Performance of Open-Source LLMs in Challenging Radiological Cases – A Benchmark Study on 4,049 Eurorad Case Reports

Su Hwan Kim [1], Severin Schramm [1], Lisa C. Adams [2], Rickmer Braren [2], Keno K. Bressem [3], Matthias Keicher [4], Claus Zimmer [1], Dennis M. Hedderich [1], Benedikt Wiestler [1,5]

[1] Department of Diagnostic and Interventional Neuroradiology, Klinikum rechts der Isar, School of Medicine and Health, Technical University of Munich, Munich, Germany
[2] Department of Diagnostic and Interventional Radiology, Klinikum rechts der Isar, School of Medicine and Health, Technical University of Munich, Munich, Germany
[3] Department of Cardiovascular Radiology and Nuclear Medicine, German Heart Center Munich, School of Medicine and Health, Technical University of Munich, Munich, Germany
[4] Computer Aided Medical Procedures, Technical University of Munich, Munich, Germany
[5] AI for Image-Guided Diagnosis and Therapy, School of Medicine and Health, Technical University of Munich, Munich, Germany

Abstract

Background

Recent advancements in large language models (LLMs) have created new ways to support radiological diagnostics. While both open-source and proprietary LLMs can address privacy concerns through local or cloud deployment, open-source models provide advantages in continuity of access, independence from commercial update cycles, and potentially lower costs.

Purpose

To evaluate the diagnostic performance of open-source LLMs on challenging radiological cases across multiple subspecialties.

Methods

We evaluated the diagnostic performance of eleven state-of-the-art open-source LLMs using clinical and imaging descriptions from 4,049 case reports in the Eurorad library. Cases spanned all radiological subspecialties and excluded those with explicit mentioning of the correct diagnoses in the case description. LLMs provided differential diagnoses based on clinical history and imaging findings. Responses were considered correct if the true diagnosis was included in the top three LLM suggestions. Llama-3-70B evaluated LLM responses, with its accuracy validated against radiologist ratings in a case subset (n = 140). Confidence intervals were adjusted based on this validation. Models were further tested on 60 non-public brain MRI cases from a tertiary hospital to assess generalizability.

Results

Llama-3-70B demonstrated superior performance (75.1 ± 1.7% correct), followed by Gemma-2-27B (63.9 ± 1.8%) and Mixtral-8x-7B (61.5 ± 1.8%). Performance varied across subspecialties, with highest accuracy across models in genital (female) imaging (59.7 ± 2.7%) and lowest in musculoskeletal imaging (47.1 ± 1.5%). Llama-3-70B's judging accuracy was 87.8% (123/140; 95% CI: 0.82 – 0.93) compared to radiologists. Similar performance results were found in the non-public dataset, where Llama-3-70B (71.7 ± 14.1%), Gemma-2-27B (53.3 ± 15.1%), and Mixtral-8x-7B (51.7 ± 15.1%) again emerged as the top models.

Conclusion

Several open-source LLMs showed promising performance in identifying the correct diagnosis based on case descriptions from the Eurorad library, highlighting their potential as decision support tools for radiological differential diagnosis in challenging, real-world cases.

Introduction

Recent advancements in artificial intelligence (AI) have transformed medical diagnostics, offering innovative tools to support clinical decision-making. One promising development is the emergence of large language models (LLMs), which excel at processing and generating natural language. In radiology, these models have demonstrated potential in various applications, including defining study protocols [1,2], performing differential diagnosis [3,4], generating reports [5,6], and extracting information from free-text reports [7,8].

However, a significant barrier to the widespread clinical adoption is data privacy. The LLMs primarily used in previous studies are proprietary, closed-source models, such as GPT-4, Claude 3, or Gemini [9–11]. Access to these models is typically provided via web-based interfaces or via application programming interfaces (API), both of which necessitate the transfer of data to third-party servers, thereby increasing the risk of unauthorized access or misuse of sensitive health information and limiting their use on patient data. While cloud-based solutions for proprietary LLMs can address some privacy concerns, they may still be subject to commercial update cycles and potentially higher long-term costs.

Open-source models offer a viable alternative enabling care institutions to retain patient data within their local infrastructure, mitigating these privacy concerns and providing continuity of access independent of commercial update cycles, which can lower costs due to their free availability. While historically open-source LLMs have underperformed in clinical decision support tasks [12,13], Meta's latest Llama-3 has shown performance levels on par with leading proprietary models in some areas, such as answering radiology board exam questions [14]. However, the diagnostic accuracy of such models in real-world clinical cases remains largely unexplored.

A well-suited resource for such an evaluation is Eurorad, a comprehensive repository of peer-reviewed radiological case reports managed by the European Society of Radiology (ESR). Eurorad serves as a valuable educational resource for radiologists, residents, and medical students, and encompasses a wide range of cases across radiological subspecialties such as abdominal imaging, neuroradiology, uroradiology and pediatric radiology [15].

Therefore, the aim of this study was to evaluate the performance of state-of-the-art open-source LLMs in radiological diagnosis using Eurorad case reports.

Methods

Data

To create a comprehensive and diverse dataset of challenging radiology cases, we automatically downloaded case report data—including “Clinical History,” “Imaging Findings,” “Final Diagnosis,” and “Section”—from the European Society of Radiology's case report library at https://eurorad.org/. All available case reports were scraped using the Python library “Scrapy” (version 2.11.2) on June 15, 2024.

To address potential data contamination concerns and assess generalizability, we further validated the performance of LLMs in a local dataset of 60 brain MRI cases. These were obtained from our local imaging database, as reported previously [4], and equally contained a brief clinical history and imaging findings. This local dataset is not publicly accessible and thus highly unlikely to have been included in the LLMs' training data.

LLM Setup

To evaluate a range of open-source large language models (LLMs), we developed a Python-based workflow utilizing the “llama_cpp_python” library (version 0.2.79). This library provides Python bindings for the widely-used “llama_cpp” software, enabling the execution of local, quantized LLMs in GGUF (GPT-Generated Unified Format). Quantization involves reducing the precision of the model's numerical weights, typically transitioning from floating-point to lower-bit representations, which results in a smaller and faster model while preserving performance. For most models, Q5_K_M was chosen as a quantization, typically offering a good balance between compression and quality. For the 70B models, a quantization factor of Q4_K_M was selected to allow full GPU offloading.

The “llama_cpp_python” library allows for detailed control over relevant hyperparameters. In our experiments, we fully offloaded the LLMs to a GPU for higher computational speed, set the temperature to 0 to ensure deterministic responses, and limited the context width to 1024 tokens, which we previously validated to accommodate all case reports and responses. We chose these settings to balance performance and reproducibility, although we acknowledge that different configurations might yield varying results. Our Python code for prompt construction, along with detailed links to all models (downloaded from https://huggingface.co/), is publicly available in our GitHub repository at https://github.com/ai-idt/os_llm_eurorad.

For this study, we included eleven open-source LLM models, which are detailed in Table 1. All experiments were conducted using an Nvidia P8000 GPU with 48GB of video memory.

Case Selection and Response Assessment

Upon review, we noted that a significant proportion of cases already contained the correct diagnosis within the “Clinical History” and “Imaging Findings” sections. Drawing inspiration from the “LLM-as-a-Judge” paradigm [16], we employed the most advanced model available at the outset of this study, Llama-3-70B, to filter out these cases. A recent study indicated that Llama-3-70B, along with GPT-4 Turbo, demonstrated the closest alignment with human evaluations [17], making it particularly suitable for this task. We prompted Llama-3-70B to assess all cases with the following instruction:

"*You are a senior radiologist. Below, you will find a case description for a patient diagnosed with [Diagnosis]. Please check if the diagnosis or any part of it is mentioned, discussed, or suggested in the case description. Respond with either 'mentioned' (if the diagnosis is included) or 'not mentioned,' and nothing else.*"

Subsequently, we prompted each of the eleven LLMs to provide three differential diagnoses along with a brief rationale for each, using the concatenated “Clinical History” and “Imaging Findings” as input:

"*You are a senior radiologist. Below, you will find information about a patient: first, the clinical presentation, followed by imaging findings. Based on this information, name the three most likely differential diagnoses, with a short rationale for each.*"

Finally, we again utilized Llama-3-70B to evaluate each LLM’s responses on a binary scale, categorizing them as either “correct” (if the correct diagnosis was among the three differential diagnoses) or “wrong.” The prompt for this evaluation was:

"*You are a senior radiologist. Below, you will find the correct diagnosis (indicated after 'Correct Diagnosis:') followed by the differential diagnoses provided by a Radiology Assistant during an exam. Please assess whether the Radiology Assistant included the correct diagnosis in their differential diagnosis. Respond only with 'correct' (if the correct diagnosis is included) or 'wrong' (if it is not).*"

Human Evaluation

In order to gain an understanding of Llama-3-70B’s performance as an LLM judge for correctness of diagnoses, three experienced radiologists (SHK, with 2 years of experience, DMH and BW, board-certified radiologists with 10 years of experience each) additionally evaluated 60 LLM responses each for correctness, of which 20 were shared between all

three reviewers to assess human inter-rater agreement. Using a total of 140 LLM responses for which both human "ground truth" and LLM judge assessments were known, we calculated the accuracy of the LLM judge (Figure 1).

Statistics

Both the LLM judge as well as human raters evaluated LLM responses on a binary scale, i.e., if the correct diagnosis was among the top 3 differential diagnoses listed by the LLM or not. From this response data, we calculated the standard error per model and category as:

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

where *p* is the proportion of correct responses, and *n* is the number of samples.

However, from our human evaluation of the LLM judge performance, we know about its inaccuracies and have to adjust the SE to account for this:

$$SE_{adj} = \sqrt{\frac{A * (1 - A)}{n} * SE^2}$$

where *A* is the accuracy of the LLM judge. The adjusted 95% Confidence Interval is then:

$$95\%\ CI = p \pm 1.96 * SE_{adj}$$

Results

Dataset

The initial dataset retrieved from the Eurorad library consisted of 8,746 case reports. Using the Llama-3-70B model, we identified 4,697 cases where the diagnosis was explicitly stated within the case description. These cases were subsequently excluded, resulting in a final dataset of 4,049 cases for analysis. This filtering process ensured that the LLMs were evaluated on genuinely challenging cases that required inference rather than simple information extraction. The dataset was primarily composed of cases from abdominal imaging (22.0%), neuroradiology (17.7%), and musculoskeletal imaging (14.3%), whereas breast imaging (2.9%) and interventional radiology (2.2%) were underrepresented (Table 2). This distribution broadly reflects the relative prevalence of different radiological subspecialties in clinical practice.

LLM Judge Performance

Based on 140 LLM responses rated by radiologists as the reference standard, Llama-3-70B exhibited an accuracy of 87.8% in classifying responses as “correct” or “incorrect” (123/140 responses; 95% CI: 0.82 – 0.93). Furthermore, in a subset of 20 responses rated by all three radiologists, the interrater agreement was found to be 100%, indicating complete consensus. This high level of agreement between Llama-3-70B and human radiologists, as well as among radiologists themselves, supports the validity of using Llama-3-70B as an automated judge for the larger dataset.

Model Performance

Across all models, the highest levels of diagnostic accuracy were achieved in genital (female) imaging (59.7 ± 2.7%), cardiovascular imaging (59.1 ± 2.3%), and abdominal imaging (58.8 ± 1.2%), whereas lower accuracy was observed in musculoskeletal (47.1 ± 1.5%) and breast imaging (47.6 ± 3.3%) (Figure 2). Granular accuracy metrics by subspecialty and model are provided in Supplement 1. Among the evaluated models, Llama-3-70B demonstrated superior diagnostic performance across all subspecialties, achieving a rate of 75.1 ± 1.7% correct responses, a considerable margin ahead of Gemma-2-27B (63.9 ± 1.8%), Mixtral-8x7B (61.5% ± 1.8%), and Meta-Llama-3-8B (58.7 ± 1.8%) (Figure 3).

In the local brain MRI dataset, comparable results were observed, with Llama-3-70B (71.7 ± 14.1%), Gemma-2-27B (53.3 ± 15.1%), and Mixtral-8x-7B (51.7 ± 15.1%) again leading the rankings (Figure 4). The consistent performance on this non-public dataset suggests that the

models' capabilities generalize beyond potentially contaminated public data, reinforcing the robustness of our findings.

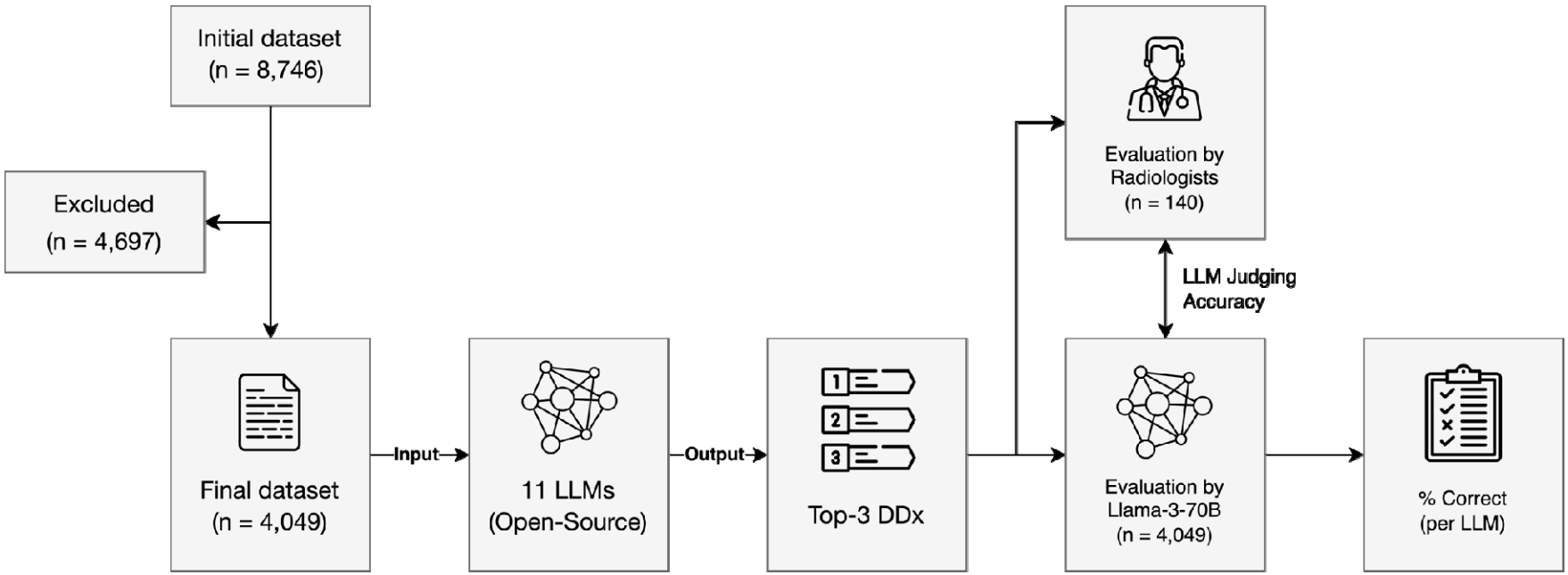

Figure 1: Study Design. A total of 4,697 cases were excluded as the true diagnosis was mentioned in the case description to be provided as LLM input. DDx: differential diagnoses.

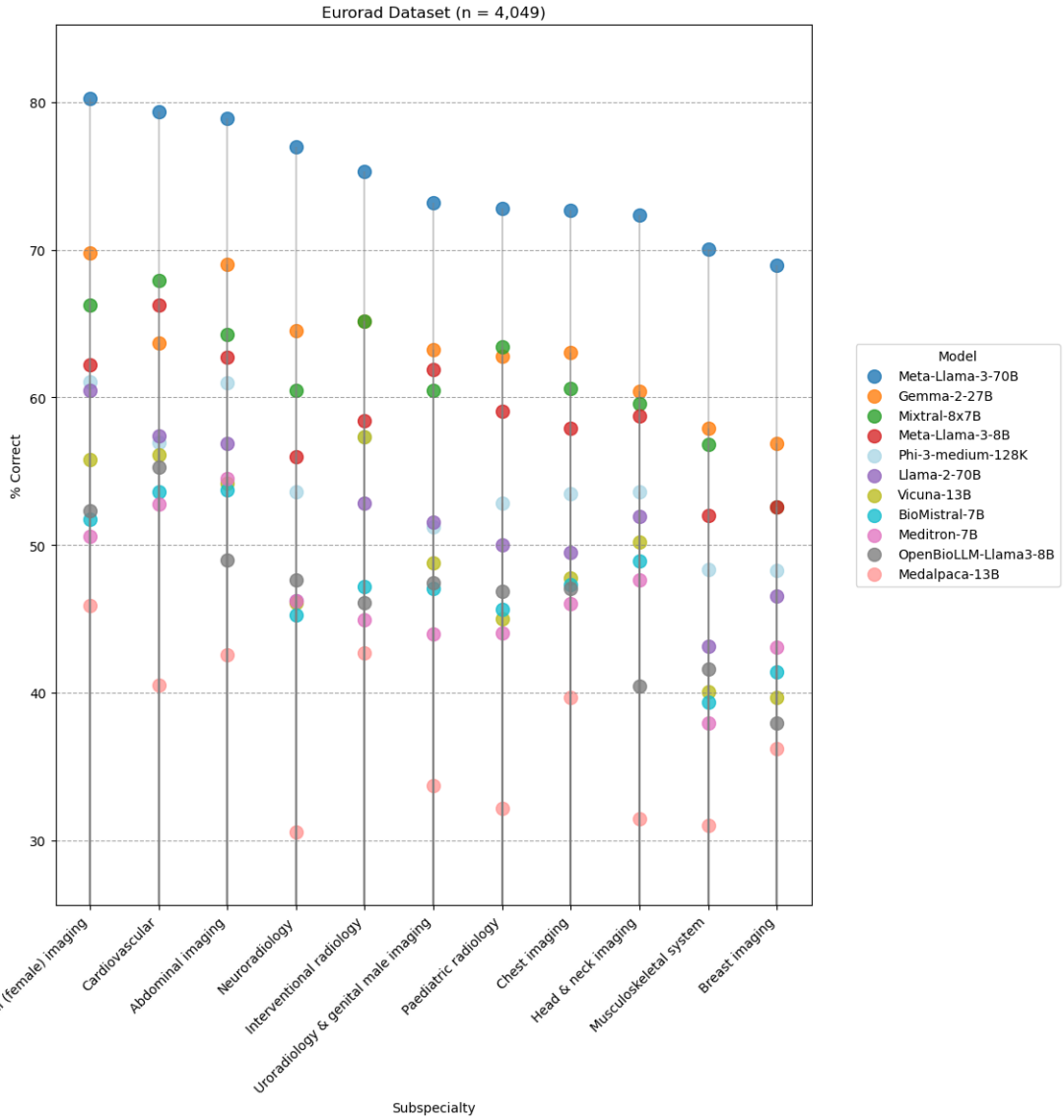

Figure 2: Model Performance by Subspecialty. Meta-Llama-3-70B demonstrated highest performance across all subspecialties.

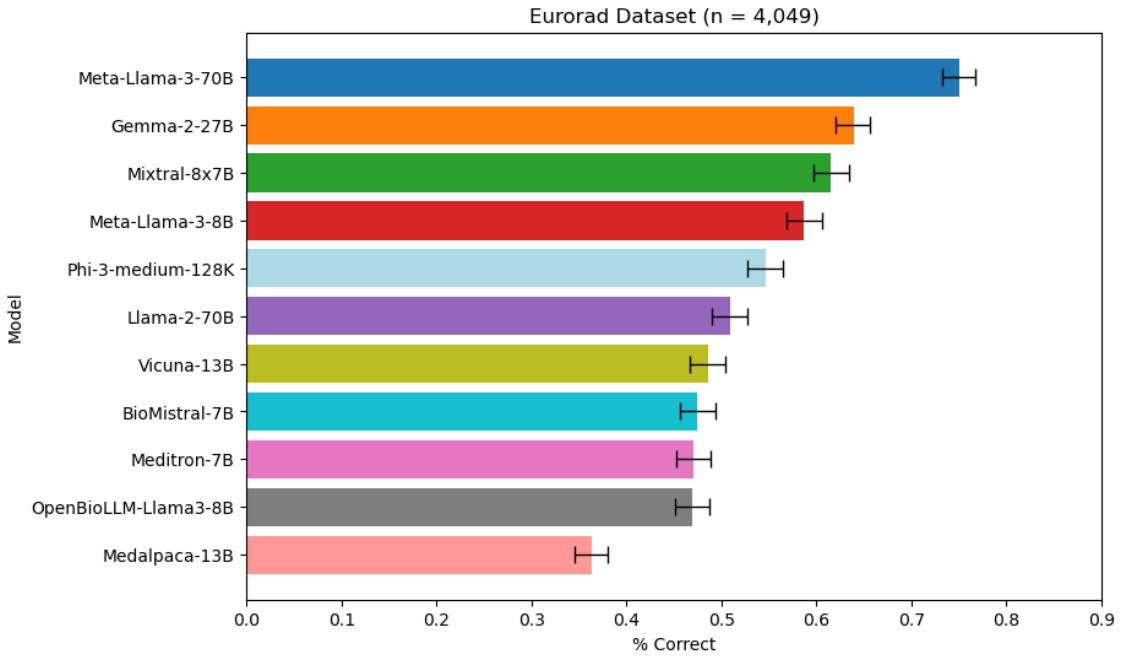

Figure 3: Performance of Open-Source LLMs in Eurorad dataset (n = 4,049). Error bars indicate adjusted 95% confidence intervals.

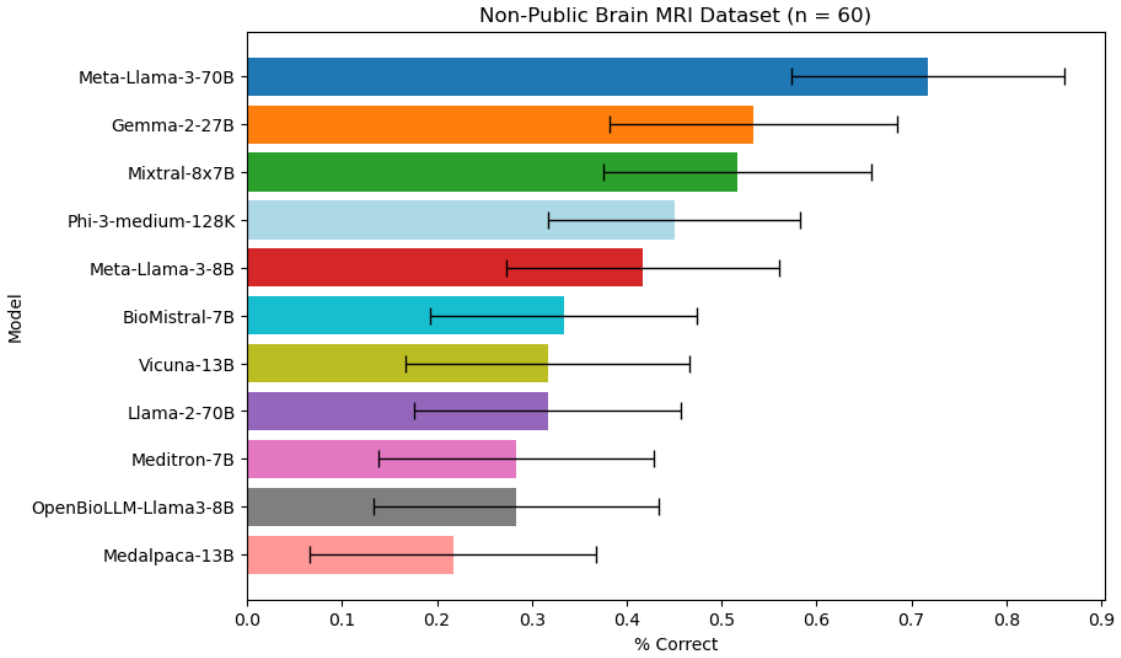

Figure 4: Performance of Open-Source LLMs in non-public brain MRI dataset (n = 60). Error bars indicate adjusted 95% confidence intervals.

Model	No. of Parameters	Link to Base Model
Meta-Llama-3-70B	70 billion	https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct
Gemma-2-27B	27 billion	https://huggingface.co/google/gemma-2-27b-it
Mixtral-8x7B	47 billion	https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1
Meta-Llama-3-8B	8 billion	https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
Phi-3-medium-128K	14 billion	https://huggingface.co/microsoft/Phi-3-medium-128k-instruct
Llama-2-70B	70 billion	https://huggingface.co/meta-llama/Llama-2-70b-chat-hf
Vicuna-13B	13 billion	https://huggingface.co/lmsys/vicuna-13b-v1.5
BioMistral-7B	7 billion	https://huggingface.co/BioMistral/BioMistral-7B
Meditron-7B	7 billion	https://huggingface.co/epfl-llm/meditron-7b
OpenBioLLM-Llama3-8B	8 billion	https://huggingface.co/aaditya/Llama3-OpenBioLLM-8B
Medalpaca-13B	13 billion	https://huggingface.co/medalpaca/medalpaca-13b

Table 1: Model details.

Subspecialty	No. Cases	Proportion
Abdominal imaging	890	22.0%
Neuroradiology	716	17.7%
Musculoskeletal system	577	14.3%
Chest imaging	406	10.0%
Paediatric radiology	320	7.9%
Uroradiology & genital male imaging	291	7.2%
Cardiovascular	237	5.9%
Head & neck imaging	235	5.8%
Genital (female) imaging	172	4.2%
Breast imaging	116	2.9%
Interventional radiology	89	2.2%
Total	4049	100.0%

Table 2: Dataset composition by subspecialty.

Discussion

In this study, we benchmarked the diagnostic performance of eleven leading open-source LLMs in a heterogeneous, challenging cohort of 4,049 peer-reviewed case reports from the Eurorad library. Meta's Llama-3-70B demonstrated superior performance, surpassing the other models across all radiological subspecialties with an overall accuracy of 75.1%. This level of performance is particularly noteworthy given the complexity and diversity of the cases included in our dataset.

These findings underscore the current dominance of Llama-3 among open-source models, consistent with its proficiency in other clinical tasks, such as answering close-ended medical questions, summarizing clinical documents, and patient education [14,18].

Importantly, this study assessed the diagnostic performance of LLMs based on real case descriptions, more accurately representing the complexities of real-life clinical decision-making than questions with pre-defined response options. This approach provides a more realistic evaluation of LLMs' potential in clinical settings, where the ability to interpret nuanced clinical information is crucial.

Our results revealed interesting variations in performance across radiological subspecialties, with higher accuracy in genital (female) imaging and lower accuracy in musculoskeletal imaging. These differences may reflect inherent complexities within each subspecialty, variations in the quality or specificity of case descriptions, or potential biases in the models' training data. Further investigation into these subspecialty-specific performance variations could provide valuable insights for targeted model improvements and clinical applications.

Interestingly, some lighter models such as Meta-Llama-3-8B exhibited strong performance, outperforming larger models with more parameters (e.g. Llama-2-70B, Vicuna-13B). This suggests that smaller, lower-cost models with nonetheless robust results are attainable, making the implementation of LLMs in resource-constrained healthcare settings more viable. The strong performance of smaller models highlights the importance of model architecture and training strategies, rather than just model size, in achieving high performance on specialized tasks.

Employing a state-of-the-art LLM model to automate the evaluation of LLM responses facilitated the large-scale analysis of thousands of cases, a scope unrealizable through manual processing. This strategy establishes a methodical benchmark for future large-scale investigations of clinical text documents.

<u>Limitations</u>

First, data contamination of LLMs cannot be definitively ruled out. Given the lack of transparency regarding the LLM training datasets, it is possible that the case reports used in this study overlap with the training data of some models. However, our complementary

assessment on a non-public brain MRI dataset revealed only a minor drop in performance, while the overall model rankings remained nearly identical.

Second, while the use of an LLM for the evaluation of LLM responses significantly enhanced the scalability of the analysis, it did so at the expense of reduced accuracy. To mitigate this limitation, we adjusted the standard error of model performance assessment based on our evaluation of Llama-3-70B's judging accuracy in a subset of the data.

Third, we did not investigate the impact of temperature settings or prompt design on LLM performance. To ensure deterministic responses, we applied a temperature of 0, but higher temperatures could potentially improve diagnostic accuracy [10]. Similarly, the optimal task-specific prompting strategy for radiological diagnosis is yet to be determined [19].

Finally, this study did not account for the influence of varying descriptions of the same case. A recent study evaluating GPT-4(V) in radiological diagnosis revealed that the image description is a major determinant of LLM accuracy [4]. The Eurorad case descriptions were written in awareness of the correct diagnosis, and their use of specific terminology or emphasis on certain image characteristics might have introduced a positive bias in LLM performance.

In conclusion, we found that several open-source LLMs demonstrate promising performance in identifying the correct diagnosis based on case descriptions from the Eurorad library, highlighting their potential as decision support tool for radiological differential diagnosis.

References

[1] Gertz RJ, Bunck AC, Lennartz S, Dratsch T, Iuga AI, Maintz D, et al. GPT-4 for Automated Determination of Radiologic Study and Protocol Based on Radiology Request Forms: A Feasibility Study. Radiology 2023;307. https://doi.org/10.1148/RADIOL.230877.

[2] Rau A, Rau S, Zöller D, Fink A, Tran H, Wilpert C, et al. A Context-based Chatbot Surpasses Radiologists and Generic ChatGPT in Following the ACR Appropriateness Guidelines. Radiology 2023;308. https://doi.org/10.1148/RADIOL.230970.

[3] Kottlors J, Bratke G, Rauen P, Kabbasch C, Persigehl T, Schlamann M, et al. Feasibility of Differential Diagnosis Based on Imaging Patterns Using a Large Language Model. Radiology 2023;308. https://doi.org/10.1148/radiol.231167.

[4] Schramm S, Preis S, Metz M-C, Jung K, Schmitz-Koep B, Zimmer C, et al. Impact of Multimodal Prompt Elements on Diagnostic Performance of GPT-4(V) in Challenging Brain MRI Cases. MedRxiv 2024:2024.03.05.24303767. https://doi.org/10.1101/2024.03.05.24303767.

[5] Mallio CA, Sertorio AC, Bernetti C, Beomonte Zobel B. Large language models for structured reporting in radiology: performance of GPT-4, ChatGPT-3.5, Perplexity and Bing. Radiologia Medica 2023;128:808–12. https://doi.org/10.1007/S11547-023-01651-4/.

[6] Doshi R, Amin KS, Khosla P, Bajaj S, Chheang S, Forman HP. Quantitative Evaluation of Large Language Models to Streamline Radiology Report Impressions: A Multimodal Retrospective Analysis. Radiology 2024;310. https://doi.org/10.1148/RADIOL.231593/.

[7] Guellec B Le, Lefèvre A, Geay C, Shorten L, Bruge C, Hacein-Bey L, et al. Performance of an Open-Source Large Language Model in Extracting Information from Free-Text Radiology Reports. Radiol Artif Intell 2024. https://doi.org/10.1148/RYAI.230364.

[8] Lehnen NC, Dorn F, Wiest IC, Zimmermann H, Radbruch A, Kather JN, et al. Data Extraction from Free-Text Reports on Mechanical Thrombectomy in Acute Ischemic Stroke Using ChatGPT: A Retrospective Analysis. Radiology 2024;311. https://doi.org/10.1148/RADIOL.232741.

[9] Katz U, Cohen E, Shachar E, Somer J, Fink A, Morse E, et al. GPT versus Resident Physicians — A Benchmark Based on Official Board Scores. NEJM AI 2024;1. https://doi.org/10.1056/AIDBP2300192.

[10] Suh PS, Shim WH, Suh CH, Heo H, Park CR, Eom HJ, et al. Comparing Diagnostic Accuracy of Radiologists versus GPT-4V and Gemini Pro Vision Using Image Inputs from Diagnosis Please Cases. Radiology 2024;312:e240273. https://doi.org/10.1148/RADIOL.240273.

[11] Sonoda Y, Kurokawa R, Nakamura Y, Kanzawa J, Kurokawa M, Ohizumi Y, et al. Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in “Diagnosis Please” cases. Jpn J Radiol 2024:1–5. https://doi.org/10.1007/S11604-024-01619-Y.

[12] Wu S, Koo M, Blum L, Black A, Kao L, Fei Z, et al. Benchmarking Open-Source Large Language Models, GPT-4 and Claude 2 on Multiple-Choice Questions in Nephrology. NEJM AI 2024;1. https://doi.org/10.1056/AIDBP2300092.

[13] Sandmann S, Riepenhausen S, Plagwitz L, Varghese J. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. Nat Commun 2024;15. https://doi.org/10.1038/S41467-024-46411-8.

[14] Adams LC, Truhn D, Busch F, Dorfner F, Nawabi J, Makowski MR, et al. Llama 3 Challenges Proprietary State-of-the-Art Large Language Models in Radiology Board–style Examination Questions. Radiology 2024;312. https://doi.org/10.1148/RADIOL.241191.

[15] Homepage | Eurorad n.d. https://eurorad.org/ (accessed August 19, 2024).

[16] Zheng L, Chiang W-L, Sheng Y, Zhuang S, Wu Z, Zhuang Y, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Adv Neural Inf Process Syst 2023;36:46595–623.

[17] Singh Thakur A, Choudhary K, Srinik Ramayapally V, Vaidyanathan S, Hupkes Meta D. Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges 2024.

[18] Liu F, Zhou H, Hua Y, Rohanian O, Thakur A, Clifton L, et al. Large Language Models in the Clinic: A Comprehensive Benchmark. MedRxiv 2024:2024.04.24.24306315. https://doi.org/10.1101/2024.04.24.24306315.

[19] Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing: Algorithm Development and Validation Study. JMIR Med Inform 2024;12:e55318. https://doi.org/10.2196/55318.