JAMA Health Forum™

Research Letter | AI in Health Policy

# Comparison of Frontier Open-Source and Proprietary Large Language Models for Complex Diagnoses

Thomas A. Buckley, BS; Byron Crowe, MD; Raja-Elie E. Abdulnour, MD; Adam Rodman, MD, MPH; Arjun K. Manrai, PhD

## Introduction

Large language models (LLMs) now perform cognitive tasks in medicine previously believed to be the sole domain of humans, including accurately answering medical multiple-choice questions,[1] demonstrating nuanced clinical reasoning,[2] and establishing a robust differential diagnosis for complex diagnostic cases.[3] Since its release in March 2023, the GPT-4 (Generative Pre-trained Transformer 4) model (hereafter, *the closed-source LLM*) created by OpenAI has been among the best-performing LLMs on medical tasks and is being incorporated into health care applications. Although open-source LLMs have been available, they have generally not performed as well as proprietary models.[4] However, it remains unclear how newer open-source models, such as the 405-billion parameter Llama 3.1 model (Meta) (hereafter, *the open-source LLM*), perform. Such open-source frontier models, named for their superior performance on benchmarks, may now be competitive alternatives to closed-source models.

+ Related article

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

## Methods

We evaluated the open-source LLM from August 6 to August 10, 2024, on 70 challenging diagnostic cases used previously to assess the closed-source LLM (case set 1).[3] To mitigate risk of memorization, we retrieved 22 cases published between January 2024 and July 2024 (case set 2), after pretraining of the open-source LLM (December 2023). Cases come from the case records of the Massachusetts General Hospital series published by the *New England Journal of Medicine*; we obtained written permission from the Massachusetts Medical Society to use these cases. The models were not permitted to search the internet. The prompt is in the eAppendix in Supplement 1. A quality score[5] was assigned to outputs independently by B.C. and A.R.; discordance was measured by linear-weighted Cohen κ, and discordant scores were reconciled through discussion. *P* values were computed using the 2-sided McNemar test and deemed statistically significant at $P < .05$. Analysis was performed in R, version 4.3.2. This study was deemed not human subjects research and did not require institutional review board oversight per the Common Rule. This study followed the STROBE reporting guideline.[6]

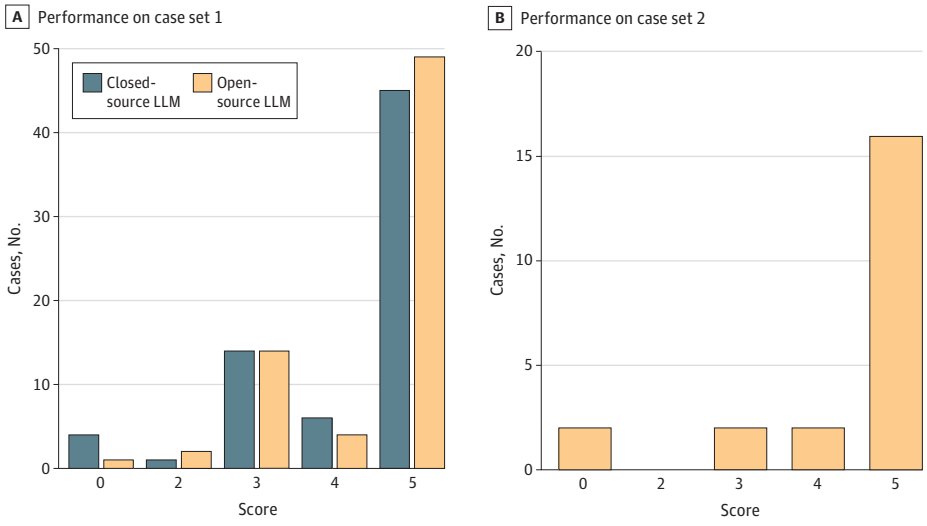
## Results

The reviewers agreed on 66% of differential quality scores (46 of 70; $\kappa = 0.39$) in the prior evaluation of the closed-source LLM and on 78% of cases (72 of 92; $\kappa = 0.69$) in this evaluation of the open-source LLM. We compared the distribution of differential diagnosis quality scores between the 2 models (**Figure**).[3,5] In case set 1, the open-source LLM included the final diagnosis in the differential for 70% of cases (49 of 70) compared with 64% (45 of 70) for the closed-source LLM ($P = .35$). The first suggestion from the open-source LLM was correct for 41% of cases (29 of 70) compared with 37% (26 of 70) for the closed-source LLM. For case set 2, the open-source LLM included the final diagnosis in the differential in 73% of cases (16 of 22) and identified the final diagnosis as the first suggestion in 45% of cases (10 of 22). Example cases in which the models diverged in diagnostic quality are shown in the **Table**.

## Discussion

For the first time, to our knowledge, an open-source LLM performed on par with GPT-4 in generating a differential diagnosis on complex diagnostic challenge cases. Our findings suggest an increasingly competitive landscape in LLM clinical decision support, and that institutions may be able to deploy high-performing custom models that run locally without sacrificing data privacy or flexibility.

**Figure. Performance of 2 Large Language Models (LLMs) on Complex Diagnostic Challenges**




Histogram comparison of differential diagnosis quality by the GPT-4 (Generative Pre-trained Transformer 4) model (closed-source LLM) and the 405-billion parameter Llama 3.1 model (open-source LLM) using the quality score metric on case set 1 (A) and case set 2 (B). The scoring used is a version adapted by Kanjee et al[3] from the original in Bond et al[5] (the original assigned inversely to the one presented here). The scale used is as follows: 5 = the actual diagnosis was suggested in the differential; 4 = the suggestions included something very close, but not exact; 3 = the suggestions included something closely related that might have been helpful; 2 = the suggestions included something related, but unlikely to be helpful; and 0 = no suggestions close to the target diagnosis.

**Table. Example Cases Using Both LLMs**

| | List of diagnoses | |
|---|---|---|
| **Final diagnosis** | **Closed-source LLM** | **Open-source LLM** |
| **A. Case 26-2022: open-source LLM identifies correct diagnosis; closed-source LLM, incorrect** | | |
| Langerhans cell histiocytosis of the lung and liver | 1. Eosinophilic granulomatosis with polyangiitis<br>2. Hypereosinophilic syndrome<br>3. Eosinophilic pneumonia<br>4. Parasitic infection (eg, strongyloidiasis or visceral larva migrans)<br>5. Disseminated fungal infection (eg, histoplasmosis or coccidioidomycosis)<br>6. Lymphoma with eosinophilic infiltration<br>7. Metastatic cancer with eosinophilic infiltration<br>8. Sarcoidosis with eosinophilic infiltration<br>9. Drug-induced eosinophilic syndrome<br>Quality score: 0 | 1. Langerhans cell histiocytosis<br>2. Metastatic cancer (eg, breast, thyroid, or colon cancer)<br>3. Sarcoidosis<br>4. Tuberculosis<br>5. Fungal infection (eg, histoplasmosis or cryptococcosis)<br>6. Parasitic infection (eg, echinococcosis or toxocariasis)<br>7. Granulomatosis with polyangiitis<br>8. Erdheim-Chester disease<br>9. Rosai-Dorfman disease<br>10. Idiopathic hypereosinophilic syndrome<br>Quality score: 5 |
| **B. Case 4-2022: closed-source LLM identifies correct diagnosis; open-source LLM, incorrect** | | |
| Neurosyphilis | 1. Cogan syndrome<br>2. Granulomatosis with polyangiitis (Wegener)<br>3. Sarcoidosis<br>4. Susac syndrome<br>5. Behcet disease<br>6. Vogt-Koyanagi-Harada syndrome<br>7. Relapsing polychondritis<br>8. Meniere disease<br>9. Lyme disease<br>10. Syphilis<br>11. Autoimmune inner ear disease<br>12. Acute viral labyrinthitis<br>Quality score: 5 | 1. Cogan syndrome<br>2. Sarcoidosis<br>3. Wegener granulomatosis (granulomatosis with polyangiitis)<br>4. Relapsing polychondritis<br>5. Rheumatoid arthritis with ocular and auditory manifestations<br>6. Lupus erythematosus with ocular and auditory manifestations<br>7. Giant cell arteritis<br>8. Polyarteritis nodosa<br>9. Churg-Strauss syndrome (eosinophilic granulomatosis with polyangiitis)<br>10. Endolymphatic sac tumor (unlikely, but possible given the hearing loss and vestibular symptoms)<br>Quality score: 2 |

Abbreviations: closed-source LLM, Generative Pre-trained Transformer 4 model; LLM, large language model; open-source LLM, 405-billion parameter Llama 3.1 model.

This study has several limitations. First, details about the training data for LLMs are not publicly released. However, the open-source LLM exhibited similar performance on 22 cases published after pretraining ended. Second, these evaluations consider a single diagnostic task in which a highly informative summary of a clinical case is provided, which may not reflect the breadth of clinical reasoning. Future evaluations should use cases from clinical settings, including electronic health records. Third, this study had limited statistical power to assess differences between LLMs. Overall, the availability of an open-source LLM that performs on par with a leading proprietary model may both broaden access and expand possibilities of fine-tuning custom LLMs for integration in clinical care settings.

## ARTICLE INFORMATION

**Accepted for Publication:** January 9, 2025.

**Published:** March 14, 2025. doi:10.1001/jamahealthforum.2025.0040



**Corresponding Author:** Arjun K. Manrai, PhD, Department of Biomedical Informatics, Harvard Medical School, 10 Shattuck St, Boston, MA 02115 (arjun_manrai@hms.harvard.edu).

**Author Affiliations:** Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts (Buckley, Manrai); Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts (Crowe, Rodman); Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, Massachusetts (Abdulnour).

**Author Contributions:** Mr Buckley and Dr Manrai had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Drs Rodman and Manrai contributed equally to this work.

*Concept and design*: Buckley, Abdulnour, Rodman, Manrai.

*Acquisition, analysis, or interpretation of data*: All authors.

*Drafting of the manuscript*: All authors.

*Critical review of the manuscript for important intellectual content:* Crowe, Abdulnour, Rodman, Manrai.

*Statistical analysis:* Buckley, Rodman, Manrai.

*Obtained funding*: Manrai.

*Administrative, technical, or material support:* Crowe, Abdulnour, Rodman.

*Supervision:* Crowe, Abdulnour, Rodman, Manrai.

**Conflict of Interest Disclosures:** Dr Crowe reported receiving personal fees from Solera Health outside the submitted work. Dr Rodman reported receiving grants from the Gordon and Betty Moore Foundation outside the submitted work. No other disclosures were reported.

**Funding/Support:** This project was supported by award K01HL138259 from the National Heart, Lung, and Blood Institute and a Harvard Medical School Dean's Innovation Award.

**Role of the Funder/Sponsor:** The funding sources had no role in the design and conduct of the study; preparation or approval of the manuscript; and decision to submit the manuscript for publication.

**Data Sharing Statement:** See Supplement 2.

## REFERENCES

1. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. *arXiv*. Preprint posted online March 20, 2023. http://arxiv.org/abs/2303.13375

2. Cabral S, Restrepo D, Kanjee Z, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians. *JAMA Intern Med*. 2024;184(5):581-583. doi:10.1001/jamainternmed.2024.0295

3. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA*. 2023;330(1):78-80. doi:10.1001/jama.2023.8288

4. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med*. 2024;30(9):2613-2622. doi:10.1038/s41591-024-03097-1

5. Bond WF, Schwartz LM, Weaver KR, Levick D, Giuliano M, Graber ML. Differential diagnosis generators: an evaluation of currently available computer programs. *J Gen Intern Med*. 2012;27(2):213-219. doi:10.1007/s11606-011-1804-8

6. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet*. 2007;370(9596):1453-1457. doi:10.1016/S0140-6736(07)61602-X

**SUPPLEMENT 1.**
**eAppendix.**

**SUPPLEMENT 2.**
**Data Sharing Statement**