Original Article

Performance Benchmarking of Open-Source Large Language Models on the Brazilian Society of Cardiology's Certification Exam

João Victor Bruneti Severino,[1,2] Matheus Nespolo Berger,[1] Pedro Angelo Basei de Paula,[1] Filipe Silveira Loures,[1] Solano Amadori Todeschini,[1] Eduardo Augusto Roeder,[1] Maria Han Veiga,[3] José Knopfholz,[2] Gustavo Lenci Marques[1,2]

Universidade Federal do Paraná,[1] Curitiba, PR – Brazil
Pontificia Universidade Catolica do Paraná,[2] Curitiba, PR – Brazil
Ohio State University Foundation,[3] Columbus, Ohio – United States of America

Abstract

Background: **Large language models (LLMs) have made a significant impact in medicine and demonstrate substantial promise for further development. However, most of the existing research has predominantly centered on English-language tasks with lower medical complexity. This underscores the importance of investigating the performance of state-of-the-art LLMs in more complex specialties, such as cardiology, and in languages beyond English, such as Portuguese.**

Objective: **This study aimed to evaluate and compare leading LLMs based on their performance on the validated cardiology knowledge assessed by the Brazilian Society of Cardiology's (SBC) Certification Exam.**

Methods: **This study conducted a comparative analysis of 23 LLMs in the context of the SBC's Certification Exam. The exam consists of 100 multiple-choice questions, 20 of which include images that cannot be processed by all LLMs. Therefore, these image-based questions were excluded from the analysis.**

Results: **Proprietary LLMs showed a varying performance, with GPT-4o achieving the highest success rate at 62.25%, followed by Claude Opus at 60.25%. In the medium-sized model category (up to 100 billion parameters), Claude Haiku reached 47.25%. Among open-source models, Llama3 70B Instruct scored 53.50% in the large model category (over 100 billion parameters), while Llama3 8B achieved 36.25% in the small model category (under 20 billion parameters).**

Conclusions: **Both proprietary and open-source LLMs underperformed on the test, failing to meet the exam's cutoff score. Although larger models generally achieved better results, some medium-sized models — such as Llama3 70B Instruct and Claude Haiku—showed noteworthy results. The LLMs lacked specialized knowledge in cardiology and faced challenges in understanding Portuguese, revealing a significant gap in current AI capabilities and emphasizing the need for improvements.**

Keywords: **Artificial Intelligence; Cardiology; Benchmarking.**

Introduction

Artificial intelligence (AI), particularly large language models (LLMs), is transforming numerous fields, including medicine.[1,2] These advanced platforms demonstrate significant potential in supporting diagnosis,[3-5] enhancing medical education,[6] and advancing research in disease management[7] and decision-making.[8] Recent studies have explored the performance of LLMs on medical proficiency tests to evaluate their practical applications.[9-13] These tests simulate real-world clinical scenarios, allowing for the assessment of the accuracy, relevance, and practical utility of LLM-generated responses in healthcare settings. However, most of these studies have focused on English[9,10,12,13] or Chinese,[14,15] with limited research addressing other languages, such as Arabic or Spanish.[16,17]

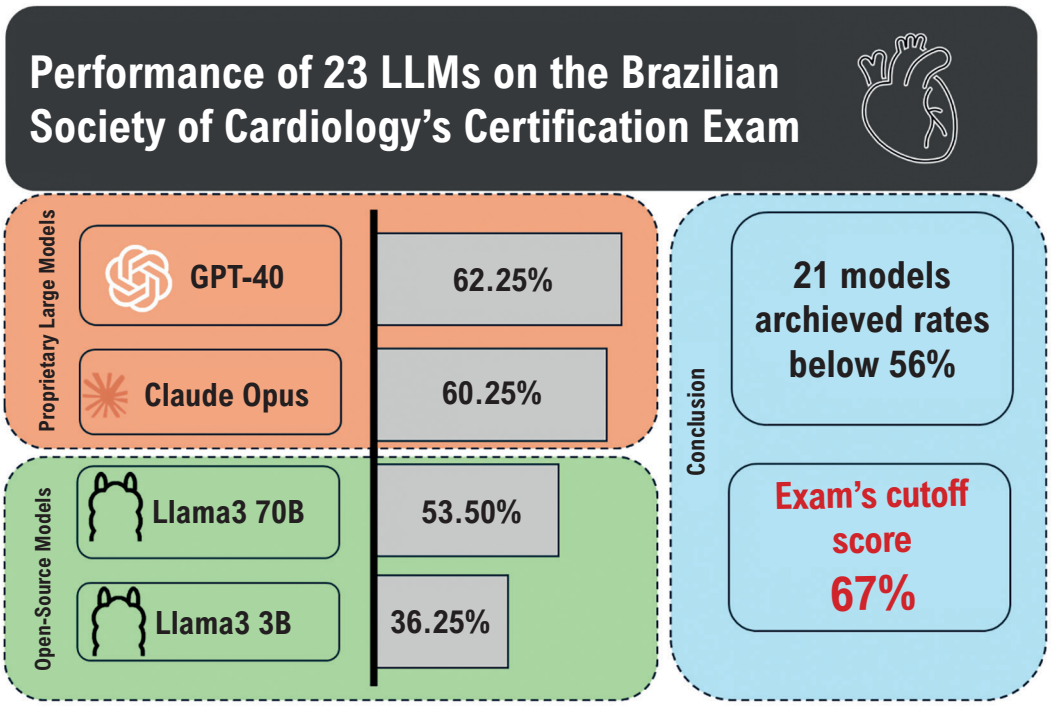
Over half of the global population, including approximately 293 million Portuguese speakers, remains underrepresented in English-centric datasets, posing a major challenge. This lack of representation risks amplifying health inequities in the application of LLMs in medicine. Health disparities are particularly prevalent in regions where English is not the primary language.

This approach underscores a critical gap in capturing the diversity of regional diseases,[18,19] particularly within the Brazilian context. In cardiology — a field requiring specialized knowledge due to the complexity and importance of managing cardiovascular diseases — there is a noticeable lack of detailed analyses on the performance of these tools.

Mailing Address: Gustavo L. Marques •
Universidade Federal do Paraná. Rua General Carneiro, 181. Postal code: 80060-000. Curitiba, PR – Brazil
E-mail: gustavolencimarques@gmail.com
Manuscript received December 4, 2024; revised manuscript March 5, 2025; accepted April 14, 2025.
Editor responsible for the review: Erito Marques

DOI: https://doi.org/10.36660/ijcs.20240231

Central Illustration: Performance Benchmarking of Open-Source Large Language Models on the Brazilian Society of Cardiology's Certification Exam



Int J Cardiovasc Sci. 2025; 38:e20240231

Performance Benchmarking of Open-Source Large Language Models on the Brazilian Society of Cardiology's Certification Exam. LLM: Large language models.

Despite global efforts to reduce these gaps, progress has been uneven, often hindered by slow advancements toward universal health coverage.[20] In this context, technology has the potential to play a pivotal role.[21] The integration of LLMs into healthcare systems could serve as a powerful tool for addressing and mitigating these inequalities.

This study aims to explore the application of LLMs in cardiology by evaluating and comparing the performance of 23 LLM platforms. The assessment was conducted using the Cardiology Certification Exam — an official certification exam for the cardiology specialist title —, administered by the Brazilian Society of Cardiology (SBC). Known for its rigor and scope, this exam evaluates the comprehensive knowledge required for clinical practice in cardiology.

The primary objective is to assess how effectively LLMs can address cardiology-related questions, focusing on their abilities in interpretation, reasoning, and decision-making compared to expert-level standards. Additionally, this study emphasizes the challenges and opportunities of deploying LLMs in Portuguese-language contexts. It aims to offer valuable insights for developing more effective AI tools to improve cardiovascular care and support healthcare professionals in non-English-speaking countries.

Methods

The Dataset

The 2023 Cardiology Certification Exam (TEC), administered annually by the SBC, grants the title of cardiologist to physicians trained in Brazil. Eligibility criteria for the examination include graduation from a medical school accredited by the Brazilian Ministry of Education (MEC) at least four years prior, registration with the Regional Medical Council (CRM), and possession of a certificate of completion from a cardiology residency program.

The examination consists of 100 multiple-choice questions covering a broad spectrum of general cardiology topics, strategically designed to assess both theoretical knowledge and practical competencies. Recognized for its rigor and comprehensiveness, the TEC assesses expertise across diverse areas in cardiology.[22] The questions are divided into three categories: memorization-based (50 questions), clinical reasoning (30 questions), and image or diagnostic exam interpretation (20 questions).

To attain certification, candidates must achieve a minimum score that ranges from 67% to 77%, depending on the strength of their professional curriculum. In 2023, approximately 1,000

candidates took the exam, with 234 earning certification—an approval rate of 23.4%.[23]

LLM

The study aimed to evaluate the performance of several proprietary LLMs and their variants, including Claude Opus, Haiku, and Sonnet; GeminiPro 1.5 and 1.0;[24] as well as GPT-4o, GPT-4, and GPT-3.5.[25] In addition, 23 open-source LLMs were assessed, including Gemma (2B, 7B),[26] Llama2 (7B, 13B, 70B),[27] Llama3 (8B, 70B, 70B Instruct),[28] Mistral (7B, 8x7B, 8x22B),[29] and Qwen (1.8B, 4B, 7B, 72B).[30] These models were selected from the Hugging Face LLM performance leaderboard,[31] resulting in a total of 23 models tested.

The models were deployed on a GPU service using LLMBox libraries.[32] To optimize performance and resource usage, larger models were quantized for testing, while smaller models were run in their full configuration. Advanced techniques such as retrieval-augmented generation (RAG)[33] were not employed in this phase but are planned for future studies.

The LLMs were classified by training parameters count, generally measured in billions. They were divided into three categories: small models (up to 10 billion parameters), medium models (up to 70 billion parameters), and large or proprietary models. Although larger models entail higher development and operational costs, they generally outperformed their smaller counterparts.

Each LLM received identical prompts containing the question ID, context, four answer choices, and the command: 'Choose the alternative from (A, B, C, D, or E) that best answers the question.' All prompts were in written Portuguese to evaluate the model's proficiency in the language. Table 1 presents an example of an original prompt in Portuguese, along with its English translation.

Metrics

Of the total 100 questions in the 2023 SBC examination, 20 included visual components and were excluded from the evaluation, as most LLMs currently lack the ability to process images. The remaining 80 questions were randomly sampled and presented to each LLM. This sampling process was repeated five times, with each sample applied independently to the models. This approach ensures that the average performance results are reliable and minimizes potential bias in the evaluation.

With 80 questions administered across all 23 models over five iterations, a total of 9,200 outputs were generated. Each output was compared to the established "ground truth" to determine whether the LLM provided a correct or incorrect

Table 1 – Example of an original Portuguese prompt and an English translated prompt

	Original Prompt	English Prompt *
"Question ID"	{00c8a70abf166c6034c658f7240bb172	
"Context"	"Em relação à síndrome coronariana aguda (SCA), assinale a alternativa correta:"	Regarding acute coronary syndrome (ACS), select the correct alternative:
"Options"	"A) Nos pacientes com disfunção renal crônica e taxa de filtração glomerular < 15 mL/min, a enoxaparina deve ser administrada com dose corrigida para 50% da dose inicial."	A) In patients with chronic renal dysfunction and a glomerular filtration rate < 15 mL/min, enoxaparin should be administered at 50% of the initial dose.
	"B) A enoxaparina está contraindicada para pacientes com mais de 150 kg."	B) Enoxaparin is contraindicated in patients weighing more than 150 kg.
	"C) Nos pacientes que receberam enoxaparina na sala de urgência, o uso de heparina não fracionada na sala de hemodinâmica pode ser considerado sem impacto no risco de sangramento."	C) In patients who received enoxaparin in the emergency room, the use of unfractionated heparin in the catheterization laboratory may be considered, without impacting the risk of bleeding.
	"D) Para pacientes em uso regular de anticoagulantes diretos, a estratégia invasiva está contraindicada nas primeiras 24 horas, independentemente do cenário clínico."	D) In patients on regular direct oral anticoagulants, an invasive strategy is contraindicated within the first 24 hours, regardless of the clinical scenario.
	"E) A utilização da via radial, quando da realização de cateterismo/angioplastia, não se mostrou efetiva em diminuir sangramento e mortalidade comparada à via femoral."	E) The use of the radial route for catheterization/ angioplasty has not proven effective in reducing bleeding and mortality compared to the femoral route.
"Ground Truth"	"B";	"B";
"Command"	"Escolha a alternativa em A) B) C) D) E) que melhor responde ao enunciado."	Choose the alternative from A), B), C), D), or E) that best answers the question.

** The English prompt was not used in the project; it serves solely to aid readers' comprehension.*

response. At no point during this study did the models have access to the ground truth.

The data analysis revealed four distinct types of outputs provided by the LLMs. The first type consisted of a simple letter indicating the selected answer. The second type included the letter along with a brief explanation. The third type featured a detailed explanation followed by the correct answer letter. The fourth type consisted of a lengthy response that lacked logical coherence with the question context or provided an answer entirely in English. Since the aim of this study is to assess the medical capabilities of multiple LLMs in the Brazilian context, answers in English were deemed unacceptable. Table 2 shows examples of these output types.

Given the large volume of 9,200 outputs, manually assessing all responses would be highly impractical for a human evaluator. To overcome this challenge, a systematic evaluation method was developed.

For types 1 and 2, an uncomplicated text comparison using a Regex method[34] was employed for classification. For types 3 and 4, given the length and complexity of the responses, a more robust classification approach was required. Two of the most advanced LLM classifiers, GPT-4o and Claude Opus, were selected for this task. These models had access only to the output and the ground truth, with the sole task of determining whether the ground truth letter appeared in the output answer. They were not provided with any information regarding which LLM generated the output or the context of the question. Each LLM classified the outputs independently; when they agreed on the correctness of the answer, the classification was considered final. In cases of disagreement, the answer was deemed incorrect due to the complexity involved. Figure 1 presents a flowchart illustrating this evaluation method.

Results

The 2023 Cardiology Certification Exam, conducted across five rounds by 23 of the world's leading LLMs in November 2024, resulted in a total of 9,200 outputs. Of these, 44.8% were classified as Type 1, 58.3% as Type 2, 11.9% as Type 3, and 10% as Type 4. All Type 4 responses were classified as incorrect based on the evaluation method. This classification directly reflects how effectively each LLM comprehended the task presented in Portuguese. Table 3 presents the number of each question types for each model.

The performance of each proprietary LLM based on our dataset is presented in Table 4. Notably, none of the models reached the exam's passing threshold of 67%. Among the proprietary models, GPT-4o achieved the highest success rate at 62.25%, followed by Claude Opus with 60,25%, as shown in the Center Illustration. In the medium-sized model category, Claude Haiku achieved a success rate of 47.25%.

The results of the open-source models are presented in Table 5. In the large-sized model category, Llama3 70B Instruct achieved a success rate of 53.50%. In contrast, Llama3 8B,

Table 2 – The four types of outputs

Answer	Original	English
Type 1	A; [A]; (A); A.; Letra A	A; [A]; (A); A.; Letter A
Type 2	D) Acrescentar fludrocortisona	D) Add fludrocortisone
Type 3	Fechamento de comunicação interatrial tipo ostium primum: A comunicação interatrial é uma anomalia congênita caracterizada por uma comunicação entre as aurículas esquerda e direita do coração. O cateterismo é uma técnica utilizada para tratar a comunicação interatrial, e o provável procedimento realizado no paciente é o fechamento da comunicação interatrial tipo ostium primum, que é a forma mais comum de fechamento da comunicação interatrial. A escolha correta é A.	Closure of an ostium primum type atrial septal defect: An atrial septal defect is a congenital anomaly characterized by a communication between the left and right atria of the heart. Catheterization is a technique used to treat atrial septal defect, and the likely procedure performed on the patient is the closure of the ostium primum type atrial septal, which is the most common form of atrial septal defect closure. The correct choice is A.
Type 4a	Presença de B4, desdobramento paradoxal de B2, sopro sistólico em foco mitral e a redução da intensidade do sopro após a manobra de Handgrip são indicações de insuficiência mitral importante.	Presence of B4, paradoxical splitting of B2, systolic murmur at the mitral focus, and reduction in the intensity of the murmur after the Handgrip maneuver are indications of significant mitral insufficiency.
Type 4b	Explicação: O paciente teve histórico de febre reumática e cardite leve aos 18 anos, mas o ecocardiograma atual não mostra lesões valvulares e apresenta boa função ventricular sistólica. Como o paciente está atualmente assintomático e os achados do ecocardiograma são normais, não há indicação para profilaxia com penicilina de acordo com as diretrizes atuais.	Explanation: The patient had a history of rheumatic fever and mild carditis at the age of 18, but the current echocardiogram shows no valvular lesions and good systolic ventricular function. Since the patient is currently asymptomatic and the echocardiogram findings are normal, there is no indication for penicillin prophylaxis according to current guidelines.

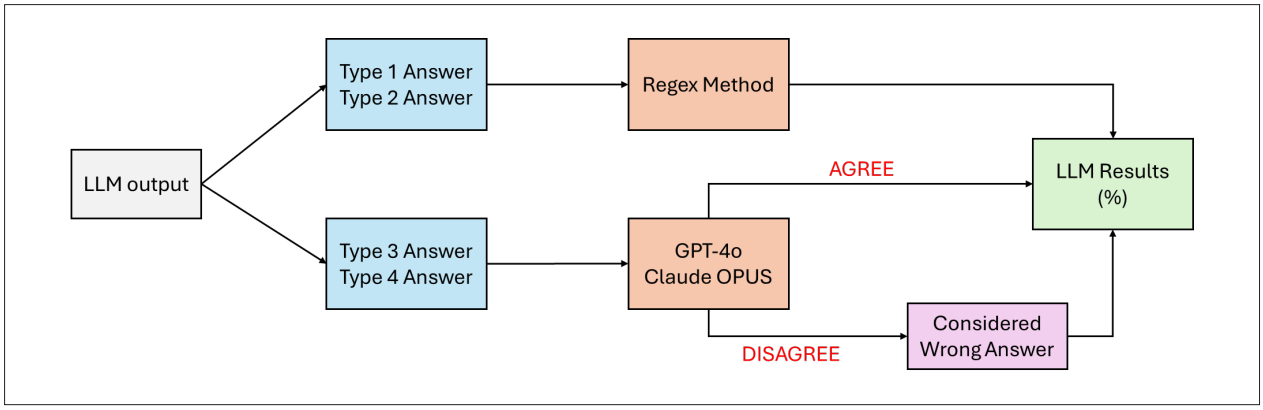


Figure 1 – *Evaluation Method flowchart.*

Table 3 – Percentage of each question types for each LLM

LLM	Type 1	Type 2	Type 3	Type 4
GPT-4o	30%	70%	0%	0%
GPT-4	100%	0%	0%	0%
GPT-3.5	4%	96%	0%	0%
Claude Opus	100%	0%	0%	0%
Claude Sonnet	19%	81%	0%	0%
Claude Haiku	47%	53%	0%	0%
Gemini 1.0	100%	0%	0%	0%
Gemini 1.5	100%	0%	0%	0%
Gemma 2B	0%	100%	0%	0%
Gemma 7B	0%	0%	0%	100%
Qwen 18B	5%	95%	0%	0%
Qwen 4B	16%	84%	0%	0%
Qwen 7B	79%	21%	0%	0%
Qwen 72B	16%	84%	0%	0%
Llama2 7B	0%	37%	45%	18%
Llama2 13B	0%	57%	43%	0%
Llama2 70B	0%	0%	43%	57%
Llama3 8B	99%	1%	0%	0%
Llama3 70B	1%	57%	42%	0%
Mistral 7B	0%	59%	41%	0%
Mistral 8x7B	1%	86%	4%	9%
Mistral 8x22B	7%	92%	1%	0%
Llama3 70B Instruct	100%	0%	0%	0%

categorized as a small-sized model, achieved a success rate of 36.25%. Despite achieving 35% accuracy, Gemma 7B only provided "A" as the answer, suggesting that the model may have been unable to interpret the questions effectively, instead defaulting to the first option it encountered.

Discussion

This study evaluated the performance of the leading 23 language models in responding to cardiology-related question in Portuguese. The results, presented in the Central Illustration and Tables 4 and 5, indicate that none of the models exceeded

Table 4 – Proprietary LLM results

LLM	Open-Source x Proprietary	Owner	Parameters Count (Billions)	Average Accuracy	Standard Deviation	CI 0.95
GPT-4o*	Proprietary	Open AI	200	62.25%	1.66%	1.45%
Gemini 1.5*	Proprietary	Google	200	51.25%	0.00%	0.00%
GPT-4*	Proprietary	Open AI	175	55.75%	1.00%	0.88%
Gemini 1.0*	Proprietary	Google	172	46.25%	0.00%	0.00%
Claude Opus*	Proprietary	Anthropic	150	60.25%	0.50%	0.44%
GPT-3.5*	Proprietary	Open AI	100	35.75%	2.32%	2.03%
Claude Sonnet*	Proprietary	Anthropic	70	36.00%	0.94%	0.82%
Claude Haiku*	Proprietary	Anthropic	20	47.25%	0.50%	0.44%

**The exact sizes of proprietary LLMs are not disclosed. Consequently, the number of parameters attributed to these models is based on estimations derived from discussions on online forums. LLM: Large language model.*

Table 5 – Proprietary models results

LLM	Open-Source x Proprietary	Owner	Parameters Count (Billions)	Average Accuracy	Standard Deviation	CI 0.95
Mistral 8x22B	Open source	Mistral AI	141	47.50%	0.00%	0.00%
Qwen 72B	Open Source	Alibaba	72	48.75%	0.79%	0.69%
Llama2 70B	Open Source	Meta AI	70	32.75%	2.67%	2.34%
Llama3 70B	Open Source	Meta AI	70	44.00%	0.50%	0.44%
Llama3 70B Instruct	Open Source	Meta AI	70	53.50%	0.50%	0.44%
Mistral 8x7B	Open Source	Mistral AI	46.7	35.75%	1.27%	1.12%
Llama2 13B	Open Source	Meta AI	13	21.50%	0.94%	0.82%
Llama3 8B	Open Source	Meta AI	8	36.25%	0.00%	0.00%
Gemma 7B	Open Source	Google	7	35.00%	0.00%	0.00%
Qwen 7B	Open Source	Alibaba	7	26.25%	0.00%	0.00%
Llama2 7B	Open Source	Meta AI	7	23.25%	0.61%	0.54%
Mistral 7B	Open Source	Mistral AI	7	30.25%	2.67%	2.34%
Qwen 4B	Open Source	Alibaba	4	33.00%	1.00%	0.88%
Gemma 2B	Open Source	Google	2	29.75%	1.22%	1.07%
Qwen 1.8B	Open Source	Alibaba	1.8	16.75%	0.61%	0.54%

LLM: Large language model.

the exam's cutoff score of 67%. This outcome is concerning, especially considering that most of these models achieve scores above 80% when tested in English and within an American medical context.[17]

It is important to note that this is a highly challenging specialist exam, with Brazilian practitioners undergoing a minimum of 10 years of rigorous medical training before taking the test. However, the results remain strikingly low, raising concerns about the current capabilities of state-of-the-art LLMs when applied to the Brazilian medical field and its specific context. Figure 2 presents a comparative graph that illustrates the accuracy of the LLMs in relation to their parameter sizes.

Considering the goal of leveraging technology to benefit a broader population, applying LLM models in public hospitals and healthcare centers in Brazil could be a promising approach. These institutions often operate with limited computational resources and face funding constraints that hinder investments in complex systems. As such, the most effective strategy would be to choose an LLM that delivers the best quality while maintaining a minimal parameter count. This would optimize

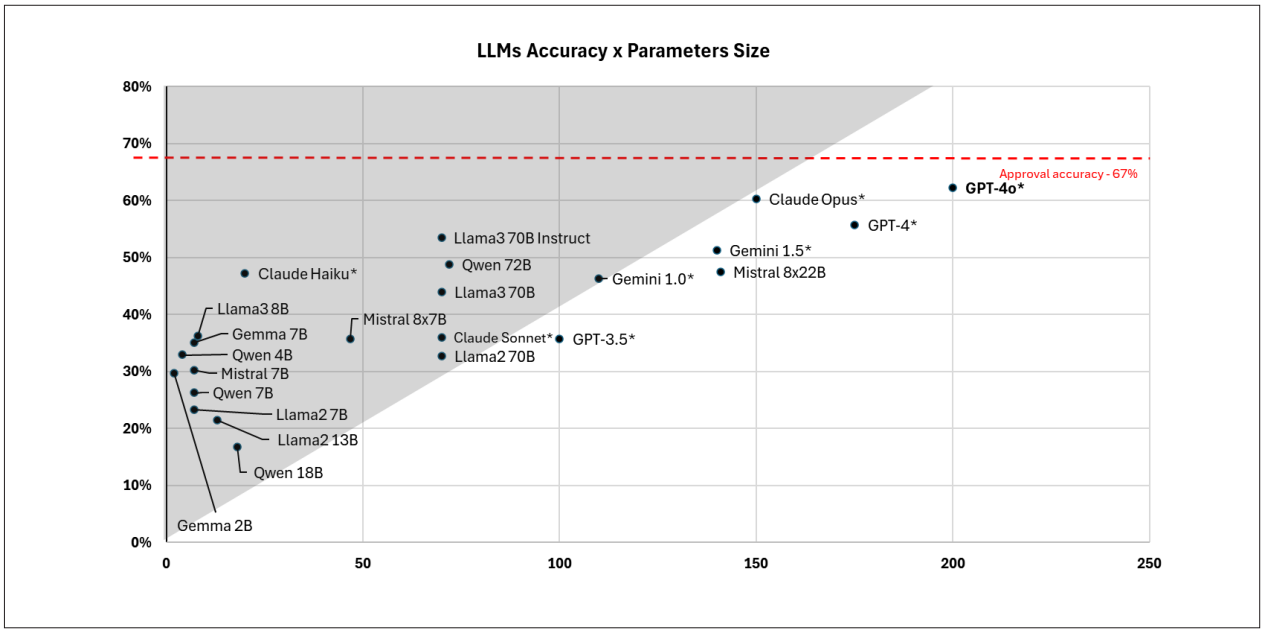


Figure 2 – *LLMs parameters size x accuracy. *Proprietary models.*

performance within available resources and ensure accessibility to advanced healthcare tools in underserved areas.

The grey area in Figure 2 highlights the LLMs that offer the best cost-benefit ratio for addressing Brazilian medical challenges. Among these, the Llama 3 70B Instruct model demonstrates a solid performance of 53.5% accuracy given its 70 billion parameters. However, the standout model is Claude Haiku, which achieves 47.25% accuracy with only 20 billion parameters. While neither model meets the exam's cutoff score, their results suggest considerable promise, particularly when considering their efficient parameter sizes.

It is important to understand the reasons behind such poor performance. Considering that output types 3 and 4 are generated when the model fails to properly interpret the prompt or command, it can be inferred from Table 3 that models such as Gemma 7B, Llama2 7B, Llama2 13B, Llama2 70B, Llama3 70B, and Mistral 7B struggled with comprehending the command in Portuguese. This indicates that these models faced significant challenges in accurately processing the task in the given language, which led to their lower performance.

Models like Gemma 7B exhibited a particular issue, where they consistently provided only the letter "A" in their answer, indicating that when they partially understood the question, they defaulted to selecting the first available option. Conversely, models like Mistral 7B generated more than 50% of their responses in English, failing to deliver appropriate outputs in Portuguese. This poses a significant limitation for applying these models in the Brazilian public health system, where English responses would be unsuitable and undermine the model's effectiveness in real-world medical scenarios.

Even when the models lack sufficient knowledge to provide a correct answer or fail to understand the question, they often attempt to generate a response, which is frequently inaccurate. In the medical field, this poses a serious risk, potentially leading to misdiagnosis or inappropriate treatment, which could potentially result in misconduct by a healthcare provider. However, some LLMs stood out by acknowledging their limitations, either stating they lacked the necessary knowledge or tools to provide the best answer. These models included all GPT models, Claude Opus, Gemini 1.5 Pro, Llama3 70B, and Mistral 8x7B.

Conclusion

Both proprietary and open-source LLMs performed poorly on the exam, failing to meet the exam's cutoff score, as can be seen in the Center Illustration. In general, although larger models tended to outperform smaller ones, some medium-sized models, such as Llama3 70B Instruct and Claude Haiku, showed relatively strong results given their size. The LLMs demonstrated a lack of specialized cardiology knowledge and struggled with interpreting and providing answers in Portuguese. This reveals a significant gap in global AI capabilities, highlighting the need for further advancements in both domain-specific knowledge and multilingual understanding.

For further studies, a cardiology exam in English could be compared to evaluate whether the primary issue lies in the lack of cardiology knowledge or in the models' ability to comprehend Portuguese. Parallel investigations are underway in other specialized medical fields, and the development of a Portuguese-specific medical LLM is being considered. However, key challenges in its development include selecting an appropriate training dataset, managing computational costs, and ensuring high-quality responses. This approach aims to pinpoint the underlying challenges and explore potential solutions for improving LLM performance in both medical domains and languages.

Author Contributions

Conception and design of the research: Severino JVB, Berger MN, Paula PAB, Loures FS, Todeschini SA, Roeder EA, Veiga MH; acquisition of data: Severino JVB, Berger MN, Paula PAB, Roeder EA, Knopfholz J, Marques GL; analysis and interpretation of the data: Severino JVB; statistical analysis: Severino JVB, Veiga MH, Marques GL; obtaining financing: Loures FS, Todeschini SA, Roeder EA, Marques GL; writing of the manuscript: Severino JVB, Marques GL; critical revision of the manuscript for intellectual content: Berger MN, Paula PAB, Loures FS, Todeschini SA, Roeder EA, Veiga MH, Knopfholz J, Marques GL.

Potential Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Sources of Funding

This study was funded by Voa Health, which covered the costs associated with the use of all closed-source language models (LLMs) and the necessary computing infrastructure required for coding. The company did not influence the results of the paper in any manner.

Study Association

This article is part of the thesis of Doctoral submitted by João Victor Bruneti Severino, from Federal University of Paraná.

Ethics Approval and Consent to Participate

This article does not contain any studies with human participants or animals performed by any of the authors.

Use of Artificial Intelligence

The authors did not use any artificial intelligence tools in the development of this work.

Research Data

All datasets supporting the results of this study are available upon request from the corresponding author. The code, prompts, and spreadsheets containing the responses from each LLM are available upon request via email. These materials have not yet been fully separated from data related to other projects and may require some time for proper organization. However, the authors have no restrictions regarding the sharing of these resources and are willing to provide them upon request.

References

1. Gupta S, Kashou AH, Smith RHS, May A, Echeverri AGM, Sueldo MD, et al. Computer-Interpreted Electrocardiograms: Impact on Cardiology Practice. Int J Cardiovasc Sci. 2024;37:e20240079. doi: 10.36660/ijcs.20240079.
2. Krakauer M, Flores M, Lima RVN, Sachetti LA. The Role of Digital Technology and New Strategies in Engagement and Adherence Among Patients with Cardiometabolic Disease. Int J Cardiovasc Sci. 2023;36:e20230126. doi: 10.36660/ijcs.20230126.
3. Unlu O, Shin J, Mailly CJ, Oates MF, Tucci MR, Varugheese M, et al. Retrieval Augmented Generation Enabled Generative Pre-Trained Transformer 4 (GPT-4) Performance for Clinical Trial Screening. medRxiv. 2024:2024.02.08.24302376. doi: 10.1101/2024.02.08.24302376.
4. Moazemi S, Vahdati S, Li J, Kalkhoff S, Castano LJV, Dewitz B, et al. Artificial Intelligence for Clinical Decision Support for Monitoring Patients in Cardiovascular ICUs: A Systematic Review. Front Med. 2023;10:1109411. doi: 10.3389/fmed.2023.1109411.
5. Göndöcs D, Dörfler V. AI in Medical Diagnosis: AI Prediction & Human Judgment. Artif Intell Med. 2024;149:102769. doi: 10.1016/j.artmed.2024.102769.
6. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med. 2023;183(6):589-96. doi: 10.1001/jamainternmed.2023.1838.
7. Fisch U, Kliem P, Grzonka P, Sutter R. Performance of Large Language Models on Advocating the Management of Meningitis: A Comparative Qualitative Study. BMJ Health Care Inform. 2024;31(1):e100978. doi: 10.1136/bmjhci-2023-100978.
8. Ebrahimian M, Behnam B, Ghayebi N, Sobhrakhshankhah E. ChatGPT in Iranian Medical Licensing Examination: Evaluating the Diagnostic Accuracy and Decision-Making Capabilities of an AI-Based Model. BMJ Health Care Inform. 2023;30(1):e100815. doi: 10.1136/bmjhci-2023-100815.
9. Bicknell BT, Butler D, Whalen S, Ricks J, Dixon CJ, Clark AB, et al. ChatGPT-4 Omni Performance in USMLE Disciplines and Clinical Skills: Comparative Analysis. JMIR Med Educ. 2024;10:e63430. doi: 10.2196/63430.
10. 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(2):1-8. doi: 10.1056/Aldbp2300092.
11. Severino JVB, Paula PAB, Berger MN, Loures FS, Todeschini SA, Roeder EA, et al. Benchmarking Open-Source Large Language Models on Portuguese Revalida Multiple-Choice Questions. BMJ Health Care Inform. 2025;32(1):e101195. doi: 10.1136/bmjhci-2024-101195.
12. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, et al. Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis. J Med Internet Res. 2024;26:e60807. doi: 10.2196/60807.
13. Altamimi I, Alhumimidi A, Alshehri S, Alrumayan A, Al-Khlaiwi T, Meo SA, et al. The Scientific Knowledge of Three Large Language Models in Cardiology: Multiple-Choice Questions Examination-Based Performance. Ann Med Surg. 2024;86(6):3261-6. doi: 10.1097/MS9.0000000000002120.
14. Tan Y, Zhang Z, Li M, Pan F, Duan H, Huang Z, et al. MedChatZH: A Tuning LLM for Traditional Chinese Medicine Consultations. Comput Biol Med. 2024;172:108290. doi: 10.1016/j.compbiomed.2024.108290.
15. Hongbin N. CBT-LLM: A Chinese Large Language Model for Cognitive Behavioral Therapy-Based Mental Health Question Answering. arXiv. 2024;1:2930-40. doi: 10.48550/arXiv.2403.16008.
16. Gangavarapu A. Multilingual Medical Language Models : A Path to Improving Lay Health Worker Effectiveness. Proc AAAI Conf. 2024;38(21):23497-9. doi: 10.1609/aaai.v38i21.30445.
17. Wang X, Chen N, Chen J, Wang Y, Zhen G, Zhang C, et al. Apollo: A Lightweight Multilingual Medical LLM towards Democratizing Medical AI to 6B People. arXiv. 2024;1:1-15. Doi: 10.48550/arXiv.2403.03640.
18. Daneshjou R, Vodrahalli K, Novoa RA, Jenkins M, Liang W, Rotemberg V, et al. Disparities in Dermatology AI Performance on a Diverse, Curated Clinical Image Set. Sci Adv. 2022;8(32):eabq6147. doi: 10.1126/sciadv.abq6147.
19. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis Bias of Artificial Intelligence Algorithms Applied

to Chest Radiographs in Under-Served Patient Populations. Nat Med. 2021;27(12):2176-82. doi: 10.1038/s41591-021-01595-0.

20. Karen M, Anderson O, Steve O, editors. The Promises and Perils of Digital Strategies in Achieving Health Equity. Washington: The National Academies Press; 2016.

21. Tangcharoensathien V, Lekagul A, Teo YY. Global Health Inequities: More Challenges, Some Solutions. Bull World Health Organ. 2024;102(2):86-86A. doi: 10.2471/BLT.24.291326.

22. Marinho GEM, Peixoto JM, Knopfholz J, Andrade MVS. Psychometric Evaluation of the Cardiology Certification Exam of the Brazilian Society of Cardiology. Arq Bras Cardiol. 2022;119(5 Suppl 1):6-13. doi: 10.36660/abc.20220355.

23. Sociedade Brasileira de Cardiologia. Prova de Titulo Especialista em Cardiologia [Internet]. São Paulo: Sociedade Brasileira de Cardiologia; 2023 [cited 2025 May 07]. Available from: https://www.portal.cardiol.br/en/cjtec/provas-anteriores/2023.

24. Islam R, Ahmed I. Gemini-the Most Powerful LLM: Myth or Truth. Commun Technol Conf. 2024: 303-8. doi:10.1109/ICTC61510.2024.10602253.

25. OpenAI. GPT-4 Technical Report [Internet]. San Francisco: OpenAI; 2023 [cited 2025 May 07]. Available from: https://cdn.openai.com/papers/gpt-4.pdf.

26. Mesnard T, Hardin C, Dadashi R, Bhupatiraju S, Pathak S, Sifre L, et al. Gemma: Open Models Based on Gemini Research and Technology. arXiv. 2024;4:1-17. doi: 10.48550/arXiv.2403.08295.

27. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv. 2307.09288v2. doi: 10.48550/arXiv.2307.09288.

28. Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, et al. The Llama 3 Herd of Models. arXiv. 2024;1:21783. doi: 10.48550/arXiv.2407.21783.

29. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, de las Casas D, et al. Mistral 7B. arXiv. 2023;1:1-9. doi: 10.48550/arXiv.2310.06825.

30. Bai J, Bai S, Chu Y, Cui Z, Dang K, Deng X, et al. Qwen Technical Report. arXiv. 2023;1:1-59. doi: 10.48550/arXiv.2309.16609.

31. Pal A, Minervini P, Motzfeldt AG, Alex B. Open Medical-LLM Leaderboard [Internet]. Hugging Face; 2024 [cited 2025 May 07]. Available from: https://huggingface.co/spaces/openlifescienceai/open_medical_llm_leaderboard.

32. Tang T, Yiwen H, Li B, Luo W, Qin Z, Sun H, et al. LLMBox: A Comprehensive Library for Large Language Models. Proc Annu Meet Assoc Comput. Linguist. 2024;3:388-99. doi: 10.18653/v1/2024.acl-demos.37.

33. Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, et al. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv. 2023;5:1-21. doi: 10.48550/arXiv.2312.10997.

34. Davis JC, Moyer D, Kazerouni AM, Lee D. Testing Regex Generalizability and Its Implications: A Large-Scale Many-Language Measurement Study. Int Conf Autom Softw Eng. 2019;34:427-39. doi:10.1109/ASE.2019.00048.