


 Search models, datasets, users...

[← Back to Articles](#)

The Open Medical-LLM Leaderboard: Benchmarking Large Language Models in Healthcare

Published April 19, 2024

[Update on GitHub](#)

▲ Upvote **181**

+175



Aaditya Ura

[aaditya](#)

Follow

guest



Pasquale Minervini

[pminervini](#)

Follow

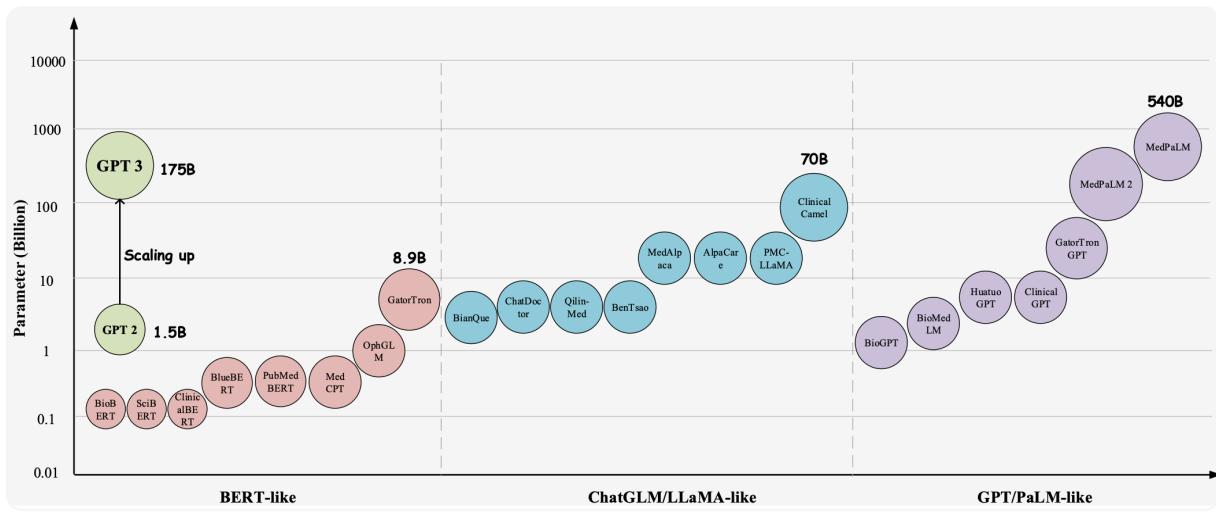
guest



Clémentine Fourrier

[clefourrier](#)

Follow



Over the years, Large Language Models (LLMs) have emerged as a groundbreaking technology with immense potential to revolutionize various aspects of healthcare. These models, such as GPT-3, GPT-4 and Med-PaLM 2 have demonstrated remarkable capabilities in understanding and generating human-like text, making them valuable tools for tackling complex medical tasks and improving patient care.

They have notably shown promise in various medical applications, such as medical question-answering (QA), dialogue systems, and text generation. Moreover, with the exponential growth of electronic health records (EHRs), medical literature, and patient-generated data, LLMs could help healthcare professionals extract valuable insights and make informed decisions.

However, despite the immense potential of Large Language Models (LLMs) in healthcare, there are significant and specific challenges that need to be addressed.

When models are used for recreational conversational aspects, errors have little repercussions; this is not the case for uses in the medical domain however, where wrong explanation and answers can have severe consequences for patient care and outcomes. The accuracy and reliability of information provided by language models can be a matter of life or death, as it could potentially affect healthcare decisions, diagnosis, and treatment plans.

For example, when given a medical query (see below), GPT-3 incorrectly recommended tetracycline for a pregnant patient, despite correctly explaining its contraindication due to potential harm to the fetus. Acting on this incorrect recommendation could lead to bone growth problems in the baby.

Question

A 24-year-old pregnant woman at 28 weeks gestation presents to the emergency department with complaints of fever with chills and pain in her knee and ankle joints for the past 2 days. She also complains of headaches and difficulty moving her neck. Further questioning reveals that she had a tick bite on her arm while gardening a few days ago. Pregnancy is progressing normally. On examination, an erythematous rash is seen on her right arm. Her obstetric examination is normal. Ultrasound of the fetus is reassuring with a normal heartbeat and no gross abnormalities. A specimen is collected to test for Lyme disease. What is the next best step for this patient?

Options

A) Ibuprofen

B) Tetracycline

C) Amoxicillin

D) Gentamicin

Physician

Amoxicillin ✓

Explanation: Amoxicillin is a safe and effective treatment for Lyme disease in pregnancy. Ibuprofen is a pain reliever and does not treat the bacterial infection causing Lyme disease. Tetracycline antibiotic is not recommended during pregnancy; it can cross the placenta and affect fetal development, including causing discoloration of the baby's teeth and potential bone growth issues. Gentamicin, a potent antibiotic generally reserved for severe infections, is not typically used for treating Lyme disease.

GPT-3.5

Tetracycline ✗

Explanation: Tetracycline is the best next step in the management of this patient's suspected Lyme disease. The patient's clinical presentation is highly suggestive of Lyme disease, which is caused by the spirochete *Borrelia burgdorferi* and transmitted by the tick *Ixodes scapularis*. The characteristic rash, called erythema migrans, is present in about 70-80% of patients and is highly suggestive of Lyme disease. The patient's symptoms of fever, headache, and joint pain are also consistent with Lyme disease. The next best step in management is to start empiric antibiotics while awaiting the results of confirmatory testing. Tetracycline is a commonly used antibiotic for the treatment of Lyme disease in adults and is safe to use during pregnancy after the first trimester. It has been shown to be effective in eradicating *B. burgdorferi* and preventing the progression of Lyme disease. Ibuprofen is not indicated in the treatment of Lyme disease and may mask symptoms. Amoxicillin is not indicated in the treatment of Lyme disease and may mask symptoms. Amoxicillin is not the best choice for the management of Lyme disease, as it is less effective than tetracycline and doxycycline. Gentamicin is not an effective treatment for Lyme disease.

To fully utilize the power of LLMs in healthcare, it is crucial to develop and benchmark models using a setup specifically designed for the medical domain. This setup should take into account the unique characteristics and requirements of healthcare data and applications. The development of methods to evaluate the Medical-LLM is not just of academic interest but of practical importance, given the real-life risks they pose in the healthcare sector.

Error

This space is experiencing an issue.

Please contact the author of the space to let them know.

The Open Medical-LLM Leaderboard aims to address these challenges and limitations by providing a standardized platform for evaluating and comparing the performance of various large language models on a diverse range of medical tasks and datasets. By offering a comprehensive assessment of each model's medical knowledge and question-answering capabilities, the leaderboard aims to foster the development of more effective and reliable medical LLMs.

This platform enables researchers and practitioners to identify the strengths and weaknesses of different approaches, drive further advancements in the field, and ultimately contribute to better patient care and outcomes.

🔗 Datasets, Tasks, and Evaluation Setup

The Medical-LLM Leaderboard includes a variety of tasks, and uses accuracy as its primary evaluation metric (accuracy measures the percentage of correct answers provided by a language model across the various medical QA datasets).

🔗 MedQA

The MedQA dataset consists of multiple-choice questions from the United States Medical Licensing Examination (USMLE). It covers general medical knowledge and includes 11,450 questions in the development set and 1,273 questions in the test set. Each question has 4 or 5 answer choices, and the dataset is designed to assess the medical knowledge and reasoning skills required for medical licensure in the United States.

Format: Q + A, multiple choice, open domain

Size (Dev/Test): 11450 / 1273

Question: A 65-year-old man with hypertension comes to the physician for a routine health maintenance examination. Current medications include atenolol, lisinopril, and atorvastatin. His pulse is 86/min, respirations are 18/min, and blood pressure is 145/95 mm Hg. Cardiac examination reveals end diastolic murmur. Which of the following is the most likely cause of this physical examination?

Answer: (A) Decreased compliance of the left ventricle (B) Myxomatous degeneration of the mitral valve (C) Inflammation of the pericardium (D) Dilatation of the aortic root (E) Thickening of the mitral valve leaflets

🔗 MedMCQA

MedMCQA is a large-scale multiple-choice QA dataset derived from Indian medical entrance examinations (AIIMS/NEET). It covers 2.4k healthcare topics and 21 medical subjects, with over 187,000 questions in the development set and 6,100 questions in the test set. Each question has 4 answer choices and is accompanied by an explanation. MedMCQA evaluates a model's general medical knowledge and reasoning capabilities.

Format: Q + A, multiple choice, open domain

Size (Dev/Test): 187000 / 6100

Question: Which of the following ultrasound findings has the highest association with aneuploidy?

Answer: (A) Choroid plexus cyst (B) Nuchal translucency **(C) Cystic hygroma** (D) Single umbilical artery

Explanation: All the above mentioned are ultrasound findings associated with increased risk of aneuploidy although the highest association is seen with cystic hygroma. Nuchal translucency and cystic hygroma are both measured in the first trimester. Trisomy 21 is the most common aneuploidy associated with increased NT and cystic hygroma while monosomy X presents as second-trimester hygroma.

🔗 PubMedQA

PubMedQA is a closed-domain QA dataset, In which each question can be answered by looking at an associated context (PubMed abstract). It consists of 1,000 expert-labeled question-answer pairs. Each question is accompanied by a PubMed abstract as context, and the task is to provide a yes/no/maybe answer based on the information in the abstract. The dataset is split into 500 questions for development and 500 for testing. PubMedQA assesses a model's ability to comprehend and reason over scientific biomedical literature.

Format: Q + A + context, multiple choice, closed domain

Size (Dev/Test): 500 / 500

Question: Double balloon enteroscopy: is it efficacious and safe in a community setting?

Context: From March 2007 to January 2011, 88 DBE procedures were performed on 66 patients. Indications included evaluation anemia/gastrointestinal bleed, small bowel IBD and dilation of strictures. Video-capsule endoscopy (VCE) was used prior to DBE in 43 of the 66 patients prior to DBE evaluation. The mean age was 62 years. Thirty-two patients were female, 15 were African-American; 44 antegrade and 44 retrograde DBEs were performed. The mean time per antegrade DBE was 107.4 ± 30.0 minutes with a distance of 318.4 ± 152.9 cm reached past the pylorus. The mean time per lower DBE was 100.7 ± 27.3 minutes with 168.9 ± 109.1 cm meters past the ileocecal valve reached. Endoscopic therapy in the form of electrocautery to ablate bleeding sources was performed in 20 patients (30.3%), biopsy in 17 patients (25.8%) and dilation of Crohn's-related small bowel strictures in 4 (6.1%). 43 VCEs with pathology noted were performed prior to DBE, with findings endoscopically confirmed in 32 cases (74.4%). In 3 cases the DBE showed findings not noted on VCE.

Answer: Yes

Long Answer: DBE appears to be equally safe and effective when performed in the community setting as compared to a tertiary referral center with a comparable yield, efficacy, and complication rate.

🔗 MMLU Subsets (Medicine and Biology)

The MMLU benchmark (Measuring Massive Multitask Language Understanding) includes multiple-choice questions from various domains. For the Open Medical-LLM Leaderboard, we focus on the subsets most relevant to medical knowledge:

- Clinical Knowledge: 265 questions assessing clinical knowledge and decision-making skills.
- Medical Genetics: 100 questions covering topics related to medical genetics.
- Anatomy: 135 questions evaluating the knowledge of human anatomy.
- Professional Medicine: 272 questions assessing knowledge required for medical professionals.
- College Biology: 144 questions covering college-level biology concepts.
- College Medicine: 173 questions assessing college-level medical knowledge.

Each MMLU subset consists of multiple-choice questions with 4 answer options and is designed to evaluate a model's understanding of specific medical and biological domains.

Format: Q + A, multiple choice, open domain

Anatomy	Size (Dev/Test): 14 / 135 Question: Which of the following controls body temperature, sleep, and appetite? Answer: (A) Adrenal glands (B) Hypothalamus (C) Pancreas (D) Thalamus
Clinical Knowledge	Size (Dev/Test): 29 / 265 Question: The following are features of Alzheimer's disease except: Answer: (A) short-term memory loss. (B) confusion. (C) poor attention. (D) drowsiness .
College Medicine	Size (Dev/Test): 22 / 173 Question: The main factors determining success in sport are: Answer: (A) a high energy diet and large appetite. (B) high intelligence and motivation to succeed. (C) a good coach and the motivation to succeed. (D) innate ability and the capacity to respond to the training stimulus .
Medical Genetics	Size (Dev/Test): 11 / 100 Question: The allele associated with sickle cell anemia apparently reached a high frequency in some human populations due to: Answer: (A) random mating (B) superior fitness of heterozygotes in areas where malaria was present (C) migration of individuals with the allele into other populations (D) a high mutation rate at that specific gene.
Professional Medicine	Size (Dev/Test): 31 / 272 Question: A 19-year-old woman noticed a mass in her left breast 2 weeks ago while doing monthly breast self-examination. Her mother died of metastatic breast cancer at the age of 40 years. Examination shows large dense breasts; a 2-cm, firm, mobile mass is palpated in the upper outer quadrant of the left breast. There are no changes in the skin or nipple, and there is no palpable axillary adenopathy. Which of the following is the most likely diagnosis? Answer: (A) Fibroadenoma (B) Fibrocystic changes of the breast (C) Infiltrating ductal carcinoma (D) Intraductal papilloma
College Biology	Size (Dev/Test): 16 / 144 Question: Which of the following is the most direct cause of polyteny in somatic cells of certain organisms? Answer: (A) RNA transcription (B) Supercoiling of chromatin (C) Chromosome replication without cell division (D) Chromosome recombination

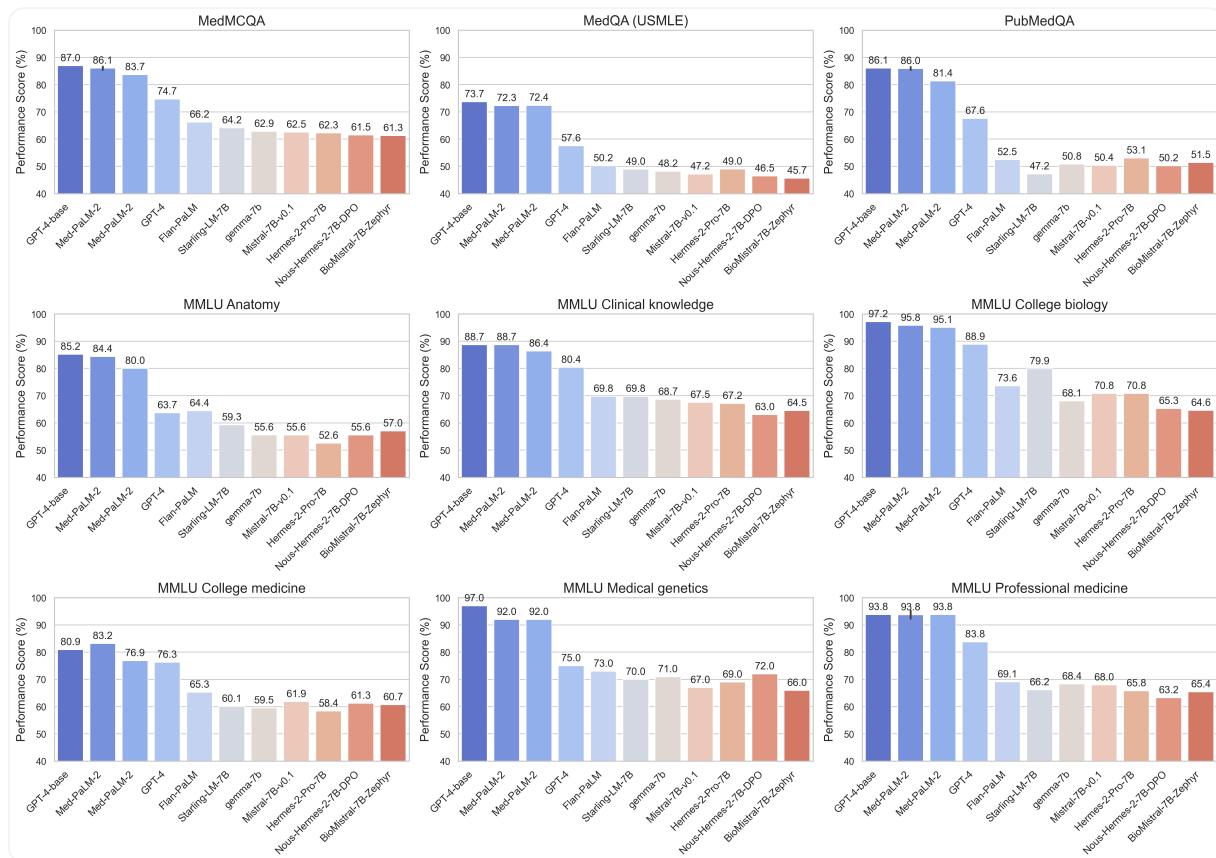
The Open Medical-LLM Leaderboard offers a robust assessment of a model's performance across various aspects of medical knowledge and reasoning.

🔗 Insights and Analysis

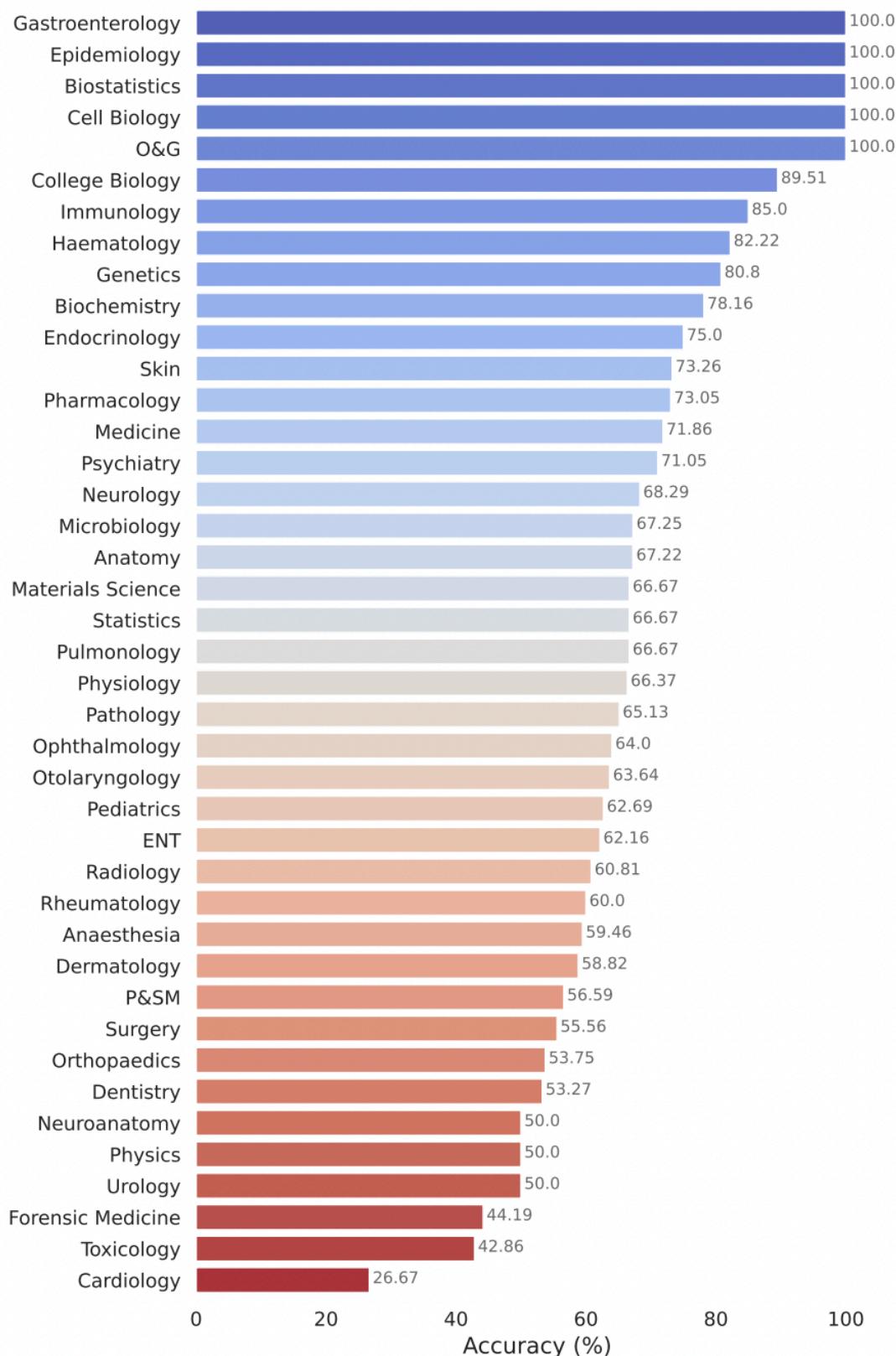
The Open Medical-LLM Leaderboard evaluates the performance of various large language models (LLMs) on a diverse set of medical question-answering tasks. Here are our key findings:

- Commercial models like GPT-4-base and Med-PaLM-2 consistently achieve high accuracy scores across various medical datasets, demonstrating strong performance in different medical domains.

- Open-source models, such as Starling-LM-7B, gemma-7b, Mistral-7B-v0.1, and Hermes-2-Pro-Mistral-7B, show competitive performance on certain datasets and tasks, despite having smaller sizes of around 7 billion parameters.
- Both commercial and open-source models perform well on tasks like comprehension and reasoning over scientific biomedical literature (PubMedQA) and applying clinical knowledge and decision-making skills (MMLU Clinical Knowledge subset).



Google's model, Gemini Pro demonstrates strong performance in various medical domains, particularly excelling in data-intensive and procedural tasks like Biostatistics, Cell Biology, and Obstetrics & Gynecology. However, it shows moderate to low performance in critical areas such as Anatomy, Cardiology, and Dermatology, revealing gaps that require further refinement for comprehensive medical application.



🔗 Submitting Your Model for Evaluation

To submit your model for evaluation on the Open Medical-LLM Leaderboard, follow these steps:

1. Convert Model Weights to Safetensors Format

First, convert your model weights to the safetensors format. Safetensors is a new format for storing weights that is safer and faster to load and use. Converting your model to this format will also allow the leaderboard to display the number of parameters of your model in the main table.

2. Ensure Compatibility with AutoClasses

Before submitting your model, make sure you can load your model and tokenizer using the AutoClasses from the Transformers library. Use the following code snippet to test the compatibility:

```
from transformers import AutoConfig, AutoModel, AutoTokenizer
config = AutoConfig.from_pretrained(MODEL_HUB_ID)
model = AutoModel.from_pretrained("your model name")
tokenizer = AutoTokenizer.from_pretrained("your model name")
```

If this step fails, follow the error messages to debug your model before submitting it. It's likely that your model has been improperly uploaded.

3. Make Your Model Public

Ensure that your model is publicly accessible. The leaderboard cannot evaluate models that are private or require special access permissions.

4. Remote Code Execution (Coming Soon)

Currently, the Open Medical-LLM Leaderboard does not support models that require `use_remote_code=True`. However, the leaderboard team is actively working on adding this feature, so stay tuned for updates.

5. Submit Your Model via the Leaderboard Website

Once your model is in the safetensors format, compatible with AutoClasses, and publicly accessible, you can submit it for evaluation using the "Submit here!" panel on the Open Medical-LLM Leaderboard website. Fill out the required information,

such as the model name, description, and any additional details, and click the submit button.

The leaderboard team will process your submission and evaluate your model's performance on the various medical QA datasets. Once the evaluation is complete, your model's scores will be added to the leaderboard, allowing you to compare its performance with other submitted models.

What's next? Expanding the Open Medical-LLM Leaderboard

The Open Medical-LLM Leaderboard is committed to expanding and adapting to meet the evolving needs of the research community and healthcare industry. Key areas of focus include:

1. Incorporating a wider range of medical datasets covering diverse aspects of healthcare, such as radiology, pathology, and genomics, through collaboration with researchers, healthcare organizations, and industry partners.
2. Enhancing evaluation metrics and reporting capabilities by exploring additional performance measures beyond accuracy, such as Pointwise score and domain-specific metrics that capture the unique requirements of medical applications.
3. A few efforts are already underway in this direction. If you are interested in collaborating on the next benchmark we are planning to propose, please join our [Discord community](#) to learn more and get involved. We would love to collaborate and brainstorm ideas!

If you're passionate about the intersection of AI and healthcare, building models for the healthcare domain, and care about safety and hallucination issues for medical LLMs, we invite you to join our vibrant [community on Discord](#).

Credits and Acknowledgments



Special thanks to all the people who helped make this possible, including Clémentine Fourrier and the Hugging Face team. I would like to thank Andreas Motzfeldt, Aryo Gema, & Logesh Kumar Umapathi for their discussion and feedback on the leaderboard during development. Sincere gratitude to Prof. Pasquale Minervini for his time, technical assistance, and for providing GPU support from the University of Edinburgh.

🔗 About Open Life Science AI

Open Life Science AI is a project that aims to revolutionize the application of Artificial intelligence in the life science and healthcare domains. It serves as a central hub for lists of medical models, datasets, benchmarks, and tracking conference deadlines, fostering collaboration, innovation, and progress in the field of AI-assisted healthcare. We strive to establish Open Life Science AI as the premier destination for anyone interested in the intersection of AI and healthcare. We provide a platform for researchers, clinicians, policymakers, and industry experts to engage in dialogues, share insights, and explore the latest developments in the field.



Open Life Science AI

🔗 Citation

If you find our evaluations useful, please consider citing our work

Medical-LLM Leaderboard

```
@misc{Medical-LLM Leaderboard,  
author = {Ankit Pal, Pasquale Minervini, Andreas Geert Motzfeldt, Aryo  
title = {openlifescienceai/open_medical_llm_leaderboard},  
year = {2024},  
publisher = {Hugging Face},  
howpublished = "\url{https://huggingface.co/spaces/openlifescienceai/o  
}
```

More Articles from our Blog

Big Bench Audio

Evaluating Audio Reasoning with Big Bench Audio

Leaderboards on the Hub

Introduction to the Open Leaderboard for Japanese LLMs

🤝 Community

ggbobao Apr 24

Hello everyone, I would like to ask how long it takes from submitting the model to completing the test



1

[Reply](#)[Edit](#)[Preview](#)

Start discussing this article

Tap or paste here to upload images

Comment

[Sign up](#) or [log in](#) to comment

[System theme](#)

Company

[TOS](#)[Privacy](#)[About](#)[Jobs](#)

Website

[Models](#)[Datasets](#)[Spaces](#)[Pricing](#)[Docs](#)