Clinical Imaging 118 (2025) 110382

Contents lists available at ScienceDirect

Clinical Imaging

journal homepage: www.elsevier.com/locate/clinimag

Imaging Physics and Informatics

Comparative diagnostic accuracy of GPT-4o and LLaMA 3-70b: Proprietary vs. open-source large language models in radiology☆

David Li[a], Kartik Gupta[a], Mousumi Bhaduri[a], Paul Sathiadoss[a], Sahir Bhatnagar[b], Jaron Chong[a,*]

[a] *Department of Medical Imaging, London Health Sciences Centre, 800 Commissioners Rd E, London, Ontario, Canada*
[b] *Department of Epidemiology, Biostatistics and Occupational Health, McGill University, 2001 McGill College, Montréal, Quebec, Canada*

ARTICLE INFO

Keywords:
Large language model
Generative artificial intelligence
Generative pre-trained transformer
Open source
Radiology

1. Introduction

The impressive capabilities of large language models (LLMs) to perform a wide range of cognitive tasks have garnered substantial attention from both the public and the radiology community.[1] Proprietary large language models (LLMs), such as ChatGPT, have dominated the spotlight due to the widespread perception that they outperform free, open-source LLMs, though they may not necessarily surpass deep learning models in certain tasks.[2] This study compares the diagnostic accuracy of GPT-4o and LLaMA 3-70b, which are representative of proprietary and open-source LLMs, respectively, due to their popularity and performance, in generating differential diagnoses for *Radiology* Diagnosis Please cases. These challenging cases capture the complexities of radiology, providing a rigorous evaluation of each model's ability to generate differential diagnoses across a diverse spectrum of pathologies.

2. Methods

This prospective study adheres to the Checklist for Artificial Intelligence in Medical Imaging and was exempt from research ethics board review due to the use of publicly available data. The *Radiology* Diagnosis Please dataset, comprising 287 cases from August 1998 to July 2023, excluding information leak cases, was utilized.[3] GPT-4o (OpenAI, USA) and LLaMA 3-70b (Meta, USA) were prompted to generate the top five differential diagnoses based on history, imaging findings, and both combined. Default hyperparameters were applied, except for temperature, which was set to 0. Three radiologists ([AUTHOR-6], 8; [AUTHOR-4], 8; [AUTHOR-3], 23 years of experience, respectively) independently evaluated the output of each LLM, with discrepancies resolved through mediated discussion. Exact McNemar tests were performed using the *statsmodels* Python package (version 0.14.2), with the null hypothesis that both LLMs have the same proportion of cases where their predictions disagree. The significance level was set at $\alpha = 0.05$.

3. Results

GPT-4o achieved diagnostic accuracies of 55/287 (19.2 %) for history, 133/287 (46.3 %) for imaging findings, and 160/287 (55.7 %) for history and imaging findings combined. LLaMA 3-70b achieved diagnostic accuracies of 58/287 (20.2 %) for history, 121/287 (42.2 %) for imaging findings, and 152/287 (53.0 %) for history and imaging findings combined. Both LLMs demonstrated comparable performance across all 10 evaluated subspecialties (Fig. 1, Table 1).

☆ Summary statement: LLaMA 3-70b achieved statistically equivalent diagnostic accuracy compared to the proprietary large language model GPT-4o on the *Radiology* Diagnosis Please differential diagnosis generation task, demonstrating substantial advancements in open-source LLM capabilities.

* Corresponding author at: Department of Medical Imaging, London Health Sciences Centre, 800 Commissioners Rd E, London, ON, Canada.
E-mail address: Jaron.Chong@lhsc.on.ca (J. Chong).

https://doi.org/10.1016/j.clinimag.2024.110382
Received 9 September 2024; Received in revised form 29 November 2024; Accepted 30 November 2024
Available online 2 December 2024

4. Discussion

Although the progressive improvement of proprietary frontier models is commonly assumed, the same cannot be said for open-source models, which often lack access to the same scale of training data, computational resources, and proprietary algorithmic innovations. However, both GPT-4o and LLaMA 3-70b demonstrated substantial diagnostic capabilities, surpassing the performance of earlier GPT generations.[4,5] While GPT-4o achieved slightly higher diagnostic accuracy compared to LLaMA 3-70b, the lack of statistical significance in the performance gap suggests equivalence in generating accurate differential diagnoses.

Open-source models offer several benefits. Freely available model weights with more permissive licensing for local inference provisioning promote greater collaboration and adoption. On-premises deployment of open-source models also addresses security and patient privacy

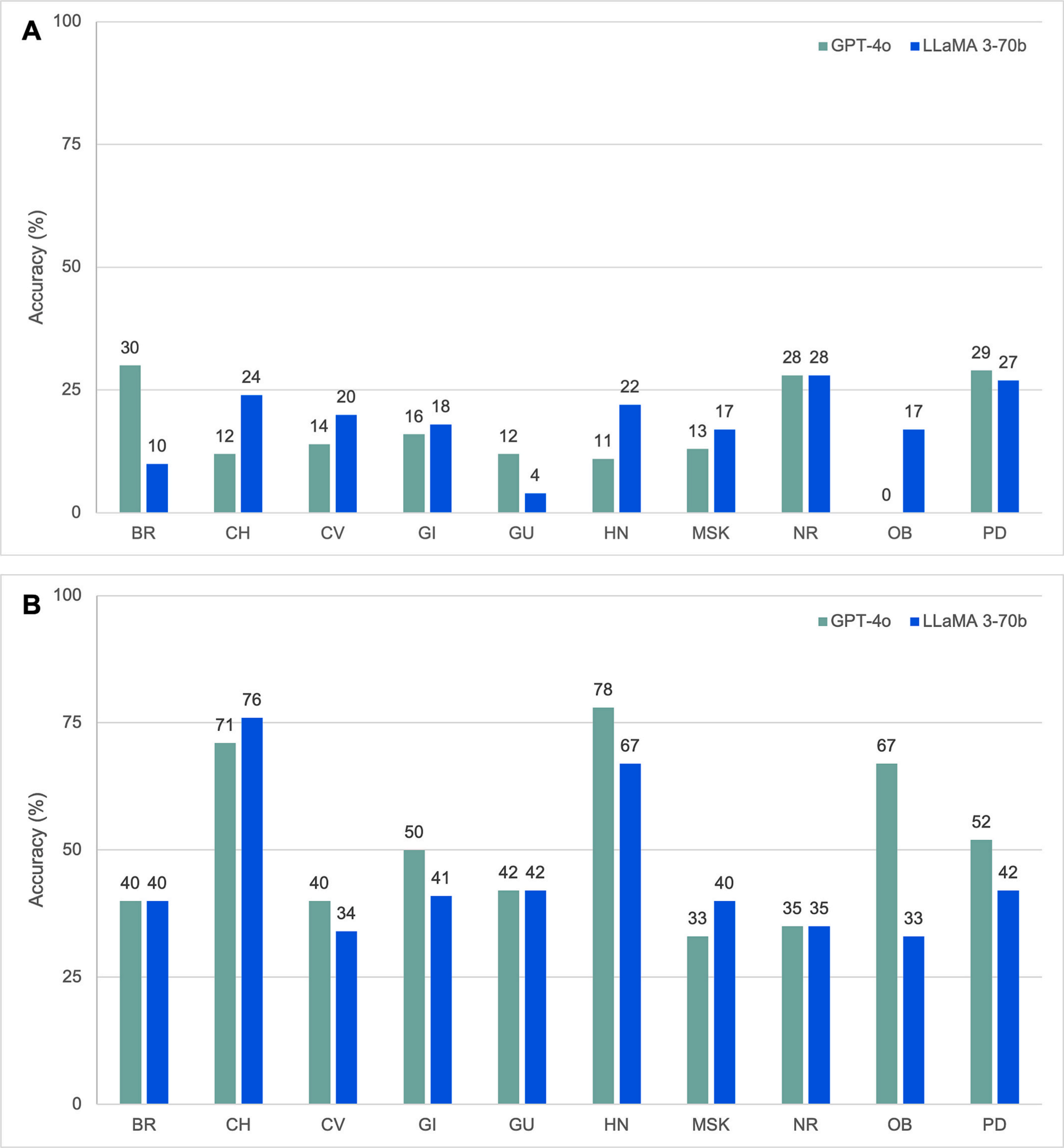

Fig. 1. Comparison of diagnostic accuracy of GPT-4o and LLaMA 3-70b on 287 *Radiology* Diagnosis Please cases using text-based inputs of A) history, B) imaging findings, and C) history and imaging findings combined. (BR: Breast, CV: Cardiovascular, CH: Chest, GI: Gastrointestinal, GU: Genitourinary, HN: Head & Neck, MSK: Musculoskeletal, NR: Neuroradiology, OB: Obstetrics, PD: Pediatric).

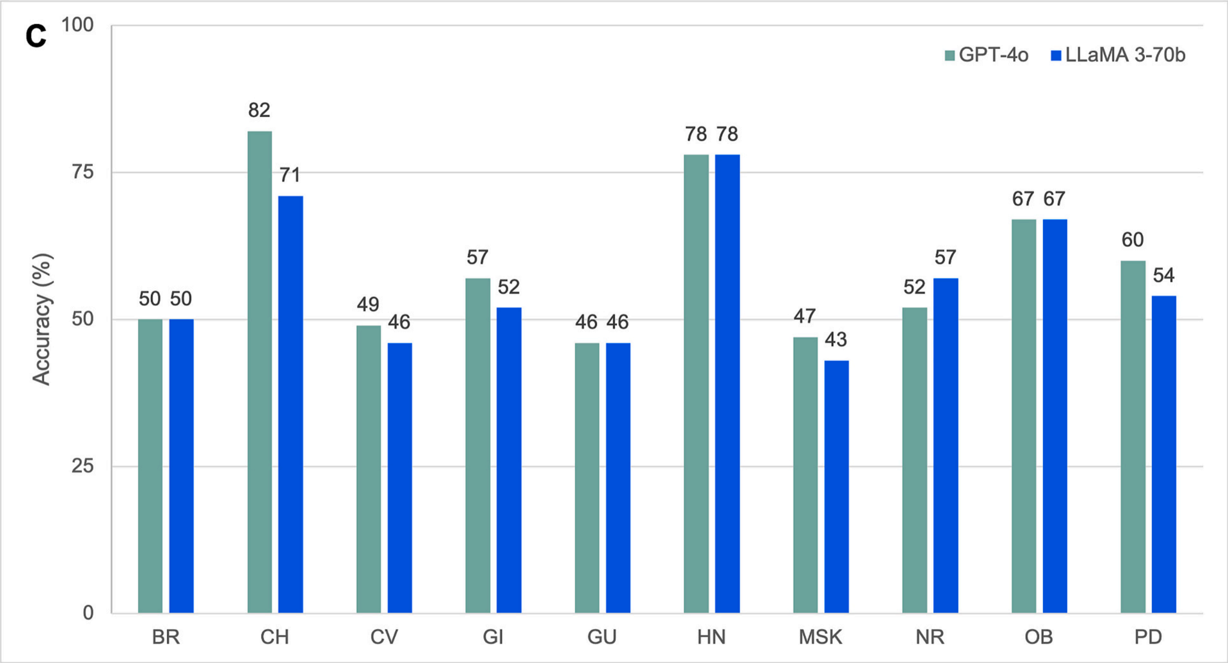

Fig. 1. (*continued*).

concerns associated with utilizing proprietary models in mission-critical clinical workflows.[6] Continued real-world benchmarking and validation efforts are essential to ensure the efficacy and reliability of LLMs in clinical practice.[7]

The primary limitation of this pilot study is the limited generalizability of its findings. By evaluating two representative LLMs, the study provides an initial comparison of proprietary and open-source models. Additionally, model performance was evaluated on a single diagnostic task, which may not encompass the full spectrum of pathologies encountered in clinical practice. Future studies comparing a larger number of models are necessary to validate these findings and to establish a more comprehensive understanding of LLM performance in radiology.[8]

Although this study focused on text-based cases, multimodal LLMs hold great promise in medical imaging. By integrating image and text analysis, these models can extract insights from both medical images and clinical data. Recent research has explored the application of multimodal LLMs for detecting radiologic findings on radiographs, revealing both their potential and associated challenges.[9,10] As multimodal LLMs continue to evolve, their integration into clinical workflows could enhance the accuracy and efficiency of radiologists. Future studies should focus on validating multimodal LLMs in clinical settings to fully realize their potential in medical imaging.

In conclusion, this pilot study challenges the prevailing assumption that proprietary LLMs outperform their open-source counterparts, providing empirical evidence of their comparable diagnostic accuracy in generating differential diagnoses for radiology cases. These findings highlight the potential for integrating high-performance open-source LLMs in clinical settings. Such a paradigm shift could lead to more accessible and cost-effective generative artificial intelligence applications in radiology, ultimately enhancing patient care.

CRediT authorship contribution statement

David Li: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kartik Gupta:** Writing – review & editing, Investigation, Formal analysis, Data curation,

Table 1
Overall and subspecialty diagnostic accuracy of GPT-4o and LLaMA 3-70b on 287 *Radiology* Diagnosis Please cases.

Subspecialty	History only			Imaging findings only			History and imaging findings		
	GPT-4o	LLaMA 3-70b	*P*-value	GPT-4o	LLaMA 3-70b	*P*-value	GPT-4o	LLaMA 3-70b	*P*-value
	55/287 (19.2)	58/287 (20.2)	0.74	133/287 (46.3)	121/287 (42.2)	0.16	160/287 (55.7)	152/287 (53.0)	0.38
Breast	3/10 (30)	1/10 (10)	0.5	4/10 (40)	4/10 (40)	1.0	5/10 (50)	5/10 (50)	1.0
Cardiovascular	2/17 (12)	4/17 (24)	0.5	12/17 (71)	13/17 (76)	1.0	14/17 (82)	12/17 (71)	0.69
Chest	5/35 (14)	7/35 (20)	0.5	14/35 (40)	12/35 (34)	0.73	17/35 (49)	16/35 (46)	1.0
Gastrointestinal	9/56 (16)	10/56 (18)	1.0	28/56 (50)	23/56 (41)	0.18	32/56 (57)	29/56 (52)	0.51
Genitourinary	3/26 (12)	1/26 (4)	0.5	11/26 (42)	11/26 (42)	1.0	12/26 (46)	12/26 (46)	1.0
Head & neck	1/9 (11)	2/9 (22)	1.0	7/9 (78)	6/9 (67)	1.0	7/9 (78)	7/9 (78)	1.0
Musculoskeletal	4/30 (13)	5/30 (17)	1.0	10/30 (33)	12/30 (40)	0.69	14/30 (47)	13/30 (43)	1.0
Neuroradiology	13/46 (28)	13/46 (28)	1.0	16/46 (35)	16/46 (35)	1.0	24/46 (52)	26/46 (57)	0.77
Obstetrical	0/6 (0)	1/6 (17)	1.0	4/6 (67)	2/6 (33)	0.5	4/6 (67)	4/6 (67)	1.0
Pediatric	15/52 (29)	14/52 (27)	1.0	27/52 (52)	22/52 (42)	0.18	31/52 (60)	28/52 (54)	0.58

Note.—Data are presented as the proportion of cases answered correctly, with percentages shown in parentheses. Each case was categorized by body system based on a review of the original case images and diagnoses. For cases involving multiple systems, the initiating body system was selected. *P*-values were derived from exact McNemar tests.

Conceptualization. **Mousumi Bhaduri:** Writing – review & editing, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Paul Sathiadoss:** Writing – review & editing, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sahir Bhatnagar:** Writing – review & editing, Validation, Methodology, Formal analysis, Conceptualization. **Jaron Chong:** Writing – review & editing, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Funding

No financial support was received for the research, authorship, and publication of this article.

Declaration of competing interest

DL, KG, MB, PS, SB: No relevant relationships.

JC: Member, Scientific Advisory Committee on Digital Health Technologies, Health Canada; Chair, Standing Committee on Artificial Intelligence, Canadian Association of Radiologists.

References

1 Shen Y, Heacock L, Elias J, et al. ChatGPT and other large language models are double-edged swords. Radiology 2023;307(2):e230163 [Jan 26].
2 Anas M, Saiyeda A, Sohail SS, Cambria E, Hussain A. Can generative AI models extract deeper sentiments as compared to traditional deep learning algorithms? IEEE Intelligent Systems 2024;39(2):5–10 [Apr 30].
3 Li D, Gupta K, Chong J. Evaluating diagnostic performance of ChatGPT in radiology: delving into methods. Radiology 2023;308(3):e232082 [Sep 19].
4 Li D, Gupta K, Bhaduri M, Sathiadoss P, Bhatnagar S, Chong J. Comparing GPT-3.5 and GPT-4 accuracy and drift in radiology diagnosis please cases. Radiology 2024; 310(1):e232411 [Jan 16].
5 Ueda D, Mitsuyama Y, Takita H, et al. Diagnostic performance of ChatGPT from patient history and imaging findings on the diagnosis please quizzes. Radiology 2023;308(1):e231040 [Jul 18].
6 Le Guellec B, Lefèvre A, Geay C, et al. Performance of an open-source large language model in extracting information from free-text radiology reports. Radiology Artificial Intelligence 2024;6(4):e230364 [May 8].
7 Xin KZ, Li D, Yi PH. Limited generalizability of deep learning algorithm for pediatric pneumonia classification on external data. Emerg Radiol 2022:1–7 [Feb 1].
8 Lone MR, Sohail SS. Comment on "Evaluation of responses to cardiac imaging questions by the artificial intelligence large language model ChatGPT". Clin Imaging 2024:114 [Oct 1].
9 Zhou Y, Ong H, Kennedy P, et al. Evaluating GPT-4V (GPT-4 with vision) on detection of radiologic findings on chest radiographs. Radiology 2024;311(2): e233270 [May 7].
10 Li D, Chong J. Laterality: a potential pitfall in applying multimodal large language models to radiology. Radiology 2024;313(2):e241421 [Nov 5].