questions (Gilson, Safranek, Huang, et al. 2023; Joly-Chevrier et al. 2023; Antaki et al., n.d.; "Performance of ChatGPT on a Radiology Board-Style Examination: Insights into Current Strengths and Limitations," n.d.). Previous studies of ChatGPT 3.5 and 4 performance on Orthopaedic Surgery In-Training Exam (OITE) written board questions have shown that ChatGPT 3.5 is capable of performing at the level of a PGY-1 resident, and ChatGPT 4 approximately between the level of a PGY-2 and PGY-3 resident (Jain et al. 2024; Hofmann, Guerra, Le, et al. 2024).

Importantly, there are currently no studies comparing the performance of other current LLMs on OITE questions to each other and orthopedic trainees. With the rapidly developing world of artificial intelligence, this provides a good opportunity to determine which LLM is most accurate on OITE performance and therefore most likely to aid in future orthopedic training. Our study aims to compare the performance of Chat GPT 3.5, Chat GPT 4, Google Gemini, and Claude on OITE questions from the 2022 OITE Exam against each other and orthopedic residents.

We hypothesize that the newer LLMs such as Open AI'S ChatGPT 4, Google's updated Gemini (formerly BARD) as well as Anthropic's Claude, which have image recognition features, will outperform ChatGPT 3.5 in a head-to-head analysis. Additionally, we postulate that these newer LLMs with image recognition will reach the accuracy of PGY-4 orthopaedic surgery residents and build on the previous success of LLMs up to the PGY-3 level. This raises philosophical questions as to what intelligence is and the utility of artificial intelligence in not only answering written board-style questions, but their utility in diagnosing, treating, and prognosticating patients with orthopaedic pathologies in the near future. If the trajectory of artificial intelligence continues, year-by-year we will continue to see improvements in its ability to answer orthopaedic questions, eventually likely reaching or exceeding the level of fellowship-trained orthopaedic surgeons in the near future.

MATERIALS AND METHODS

A complete set of questions from the OITE 2022 exam was inputted into four different LLMs: ChatGPT 3.5, ChatGPT 4, Google Gemini, and Claude. A simple prompt explaining the goal of the questioning was inputted prior to asking the question, for each question. The prompt was as follows: "A question from the Orthopaedic In-Training Exam will be provided to you along with 4 answer choices. Select only one correct answer along with an explanation & source of the explanation." The chat was reset each time before entering the prompt again for a new question.

The answers from each LLM were recorded, tabulated, and the percentage correct was calculated and compared against postgraduate year one through five (PGY-1 to PGY-5) orthopaedic surgery residents from ACGME-accredited residency programs. Three different types of questions were asked of four AI tools. Type A questions related to knowledge and recall of facts, Type B questions involved diagnosis and analysis of information, and Type C questions focused on the evaluation and management of diseases, requiring knowledge and reasoning to develop treatment plans. An example of a Type B question with an image and response from Claude is shown in Figures 1 and 2.

This example question was selected because imaging other than plain film x-rays seem to be harder for LLMs and trainees to interpret at a general level and in this case, Claude was able to accurately interpret the information, give the correct answer, and cite a verified source.

Data regarding resident performance nationally was obtained from the American Academy of Orthopaedic Surgeons OITE 2022 technical report. The OITE written board questions utilized in this study are copyrighted by the American Academy of Orthopaedic Surgeons (AAOS) ResStudy tool and are not publicly accessible.

STATISTICAL SUMMARY METHOD

The collected data were summarized using SPSS (version 26) through the following steps:

Data Entry: Different items were input into SPSS, including variables for Question ID, AI Tool, presence of an image in the question, the type of question, and whether the answer was correct or incorrect.

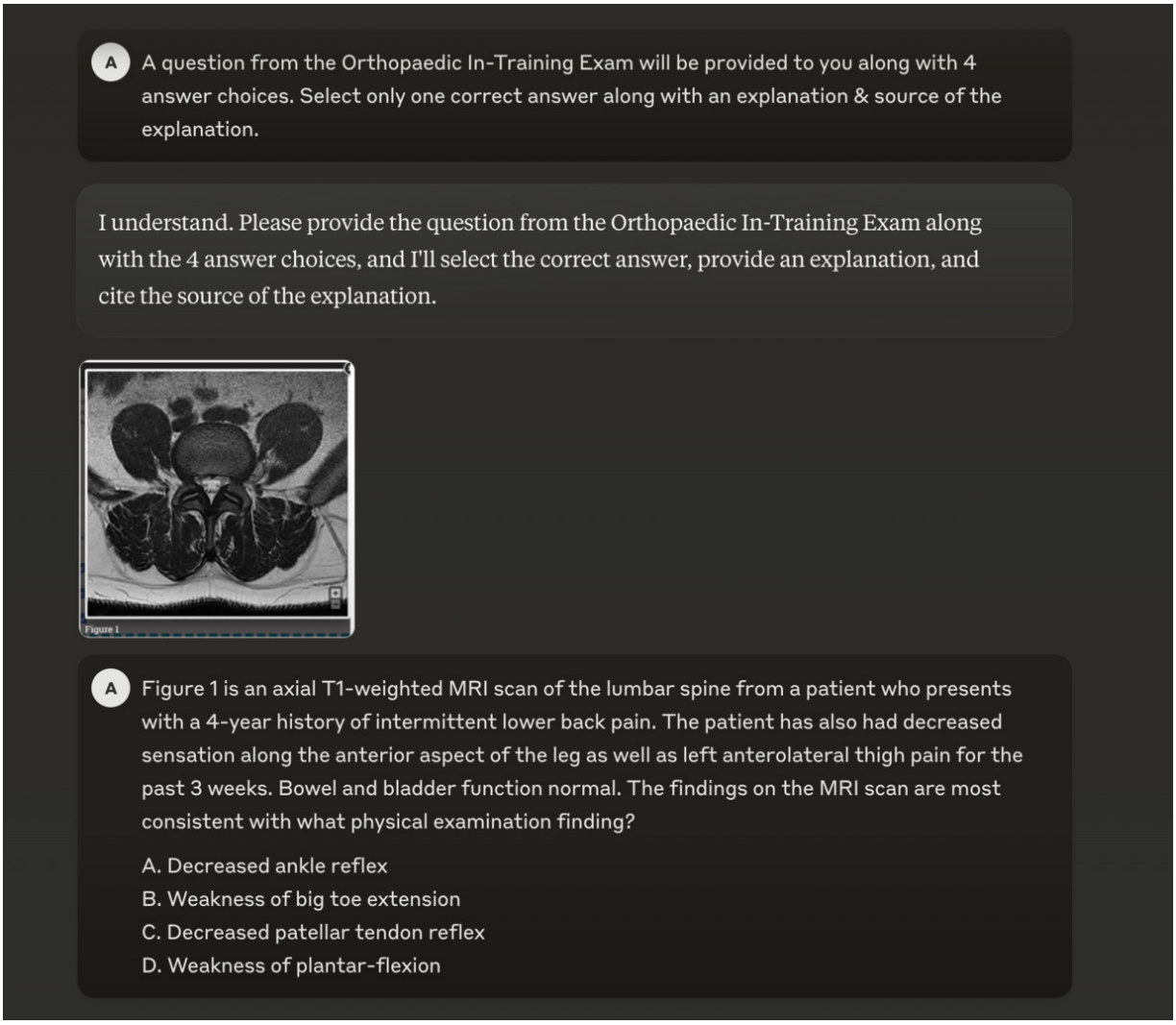

Figure 1.

Descriptive Statistics: Percentages of correct answers for each AI tool were calculated to summarize performance.

RESULTS

The breakdown of the question domain and percentages is reported in Table 1.

A total of 206 questions were asked of the AI tools: 65 (31.6%) were Type A questions, 42 (20.4%) were Type B questions, and 92 (44.7%) were Type C questions. The type for 7 (3.4%) questions was not determined (Table *2*). Table *3* represents the average scores of orthopaedic surgery residents nationally.

Gemini was the most accurate AI tool, correctly answering 144 questions (69.9%). Claude followed closely with 140 correct answers (68%), while ChatGPT 4 correctly answered 129 questions (62.6%). ChatGPT 3.5 correctly answered 120 questions (58.3%) (Table *4*).

Of all questions, 80 (38.8%) questions included images, while the remaining 126 (61.2%) did not. The AI tools showed a higher rate of correct answers for questions without images compared to those with images (65.9% vs. 34.1%) (Table *5*).

Statistical summary indicated that the difference was primarily due to ChatGPT 3.5, which had the weakest performance. In terms of questions with images, Google Gemini was the most accurate AI tool with 60% correct answers, followed by Claude, ChatGPT 4 (each with 57.5%), and ChatGPT 3 (52.5%) (Table *6*).

Gemini also performed the best on questions without images, with 76.2% correct answers, followed by Claude (75.2%), ChatGPT 4 (65.9%), and ChatGPT 3.5 (61.9%) (Table *7*).

Type A Questions: Gemini was the most accurate AI tool with 76.9% correct answers, followed by Claude (69.2%). ChatGPT 4 and ChatGPT 3.5 correctly answered 60.0% and 56.9% of questions, respectively (Table *8*).

Type B Questions: Claude was the most accurate AI tool with 70.7% correct answers, followed by ChatGPT 3.5 (69.0%). Gemini and ChatGPT 4 correctly answered 61.9% and 54.8% of questions, respectively (Table *8*).

Type C Questions: Gemini was the most accurate AI tool with 67.4% correct answers, followed by ChatGPT 4 (66.3%). Claude and ChatGPT 3.5 correctly answered 65.2% and 52.2% of questions, respectively (Table *8*).

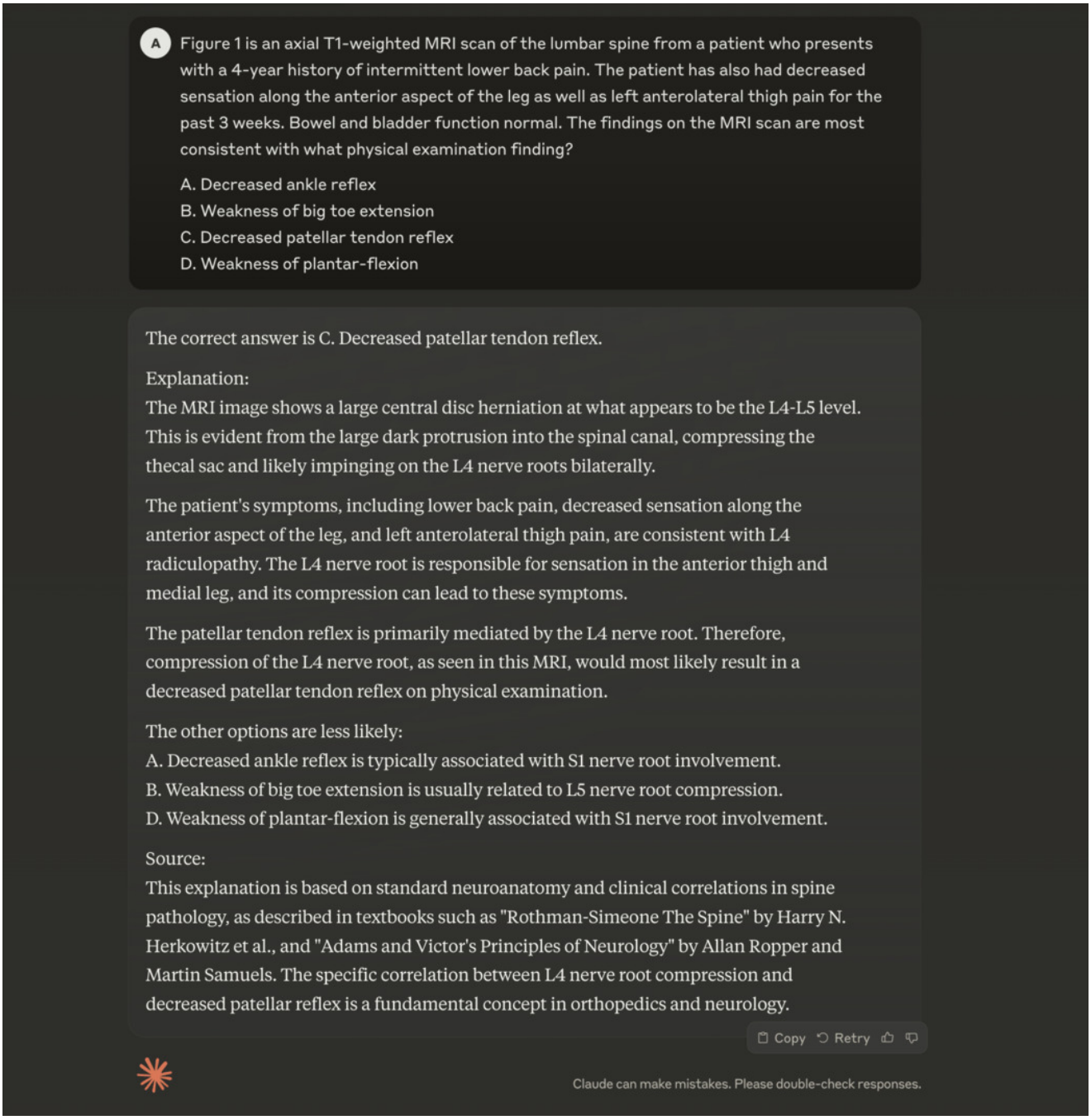

Figure 2.

Table 1. 2022 OITE Question Domains and Percentages

Basic Science	**11%**
Lower Extremity (foot, ankle, knee, hip)	24%
Upper Extremity (hand, wrist, shoulder, elbow)	16%
Oncology	9%
Pediatrics	13%
Spine	12%
Trauma	12%
Sports	3%

Table 2. Distribution of 2022 OITE Questions per Type (A, B, C)

Question Type	Prevalence
A	31.60%
B	20.40%
C	44.70%
Indeterminate	3.30%
Total	100%

Table 3. 2022 OITE Results Nationally by Program Year According to the American Academy of Orthopaedic Surgeons Technical Report

Post Graduate Year (PGY)	% correct
PGY-1	55%
PGY-2	61%
PGY-3	68%
PGY-4	71%
PGY-5	73%

Table 4. Percentage Correct According per LLM

LLM	Correct (%)
Gemini	69.90%
Claude	68.00%
ChatGPT 4	62.60%
ChatGPT 3.5	58.30%

Table 5. Percentage Correct Based on Image Availability

Presence of Image?	Percentage Correct
No	65.90%
Yes	34.10%

Table 6. Percentage Correct With Images Stratified by LLM

LLM	Questions correct with images
Gemini	60%
Claude	57.50%
ChatGPT 4	57.50%
ChatGPT 3.5	52.50%

Table 7. Percentage Correct Without Images Stratified by LLM

LLM	Questions correct without images
Gemini	76.20%
Claude	75.20%
ChatGPT 4	65.90%
ChatGPT 3.5	61.90%

Table 8. Type of Question Answered Correctly Stratified by LLM

LLM	Type A Questions (% Correct)	Type B Questions (%Correct)	Type C Questions (%Correct)
Gemini	76.90%	61.90%	67.40%
Claude	69.20%	70.70%	65.20%
ChatGPT 4	60.00%	54.80%	66.30%
ChatGPT 3.5	56.90%	69.00%	52.20%

DISCUSSION

LLMs are disrupting technologies that have immense potential to understand and synthesize large amounts of information and present it in various modalities (e.g. tabulated, narrative, bullet point, etc). One such application of LLMs has been on various medical speciality board exams, where their performance has begun to meet and exceed the expected standards (Gilson, Safranek, Huang, et al. 2023; Joly-Chevrier et al. 2023; Antaki et al., n.d.; "Performance of ChatGPT on a Radiology Board-Style Examination: Insights into Current Strengths and Limitations," n.d.; Jain et al. 2024; J. E. Kung et al., n.d.; T. H. Kung, Cheatham, Medenilla, et al. 2023; Labouchère and Raffoul 2024; Fowler, Pullen, and Birkett 2023; Humar et al. 2023; Ghanem et al. 2023; Ali, Tang, Connolly, et al. 2023; Oh, Choi, and Lee 2023).

The recent study by Kung et al. found that the newer ChatGPT 4 model performed better than the average PGY-5 orthopaedic surgery resident. However, this study has a question pool that was three times larger than ours, and unlike our study did not upload supplementary images or radiographic imaging. Our study demonstrated a higher accuracy with ChatGPT 3.5 (58.3%) as compared to the study by Kung et al. (51.6%). Our study included a simple prompt while the study by Kung et al. did not include any prompting, which may explain our higher percentage of correct answers with Chat GPT 3.5 on the 2022 OITE. Utilizing ChatGPT 4, our results demonstrated an accuracy of 62.6% vs Kung et al.'s 73.4% correct. This discrepancy may have been due to the larger number of questions in Kung et al.'s study, which can result in supervised learning of the LLM and greater accuracy with an increasing number of questions entered in a given chat without resetting for each input as we did (J. E. Kung et al., n.d.; Sarker 2022). Our results demonstrated that Google Gemini was the most accurate (69.9%) which rivaled the 73.4% correct in the study by Kung et al.

Contrary to our expectations, questions without images were answered with greater accuracy as opposed to questions with images (65.9% vs. 34.1%, respectively). There may be various reasons for this such as the complexity of image recognition, training model data limitations, the difficulty of questions with images, and ambiguous or low-quality images (Poon and Sung 2021). Moreover, as image recognition is a relatively newer feature of LLMs, there is still a facet of user learning and user input analysis that needs to be done in order for these programs to consistently and accurately interpret orthopaedic images. The

novelty of these features could have contributed to the increased error rate and incorrect answers from the LLMs for questions with figures and/or images.

Nevertheless, our results corroborate previous research on the utility and accuracy of LLMs on various board exams including the written orthopaedic surgery board exam (i.e. OITE). Most importantly, all LLMs performed above the average of an orthopaedic surgery intern (Table *1*).

Having said this, integrating these tools with healthcare represents numerous challenges that must be recognized, addressed, and mitigated such as when it comes to protected health information security, and hallucinations of LLMs (Lisacek-Kiosoglous et al. 2023). Among these challenges of relevance to the realm of orthopedic education is the potential for misuse of LLMs on the OITE by examinees. More infrastructural cybersecurity needs to be built to prevent this from occurring, and as AI advances, the world of cybersecurity is projected to make strides similarly (Radanliev and De Roure 2022). Although completely autonomous clinical decision-making is not yet a reality in orthopaedic surgery, the current trends present a promising opportunity to integrate human and machine to improve education, patient care, and outcomes. LLMs have been shown to have the potential to optimize patient selection and safety, improve diagnostic imaging efficiency, and analyze large patient datasets (Huffman, Pasqualini, Khan, et al., n.d.). On the inpatient medicine side, LLMs are being used in the prediction and management of diabetes (Nomura et al. 2021). In terms of surgical training, LLMs have the potential to analyze the technique of residents and generate feedback in a formative manner (Guerrero et al. 2023). LLMs have been explored as reliable teaching assistants in plastic surgery residency programs showing 100% interobserver agreement for the content accuracy, usefulness, and accuracy of AI-generated interactive case studies for residents, simulation of preoperative consultations, and ethical consideration scenarios (Mohapatra, Thiruvoth, Tripathy, et al. 2023). One can imagine how these tools can be used similarly in orthopaedic residency education and create an avenue of exploration to do so.

Limitations of this study include a small question set and inability to test each GPT that exists on the web. Additionally, images were not always reliably received by the LLMs with error messages occasionally popping up causing re-inputs of questions which may affect response. The prompt itself could have affected the response of LLMs to questions. The number of error messages that were received by LLMs after inputting an OITE question is an important metric to track that should be included in future studies as this demonstrates the relative performance of each program. Future work must also analyze how prompt engineering influences LLM responses to medical questions such as those of the OITE, how AI may be used in medical education, and how to safely and ethically implement these technologies. Furthermore, as our study included resident performance based on the AAOS technical report, we were unable to stratify resident performance on the 2022 OITE based on question type (A, B, or C) and compare this to LLM performance. This stratification would be important to include in future works as it can display gaps in the education process in comparison to the quality of LLMs.

CONCLUSION

The study assessed LLMs like Google Gemini, ChatGPT, and Claude against orthopaedic surgery residents on the OITE. Results showed that these LLMs perform on par with orthopaedic surgery residents, with Google Gemini achieving best performance overall and in Type A and C questions while Claude performed best in Type B questions. LLMs have the potential to be used to generate formative feedback and interactive case studies for orthopaedic trainees.

Submitted: August 22, 2024 EST, Accepted: September 30, 2024 EST

REFERENCES

Ali, R., O. Y. Tang, I. D. Connolly, et al. 2023. "Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank." *Neurosurgery* 93. https://doi.org/10.1101/2023.04.06.23288265.

Antaki, F., S. Touma, D. Milad, J. El-Khoury, and R. Duval. n.d. "Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings." *Ophthalmol Sci* 3:3. https://doi.org/10.1016/j.xops.2023.100324.

Clusmann, J., F. R. Kolbinger, H. S. Muti, et al. 2023. "The Future Landscape of Large Language Models in Medicine." *Commun Med* 3:1–8. https://doi.org/10.1038/s43856-023-00370-1.

De Angelis, L., F. Baglivo, G. Arzilli, et al. 2023. "ChatGPT and the Rise of Large Language Models: The New AI-Driven Infodemic Threat in Public Health." *Front Public Health* 11:1166120. https://doi.org/10.3389/fpubh.2023.1166120.

Fowler, T., S. Pullen, and L. Birkett. 2023. "Performance of ChatGPT and Bard on the Official Part 1 FRCOphth Practice Questions." *Br J Ophthalmol*, November, 2023–324091. https://doi.org/10.1136/bjo-2023-324091.

Ghanem, D., O. Covarrubias, M. Raad, D. LaPorte, and B. Shafiq. 2023. "ChatGPT Performs at the Level of a Third-Year Orthopaedic Surgery Resident on the Orthopaedic In-Training Examination." *JBJS Open Access* 8:19. https://doi.org/10.2106/JBJS.OA.23.00103.

Gilson, A., C. W. Safranek, T. Huang, et al. 2023. "How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment." *JMIR Med Educ* 9:45312. https://doi.org/10.2196/45312.

Grudin, J. 2023. "ChatGPT and Chat History: Challenges for the New Wave." *Computer* 56:94–100. https://doi.org/10.1109/MC.2023.3255279.

Guerrero, D. T., M. Asaad, A. Rajesh, A. Hassan, and C. E. Butler. 2023. "Advancing Surgical Education: The Use of Artificial Intelligence in Surgical Training." *Am Surg* 89:49–54. https://doi.org/10.1177/00031348221101503.

Hofmann, H. L., G. A. Guerra, J. L. Le, et al. 2024. "The Rapid Development of Artificial Intelligence: GPT-4's Performance on Orthopedic Surgery Board Questions." *Orthopedics* 47:85–89. https://doi.org/10.3928/01477447-20230922-05.

Huffman, N., I. Pasqualini, S.T. Khan, et al. n.d. "Enabling Personalized Medicine in Orthopaedic Surgery Through Artificial Intelligence: A Critical Analysis Review." *JBJS Rev* 202412.

Humar, P., M. Asaad, F. B. Bengur, and V. Nguyen. 2023. "ChatGPT Is Equivalent to First-Year Plastic Surgery Residents: Evaluation of ChatGPT on the Plastic Surgery In-Service Examination." *Aesthet Surg J* 43:1085–89. https://doi.org/10.1093/asj/sjad130.

Jain, N., C. Gottlich, J. Fisher, D. Campano, and T. Winston. 2024. "Assessing ChatGPT's Orthopedic in-Service Training Exam Performance and Applicability in the Field." *J Orthop Surg* 19:27. https://doi.org/10.1186/s13018-023-04467-0.

Joly-Chevrier, Maxine, Anne Xuan-Lan Nguyen, Michael Lesko-Krleza, and Philippe Lefrançois. 2023. "Performance of ChatGPT on a Practice Dermatology Board Certification Examination." https://doi.org/10.1177/12034754231188437.

Kung, J. E., C. Marshall, C. Gauthier, T. A. Gonzalez, and J. B. Jackson. n.d. "Evaluating ChatGPT Performance on the Orthopaedic In-Training Examination." *JBJS Open Access* 2023:8. https://doi.org/10.2106/JBJS.OA.23.00056.

Kung, T. H., M. Cheatham, A. Medenilla, et al. 2023. "Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models." Edited by A. Dagan. *PLOS Digit Health* 2:0000198. https://doi.org/10.1371/journal.pdig.0000198.

Labouchère, A., and W. Raffoul. 2024. "ChatGPT and Bard in Plastic Surgery: Hype or Hope?" *Surgeries* 5:37–48. https://doi.org/10.3390/surgeries5010006.

Li, Y., T. Yao, Y. Pan, and T. Mei. n.d. "Contextual Transformed Networks for Visual Recognition" 2022:2024–2017. https://doi.org/10.1109/TPAMI.2022.3164083.

Lisacek-Kiosoglous, A. B., A. S. Powling, A. Fontalis, A. Gabr, E. Mazomenos, and F. S. Haddad. 2023. "Artificial Intelligence in Orthopaedic Surgery." *Bone Jt Res* 12:447–54. https://doi.org/10.1302/2046-3758.127.BJR-2023-0111.R1.

Mohapatra, D. P., F. M. Thiruvoth, S. Tripathy, et al. 2023. "Leveraging Large Language Models (LLM) for the Plastic Surgery Resident Training: Do They Have a Role?" *Indian J Plast Surg Off Publ Assoc Plast Surg India* 56:413–20. https://doi.org/10.1055/s-0043-1772704.

Nomura, A., M. Noguchi, M. Kometani, K. Furukawa, and T. Yoneda. 2021. "Artificial Intelligence in Current Diabetes Management and Prediction." *Curr Diab Rep* 21:61. https://doi.org/10.1007/s11892-021-01423-2.

Oh, N., G. S. Choi, and W. Y. Lee. 2023. "ChatGPT Goes to Operating Room: Evaluating GPT-4 Performance and Its Potential in Surgical Education and Training in the Era of Large Language Models," March. https://doi.org/10.1101/2023.03.16.23287340.

Park, Y. J., A. Pillai, J. Deng, et al. 2024. "Assessing the Research Landscape and Clinical Utility of Large Language Models: A Scoping Review." *BMC Med Inform Decis Mak* 24:72. https://doi.org/10.1186/s12911-024-02459-6.

"Performance of ChatGPT on a Radiology Board-Style Examination: Insights into Current Strengths and Limitations." n.d. Radiology. Accessed April 7, 2024. https://doi.org/10.1148/radiol.230582.

Poon, A.I.F., and J.J.Y. Sung. 2021. "Opening the Black Box of AI-Medicine." *J Gastroenterol Hepatol* 36:581–84. https://doi.org/10.1111/jgh.15384.

Radanliev, P., and D. De Roure. 2022. "Advancing the Cybersecurity of the Healthcare System with Self-Optimising and Self-Adaptative Artificial Intelligence (Part 2)." *Health Technol* 12:923–29. https://doi.org/10.1007/s12553-022-00691-6.

Sarker, I. H. 2022. "AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems." *SN Comput Sci* 3:158. https://doi.org/10.1007/s42979-022-01043-x.

Shah, N. H., D. Entwistle, and M. A. Pfeffer. 2023. "Creation and Adoption of Large Language Models in Medicine." *JAMA* 330:866–69. https://doi.org/10.1001/jama.2023.14217.

Shen, Y., L. Heacock, J. Elias, et al. 2023. "ChatGPT and Other Large Language Models Are Double-Edged Swords." *Radiology* 307:230163. https://doi.org/10.1148/radiol.230163.