Open access Implementer report

BMJ Health & Care Informatics

Benchmarking open-source large language models on Portuguese Revalida multiple-choice questions

João Victor Bruneti Severino ,[1,2] Pedro Angelo Basei de Paula,[1] Matheus Nespolo Berger,[1] Filipe Silveira Loures,[3] Solano Amadori Todeschini,[3] Eduardo Augusto Roeder,[1,3] Maria Han Veiga,[4] Murilo Guedes,[2] Gustavo Lenci Marques[1,2,3]

To cite: Bruneti Severino JV, Basei de Paula PA, Berger MN, *et al*. Benchmarking open-source large language models on Portuguese Revalida multiple-choice questions. *BMJ Health Care Inform* 2025;**32**:e101195. doi:10.1136/bmjhci-2024-101195

Received 09 July 2024
Accepted 12 February 2025

[1]Federal University of Parana, Curitiba, Brazil
[2]Pontifical Catholic University of Parana, Curitiba, Brazil
[3]Voa Health, Belo Horizonte, Brazil
[4]Mathematics, Ohio State University, Columbus, Ohio, USA

Correspondence to
Gustavo Lenci Marques;
gustavomarques@ufpr.br

ABSTRACT

Objective The study aimed to evaluate the top large language models (LLMs) in validated medical knowledge tests in Portuguese.

Methods This study compared 31 LLMs in the context of solving the national Brazilian medical examination test. The research compared the performance of 23 open-source and 8 proprietary models across 399 multiple-choice questions.

Results Among the smaller models, Llama 3 8B exhibited the highest success rate, achieving 53.9%, while the medium-sized model Mixtral 8×7B attained a success rate of 63.7%. Conversely, larger models like Llama 3 70B achieved a success rate of 77.5%. Among the proprietary models, GPT-4o and Claude Opus demonstrated superior accuracy, scoring 86.8% and 83.8%, respectively.

Conclusions 10 out of the 31 LLMs attained better than human level of performance in the Revalida benchmark, with 9 failing to provide coherent answers to the task. Larger models exhibited superior performance overall. However, certain medium-sized LLMs surpassed the performance of some of the larger LLMs.

INTRODUCTION

The emergence of large language models (LLMs) has prompted discussions on their potential in the medical field. These advanced models demonstrate significant potential in areas such as disease management,[1] decision-making[2] and medical research.[3] Despite their promising capabilities, existing research predominantly concentrates on datasets in Chinese[4] and English,[5] with limited attention given to multilingual models[6] and less commonly spoken languages. This poses a substantial problem as over half of the global population, including around 293 million Portuguese speakers, is not represented in English-centric datasets, potentially leading to health inequities in the deployment of LLMs in medicine. Global inequities in medicine are widespread, particularly in countries where English is not the primary language. Despite international efforts to reduce health disparities, progress has been uneven and often hindered by the slow advancement towards universal health coverage.[7] In this context, technology can play a crucial role in addressing these disparities.[8] Therefore, deploying LLMs in healthcare could be a powerful tool to help mitigate the existing inequalities. Given the considerable variability of medical knowledge across diverse cultural contexts, particularly evident in language diversity, this study seeks to develop a benchmark, specifically in Portuguese, for assessing the medical knowledge of the top 31 LLMs in a non-English and non-Chinese scenario.

METHODS

Dataset

The Revalida examination in Brazil is conducted each semester, and its primary objective is to evaluate the competency of physicians who have obtained their medical degrees from foreign institutions. It has an approval rate of approximately 15%–20%.[9] The exam's cut-off score is adjusted annually based on its difficulty. From 2020 to 2023, the cut-off scores were 61%, 60%, 66% and 67%, respectively. Each question is structured around a clinical scenario, followed by a prompt requiring the selection of the correct response from four provided choices. We assemble a dataset consisting of 399 questions extracted from the multiple-choice stage of the Revalida examination conducted between the years 2020 and 2023. Questions that included tables or images were excluded from the due to the inherent complexities involved in interpreting such formats using language models.

Large Language Model

The research sought to assess the competency of prominent proprietary LLMs and their variations, namely Claude Opus, Haiku and Sonnet; Gemini Pro 1.5 and 1.0 as well as GPT-4o, GPT-4 and 3.5. Additionally, 23 open-source LLMs were evaluated: Apollo (1B, 2B, 6B and 7B), Gemma (2B and 7B), Llama 2 (7B, 13B and 70B), Llama 3 (8B, 70B and 70B instruct), Meditron (7B and 70B), Mistral (7B, 8×7B and 8×22B), Qwen (1.8B, 4B, 7B and 72B) and Yi (6B and 34B). These models were included based on the leaderboard of performing LLMs provided by Hugging Face in september 2024,[10] resulting in a total of 31 models. A GPU service hosted the models, using vLLM libraries. Larger models were quantised for testing, while smaller models were used in full. Advanced methods like RAG were not used but will be explored in future research. Each LLM received identical prompts during evaluation, comprising only the question statement, four answer options and the command 'choose the only correct alternative'. These models can be categorised based on the number of training parameters (size), typically quantified in billions. Larger LLMs have a higher development and operational cost, and they generally exhibit superior performance compared with smaller models. In this article, we classify the models into small (up to 10B), medium (up to 70B), large and proprietary.

Metrics

The evaluation process generated over 8000 outputs. Manual evaluation was deemed impractical. To address this challenge, a script was developed to compare the outputs against the ground truth for each question. For outputs that consisted of a single letter, a basic text comparison method was employed to assess the similarity between the answer and the ground truth. In cases where the output contained text and the chosen letters, the previous method was inadequate for reliable evaluation. Therefore, for such instances, we used GPT-4 and Claude Opus to classify the text output comparing with the ground truth. Both models are prompted with only pairs of answers and ground truth, without knowledge of the model that produced the answer, other alternatives, or the question statement, rendering the evaluation task into a simple comparison task. When both GPT-4 and Opus agreed with the answer, we considered the classification correct. Outputs without agreement were excluded from the study.

RESULTS

Table 1 displays the performance of each LLM based on our dataset. We evaluated each LLM by running all 399 questions five times to account for LLM randomness, thereby achieving a 99% CI (based on SD). The results exclude all Apollo, Meditron, Yi and Gemini 1.5 Pro models due to their lack of coherence in responses. These models produced outputs without meaningful connections to the questions asked. Among the open-source small models, Llama 3 8B achieved the highest success rate of 53.9%. In the category of medium-sized models, Mixtral 8×7B achieved a success rate of 63.7%. Transitioning to the large-scale models, Llama 3 70B instruct demonstrated a success rate of 77.5%. Among the proprietary models, GPT-4o achieved a score of 86.8%, while Claude Opus achieved 83.8% success.

DISCUSSION

This study examined the performance of the top 31 LLMs in responding to Portuguese questions within a medical context. The results indicated that ten models exceeded the highest exam's cut-off and human average score of 67%, including ChatGPT-4o, which achieved a score of 86.8%. Additionally, 12 models scored below the human average, and 9 models were unable to generate coherent answers for the proposed tasks. It is important to note that these 31 models represent the highest performing models in test taking. It is notable that all responses from proprietary models consisted of single letters, allowing for straightforward text comparison to evaluate the models' outputs. Similarly, the larger open-source models predominantly provided single-letter responses. Conversely, smaller models often generated more complex text responses, suggesting a failure to accurately interpret the command prompts. Ultimately, 227 out of 8778 questions were excluded from the results due to a lack of agreement between ChatGPT-4 and Claude Opus in the correction process. Among the open-source models, Llama 3 70B achieved the highest performance. Along with Qwen1.5 72B and Mixtral 8×22, these open-source models outperformed the human test takers. As expected, the proprietary larger models, GPT-4o and Claude Opus, exhibited the best performance. Additionally, Gemini Pro 1.0 and the other Claude models also outperformed the human average. The companies behind these models do not disclose the number of training parameters, making it challenging to analyse the performance of each LLM relative to its size. However, it can be estimated that both GPT-4o and Claude Opus are larger models compared with the others. Although the smaller models could not compete with the larger ones, Llama 3 8B and Claude Haiku demonstrated impressive performance relative to their training sizes. Notably, Claude Haiku, with approximately 20 billion parameters, surpassed the human average.

CONCLUSION

Both proprietary and open-source LLMs have achieved satisfactory performance on a standardized national test evaluating medical knowledge among physicians in Brazil, often surpassing the human test takers. In general, although larger LLMs tended to perform better, some medium-sized LLMs (Llama 3 70B and 70B instruct, Claude Haiku and Claude Sonnet) were competitive, outperforming some of the larger LLMs.

Table 1 Performance of each LLM

LLM	Proprietary × open	Parameters (Billions)	Questions answered	Without answer	Average accuracy	CI 0.99
GPT-4o*	Proprietary	200	399	0	**86.8%**	± 0.85%
GPT-4*	Proprietary	175	399	0	**84.7%**	± 1.83%
Gemini 1.0 Pro*	Proprietary	172	399	0	**69.2%**	± 0.00%
Claude Opus*	Proprietary	150	399	0	**83.8%**	± 0.28%
Mixtral 8×22B	Open source	141	397	2	**71.4%**	± 0.56%
GPT-3*	Proprietary	100	399	0	59.6%	± 1.02%
Qwen1.5 72B	Open source	72	399	0	**67.7%**	± 0.00%
Claude Sonnet*	Proprietary	70	398	1	**75.4%**	± 0.00%
Llama 3 70B	Open source	70	394	5	**75.0%**	± 0.65%
Llama 2 70B	Open source	70	390	9	51.2%	± 0.37%
Llama 3 70B instruct	Open source	70	399	0	**77.5%**	± 0.23%
Mixtral 8×7B	Open source	46.7	386	13	63.7%	± 0.27%
Claude Haiku*	Proprietary	20	399	0	**73.2%**	± 0.00%
Qwen1.5 18B	Open source	18	391	8	34.3%	± 0.24%
Llama 2 13B	Open source	13	355	44	46.4%	± 0.48%
Llama 3 8B	Open source	8	399	0	53.9%	± 0.00%
Gemma 7B	Open source	7	335	64	30.7%	± 0.00%
Llama 2 7B	Open source	7	371	28	31.8%	± 0.28%
Mistral 7B	Open source	7	384	15	48.8%	± 0.40%
Qwen1.5 7B	Open source	7	398	1	47.1%	± 0.14%
Qwen1.5 4B	Open source	4	396	3	41.9%	± 0.44%
Gemma 2B	Open source	2	365	34	37.2%	± 0.98%
Apollo (1B, 2B, 6B and 7B)	Open source	1, 2, 6, 7	**0**	399	0.00%	± 0.00%
Meditron (7B and 70B)	Open source	7, 70	**0**	399	0.00%	± 0.00%
Gemini 1.5 Pro*	Proprietary	200	**0**	399	0.00%	± 0.00%
Yi (6B and 34B)	Open source	6, 34	**0**	399	0.00%	± 0.00%

Bold type indicates they score above the human average.
*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.
LLMs, large language models.

The Portuguese benchmark tool is now implemented and available for use by the scientific community. For future investigations, it would be important to compare the performance of the same LLMs on the Revalida Benchmark with English-written benchmarks. This comparison would enable a thorough analysis to determine if there is a bias in the advancement of LLMs outside the English and Chinese contexts. Finally, it would be valuable to investigate the impact of methods, such as RAG, on the LLM models used in this study when applied to the Revalida benchmark.

Contributors GLM is the guarantor of the study. JVBS designed the research conception, conducted the data analysis and interpretation of the trials, and was responsible for drafting and revising the manuscript. PABdP and MNB contributed to designing the manuscript structure, supported the writing process and assisted with the interpretation of the data. FSL, SAT and EAR developed the software infrastructure for the project, including automating large language models (LLMs), benchmarks, evaluation methods and resolving code errors throughout the project. They also contributed to the revision of the manuscript. MHV contributed to the conceptualisation of the manuscript design and writing and was responsible for the mathematical analysis of the data. MHG led the manuscript design conception and acted as a reviewer for the writing. GLM, as the guarantor of the paper, oversaw the manuscript design, revisions and data analysis while monitoring the project's overall progress. Several LLMs were used in this study to analyse their capability when resolving Portuguese medical questions like ChatGPT, Claude, Gemini and others.

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.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study used publicly available Revalida (exam) questions accessible to the general community in Brazil. No real cases or data involving real individuals were used. Therefore, ethical approval was not required for this research.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. We used the data from the Revalida examination in Brazil for the years 2020–2023, which is freely available online for anyone.

ORCID iD

João Victor Bruneti Severino http://orcid.org/0000-0002-8649-6494

REFERENCES

1. Fisch U, Kliem P, Grzonka P, *et al*. Performance of large language models on advocating the management of meningitis: a comparative qualitative study. *BMJ Health Care Inform* 2024;31:e100978:1–5:.
2. Ebrahimian M, Behnam B, Ghayebi N, *et al*. 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:e100815:1–6:.
3. Roberts RH, Ali SR, Hutchings HA, *et al*. Comparative study of ChatGPT and human evaluators on the assessment of medical literature according to recognised reporting standards. *BMJ Health Care Inform* 2023;30:e100830:1–5:.
4. Tan Y, Zhang Z, Li M, *et al*. MedChatZH: A tuning LLM for traditional Chinese medicine consultations. *Comput Biol Med* 2024;172:108290.
5. Wu S, Koo M, Blum L, *et al*. Benchmarking Open-Source Large Language Models, GPT-4 and Claude 2 on Multiple-Choice Questions in Nephrology. *NEJM AI* 2024;1:1–8.
6. Wang X, Chen N, Chen J, *et al*. Apollo: An Lightweight Multilingual Medical LLM towards Democratizing Medical AI to 6B People. 2024;1–20. Available: http://arxiv.org/abs/2403.03640
7. Karen M, Anderson O, Steve O. The promises and perils of digital strategies in achieving health equity. In: *The promises and perils of digital strategies in achieving health equity*. 2016.
8. Tangcharoensathien V, Lekagul A, Teo YY. Global health inequities: more challenges, some solutions. *Bull World Health Organ* 2024;102:86–86A.
9. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, Inep. Painel Revalida. 2024. Available: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/inep-data/painel-revalida
10. Pal A, Minervini P, Motzfeldt AG, *et al*. Open medical-LLM leaderboard. Hugging Face; 2024. Available: https://huggingface.co/spaces/openlifescienceai/open_medical_llm_leaderboard