Performance of Generative Artificial Intelligence Tools on Brazilian Medical Degree Revalidation Exam

Maurício M. Maia
mauricio.maia@gmail.com

***Abstract*—The rapid advancements in Large Language Models (LLMs) and generative artificial intelligence (AI) tools have shown promising results in various domains, including medicine. However, their performance in the context of country-specific medical exams remains largely unexplored. This study evaluates the performance of four prominent generative AI tools on the Brazilian Medical Degree Revalidation Exam.**

The results demonstrate a clear hierarchy in the overall performance of the AI tools, however, most tools showed a significant decrease in accuracy when answering Brazil-specific questions.

Our findings highlight the importance of considering local context and country-specific knowledge when developing and evaluating AI tools for medical applications.

I. Introduction

The emergence of Large Language Models (LLMs) in recent years has opened up new possibilities for artificial intelligence (AI) in specialized domains like medicine [1]. Several studies have demonstrated the capabilities of LLMs on medical question answering tasks [2], [3], [4], [5], [6], with state-of-the-art models surpassing human experts with accuracies greater than 90% on multiple-choice exams [7].

The release of conversational generative AI tools built on top of these LLMs, like ChatGPT, has broadened the access to these powerful models to medical practitioners, educators, and students [8], [9].

However, the majority of research on LLMs and AI tools in medicine has focused on the English language. There has been less investigation into their performance in other languages and their adherence to the medical guidelines and protocols of countries outside the United States [10]. This is an important consideration, as healthcare systems and practices can vary significantly between nations.

In Brazil, medical care is provided through the Unified Health System (SUS), a universal and free healthcare system. The Clinical Protocols and Therapeutic Guidelines used throughout the SUS are established by the federal government under the Health Ministry. Doctors who obtained their medical degree abroad, whether foreign nationals or Brazilians who studied in another country, must pass the Medical Degree Revalidation Exam (REVALIDA) in order to legally practice medicine in Brazil. The exam, held each semester, assesses candidates' skills, competencies and knowledge necessary for professional practice appropriate to the principles and needs of SUS. In its latest edition, the exam had 10,080 registered candidates.

To evaluate the applicability of AI tools to the Brazilian medical context, we tested the performance of four prominent systems: OpenAI's ChatGPT 3.5 and ChatGPT 4, Google's Gemini 1.0, and Maritaca AI's Maritalk Sábia2-Medium (a Portuguese language model [11]) on questions from the REVALIDA exam administered in the first semester of 2024 (REVALIDA 2024.1).

This paper presents the methodology and results of our study comparing the performance of generative AI tools on the REVALIDA 2024.1 exam. We discuss the implications of our findings for the potential use of these technologies in Brazil. Finally, we highlight the limitations of the current study and propose directions for future research in this area.

TABLE I
Evaluated Tools

Tool	Browsing Capability	Training Data Up To
ChatGPT 4	Yes	December 2023
ChatGPT 3.5	No	December 2023
Gemini 1.0	Yes	February 2023
Maritalk Sábia2-Medium	No	Mid 2023

II. Related Works

Research on the medical domain capabilities of LLMs and AI tools in the Brazilian context is currently limited.

ChatGPT 3.5 was evaluated with the Brazilian Council of Ophthalmology Board Examination correctly answering 41.46% of the questions [12], and with the Brazilian College of Radiology annual resident evaluation test, answering 53.3% of the questions correctly [13].

ChatGPT 4, a more advanced model, was tested on the REVALIDA 2022 exam. The results showed an accuracy of 87.7% [14].

A recent study evaluated seventeen different models, including GPT-4 Turbo, GPT-3.5 Turbo, and Sabiá-2 Medium, using the REVALIDA 2023.2 exam [11]. The models achieved accuracies of 84.0%, 57.0%, and 73.0%, respectively. The same study also tested these models on two Brazilian Medical Residency Entrance Exams, with GPT-4 Turbo, GPT-3.5 Turbo, and Sabiá-2 Medium scoring 64.2%, 35.8%, and

57.6%, respectively. These results demonstrate the variability in performance across different models and exam types, emphasizing the importance of comprehensive evaluations.

To our knowledge, no study has comprehensively evaluated and compared generative AI tools available in the medical domain for Brazil. Our work aims to address this gap by conducting a thorough evaluation of four prominent tools available in Brazil.

III. Methodology

A. Dataset

The dataset for this study was obtained from the objective portion of the Brazilian Medical Degree Revalidation Exam administered in the first semester of 2024 (REVALIDA 2024.1). The exam is publicly available on the website of the National Institute of Educational Studies and Research Anísio Teixeira (INEP) as a PDF document. The REVALIDA 2024.1 exam consists of 100 multiple choice questions, each with four alternative answers.

The questions were automatically extracted from the official PDF document and then subjected to manual review. During the review process, each question was annotated for the presence of images (n=4) and tables (n=6).

To further categorize the questions, we employed GPT-4-Turbo to classify each item according to REVALIDA's official Reference Matrix. This matrix outlines the key areas of medical knowledge and competencies assessed by the exam. In addition to the Reference Matrix categorization, we annotated each question regarding the requirement of specific knowledge about Brazil (n=17). See figure 1 for examples.

The final dataset used in this study consisted of 91 questions. We excluded questions that contained images (n=4) and those that were nullified after the exam (n=5).

B. Evaluation

To evaluate the performance of the generative AI tools on the REVALIDA 2024.1 exam, we manually entered each question into independent chat sessions on four platforms: OpenAI's ChatGPT 3.5 and ChatGPT 4, Google's Gemini 1.0, and Maritaca AI's MaritalkSábia2-Medium. This approach ensured that each tool processed the questions independently, without any influence from previous interactions.

We used a standard prompt format in Portuguese to maintain consistency across all the AI tools. The prompt included three key components: the context, the question itself, and the alternative answers. (Figure 2). More advanced prompt techniques were not applied to avoid involuntary optimization for a specific tool [15].

After submitting each question to the AI tools, we manually reviewed the generated answers to extract the selected alternative choice.

IV. Results

ChatGPT-4 outperformed its counterparts with an overall accuracy of 84%, a result 21.74% higher than the second best performance by ChatGPT 3.5 (Table II).

Upon taking over the coordination of a family health team, a family and community physician notices that his team does not use criteria for establishing the flow of scheduling home visits. In light of this situation, he decides to promote a discussion circle to raise awareness among his team about the need to organize criteria for defining home visits as an approach to the individual in their family and community context.
In this case, **which principle of the Unified Health System (SUS)** best fits the strategy adopted for improving the team's work?

A. Equity.
B. Autonomy.
C. Comprehensiveness (Integralidade).
D. Universality.

A 35-year-old woman, nulligravida, who has been using oral contraceptives for 10 years, complains of breast pain (mastalgia) for the past 10 days. She reports that her last menstrual period was 25 days ago. In her family history, she mentions an aunt who is 65 years old and has been treating breast cancer for 3 years. The patient's body mass index (BMI) is 32 kg/m², and the physical examination of her breasts does not reveal any abnormalities.
According to the guidelines of the Ministry of Health, the follow-up for this patient should include:

A. Annual breast magnetic resonance imaging.
B. Biennial breast ultrasound after the age of 40.
C. Annual clinical breast examination and mammography.
D. Annual clinical examination and biennial mammography after the age of 50.

Fig. 1. Example of questions from REVALIDA 2024.1 requiring specific knowledge about Brazil (translated from Portuguese).

```
Context
{context}

Question
{question}

Alternatives
{alternatives}
```

Fig. 2. Translated prompt template (originally submitted in Portuguese).

When considering questions that required Brazil-specific knowledge, most tools presented a substantial decrease in accuracy (Table III). Maritalk Sábia2-Medium outperformed ChatGPT 3.5 by 17.14% in this subset. This finding suggests that models trained on language-specific corpora may have an advantage in capturing the nuances and intricacies of medical knowledge specific to a particular country or region, in accordance with other studies in the medical domain [16], [17].

Gemini 1.0, developed by Google, proved to be an exception to the trend of decreased accuracy on Brazil-specific questions. It experienced a relatively small decline of 5.36% compared to its overall performance. We hypothesize that Gemini 1.0's ability to browse the web and retrieve relevant information from online sources may have contributed to its resilience in handling country-specific questions. By accessing up-to-date and localized information, Gemini 1.0 could potentially compensate for gaps in its pre-trained knowledge.

Another finding is the overall increase in accuracy for questions that included tables (Table IV). We theorize that questions with tables often provide a self-contained context, presenting all the necessary information required to arrive at the correct answer, reducing the need for the models to rely on their previous encoded knowledge. Furthermore, the structured nature of tables could help reduce ambiguity compared to unstructured text.

TABLE II
OVERALL ACCURACY

Tool	Accuracy (n=91)
ChatGPT 4	**0.84**
ChatGPT 3.5	0.69
Maritalk Sábia2-Medium	0.68
Gemini 1.0	0.56

TABLE III
ACCURACY FOR BRAZIL-SPECIFIC QUESTIONS

Tool	Brazil (n=17)	Change %
ChatGPT 4	**0.53**	−36.90
ChatGPT 3.5	0.35	−49.28
Maritalk Sábia2-Medium	0.41	−39.71
Gemini 1.0	**0.53**	−5.36

TABLE IV
ACCURACY BY TABLE PRESENCE

Tool	With Table (n=6)	Change %
ChatGPT 4	**1.00**	+19.05
ChatGPT 3.5	0.83	+20.29
Maritalk Sábia2-Medium	0.83	+22.06
Gemini 1.0	0.67	+19.64

V. CONCLUSION

Our results offer valuable insights into the strengths and limitations of generative AI tools in the context of non-English medical licensing exams, particularly in Brazil.

While ChatGPT-4 demonstrated superior overall performance, our analysis also uncovers a significant gap in the ability of these tools to provide accurate answers when confronted with questions that require Brazil-specific medical knowledge. The decreased performance of most tools in this domain underscores the challenges faced by AI models in capturing the nuances and complexities of country-specific medical information, such as national health guidelines, protocols, and cultural factors that influence healthcare practices.

The performance of Maritalk Sábia2-Medium, a model trained on Portuguese texts, in outperforming ChatGPT 3.5 on Brazil-specific questions suggests that language-specific training could be a promising approach to enhance the accuracy of AI tools in handling local medical contexts. The small relative decline of Gemini 1.0 suggests that enabling AI tools to access up-to-date and localized content is another promising avenue.

Additionally, the positive impact of tables on accuracy underscores the importance of considering the format and structure of exam questions when evaluating and optimizing AI-based assessment tools.

VI. LIMITATIONS AND FUTURE WORKS

As we focused solely on assessing whether the AI tools chose the correct alternative, without considering the explanations they may have produced, future studies should aim to evaluate the clinical reasoning and ability to provide sound justifications for the answers.

Although REVALIDA's structure in evaluating candidates' adherence to Brazilian health protocols and guidelines was valuable in measuring AI tools, more work needs to be done in evaluation datasets for Brazilian medicine.

REFERENCES

[1] V. Liévin, C. E. Hother, A. G. Motzfeldt, and O. Winther, "Can large language models reason about medical questions?" *arXiv*, Dec 2023.

[2] K. Singhal, E. Azizi, E. M. Bender, E. V. Bernstam, S. Bhat, R. Chandra, E. Choi, M. D. Choudhury, J. A. Dunnmon, J. A. Fries *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, Aug 2023.

[3] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of gpt-4 on medical challenge problems," *arXiv*, Apr 2023.

[4] K. Singhal, E. Azizi, E. M. Bender, E. V. Bernstam, S. Bhat, R. Chandra, E. Choi, M. D. Choudhury, J. A. Dunnmon, J. A. Fries *et al.*, "Towards expert-level medical question answering with large language models," *arXiv*, May 2023.

[5] H. Nori, E. Choi, J. A. Dunnmon, J. A. Fries, D. Ganguli, M. Ghassemi, E. Horvitz, S. C. Huang, M. Komorowski, C. P. Langlotz *et al.*, "Can generalist foundation models outcompete special-purpose tuning? case study in medicine," *arXiv*, Nov 2023.

[6] A. Pal and M. Sankarasubbu, "Gemini goes to med school: Exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations," *arXiv*, Feb 2024.

[7] H. Zhou, Y. Jiang, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang *et al.*, "A survey of large language models in medicine: Progress, application, and challenge," *arXiv*, May 2024.

[8] A. Gilson, L. A. Celi, D. M. Maslove, R. O. Deliberato, B. Rush, E. Lehman, E. Jaffe, C. A. Umscheid, J. S. Haas, C. Safran *et al.*, "How does chatgpt perform on the united states medical licensing examination (usmle)? the implications of large language models for medical education and knowledge assessment," *JMIR Med. Educ.*, vol. 9, no. 1, p. e45312, Feb 2023.

[9] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz, C. Honculada *et al.*, "Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models," *PLOS Digit. Health*, vol. 2, no. 2, p. e0000198, Feb 2023.

[10] Y. Jin, M. Chandra, G. Verma, Y. Hu, M. De Choudhury, and S. Kumar, "Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries," *arXiv*, Oct 2023.

[11] T. S. Almeida, H. Abonizio, R. Nogueira, and R. Pires, "Sabiá-2: A new generation of portuguese large language models," *arXiv*, Mar 2024.
[12] M. C. Gobira, R. C. Moreira, L. F. Nakayama, C. V. S. Regatieri, E. Andrade, and B. Rubens, "Performance of chatgpt-3.5 answering questions from the brazilian council of ophthalmology board examination," *Pan-Am. J. Ophthalmol.*, vol. 5, no. 1, p. 17, May 2023.
[13] C. A. Leitão, G. L. d. O. Salvador, L. M. Rabelo, and D. L. Escuissato, "Performance of chatgpt on questions from the brazilian college of radiology annual resident evaluation test," *Radiol. Bras.*, vol. 57, p. e20230083, Apr 2024.
[14] M. Gobira, L. F. Nakayama, R. Moreira, E. Andrade, C. V. S. Regatieri, and R. Belfort Jr., "Performance of chatgpt-4 in answering questions from the brazilian national examination for medical degree revalidation," *Rev. Assoc. Médica Bras.*, vol. 69, p. e20230848, Sep 2023.
[15] N. Alzahrani, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang *et al.*, "When benchmarks are targets: Revealing the sensitivity of large language model leaderboards," *arXiv*, Feb 2024.
[16] P. Qiu, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang *et al.*, "Towards building multilingual language model for medicine," *arXiv*, Feb 2024.
[17] X. Wang, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang, Z. Jiang *et al.*, "Apollo: An lightweight multilingual medical llm towards democratizing medical ai to 6b people," *arXiv*, Mar 2024.