

A Multi-Chatbot Evaluation Framework for Knee MRI Diagnosis Assistance

Serhat Uzunbayir¹, Beyza Altuner², Deniz Eren Terziler¹, Cinar Gedizlioglu²

¹Software Engineering Department

²Computer Engineering Department

Izmir University of Economics

Izmir, Türkiye

Abstract—Knee injuries, and in particular abnormalities of the Anterior Cruciate Ligament (ACL) and the meniscus, are diagnosed frequently using MRI scans. Although MRI interpretations typically require expert knowledge, that expertise may not always be accessible. Recently, researchers have begun using Large Language Models (LLMs) in the medical domain, applied to assist with diagnostic interpretative tasks. Here, we investigate the potential for LLM-based chatbots to assist and augment the reasoned diagnostic interpretation of knee MRI images. Specifically, we report our comparisons across chatbot diagnostics including ChatGPT-4o, Gemini 2.5 Flash, and Claude Sonnet 4, to see if they can annotate ACL injury, meniscal tear, and abnormality of any type. Using visual MRI slices as input, we evaluated the interpretations produced by multimodal capable chatbots against the ground truth data labelled by professional radiologists. Our findings illustrate each chatbot model's relative strengths and weaknesses in medical imaging analysis that contribute evidence towards supporting the development of AI-augmented workflows for medical imaging and radiology.

Keywords—medical decision-making, ai, chatbot, healthcare, medical imaging, large language models.

I. INTRODUCTION

Magnetic Resonance Imaging (MRI) is the gold standard diagnostic tool for soft tissue structures of the knee joint, including the Anterior Cruciate Ligament (ACL) and meniscus. Timely and accurate interpretations of knee MRI scans can lead to injuries that will limit somebody's ability to move, quality of life, and ultimately function of the joint in the long term. However, radiological resources may not always be available, particularly in low-resource contexts. This has created excitement to develop and implement Artificial Intelligence (AI) tools to assist with clinical decision making.

Large Language Models (LLMs) and multimodal AI chatbots have grown increasingly powerful over the past few years. They are able to process text and visual information simultaneously. Models like GPT-4o, Deepseek, and Kimi-K2 have demonstrated some success in various fields, including health care [1]. These technology have the potential to help augment diagnostic activities in radiology by interpreting medical images with clinical text; however, their performance has yet to be fully explored in narrower applications (e.g., musculoskeletal imaging).

The objective of this study is to assess and compare the performance of various leading AI chatbot models to interpret knee MRI images on three separate diagnostic criteria: the detection of general abnormalities, identification of ACL injuries, and identification of meniscal tears.

We will carefully examine how and if each chatbot interprets standardized MRI inputs, and will specifically look to answer two research questions:

RQ1: Which AI chatbot provides the most accurate assessment of knee MRI images for detecting general abnormalities, ACL injuries, and meniscal tears?

RQ2: How do AI chatbot diagnostics compare with models trained to handle MRI images?

We contribute a unified evaluation framework that feeds standardized knee MRI images into chatbot systems and benchmark their performance across three core diagnostic tasks. Our comparative analysis covers three state-of-the-art chatbots -ChatGPT-4o, Gemini 2.5 Flash, and Claude Sonnet 4- and a CNN baseline. The results reveal the current strengths and limitations of LLMs in clinical imaging tasks, along with broader implications for ethical deployment in real-world workflows.

II. RELATED WORK

In this section we summarize the literature on artificial intelligence and its applications, particularly deep learning and large language models, in interpreting knee MRI images.

One popularly reported method has been using Convolutional Neural Networks (CNN) trained on knee MRI datasets, like the MRNet dataset [2], and annotated by expert radiologists [3]. These models can identify specific forms of abnormality like ACL tears and meniscal injuries fairly accurately and in some cases are even superior. CNNs are able to learn very specific, localized patterns and although are suitable to segmentation tasks like the one we discussed, CNNs are limited in the generalizability, interpretability, and contextual reasoning for specific clinical cases that constitute our gaps. This has led researchers to evaluate alternatives and more flexible forms of computational AI models.

LLMs increasingly being used in healthcare particularly destructive analysis, clinical decision support, and medical



imaging [4] [5]. With the necessary adjustments made for radiology, these foundations demonstrated versatility beyond natural language processing, with inclinations towards radiology report generation, protocol recommendation, and even multimodal image interpretation [4] [6]. In recent literature, vision-language foundation models have helped connect imaging findings and clinical context. This offers promise for improved levels of diagnostic reasoning and is an active area for future study [7].

For example, [6] assessed ChatGPT and similar models in image-based radiologic interpretation tasks, rating them on instruction-following ability, diagnostic relevance, and consistency. Their results show advantages for structured output generation, but also demonstrate lingering concerns like the frequent occurrence of hallucinations and a lack of robustness in novel situations. In related work, [8] evaluated a number of LLMs (Claude, Bing, and Perplexity) for radiologic decision-making for imaging pulmonary embolism, relative to the American College of Radiology Appropriateness Criteria.

Comparative analyses such as those in [9] and [10] show notable differences in LLM performance depending on task design, medical complexity, and contextual detail. While Claude and Gemini have shown strong contextual understanding, ChatGPT often excels in structured output and consistency. These underscore the need for targeted evaluation frameworks.

Furthermore, useful literature reviews have looked at larger implications of LLMs in clinical workflows. For instance, [11] reviews how LLMs integrated into workflow in radiology, identifying advantages of images LLMs including automated structured reporting and potential workflow efficiencies. In a similar vein, [12] survey a scalable data-driven approaches, including low-rank approximations, to improve medical imaging pipeline efficiency.

Even with these advancements, there have been limited, comparisons unique to multimodal LLMs for interpreting MRI images, and knee injuries or conditions like ACL and meniscus tears. There have been prior evaluations of multimodal LLMs, but they were limited primarily to chest images, or general radiology tasks [4] [7]. Therefore, the number of studies related to applications associated with musculoskeletal imaging is limited. This research provides related contribution by evaluating three LLMs with high-quality visual input capabilities, ChatGPT-4o, Gemini 2.5 Flash, and Claude Sonnet 4, and their ability to examine knee MRI images and demonstrate abnormalities. The findings provide insights into their interpretive accuracy, diagnostic confidence, and overall feasibility for integration into AI-augmented radiology workflows.

III. METHODOLOGY

In this section, we outline the methodology that we developed to assess the performance of different multimodal chatbot models on knee MRI interpretations.

A. Dataset

We used the publicly available **MRNet dataset** by Stanford University, which is comprised of knee MRI scans obtained

from clinical exams. The dataset contains **1,370 MRI studies** from **1,088 patients** and each study consists of three standard MRI planes: *sagittal*, *coronal*, and *axial*. Each exam is annotated with three binary labels indicating if there was a title presence or absence of ACL tears, meniscal tears, and general abnormalities.

Selectively, we chose a representative subset of 100 studies from the MRNet dataset, with an equal split between positive and negative cases across the three diagnostic categories, and to prepare the data for chatbot evaluation, we aggregated all slices from each anatomical plane (sagittal, coronal, axial) per study. Thus, for each case we created three composite images and a composite image for each plane. Each of these composite images were utilized as the input.

B. Baseline Model for Comparison

To provide a basis for comparison with the performance of chatbots empowered by LLM, we constructed a baseline Deep Learning model using a Convolutional Neural Network (CNN). The architecture for the baseline followed the MRNet approach, consisting of 3 separate 2D CNNs which are trained separately on each MRI plane and then cohort late-fusion model for the final prediction across the three target labels.

Each per-plane network comprises five convolutional blocks with increasing channel dimensions ($8 \rightarrow 16 \rightarrow 32 \rightarrow 64 \rightarrow 128$). All convolutional layers use a kernel size of 3×3 , except the first layer which uses a 5×5 kernel.

In each block, we applied ReLU activation functions and 2×2 max-pooling, which was ultimately followed by an adaptive max-pooling that transformed the output to a 1×1 spatial map. A final fully connected layer was used to map the resulting 128-dimensional feature vector to 3 binary outputs corresponding to our diagnostic targets; general abnormality, ACL tear, and meniscal tear.

The CNN model was trained using training split of the dataset. Performance was evaluated using the same diagnostic metrics and samples as the chatbot models.

C. Chatbot Models

We examined the diagnostic performance of three current, general-purpose multimodal chatbot models: ChatGPT-4o (OpenAI) [13], Gemini 2.5 Flash (Google DeepMind) [14], and Claude Sonnet 4 (Anthropic) [15]. These models were selected based on ease of access, recent launch into general circulation and broad public accessibility. Each is capable of processing both images and text, making them suitable candidates for evaluating performance on medical imaging tasks without domain-specific fine-tuning. Our aim was to explore how general-purpose multimodal models, without domain-specific medical fine-tuning, perform on specialized tasks such as knee MRI interpretation.

Each model was provided with three composite images per case patient case. The models were prompted to deliver diagnostic predictions for **ACL tears**, **meniscal tears**, and **general abnormalities**. In addition to the predicted diagnoses, we

requested **textual explanations** supporting their conclusions. This protocol enables a multi-dimensional assessment of each model's diagnostic capability, interpretability, and reliability in the medical imaging domain.

D. Prompt Engineering and Input Formatting

To standardize the evaluation across models, we applied a uniform prompt format that includes a brief clinical context and a directive to provide a diagnosis. Visual inputs were uploaded to the chatbot interfaces where image capabilities were available. For each case, the prompt structure was as follows:

You are a radiologist specializing in musculoskeletal MRI interpretation. Here are three MRI slices of a human knee in different planes: *Sagittal view, Coronal view, Axial view*. Please analyze these images and respond to the following clinical classification questions. Use the exact format shown below.

- 1) Do you see any abnormality overall? Answer yes/no and explain. What abnormalities (if any) are visible?
- 2) Do you see any signs of an ACL tear? Answer yes/no and explain.
- 3) Do you see any signs of a meniscal tear? Answer yes/no and explain.

Example Format:

- Abnormal: ____
 - ACL Tear: ____
 - Meniscal Tear: ____
- 4) Briefly summarize the imaging features that support your conclusions.

This structure ensured consistency in model queries, minimized ambiguity, and allowed fair comparative analysis.

E. Evaluation Protocol

We developed an comprehensive **evaluation protocol** for assessing comprehensive chatbot model performance, which will consider:

- **Diagnostic Accuracy:** Comparison of chatbot output against the ground truth labelled outputs for ACL, meniscus, and aberration detection which will include performance metrics including *precision, recall, F1-score*, and *accuracy*.
- **Comparative Analysis:** A cross-model analysis to find which chatbot represents the best trade-off between diagnostic accuracy, reliability of confidence, and quality of explanation.

IV. RESULTS

For our analysis of diagnostic performance, we examined the mean classification scores for the three tasks: ACL tear, meniscal tear, and general abnormality detection. The comparisons were between the CNN model and the three AI chatbots: ChatGPT-4o, Gemini 2.5 Flash, Claude Sonnet 4

using independent-samples t-tests. Reported *p*-values reflect two-tailed significance testing with a threshold of $p < 0.05$.

A. Total Diagnostic Performance

The CNN model achieved the highest overall diagnostic performance ($M = 2.39$, $SD = 0.57$). ChatGPT-4o ($M = 1.71$, $SD = 0.92$; $p < 0.001$), Gemini ($M = 1.55$, $SD = 0.89$; $p < 0.001$), and Claude ($M = 1.39$, $SD = 0.91$; $p < 0.001$) all performed significantly worse than the CNN.

There was no statistically significant difference between ChatGPT-4o and Gemini or between Gemini and Claude. However, ChatGPT-4o significantly outperformed Claude ($p = 0.014$).

Table 1 shows the performance metrics supported these findings. The CNN achieved the highest overall accuracy (0.80), recall (0.78), precision (0.77), and F1-score (0.77). ChatGPT-4o followed by balanced recall (0.79) and F1 (0.62) performed better than Gemini (F1 = 0.54) and Claude (F1 = 0.59).

B. ACL Tear Detection

The CNN again led in ACL detection performance ($M = 0.87$, $SD = 0.34$). ChatGPT-4o ($M = 0.48$, $SD = 0.50$; $p < 0.001$), Gemini ($M = 0.48$, $SD = 0.50$; $p < 0.001$), and Claude ($M = 0.25$, $SD = 0.43$; $p < 0.001$) all performed significantly worse than the CNN.

In terms of inter-LLM analysis, both ChatGPT-4o and Gemini significantly outperformed Claude ($p < 0.001$).

In terms of metrics, Claude had higher recall (0.92) than Gemini (0.67) and ChatGPT-4o (0.75), but lower precision (GPT: 0.16; Gemini: 0.14; Claude: 0.13) which resulted in comparatively low F1-scores (GPT: 0.26; Gemini: 0.24; Claude: 0.23). The CNN had the worst recall (0.08) but the best precision (0.33).

C. Meniscal Tear Detection

The CNN once more produced performance metrics with F1 = 0.61. GPT (F1 = 0.60) and Claude (F1 = 0.56) performed closely behind while Gemini slightly lagged behind (F1 = 0.51). Claude had the highest recall (0.86) but the lowest precision (0.41).

D. General Abnormality Detection

For abnormality detection, the CNN performed best ($M = 0.83$, $SD = 0.38$). GPT ($M = 0.69$, $SD = 0.47$; $p = 0.02$), Gemini ($M = 0.57$, $SD = 0.50$; $p < 0.001$), and Claude ($M = 0.71$, $SD = 0.46$; $p = 0.044$) all scored lower.

Differences between GPT and other chatbots were not statistically significant: GPT vs. Gemini ($p = 0.08$), GPT vs. Claude ($p = 0.76$). However, Claude underperformed Gemini ($p = 0.039$).

The F1-scores for Claude were highest (0.83), followed closely by GPT (0.80) and then Gemini (0.70). The best F1 overall was the CNN (0.90). GPT had good balance across recall (0.78) and precision (0.83). Both Gemini and Claude had slightly higher precision (0.79), but Claude had the best recall (Claude = 0.86 and Gemini = 0.63).

TABLE I: PERFORMANCE COMPARISON OF MODELS ACROSS ACL, MENISCUS, AND ABNORMAL CATEGORIES

Model	ACL				MENISCUS				ABNORMAL				AVERAGE			
	Acc	Recall	Precision	F1	Acc	Recall	Precision	F1	Acc	Recall	Precision	F1	Acc	Recall	Precision	F1
ChatGPT-4o	0.48	0.75	0.16	0.26	0.54	0.83	0.47	0.60	0.69	0.78	0.83	0.80	0.57	0.79	0.51	0.62
Gemini 2.5 Flash	0.48	0.67	0.14	0.24	0.50	0.62	0.43	0.51	0.57	0.63	0.79	0.70	0.52	0.63	0.47	0.54
Claude Sonnet 4	0.25	0.92	0.13	0.23	0.43	0.86	0.41	0.56	0.71	0.86	0.79	0.83	0.46	0.87	0.45	0.59
CNN Model	0.87	0.08	0.33	0.13	0.69	0.57	0.65	0.61	0.83	0.99	0.83	0.90	0.80	0.78	0.77	0.77

V. DISCUSSION

This research investigated the differential diagnostic abilities of ChatGPT-4o, Gemini 2.5 Flash, and Claude Sonnet 4 vs a CNN which was trained on knee MRI data. The evaluation focused on their performance across multiple diagnostic tasks involving musculoskeletal imaging. The findings answer two previously asked **RQs**:

- **Answer to RQ1:** ChatGPT-4o showed the overall best consistent performance among the LLMs.
- **Answer to RQ2:** All of the chatbots performed at a significantly lower level than the CNN for all tasks.

ChatGPT-4o demonstrated the most consistent and higher performance across recall and precision particularly with general abnormality detection and therefore, it may be the most reliable LLM for general screening purposes.

Claude's high recall in the ACL and meniscus tasks indicates it takes an aggressive positive classification stance, but its low precision must be qualified by the potential for high false positivity rates. This may be useful for triage systems potentially but would need to be followed by human verification to reduce the risk of overdiagnosis.

Currently, as found in the LMArena [1] vision benchmark results, GPT has a better understanding of vision than the Gemini 2.5 Flash and Claude Sonnet 4 models. The results of our findings are consistent with the benchmark results. This alignment reinforces the credibility of both the benchmark and our experimental methodology.

The results also indicate that LLMs are not yet reliable as primary diagnostic systems but may assist in supportive or preliminary roles in clinical workflows. Furthermore, systems tailored to processing MRI images appear to offer clear advantages over general-purpose LLMs.

The use of AI chatbots in clinical care can speed up and standardize diagnoses, but it also brings ethical risks. The biased training data can yield treatment bias and a loss of patient trust, and many AI models have "black-box" natures that obscure how decisions are reached making it difficult to assign accountability or gain a genuinely informed consent [10]. To address this, we must mandate that (1) data sets were diverse and fair, (2) models were assessed for bias and accuracy before use, and (3) AI approaches produce documents demonstrating how decisions are made, i.e., provide an explanation. In public facing examples, the MRNet dataset provides decomposed MRIs for knee images, 1,370 anonymous knees from Stanford each with available for research, which allows researchers compare models simultaneously and transparently. With solid data rules and respect for intellectual property, we can use AI in healthcare in a fair and clear way.

VI. CONCLUSION

LLMs offer a flexible, multimodal interface and language reasoning capabilities. GPT's balanced metrics highlight their utility in broad diagnostic screening. Nevertheless, LLMs still exhibit sensitivity to calibration and hallucinations. For example, although the Claude had high legibility control (ACL) recall, it exhibited low precision resulting in a high false positive rate which may be problematic in clinical workflows.

In sum, LLMs may be useful in a triage, education or decision support role - particularly in resource poor or volume rich environments - but should be applied with caution.

REFERENCES

- [1] LM Arena, "Leaderboard overview," <https://lmarena.ai/leaderboard/>, 2025. [Online; accessed 4-August-2025].
- [2] N. Bien, P. Rajpurkar, R. L. Ball, J. Irvin, A. Park, E. Jones, M. Bereket, B. N. Patel, K. W. Yeom, *et al.*, "Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of mrnet," *PLoS medicine*, vol. 15, no. 11, 2018.
- [3] A. Botnari, M. Kadar, and J. M. Patrascu, "A comprehensive evaluation of deep learning models on knee mris for the diagnosis and classification of meniscal tears: A systematic review and meta-analysis," *Diagnostics (Basel)*, vol. 14, p. 1090, May 2024.
- [4] Z. Liu, T. Zhong, and *et al.*, "Radllm: A comprehensive healthcare benchmark of large language models for radiology," 2023.
- [5] L. Pan, Z. Zhao, Y. Lu, K. Tang, L. Fu, Q. Liang, and S. Peng, "Opportunities and challenges in the application of large artificial intelligence models in radiology," *Meta-Radiology*, vol. 2, no. 2, p. 100080, 2024.
- [6] T. A. D'Antonoli, A. Stanzione, C. Bluethgen, F. Vernuccio, L. Ugga, M. E. Klontzas, R. Cuocolo, R. Cannella, and B. Koçak, "Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions," *Diagnostic and Interventional Radiology*, vol. 30, no. 2, p. 80, 2024.
- [7] S. Lee, W. J. Kim, J. Chang, and J. C. Ye, "Llm-cxr: instruction-finetuned llm for cxr image understanding and generation," *arXiv preprint arXiv:2305.11490*, 2023.
- [8] V. K. Uppalapati and D. S. Nag, "A comparative analysis of ai models in complex medical decision-making scenarios: evaluating chatgpt, claude ai, bard, and perplexity," *Cureus*, vol. 16, no. 1, 2024.
- [9] P. K. Sarangi, S. Datta, M. S. Swarup, S. Panda, D. S. K. Nayak, A. Malik, A. Datta, and H. Mondal, "Radiologic decision-making for imaging in pulmonary embolism: accuracy and reliability of large language models—bing, claude, chatgpt, and perplexity," *Indian Journal of Radiology and Imaging*, vol. 34, no. 04, pp. 653–660, 2024.
- [10] S. Srivastav, R. Chandrakar, S. Gupta, V. Babhulkar, S. Agrawal, A. Jaiswal, R. Prasad, M. B. Wanjari, V. P. Babhulkar, S. Agarwal, *et al.*, "Chatgpt in radiology: the advantages and limitations of artificial intelligence for medical imaging diagnosis," *Cureus*, vol. 15, no. 7, 2023.
- [11] Z. Liu, T. Zhong, and *et al.*, "Evaluating large language models for radiology natural language processing," 2023.
- [12] S. Hamlomo, M. Atemkeng, Y. Brima, C. Nunhokee, and J. Baxter, "A systematic review of low-rank and local low-rank matrix approximation in big data medical imaging," *Neural Computing and Applications*, pp. 1–56, 2025.
- [13] OpenAI, J. Achiam, and *et al.*, "Gpt-4 technical report," 2024.
- [14] G. Comanici, E. Bieber, and *et al.*, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," 2025.
- [15] Anthropic, "Introducing claude 4," <https://www.anthropic.com/news/claude-4>, May 2025.