Impact of Imaging Content on LLM Performance in Orthopaedic Questions

Gnaneswar Chundi, Abhiram Dawar, Syed A Sarwar, Sanjiv Prasad, Irfan H Ahmed, Michael M Vosbikian

INTRODUCTION:

Radiographic interpretation is a fundamental skill in orthopaedic diagnosis and management, and proficiency in reading medical images such as X-rays, CT scans, and MRIs is critical for effective clinical decision-making. While large language models (LLMs) have demonstrated promising capabilities in text-based medical education and reasoning, their performance on multimodal tasks involving visual data remains underexplored. This study investigates the impact of image inclusion on LLM accuracy in orthopaedic question answering, aiming to assess the current limitations in multimodal comprehension and highlight areas for improvement.

METHODS:

A total of 2,906 board-style questions were sourced from the American Academy of Orthopaedic Surgeons (AAOS) ResStudy Question Bank. Each question was classified as either “text-only” or “image-based,” depending on the presence of associated clinical or radiographic images. Questions covered a range of orthopaedic disciplines, including trauma, spine, sports, and basic science. Multiple state-of-the-art LLMs were evaluated using a standardized prompt format that included the question stem, answer choices, and accompanying images when available. Each model was asked to return a letter answer, a confidence score (0–100%), and a brief justification. Accuracy was determined by comparing model responses to the ResStudy official answer key. A paired t-test was used to compare model performance on text-only versus image-based questions.

RESULTS:

Results demonstrated a consistent performance deficit across all models when answering image-based questions. On average, the inclusion of images led to a 10.9% decline in accuracy ($t = 17.77$, $p < 0.001$), a highly statistically significant finding. Even the best-performing model, GPT-4o, which integrates vision and text capabilities, exhibited a noticeable drop in performance when visual input was required. This suggests that, while capable of parsing detailed written information, current LLMs remain limited in their ability to interpret and reason through complex visual inputs like radiographs.

DISCUSSION AND CONCLUSION:

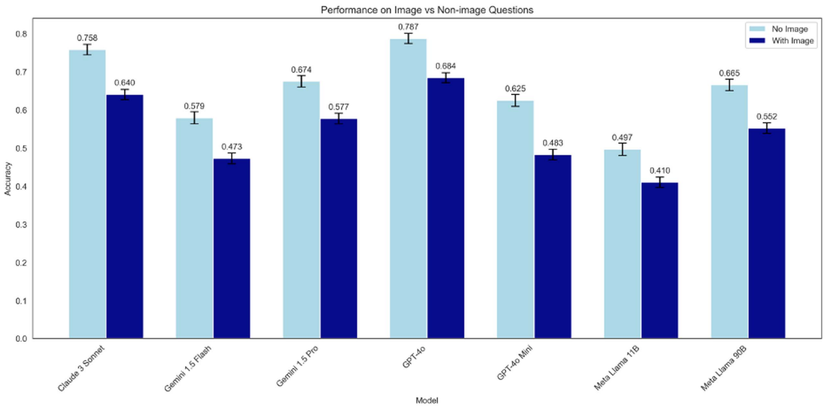
These findings underscore the current shortcomings of LLMs in handling clinical images and emphasize the need for targeted improvements in multimodal model training. Specifically, enhancing image interpretation capabilities through pretraining on annotated orthopaedic imaging datasets, incorporating diagnostic features into vision-language embedding architectures, and refining visual reasoning modules are key priorities for future development. The underperformance on image-based questions may stem from several challenges, including inadequate spatial reasoning, limited exposure to radiographic image-text pairings during training, and difficulty integrating visual cues with clinical context.

As image interpretation plays an essential role in orthopaedics—particularly in trauma, spine, and tumor evaluation—these limitations represent a barrier to real-world clinical integration of LLMs. Nonetheless, this performance gap also highlights an important opportunity for innovation. By integrating convolutional neural networks (CNNs), vision transformers (ViTs), or hybrid architectures into future multimodal LLMs, developers can better equip models to process visual information and improve alignment with human diagnostic processes.

From a practical standpoint, the findings caution against deploying current LLMs in image-dependent diagnostic workflows without human oversight. However, they also support the role of LLMs as potential educational tools, especially when used in conjunction with expert guidance to explain key radiographic findings or simulate interpretive reasoning. For medical educators, this also raises the possibility of incorporating AI-based feedback loops into radiographic training environments, helping students and residents identify interpretive pitfalls.

Ultimately, improving LLM performance on image-based tasks will require not only algorithmic advancements but also collaborative efforts between computer scientists, radiologists, and orthopaedic educators to curate high-quality multimodal datasets and define clinical benchmarks for safe and effective AI use. In conclusion, this study highlights a significant and measurable gap in the visual processing capabilities of current LLMs. Bridging this gap will be essential to unlocking their full potential in orthopaedic diagnostics and medical education, and ensuring that AI tools contribute meaningfully to accurate, equitable, and efficient patient care.

Figure 2: Comparison of Model Performance on Questions with and without Images: Bars depict models' performance on questions stratified by image vs. no-image. Error bars indicate standard error of measurement.


Performance on Image vs Non-image Questions
No Image
With Image
Accuracy
Claude 3 Sonnet
Gemini 1.5 Flash
Gemini 1.5 Pro
GPT-4o
GPT-4o Mini
Meta Llama 11B
Meta Llama 90B
Model