Class of LLMs: Benchmarking Large Language Models on the Brazilian National Medical Examination

João Vitor Mariano Correia[1], **Pedro Henrique Alves de Castro**[2], **Gabriel Lino Garcia**[1], **Pedro Henrique Paiola**[1], **João Paulo Papa**[1]
[1]Department of Computing, Faculty of Sciences, São Paulo State University
[2]Department of Medical Sciences, Nove de Julho University

Abstract

The evaluation of Large Language Models (LLMs) in medicine has predominantly relied on English-language benchmarks aligned with North American clinical guidelines, limiting their applicability to other healthcare systems. In this paper, we evaluate twenty-two proprietary and open-weight LLMs on the 2025 National Examination for the Evaluation of Medical Training (ENAMED), a high-stakes, government-standardized assessment used to evaluate medical graduates in Brazil. The benchmark comprises 90 multiple-choice questions grounded in Brazilian public health policy, clinical practice, and Portuguese medical terminology, and is released as an open dataset. Model performance is measured using both standard accuracy and the official Item Response Theory (IRT) framework employed by ENAMED, enabling direct comparison with human proficiency thresholds. Results reveal a clear stratification of model capabilities: proprietary frontier models achieve the highest performance, whereas many open-weight and smaller-domain-adapted models fail to meet the minimum proficiency criterion. Across comparable scales, large generalist models consistently outperform specialized medical fine-tunes, suggesting that general reasoning capacity is a stronger predictor of success than narrow domain adaptation in this setting. These findings establish ENAMED as a rigorous benchmark for evaluating medical LLMs in Portuguese and highlight both the potential and current limitations of such models for educational assessment.

1 Introduction

The integration of Artificial Intelligence into clinical practice has advanced Large Language Models (LLMs) from experimental systems to evaluated tools for tasks like summarization and question answering. While frontier models now pass global, English-language benchmarks such as the USMLE (Nori et al., 2023; Singhal et al., 2023), these assessments ignore the organizational structure of the **Brazilian Unified Health System** (SUS), regional epidemiology, and the Federal Council of Medicine (CFM) professional norms.

In 2025, Brazil introduced the **National Examination for the Evaluation of Medical Training (ENAMED)**, a **centralized** assessment consolidating undergraduate and residency evaluations. Aligned with National Curricular Guidelines (DCNs), ENAMED uniquely prioritizes **public health policies** and primary care. Its inaugural administration yielded **unsatisfactory** institutional scores, providing a rigorous, government-standardized context for evaluating AI preparedness and clinical reasoning in Brazilian medical education.

This work evaluates the performance of general-purpose and domain-adapted Large Language Models on ENAMED. Our results show that recent generalist models consistently outperform several specialized medical models. The main contributions of this study are twofold: (i) a systematic, domain-level analysis of LLM behavior in a national medical assessment setting, highlighting both their potential and the challenges of deploying such models in context-specific healthcare environments, and (ii) the release of a structured dataset derived from ENAMED 2025, enabling reproducible evaluation and future research on LLM performance in Brazilian medical education.

2 Related Work

Early evaluations of medical LLMs relied primarily on English-language benchmarks, including general-purpose and biomedical question-answering datasets such as MMLU (Hendrycks et al., 2021), PubMedQA (Jin et al., 2019), and MedQA (Jin et al., 2021). Although these datasets effectively assess biomedical knowledge and multi-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 101–111
April 13-16, 2026

step reasoning, they reflect predominantly English-language, high-resource clinical environments. This limitation has motivated the development of Portuguese-language medical resources. SemClinBR (Oliveira et al., 2022) and BRATECA (Dias and Ulbrich, 2022) established foundational corpora for clinical NLP in Brazilian Portuguese, supporting domain adaptation efforts. Subsequent studies developed and evaluated Portuguese-adapted medical LLMs, demonstrating the feasibility of regional specialization (de Souza Pinto et al., 2024; Paiola et al., 2024).

More recently, benchmark construction has shifted toward examination-based evaluation. HealthQA-BR and related initiatives (D'addario, 2025; Garcia et al., 2025) introduced large-scale benchmarks derived from Brazilian national licensing and residency examinations, revealing substantial variability in model performance across specialties and professional domains. These works underscore that high aggregate accuracy may mask domain-specific weaknesses, particularly in locally grounded regulatory and administrative knowledge.

Our study extends this line of research by introducing a structured version of the 2025 ENAMED examination and systematically benchmarking a broad spectrum of proprietary and open-weight LLMs on this high-stakes assessment. Unlike prior evaluations that report aggregate accuracy alone, we contextualize model performance using the official Item Response Theory framework employed for human examinees, enabling direct comparability with institutional proficiency thresholds. By combining structured dataset release, large-scale comparative benchmarking, and psychometric alignment, this work provides an updated assessment of the current stage of LLM capability in a national, high-stakes medical evaluation setting.

3 Methodology

To evaluate model performance, we constructed a structured dataset based on the 2025 administration of the ENAMED examination. The dataset construction process comprised three phases: data acquisition, automated extraction with human review, and multimodal adaptation[1].

Data Acquisition and Extraction Source materials were obtained from the official INEP repository,[2], consisting of the examination booklet and official answer key in PDF format. A rule-based parsing pipeline converted the unstructured PDFs into machine-readable form by isolating question statements and alternatives, removing document artifacts such as headers and page numbers, and aligning correct answers with the official key.

Data Curation and Cleaning A human-in-the-loop review was conducted to correct residual extraction errors. Answer alternatives fragmented during parsing were manually realigned, clinical tables were converted to Markdown to improve tokenizer compatibility, and questions annulled by the examination board or excluded from official scoring were removed to ensure consistency with evaluation criteria.

Reference Matrix Classification For domain-level analysis, each question was mapped to seven competency areas from the Common Reference Matrix (Ministério da Educação (MEC) and INEP, 2025). Lacking official labels, we employed a model-as-judge consensus (majority vote from `Gemini 3 Pro`, `GPT-5`, and `Sabiá 4`). This yielded almost perfect agreement (Fleiss' $\kappa = 0.82$ (Landis and Koch, 1977)), with the highest pairwise concordance between GPT-5 and Gemini 3 Pro ($\kappa = 0.907$), followed by GPT-5 and Sabiá 4 ($\kappa = 0.798$).

Multimodal Processing A subset of questions ($n = 3$) required interpretation of clinical images. To enable evaluation with text-only models, textual descriptions of the visual stimuli were generated using the `gemini-3-pro-preview` model. These descriptions were reviewed by a medical student to improve clinical clarity, primarily correcting photographic artifacts, anatomical imprecision, and misleading surface attributes. Importantly, this intermediate review step does not constitute a validation of the generated descriptions; the absence of assessment by board-certified clinicians remains a limitation outside the scope of this study.

Final Dataset Statistics After cleaning and filtering, the final dataset comprises 90 multiple-choice questions. Each item is represented in JSON format and includes the question text, four answer options,

[1]The dataset is publicly available at: https://huggingface.co/datasets/recogna-nlp/enamed-2025.

[2]https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enamed/provas-e-gabaritos

the ground-truth label, and, where applicable, a generated image description.

4 Experimental Setup

4.1 Computational Infrastructure

All local inference experiments were conducted on a computing node equipped with a single NVIDIA H100 Tensor Core GPU. This hardware configuration supports high-throughput inference for open-weight models through large memory bandwidth and Tensor Core acceleration, being executed using the `HuggingFace Transformers` library. Proprietary models were accessed via their respective official APIs. All models were evaluated in a zero-shot setting using the standardized prompt template detailed in Appendix A.

4.2 Model Selection

To evaluate a range of current Large Language Models in medical reasoning, a cohort of 22 models was selected. The models were stratified into three categories based on access modality and adaptation strategy.

Proprietary Frontier Models These models represent the state of the art (SOTA) in reasoning and knowledge retrieval, trained on massive multilingual corpora and accessible via an API. The **Gemini 3** family (Google) was evaluated, including `gemini-3-pro` and `gemini-3-flash`, both based on mixture-of-experts architectures with multimodal capabilities. The **GPT-5** family (OpenAI) was evaluated, including `gpt-5` and its smaller variants, `gpt-5-mini` and `gpt-5-nano`, to examine the relationship between model scale and performance. We also included proprietary models pre-trained specifically on Portuguese data, rather than merely fine-tuned. We selected the **Sabiá 4 Family** (Maritaca AI), a suite of models specialized for the Lusophone context. We evaluated `Sabia 4`, designed for complex reasoning tasks, and `Sabiazinho 4`, optimized for low-latency applications. These models provide a reference point for evaluating the impact of Portuguese-language specialization relative to large general-purpose models.

Open-Weight Generalist Models We selected a range of state-of-the-art open-source models to evaluate the capabilities of accessible AI on local hardware without any specific medical adaptation. This cohort includes the **Qwen 3** family (Yang et al., 2025), known for strong multilingual performance, where we tested the 14B, 8B, and 4B parameter variants to observe scaling laws. We also included **Phi-4** (Abdin et al., 2024), a compact reasoning model (15B and 4B) trained heavily on high-quality synthetic data, and **Llama 3.1** (Grattafiori et al., 2024), specifically the 8B parameter variant, which serves as the standard baseline for general-purpose open LLMs.

Open-Weight Adapted Models This category encompasses open models that have undergone post-training adaptation for specific domains to test the efficacy of fine-tuning versus pre-training. In the medical domain, we evaluated **MedGemma** (Sellergren et al., 2025), built upon the Gemma 3 architecture (27B and 4B) and fine-tuned on biomedical corpora such as PubMed and MIMIC-III, as well as `MMed-Llama-3` (Qiu et al., 2024), an 8B model further pre-trained on the MMedC multilingual medical corpus. Regarding Portuguese and clinical adaptations, we assessed the **Bode Family** (Paiola et al., 2025), including `Bode 3.1-8B` (general Portuguese adaptation) and `DrBode-240k` (medical fine-tuning on Brazilian clinical cases). Finally, to benchmark progress over previous generations, we included legacy baselines such as `Sabiá-7B` (Pires et al., 2023)(based on Llama 1) and `Clinical-BR-LlaMA-2-7B` (de Souza Pinto et al., 2024).

4.3 Evaluating Measures

Model performance was evaluated using both standard classification metrics and the official psychometric framework adopted by ENAMED. The latter is based on Item Response Theory (IRT) and enables estimation of model proficiency (θ) on the same latent scale used to assess human examinees.

Standard Classification Metrics We model the task as a four-way multiple-choice classification problem, where each model predicts an answer $\hat{y}_i \in \{A, B, C, D\}$ for item i with ground-truth label y_i. We report Accuracy and Macro-F1, defined respectively as $\frac{1}{N}\sum_i \mathbb{I}(\hat{y}_i = y_i)$ and the unweighted mean of per-class F1 scores over $\{A, B, C, D\}$. Divergences between Accuracy and Macro-F1 are interpreted as indicators of asymmetric class behavior.

Psychometric Evaluation Beyond standard classification metrics, performance was evaluated using the official ENAMED psychometric framework based on the Rasch (1PL) Item Response Theory

model, enabling direct comparison with human examinees. In the Rasch model, each item i is characterized by a difficulty parameter b_i, defined as the proficiency level at which the probability of a correct response is 50%. For a model j with latent proficiency θ_j, the probability of correctly answering item i is:

$$P(U_{ij} = 1 \mid \theta_j, b_i) = \frac{1}{1 + e^{(b_i - \theta_j)}}. \quad (1)$$

Both θ_j and b_i are defined on the same logit scale. We used the official item difficulty parameters provided by INEP, without recalibration, consistent with the examination's 1PL scaling framework.

Model proficiency was estimated using the IRT True-Score (TS) estimator adopted by INEP. For each model, the observed raw score was mapped to the corresponding latent proficiency value by inverting the test characteristic curve implied by the Rasch model parameters. Estimation was conducted over the standard interval $[-4, 4]$, consistent with official technical documentation. The latent scale ($\mu = 0$, $\sigma = 1$) was then linearly transformed to the ENAMED reporting scale. Following the Modified Angoff procedure, the minimum proficiency threshold corresponds to $\theta = -0.40$, equivalent to a score of 60.0.

Institutional Concept (Enade Concept) We evaluate the models collectively using the **Enade Concept**, the 1–5 categorical rating employed by the Brazilian Ministry of Education to assess medical schools. The concept is determined by the proportion of evaluated models meeting the official proficiency threshold, which is mapped to discrete levels: Level 1 ($< 40\%$), Level 2 (40%-59%), Level 3 (60%-74%), Level 4 (75%-89%), and Level 5 ($\geq$ 90%). This metric summarizes the proportion of evaluated models meeting the proficiency threshold relative to institutional benchmarks used in medical education.

5 Results

Table 1 presents the evaluation metrics on the ENAMED dataset. The models are ranked by global accuracy, revealing distinct performance tiers driven by model scale, architecture, and training methodology.

The results reveal a clear stratification of capabilities across model classes. Proprietary frontier architectures, particularly the **Gemini-3** and **GPT-5** families, consistently achieved the highest performance on the medical benchmark, forming a distinct upper tier. `Gemini-3-pro` attained the best overall accuracy (98.89%), followed closely by `GPT-5` and `Gemini-3-flash`, all of which exhibited near-ceiling performance. Notably, the Portuguese-centric proprietary model `Sabiá 4` achieved 93.33% accuracy, surpassing `GPT-5-mini` and competing with other high-performing proprietary variants. Accuracy and Macro-F1 were nearly identical across models, indicating minimal class imbalance effects.

A second performance tier comprises primarily efficiency-oriented proprietary variants and regionally optimized models. `GPT-5-mini` (91.11%) and the Brazilian-specialized `Sabiazinho 4` (87.78%) demonstrated strong performance, with `Sabiazinho 4` marginally outperforming `GPT-5-nano` (86.67%). While these models also outperformed the evaluated open-weight generalist baselines, such as `Qwen3-14b` (81.11%), this comparison should be understood as conditional on the model scales examined in this study. Specifically, the open-weight results reflect mid-scale configurations rather than the largest available variants. Accordingly, the observed dominance of proprietary models is robust within the evaluated regime, but should not be interpreted as a definitive comparison against the full upper bound of open-weight architectures.

5.1 Impact of Model Scaling

The evaluation reveals consistent performance stratification within individual model families, indicating that reductions in model capacity or deployment class are associated with measurable losses in medical reasoning performance. Importantly, these observations reflect intra-family trends under the configurations evaluated, rather than a universal comparison across all possible model scales.

Within the proprietary **GPT-5** family, a clear tiering is observed across efficiency-oriented variants. Performance decreases from 97.78% for `GPT-5` to 91.11% for `GPT-5-mini` and further to 86.67% for `GPT-5-nano`. Although architectural details and parameter counts are not publicly disclosed, this pattern suggests systematic trade-offs between deployment efficiency and reasoning capability within a single proprietary model lineage.

In contrast, open-weight families exhibit sharper performance degradation as model size decreases. For the **Qwen3** family, accuracy declines substan-

Model	Raw Score	Accuracy	Macro-F1	ENAMED Score	Proficiency
Gemini 3 Pro Preview	**89**	**0.9889**	**0.9886**	**137.19**	Proficient
Gemini 3 Flash Preview	88	0.9778	0.9773	131.87	Proficient
GPT-5	88	0.9778	0.9778	131.87	Proficient
Sabia 4	84	0.9333	0.9325	110.34	Proficient
GPT-5-mini	82	0.9111	0.9109	104.24	Proficient
Sabiazinho 4	79	0.8778	0.8779	97.13	Proficient
GPT-5-nano	78	0.8667	0.8676	95.11	Proficient
Qwen 3 14B	73	0.8111	0.8123	86.54	Proficient
MedGemma 27B	69	0.7667	0.7677	80.87	Proficient
Phi 4 15B	68	0.7556	0.7527	79.56	Proficient
Gemma 3 12B	62	0.6889	0.6873	72.31	Proficient
LLaMA 3.1 8B	60	0.6667	0.6660	70.07	Proficient
Qwen 3 8B	59	0.6556	0.6520	68.97	Proficient
Bode 3.1 8B	56	0.6222	0.6212	65.74	Proficient
MedGemma 4B	54	0.6000	0.5990	63.64	Proficient
Qwen 3 4B	53	0.5889	0.5892	62.60	Proficient
MMed-Llama-3-8B	46	0.5111	0.4653	55.43	Not Proficient
Gemma 3 4B	43	0.4778	0.4648	52.37	Not Proficient
DrBode 240k	42	0.4667	0.4616	51.35	Not Proficient
Phi 4 Mini 4B	40	0.4444	0.4230	49.29	Not Proficient
Sabiá-7B	32	0.3556	0.3455	40.75	Not Proficient
Clinical-BR-LlaMA-2-7B	Failed to perform				

Table 1: Performance of Large Language Models on the ENAMED 2025 examination. Models are ranked by the official psychometric score (ENAMED Score).

tially from 81.11% in the 14B model to 65.56% and 58.89% in the 8B and 4B variants, respectively. A similar trend is observed in the **Phi-4** family, where the 15B model achieves 75.56% accuracy, while the 4B variant drops to 44.44%. These results indicate that conventional size-reduction strategies in open-weight models are associated with pronounced losses in diagnostic performance under the evaluated conditions.

We limited the analysis of open-weight models to configurations that could be executed within available GPU resources. Consequently, the observed scaling trends do not reflect the full performance potential of larger open-weight architectures, such as higher-parameter variants of the Qwen or LLaMA families. Within these constraints, proprietary models consistently maintain higher performance across multiple deployment tiers, whereas open-weight models exhibit steeper degradation as scale decreases.

5.2 Generalist vs. Domain-Adapted Models

The results reveal that performance differences between generalist and domain-adapted models are strongly mediated by underlying architecture and pre-training regime, rather than by domain specialization alone. In particular, the `Qwen3` family consistently outperforms both `Gemma` and `Phi` models at comparable or smaller parameter scales, suggesting that architectural design choices and multilingual pre-training confer a substantial advantage in medical reasoning tasks conducted in Portuguese.

This effect is evident in the comparison between `MedGemma-27B` and `Qwen3-14B`. Despite having nearly twice the parameter count and being explicitly fine-tuned for the medical domain, `MedGemma-27B` (76.67%) is outperformed by the generalist `Qwen3-14B` (81.11%). However, domain adaptation is not without benefit: `MedGemma-4B` (60.00%) exhibits a measurable improvement over its base counterpart, `Gemma-3-4B` (47.78%), indicating that medical fine-tuning can partially compensate for architectural and scale limitations, particularly in smaller models.

A contrasting pattern emerges among older, domain-adapted Portuguese and medical models. `Clinical-BR-LLaMA-2-7B`, `Sabiá-7B`, and `DrBode-240K` are all based on pre-2023 architec-

tures, which lack the instruction-following fidelity, multilingual robustness, and reasoning capacity of more recent model families. These architectural constraints appear to outweigh the benefits of language or domain adaptation, leading to substantially degraded performance. In some cases, most critically in `Clinical-BR-LLaMA-2-7B`, these models fail to reliably follow task instructions or produce valid answer selections, rendering them unsuitable for this benchmark.

Overall, these findings indicate that domain or language specialization alone is insufficient to overcome limitations imposed by outdated architectures or weaker pre-training regimes. Modern architectural advances and strong general reasoning capabilities remain prerequisites for effective domain adaptation in medical NLP tasks.

5.3 Institutional Performance Assessment

To provide a holistic assessment of the current state of Generative AI in medicine, we applied the official Enade Concept methodology to our set of evaluated models, treating them as a single "graduating cohort" of medical students. As shown in Table 1, 16 out of the 22 evaluated models achieved the minimum proficiency threshold, with one of them failing to perform and being taken out of the evaluation. According to the official conversion scale, where Level 1 represents $< 40\%$ proficiency and Level 5 represents $\geq 90\%$, our "LLM Class of 2025" falls within the $75\% - 90\%$ range.

Consequently, the aggregate performance of current Large Language Models achieves Enade Concept 4. This corresponds to a high-quality performance level under official criteria. It is critical to note that this equivalence is strictly confined to psychometric performance under the ENAMED evaluation framework and does not imply equivalence in supervised clinical practice, ethical judgment, or patient interaction competencies.

5.4 Error Analysis and Qualitative Evaluation

An item-level analysis reveals that model failures are not uniformly distributed but cluster around specific hard questions.

Alignment with Human Psychometrics To assess whether the difficulty perceived by LLMs aligns with human psychometric benchmarks, we analyzed the correlation between the official Item Difficulty (b) and the aggregated model accuracy (Figure 1). We observed a moderate negative correlation ($\rho = -0.4698$, $p < 0.001$). While the directionality indicates that models generally perform worse on items with higher discrimination parameters, the magnitude of the correlation indicates a pronounced alignment gap. Unlike human candidates, whose performance degrades linearly with item complexity, LLMs exhibit a jagged difficulty curve, often solving hard memorization-based items while failing easy context-dependent questions.

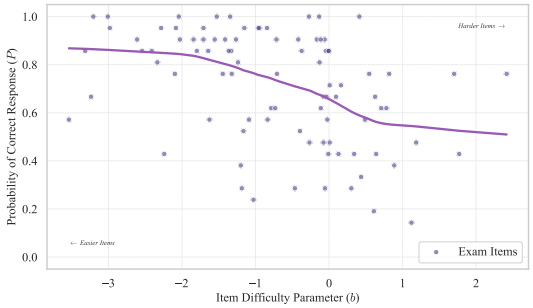

Figure 1: Empirical Item Characteristic Curve comparison. The scatter plot illustrates the probability of correct response for the models as a function of the official item difficulty (b). The trend line indicates a moderate negative correlation ($\rho = -0.4698$), highlighting the divergence between human and machine perceptions of difficulty.

Consensus Errors A significant finding is the presence of consensus errors, where models uniformly hallucinate the same incorrect procedure (e.g., Q61, Q20, Q23). The most prominent example is Question 83, which deals with cervical cancer screening guidelines (see Appendix Table 3 for the full clinical vignette). In this scenario, 100% of the failing models selected "Colposcopy" over the correct protocol of repeating the exam. This error highlights a critical misalignment between general medical knowledge and specific Brazilian public health protocols. While international or private practice guidelines might suggest immediate investigation for LSIL, the *Brazilian Guidelines for Cervical Cancer Screening* (Ministry of Health/SUS) explicitly recommend a conservative approach for LSIL in this age group. The models exhibited a bias for action, preferring an interventional procedure over the correct watchful waiting protocol, likely driven by training data dominated by diverse international guidelines rather than localized SUS protocols.

Multimodal Disparity While most image-based questions (e.g., Q4, Q5, Q53) were solved by nearly all models, Question 96 emerged as a significant outlier. For this item, the text description and the accompanying image clearly describe a *soft, painful* ulcer with irregular borders (characteristic of Chancroid). However, models frequently predicted Syphilis, which typically presents as a *hard, painless* chancre. This suggests a frequency bias in multimodal reasoning: the models likely ignored the fine-grained visual/textual semiotics provided in the vignette and defaulted to the most common statistical cause of genital ulcers (Syphilis). Possibly, current multimodal LLMs may struggle to ground specific visual features when they contradict a strong prior probability heuristic.

The Solved Subset Approximately 15% of the exam (e.g., Q1, Q8, Q77) achieved a difficulty score of 0.0, being answered correctly by every single model. A qualitative review reveals that these questions often pertain to medical ethics, the humanities, or highly standardized trauma protocols. For instance, Question 77 assesses cultural competence in the treatment of Indigenous populations (Tikuna ethnicity). All models correctly identified the need to "recognize the knowledges and practices" of the population. This uniform success likely stems from the extensive Reinforcement Learning from Human Feedback (RLHF) applied to modern LLMs, which heavily penalizes insensitivity and rewards culturally competent, empathetic responses. Similarly, questions such as Q93 (Chemical Eye Burn), which requires immediate irrigation, constitute algorithmic medical knowledge in which the standard of care is universal and unambiguous, thereby minimizing the risk of hallucination.

5.5 Domain-Specific Performance

Performance stratified by medical domain reveals substantial heterogeneity that is obscured by aggregate accuracy. Table 2 reports mean accuracy and dispersion across the seven competency areas defined by the ENAMED reference matrix.

General Surgery and Mental Health exhibit the highest average performance across models, whereas Pediatrics constitutes the most challenging domain, with a mean accuracy below 60%. This pattern is consistent across both proprietary and open-weight architectures, suggesting that pediatric reasoning and presentation may pose systematic challenges for current LLMs rather than model-specific weaknesses.

Medical Specialty	**Accuracy** ($\mu \pm \sigma$)	**Nº Items**
General Surgery	0.791 ± 0.408	11
Mental Health	0.779 ± 0.416	10
Family and Community Medicine	0.755 ± 0.431	13
Medical Clinic	0.752 ± 0.432	27
Collective Health	0.689 ± 0.464	6
Gynecology and Obstetrics	0.653 ± 0.476	16
Pediatrics	0.580 ± 0.494	17

Table 2: Model performance by medical domain on the ENAMED 2025 dataset.

Beyond average accuracy, domain-wise variance exposes marked differences in model reliability. Frontier models such as `GPT-5` and `Gemini 3 Pro`, and `Gemini 3 Flash` display comparatively low dispersion across specialties, indicating stable generalist behavior even in lower-performing domains. In contrast, several open-weight and regionally adapted models exhibit highly uneven performance profiles, achieving near-ceiling accuracy in some domains while failing substantially in others.

This effect is particularly pronounced in Collective Health. Brazilian-trained models, including `Sabiá 4` and `Sabiazinho 4`, consistently outperform larger international models in this domain, reflecting the heavy reliance of ENAMED public health items on SUS-specific legislation and administrative frameworks. Models lacking localized pre-training, such as `Phi-4` and `MedGemma`, show sharp performance degradation despite strong results in clinically oriented domains.

These findings indicate that while clinical reasoning abilities transfer relatively well across languages and health systems, institutional and legal medical knowledge remains highly localized. Consequently, robust performance on national licensing examinations requires either targeted regional pre-training or explicit integration of local knowledge sources, particularly for deployment in public health and policy-sensitive settings.

5.6 Environmental Impact

To ensure that the pursuit of high-performance medical AI aligns with ecological sustainability, we estimated the carbon footprint of our evaluation using the Machine Learning Impact calculator (Lacoste et al., 2019). Experiments were conducted on a single NVIDIA H100 GPU within the Brazilian National Interconnected System, which has a grid emission factor of 0.0461 $kgCO_2/kWh$ due to the

predominance of hydroelectric and other renewable sources.

The total estimated footprint for the full benchmark was ≈ 0.3 kgCO_2eq. This minimal impact demonstrates that high-stakes medical evaluation need not entail high environmental costs. By leveraging Brazil's low-carbon energy matrix, we show that hosting open-weight models locally offers a sustainable alternative to carbon-intensive querying of global API endpoints, demonstrating that medical proficiency can be achieved without compromising environmental stewardship.

6 Conclusion

This study presented a comprehensive evaluation of twenty-two Large Language Models on the inaugural 2025 ENAMED, a high-stakes benchmark unifying undergraduate assessment and residency selection in Brazil. By testing a diverse cohort ranging from proprietary frontier models to open-weight and domain-specific architectures, we assessed the readiness of these systems to interpret complex clinical vignettes, adhere to Portuguese terminological nuances, and navigate the specific public health guidelines of the Unified Health System (SUS).

The results indicate a clear stratification of capabilities, with proprietary models demonstrating near-ceiling performance on the majority of textual items, although residual systematic errors remain. The proprietary frontier models, specifically Gemini-3-pro, GPT-5, and Sabia 4, achieved accuracies of 98.89%, 97.78%, and 93.33%, respectively, approaching the theoretical ceiling of the examination. This performance far exceeds the threshold of unsatisfactory performance noted in a significant portion of medical universities. These results position LLMs primarily as tools for benchmarking and controlled educational support, rather than for autonomous clinical decision-making. Despite high aggregate performance, the presence of systematic and consensus errors, coupled with domain-specific variability, precludes reliable unsupervised use in real-world medical settings. Furthermore, current LLMs lack clear accountability mechanisms, reinforcing that their outputs must remain under human supervision, particularly in high-stakes clinical contexts.

Critically, our findings challenge the prevailing assumption that domain-specific fine-tuning is strictly necessary for medical proficiency. Generalist models showed capable of outperforming medically adapted counterparts when parameter count was held constant. This suggests that for standardized examinations, the reasoning capabilities derived from pre-training outweigh the benefits of targeted biomedical adaptation on smaller architectures. Additionally, the environmental analysis highlights a distinct advantage for the Brazilian ecosystem: while proprietary APIs offer peak accuracy, local inference of open-source models benefits from Brazil's low-carbon energy grid, providing a sustainable path to national technological independence.

However, the transition from examination passing to clinical utility is not without risks. The identification of consensus errors, in which distinct model families confidentially generated the same incorrect distractor, suggests latent biases or shared misconceptions in pre-training corpora that require auditing. Finally, performance heterogeneity across domains (Table 2) indicates that LLMs should not be assumed to generalize uniformly across all areas of medicine, particularly in policy and region-specific contexts.

Limitations

The primary limitation of this study lies in the multimodal evaluation methodology. Because there is no board-certified medical validation of AI-generated descriptions of clinical images, performance on visual questions serves as a proxy rather than a definitive assessment of visual diagnostic capability. Future work should prioritize end-to-end multimodal evaluation using raw image inputs and expand the error analysis to qualitative audits of "consensus failures" to ensure these systems are robust enough for real-world deployment in the Brazilian healthcare context

We also note the relatively small sample size ($N = 90$), given that ENAMED 2025 is the inaugural edition of the examination. While this limited volume restricts the granularity of subgroup analyses and results in wider statistical confidence intervals, it simultaneously ensures a high degree of data sanity. Future work should prioritize expanding this dataset as subsequent editions of ENAMED are released, allowing for longitudinal tracking of LLM performance. Additionally, the development of end-to-end multimodal evaluation pipelines is necessary to rigorously assess visual diagnostic capabilities without reliance on text-based proxies.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Andrew Maranhão Ventura D'addario. 2025. Healthqa-br: A system-wide benchmark reveals critical knowledge gaps in large language models. *Preprint*, arXiv:2506.21578.

João Gabriel de Souza Pinto, Andrey Rodrigues de Freitas, Anderson Carlos Gomes Martins, Caroline Midori Rozza Sawazaki, Caroline Vidal, and Lucas Emanuel Silva e Oliveira. 2024. Developing resource-efficient clinical llms for brazilian portuguese. In *Proceedings of the 34th Brazilian Conference on Intelligent Systems (BRACIS)*. In press.

Henrique Dias and Ana Helena Dias Pereira dos Ulbrich. 2022. BRATECA (Brazilian Tertiary Care Dataset): a Clinical Information Dataset for the Portuguese Language. *PhysioNet*. Version 1.1.

Gabriel Lino Garcia, Joao Renato Ribeiro Manesco, Pedro Henrique Paiola, Pedro Henrique Crespan Ribeiro, Ana Lara Alves Garcia, and Joao Paulo Papa. 2025. A Step Forward for Medical LLMs in Brazilian Portuguese: Establishing a Benchmark and a Strong Baseline . In *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 214–219, Los Alamitos, CA, USA. IEEE Computer Society.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174. PMID: 843571.

Ministério da Educação (MEC) and INEP. 2025. Portaria n° 478, de 18 de julho de 2025: Dispõe sobre a implementação da matriz de referência comum para a avaliação da formação médica. Diário Oficial da União. Accessed: 2026-02-07.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Lucas Emanuel Silva e Oliveira, Ana Carolina Peters, Adalniza Moura Pucca Da Silva, Caroline Pilatti Gebeluca, Yohan Bonescki Gumiel, Lilian Mie Mukai Cintho, Deborah Ribeiro Carvalho, Sadid Al Hasan, and Claudia Maria Cabral Moro. 2022. Semclinbr-a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical nlp tasks. *Journal of Biomedical Semantics*, 13(1):13.

Pedro Henrique Paiola, Gabriel Lino Garcia, João Vitor Mariano Correia, João Renato Ribeiro Manesco, Ana Lara Alves Garcia, and João Paulo Papa. 2025. The bode family of large language models: Investigating the frontiers of llms in brazilian portuguese. *Journal of the Brazilian Computer Society*, 31(1):917–938.

Pedro Henrique Paiola, Gabriel Lino Garcia, João Renato Ribeiro Manesco, Mateus Roder, Douglas Rodrigues, and João Paulo Papa. 2024. Adapting llms for the medical domain in portuguese: A study on fine-tuning and model evaluation. *Preprint*, arXiv:2410.00163.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. In *Intelligent Systems*, pages 226–240, Cham. Springer Nature Switzerland.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Preprint*, arXiv:2402.13963.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, and 12 others. 2023. Towards expert-level medical question answering with large language models. *Preprint*, arXiv:2305.09617.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Evaluation Prompt

To enhance reproducibility, we provide the Python function used to generate the zero-shot prompts. The prompt was designed in Portuguese to match the examination language and enforces a strict JSON output format to facilitate automated parsing.

```
Você é um médico especialista prestando
o Exame Nacional de Avaliação da
Formação Médica. Leia a questão
abaixo cuidadosamente e identifique a
alternativa correta.

QUESTÃO: { enunciado }

{% if img_descricao %}
DESCRIÇÃO DA IMAGEM: { img_descricao }
{% endif %}

ALTERNATIVAS:
(A) alternativas['A']
(B) alternativas['B']
(C) alternativas['C']
(D) alternativas['D']

INSTRUÇÕES:
1. Analise o caso clínico e a descrição
da imagem (se houver).
2. Responda APENAS com o formato JSON
contendo a letra da alternativa correta.
3. Não forneça explicações, apenas a
resposta.

Formato de Resposta: {"answer": "A"}
```

Figure 2: The zero-shot prompt template used for the evaluation, presented in Jinja2 syntax. The light purple background distinguishes the input prompt from the academic prose. The prompts are in Portuguese to align with the examination language.

B Qualitative Analysis of Representative Items

The following table presents a subset of the analyzed items, categorized by the error phenomenology observed in the Large Language Models. The "Model Consensus" column indicates the incorrect option most frequently selected by the models (Distractor), while "Ground Truth" indicates the official correct answer.

ID	Clinical Vignette	Model Prediction	Correct
	Consensus Errors (Systematic Misconceptions)		
83	Mulher de 32 anos, sexualmente ativa, comparece à consulta com o médico de família e comunidade para realização do seu primeiro exame preventivo. O médico realiza a coleta de citologia oncótica. Após 3 semanas, a paciente retorna com o resultado "presença de lesão intraepitelial de baixo grau". Considerando esse resultado, qual é a conduta adequada do médico?	Encaminhar para a realização de colposcopia	Repetir o exame citopatológico em 6 meses
61	Mulher de 35 anos, diabética, com laqueadura tubária bilateral, procurou atendimento médico com queixa de prurido genital e disúria terminal, com 7 dias de evolução. Recentemente, fez uso de antibiótico para tratamento de abscesso dental. Ao exame especular, notava-se edema vulvar, hiperemia, fissura, corrimento esbranquiçado e teste das aminas negativo. Com base no agente etiológico mais provável, o tratamento é	metronidazol, 1 aplicador, via vaginal, por 10 noites.	miconazol, 1 aplicador, via vaginal, por 7 noites.
20	Mulher travesti de 28 anos, profissional do sexo, comparece à Unidade Básica de Saúde (UBS) em demanda espontânea. Relata relações sexuais frequentes com diferentes parceiros, com uso inconsistente de preservativos, principalmente durante relações anais receptivas. Há 2 dias teve uma relação sexual desprotegida com um cliente que se recusou a usar camisinha. Nunca utilizou medicamento para profilaxia pré-exposição (PrEP) ou pós-exposição (PEP) à infecção pelo HIV. Considerando que a paciente está assintomática no momento, qual a melhor estratégia de prevenção?	Oferecer teste rápido para HIV e sífilis; prescrever PrEP de início imediato; orientar sobre as vacinas disponíveis no SUS para seu grupo populacional.	Realizar testagem rápida para HIV e sífilis; prescrever PEP mediante resultado não reagente para HIV e programar início da PrEP após término da PEP.
	Ambiguity & Entropy		
96	Homem de 30 anos chega para consulta em Unidade Básica de Saúde (UBS) devido à astenia e úlcera no pênis. Trabalha como profissional do sexo e nem sempre faz uso de preservativo. Há cerca de 3 meses, vem notando emagrecimento (10 kg no período), astenia, febre baixa sem horário fixo e, há 1 semana, observou o aparecimento de úlcera dolorosa no pênis. Nega secreção uretral. Ao exame físico, apresenta-se emagrecido, com uma lesão ulcerada com bordas elevadas sem secreção de aproximadamente 3 centímetros logo abaixo da glande, rasa e de base mole, além de linfonodomegalia inguinal direita, com sinais inflamatórios, sem fistulização	Veneral Disease Research Laboratory (VDRL); reagente; benzilpenicilina benzatina 1,2 milhão de unidades, intramuscular, dose única.	microscopia de esfregaço do fundo da úlcera; Gram negativos agrupados em correntes; azitromicina 500 mg, via oral, 2 comprimidos em dose única.
17	Paciente de 20 anos, sexo masculino, vítima de colisão "automóvel a muro", sem cinto de segurança, é atendido ainda na cena pelo Serviço Móvel de Atendimento de Urgência (SAMU). Exame físico: paciente torporoso; saturação de O2 de 60%, em ar ambiente; frequência respiratória de 28 irpm; frequência cardíaca de 112 bpm; pressão arterial de 90 x 50 mmHg. Desvio da traqueia para a direita, turgência de veias jugulares, hipofonese de bulhas cardíacas e diminuição acentuada do murmúrio vesicular à esquerda. Qual é a conduta adequada no atendimento pré-hospitalar?	Reposição volêmica.	Toracocentese
	Solved (Ceiling Effect)		
77	Uma equipe de saúde da família realiza atendimento itinerante a comunidades ribeirinhas e aldeias indígenas na Região Amazônica. Em visita, uma médica recém-chegada observa que uma mulher ribeirinha evita contato visual durante a consulta e responde às perguntas apenas com monossílabos. Em outra situação, um indígena da etnia Tikuna não aceita ser atendido sozinho e insiste na presença de um pajé da comunidade.	promover espaços formativos para a equipe assistencial, reconhecendo saberes e práticas das populações atendidas.	
93	Paciente de 45 anos atendida na Unidade Básica de Saúde (UBS) com dor ocular. Referiu que estava realizando limpeza doméstica com alvejante e deixou atingir o olho, acidentalmente. Ao exame físico, foi observada presença de hiperemia intensa com opacidade da córnea e queimadura química da pálpebra superior do olho direito. Qual é o correto manejo da paciente?	Lavagem ocular com solução fisiológica e avaliação imediata do especialista.	

Table 3: Qualitative analysis of representative items. Columns 3 and 4 show the divergence between the consensus of the models (Distractor) and the official ground truth.