Developing Resource-Efficient Clinical LLMs for Brazilian Portuguese

João Gabriel de Souza Pinto[1,2][0009−0006−7133−3074], Andrey Rodrigues de Freitas[1][0009−0008−6511−3248], Anderson Carlos Gomes Martins[3][0009−0001−4637−0301], Caroline Midori Rozza Sawazaki[1][0009−0008−2251−1131], Caroline Vidal[1][0009−0003−6904−2719], and Lucas Emanuel Silva e Oliveira[1,2][0000−0003−1811−5087]

[1] Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba (PR), Brasil
[2] Comsentimento NLP Lab, São Paulo (SP), Brasil
http://www.comsentimento.com.br
[3] Instituto Federal de Goiás (IFG), Luziânia (GO), Brasil

Abstract. In this study, we developed and evaluated two medical large language models, Clinical-BR-LlaMA-2-7B and Clinical-BR-Mistral-7B-v0.2, specifically designed for Brazilian Portuguese. Utilizing the Low-Rank Adaptation (LoRA) technique, our models achieved significant improvements in generating synthetic clinical text, particularly in terms of Authenticity of Format and Structure, Spelling Accuracy, and Clinical Coherence. The evaluation, conducted by medical students using a 5-point Likert scale, demonstrated the effectiveness of our approach compared to baseline models. The scores indicate superior performance compared to baseline models such as LlaMA-2-7B and Mistral-7B-v0.2. Our results suggest that these resource-efficient models can effectively generate clinically relevant text, maintaining high standards of structure, accuracy, and coherence. Future work will focus on expanding datasets, refining evaluation protocols, and enhancing model robustness to further improve performance across various medical tasks.

Keywords: Large Language Models · Medical NLP · Generative AI

1 INTRODUCTION

Foundation models have gained significant attention for their strong ability to process various data types, particularly text. Large Language Models (LLMs), a subset of these architectures, are central in natural language processing (NLP) research due to their capability to generate coherent and contextually relevant text. This makes them instrumental in applications from automated content creation to conversational agents. The versatility and scalability of LLMs highlight their potential to revolutionize interactions with digital information, driving advancements in AI.

In the medical field, LLMs enhance the efficiency and accuracy of clinical decision-making processes[5]. These models excel in interpreting complex medical language and extracting relevant information from unstructured texts[29], aiding in the understanding of patient histories and treatment outcomes[29]. LLMs streamline administrative procedures and support diagnostic/therapeutic decisions[28], potentially improving patient outcomes and operational efficiencies[5]. For example, LLMs like Med-PaLM can achieve passing scores on medical licensing exams, showcasing their potential in clinical knowledge and question-answering tasks[29]. They also extract structured information from clinical notes and reports in various languages, enhancing diagnostic accuracy and aiding in health-related information dissemination[28].

The lack of resources in languages other than English is a persistent issue in NLP research, and the development of LLMs is no exception. This gap is further worsened by the substantial computational power required to train LLMs from scratch, a resource predominantly accessible to large corporations. It is essential to find ways to make these technologies accessible to the research and academic community by leveraging publicly available multilingual models and adapting them to specific contexts.

In this work, we focus on creating a useful Medical LLM for the Brazilian Portuguese language (pt-BR) using minimal computational resources. Our methodology involved continuing the pre-training (or unsupervised fine-tuning) of LlaMA-2-7b and Mistral-7b-v0.2 base models using three different datasets of clinical narratives. We applied the LoRA (Low-Rank Adaptation) technique to enhance the efficiency of this process. The new models, called Clinical-BR-LlaMA-2-7B and Clinical-BR-Mistral-7B-v0.2, were evaluated on their ability to generate synthetic clinical data, and results showed that our models outperformed other baseline models in generating clinically relevant data.

This initiative is part of a collaborative project between HAILab and Comsentimento, named MED-LLM-BR[1], which aims to develop multiple medical LLMs for pt-BR, including base models and task-specific models, with different sizes. To the best of our knowledge, this is the first clinical LLM model publicly available for the Portuguese language.

Our work addresses the specific needs and gaps in the Brazilian Health Informatics community, providing a valuable resource for clinicians and researchers to utilize in their healthcare AI projects. By overcoming linguistic and technical challenges, we contribute to the broader field of natural language processing and medical informatics, highlighting the significance of domain-specific knowledge in medical language processing.

2 RELATED WORK

The development of Large Language Models (LLMs) has revolutionized natural language processing (NLP), especially in the medical field. This section reviews advancements in Transformer architectures, the importance of transfer learning and fine-tuning, and computational strategies for optimizing these models. We examine key initiatives in other languages, particularly Portuguese, and highlight

[1] https://github.com/HAILab-PUCPR/MED-LLM-BR/

prominent medical LLMs to provide context for our resource-efficient Medical LLM for Brazilian Portuguese.

The advent of Transformers, introduced by Vaswani et al.[26], revolutionized the field of natural language processing (NLP). This architecture uses self-attention mechanisms to handle long-range dependencies in text, significantly improving performance over previous recurrent neural network (RNN) and convolutional neural network (CNN) models.

Transformers can be configured as encoder or decoder models, supporting both generative and non-generative tasks. Encoder models, like BERT[7], focus on understanding and encoding input text into a dense representation. Decoder models, such as GPT-3[3], generate text based on an input sequence. Encoder-decoder models, like the original Transformer[26], combine both functionalities and are effective for tasks like translation and summarization.

In the context of transfer learning, it's crucial to distinguish between base models and task-specific models. Base models, like BERT and GPT-3, are pre-trained on large corpora to learn general language representations. These models can then be fine-tuned on specific tasks, such as sentiment analysis or named entity recognition, to create task-specific models that retain the general language understanding while being optimized for particular tasks.

Training and fine-tuning large language models pose significant computational challenges. Parameter Efficient Fine-Tuning (PEFT) methods, like Low-Rank Adaptation (LoRA)[11], introduce trainable low-rank matrices into each layer of a pre-trained model, reducing trainable parameters while maintaining performance and lowering costs. Models like LlaMA 2[25] and Mistral[13] are designed for efficiency in training and inference. LlaMA 2 and Mistral 7B models offer superior performance and resource efficiency, with LlaMA 2 excelling in benchmarks and being accessible for various applications due to its smaller size. Mistral 7B's Mixtral architecture uses expert mixture techniques to enhance performance and scalability, making high-performance LLMs more accessible and practical for specialized tasks.

By utilizing PEFT methods, it is possible to adapt large models like those aimed at pt-BR more efficiently, enabling broader application despite computational constraints. These techniques are crucial for developing specialized models in resource-constrained environments, ensuring that advancements in NLP are accessible and applicable across different languages and domains.

While most large language models have been developed for English, significant efforts have been made to create models for other languages. For example, mBERT (multilingual BERT) handles 104 languages by training on a multilingual corpus[27], and XLM-R[6] extends this idea by pre-training on a massive multilingual dataset.

There are initiatives to develop models for pt-BR, such as Sabiá[20], Sabiá-2[1] and Bode[9]. However, relatively few models exist for this language, despite the large portuguese-speaking population (almost 250 million speakers).

In the medical field, models like BioBERT[16] and ClinicalBERT[12]are fine-tuned on domain-specific corpora to capture the unique terminology and context of medical texts. Med-BERT[21], trained on electronic health records (EHRs), predicts patient outcomes and assists in clinical decision-making.

Meditron and BioMistral have shown significant performance gains by extending pre-training on curated medical corpora[4] and[15]. GatorTronGPT, trained from scratch using 277 billion words of mixed clinical and English text with a GPT-3 architecture, improves biomedical NLP for medical research[19]. OpenBioLLM-70B leverages the Meta-Llama-3-70B-Instruct model to achieve state-of-the-art performance on various biomedical tasks[2].

In the pt-BR medical context, the two main resources are the BioBERTpt[23] and CardioBERTpt[22]. BioBERTpt is a biomedical LLM, fine-tuned from multilingual BERT using clinical and biomedical data, showing promising results in clinical text tasks like Named Entity Recognition and Negation detection. While CardioBERTpt was fine-tuned on clinical data specifically from Cardiology. The only available generative model for pt-BR is GPT2-bio-pt[24], which was trained exclusively on biomedical data from medical articles. The model has several limitations such as to be based on a model trained on automatically translated text, absence of clinical narratives knowledge and small context window (512 tokens).

3 METHOD

This chapter describes our methodology divided into three key steps: Data Acquisition, Model Architecture and Training, and Experimental Setup. Each subsection details the processes and techniques employed to build and evaluate our models.

3.1 Data acquisition

Our project combined data from three distinct clinical datasets, totaling 2.4GB of text and 309 151 121 tokens. The first one comes from the same data sources used in the SemClinBr project[18], which are composed of 2 100 546 clinical narrative entries from multiple Brazilian Hospitals. The dataset contains diverse document types (e.g., discharge summaries, ambulatory notes, nursing notes) and medical specialties (e.g., cardiology, nephrology, endocrinology). The electronic health record (EHR) data used in the study were de-identified and approved by the PUCPR Research Ethics Committee, certificate of presentation for ethical appreciation number 51376015.4.0000.0020.

The BRATECA dataset[8] was collected as well, and it is composed of 73 040 admission notes from 10 Brazilian Hospitals and associated with multiple medical departments (e.g., obstetrics, surgery, emergency, COVID-19, intensive care, ambulatory). All the data was anonymized and granted ethical approval by the National Research Ethics Committee under the number 46652521.9.0000.5530. The dataset is available under PhysioNet Credentialed Health Data Use[10].

Finally, we utilized the data used in Lopes et al. work[17], which consists mostly of neurology clinical cases collected from medical journals written in

European Portuguese. The dataset contains 3678 medical texts and it is publicly available[2].

3.2 Model Architecture and Training

In this study, we fine-tuned the LlaMA-2 and Mistral base models using the LlaMA-Factory framework[30], focusing on 7 billion parameter (7B) models to minimize computational resources. The models were trained over 2 epochs with a learning rate of 2e-5, using Low-Rank Adaptation (LoRA) for efficient weight updates. LoRA introduces trainable low-rank matrices into each layer, reducing trainable parameters and GPU memory requirements.

To further optimize memory and computational usage, we applied LoRA with 16-bit precision on the `q_proj` and `v_proj` projections, setting LoRA `R` to 8, LoRA `Alpha` to 16, and LoRA `Dropout` to 0.1. We used the AdamW optimizer with settings $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to balance rapid convergence with stability during training. The main difference between the two fine-tuned models was the `max_position_embeddings` parameter configuration: 4096 for LlaMA-2 and 32768 for Mistral.

Training was conducted on Google Cloud Platform (GCP) with an NVIDIA Tesla A100 GPU, 12 vCPUs, and 85 GB of RAM. The resulting models, Clinical-BR-LlaMA-2-7B and Clinical-BR-Mistral-7B-v0.2, had their Training Loss shown in Figure 1. We used 2.4 GB of clinical data for this study. Training the Clinical-BR-LlaMA-2-7B model involved 309,151,121 tokens and took 47 hours and 56 minutes, costing R$ 1250.94. The Clinical-BR-Mistral-7B-v0.2 model took 50 hours and 18 minutes, costing R$ 1202.73.

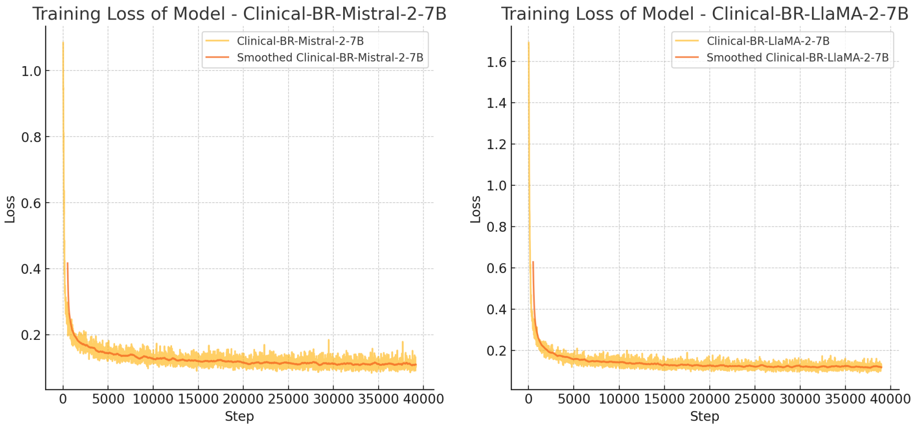


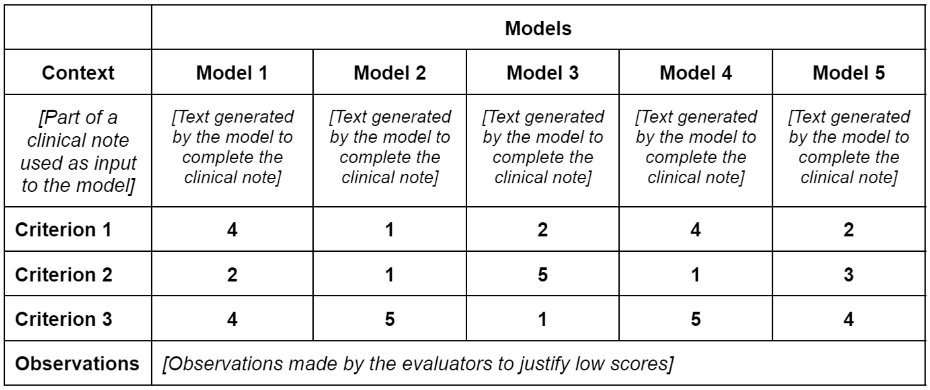
Fig. 1. Training Loss of Models

[2] https://github.com/fabioacl/PortugueseClinicalNER

3.3 Experimental setup

The models trained in this work are base models, trained unsupervised on a text corpus to learn clinical language representations. They serve as foundations for further fine-tuning for specific tasks. Evaluating base models typically involves benchmarks for tasks like language understanding, text generation, and question answering. However, these benchmarks can be imprecise, often relying on standard metrics that may not capture the nuances of language or generalization to unseen data. They might not reflect real-world performance, especially in specialized fields like healthcare, where domain-specific knowledge is crucial. For instance, most medical LLMs are evaluated on English Question Answering benchmarks like MedQA[14], which do not measure the model's ability to interpret clinical data in electronic medical records, the focus of this work.

To overcome limitations, we used the models to generate synthetic clinical text and evaluated their performance on three criteria using a 5-point Likert scale (Figure 2): Authenticity of Format and Structure, Spelling Accuracy, and Clinical Coherence.



	Models				
Context	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**
[Part of a clinical note used as input to the model]	*[Text generated by the model to complete the clinical note]*	*[Text generated by the model to complete the clinical note]*	*[Text generated by the model to complete the clinical note]*	*[Text generated by the model to complete the clinical note]*	*[Text generated by the model to complete the clinical note]*
Criterion 1	**4**	**1**	**2**	**4**	**2**
Criterion 2	**2**	**1**	**5**	**1**	**3**
Criterion 3	**4**	**5**	**1**	**5**	**4**
Observations	*[Observations made by the evaluators to justify low scores]*				

Fig. 2. Assessment card example. The evaluators had to fill the scores for each criterion, based on the output of each model

Criterion 1 - Authenticity of Format and Structure: Evaluated if the model's output adhered to clinical document norms, including layout and sectioning. The highest score was given to models that perfectly matched the standard.

Criterion 2 - Spelling Accuracy: Focused on linguistic correctness, including spelling and medical terminology in pt-BR. Models lost points for incorrect grammar or generating text in a different language.

Criterion 3 - Clinical Coherence: Assessed if the content logically correlated with the clinical history, focusing on relevance and precision. This criterion involved more subjectivity due to varying evaluator backgrounds.

We selected 100 random clinical notes, not part of the training corpus, and used excerpts as input for five models to complete the notes. This aimed to measure the models' proficiency in producing consistent clinical notes. Medical students with experience in clinical notes evaluated the generated texts from our models (Clinical-BR-LlaMA-2-7B, Clinical-BR-Mistral-7B-v0.2) and three reference models (LlaMA-2-7B, Mistral-7B-v0.2, Sabia-7B). They assigned scores from 1 to 5 on a Likert scale for each of the three criteria. The evaluation was blind; students did not know which model generated each note.

LlaMA-2-7B and Mistral-7B-v0.2 served as benchmarks to measure improvements from our training. Sabia-7B, a Portuguese-trained model, was included for comparison. This selection allowed a thorough performance analysis across benchmarks. GPT2-bio-pt was excluded due to its limited context size, which is inadequate for clinical text generation.

4 RESULTS

This section presents the evaluation outcomes of Clinical-BR-LlaMA-2-7B and Clinical-BR-Mistral-7B-v0.2 models, alongside three baselines: LlaMA-2-7B, Mistral-7B-v0.2, and Sabia-7B. The evaluation focused on Authenticity of Format and Structure, Spelling Accuracy, and Clinical Coherence, using a 5-point Likert scale. We analyze the average scores for each criterion, the frequency each criterion was met, and perform an error analysis. This helps us understand the strengths and weaknesses of each model, highlight improvements in our fine-tuned models, and identify areas for future enhancement.

4.1 Clinical Text Generation

The average score of the models for each criterion is presented in Table 1 and stacked in Figure 3. In addition, Table 2 presents the frequencies of score for each criterion in the evaluation of each model. In Table 3, we can see some examples of clinical notes generated with our models that were successful in the three evaluation metrics used (scores greater than 4 in all metrics), that is, they are cohesive texts with regard to structure, spelling and clinical coherence.

Authenticity of Format and Structure: the models were assessed on their adherence to the format and structure typical of clinical documents. Sabia-7B and Mistral-7B-v0.2 scored 3.84 and 3.85, respectively, indicating moderate adherence. LlaMA-2-7B demonstrated better consistency with a score of 4.45. Our models, Clinical-BR-LlaMA-2-7B and Clinical-BR-Mistral-2-7B, achieved the highest scores of 4.60 and 4.62, respectively. These models also had a significant number of perfect scores (5), reflecting their superior performance in maintaining document authenticity.

Spelling Accuracy: despite Sabia-7B not achieving above 4 score average in Criteria 1 and 3, it demonstrated excellent performance in orthographic correctness, likely due to its training on a large corpus of pt-BR data. In contrast, the Mistral-7B-v0.2 model showed significant difficulties, often generating random texts or words in English. LlaMA-2-7B, although not trained with a large Portuguese corpus, managed to perform well in this criterion. As indicated by the graph, our adjusted LlaMA model (i.e., Clinical-BR-LlaMA-2-7B) further improved upon the already good performance of LlaMA-2-7B. The Mistral-7B-v0.2 model showed a remarkable improvement, with its base model scoring an average of 3.82 and achieving an increase to 4.69, including 30 additional perfect scores (5) in the frequency distribution, indicating excellent performance.

Table 1. Average score of the models in the three criteria evaluated. In bold the models with the best score for each criterion.

Criteria	Mistral-7B-v0.2	Sabia-7B	LlaMA-2-7B	Clinical-BR-LlaMA-2-7B	Clinical-BR-Mistral-7B-v0.2
Authenticity of Format and Structure	3.85	3.84	4.45	**4.62**	4.60
Spelling Accuracy	3.82	4.56	4.39	4.59	**4.69**
Clinical Coherence	3.52	3.54	4.19	4.45	**4.46**

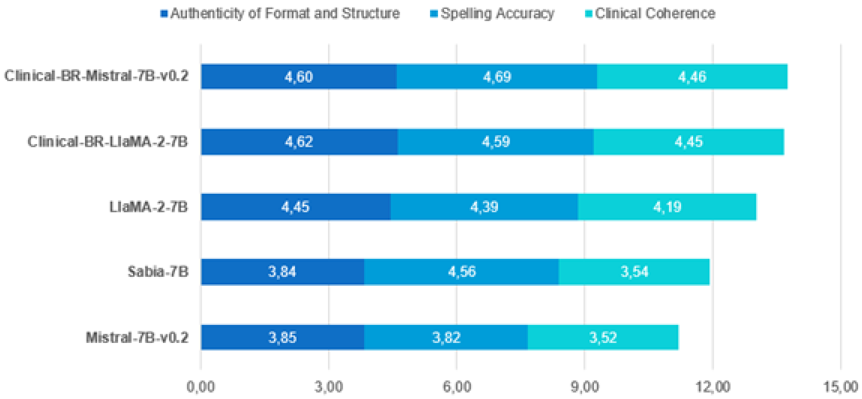


Fig. 3. Stacked average score of the models across the three evaluated criteria

Table 2. Score frequencies for each evaluated model. C1, C2 and C3 stands for criterion 1, 2 and 3 respectively.

LlaMA-2-7B				**Clinical-BR-LlaMA-2-7B**			
Score	**C1**	**C2**	**C3**	**Score**	**C1**	**C2**	**C3**
1	3	2	2	**1**	1	0	1
2	7	7	8	**2**	3	0	0
3	5	4	14	**3**	4	4	17
4	19	29	29	**4**	21	35	22
5	72	64	53	**5**	77	67	66
Mistral-7B-v0.2				**Clinical-BR-Mistral-2-7B**			
Score	**C1**	**C2**	**C3**	**Score**	**C1**	**C2**	**C3**
1	8	6	9	**1**	1	0	1
2	12	20	12	**2**	4	1	1
3	14	6	28	**3**	4	1	10
4	26	29	28	**4**	20	29	31
5	46	45	29	**5**	77	75	63

Sabia-7B			
Score	**C1**	**C2**	**C3**
1	12	2	11
2	10	3	12
3	13	5	24
4	21	23	29
5	50	73	30

Table 3. Examples of successfully generated clinical notes

#	**Model**	**Input context**	**Output**
1	Clinical-BR-LlaMA-2-7B	paciente expressa preocupações com a possibilidade de complicações metabólicas devido à SOP como diabetes tipo 2 e dislipidemia. Os exames de glicose em jejum e perfil lipídico apresentam alterações indicativas de resistência à insulina e dislipidemia. O tratamento proposto envolve a abordagem multifacetada da SOP.	O paciente expressou preocupações com a possibilidade de complicações metabólicas devido ao tratamento que serão abordadas com o paciente. Ao final do tratamento o paciente será acompanhado com o objetivo de monitorar o estado de saúde e retornar para consulta se necessário
2	Clinical-BR-Mistral-2-7B	Paciente tabagista crônico (20 anos de tabagismo) hipertenso há 10 anos com histórico de angina estável e episódios de dor torácica recorrentes nos últimos 6 meses. Realizou	angiografia coronária em 2007 com diagnóstico de estenose de 70% na origem da artéria esquerda e de 90% na artéria circunflexa. Em 2008 realizou angioplastia com implante de stent em ambas artérias.

Clinical Coherence: in terms of Clinical Coherence, Sabia-7B scored 3.54 and Mistral-7B-v0.2 scored 3.52, both struggling to maintain logical correlation within the clinical context. LlaMA-2-7B scored 4.19, indicating a better, though not optimal, performance. The models Clinical-BR-LlaMA-2-7B and Clinical-BR-Mistral-2-7B again led with scores of 4.45 and 4.46, respectively, demonstrating their ability to generate clinically coherent content. Frequency analysis revealed that these models had the highest number of perfect scores (5), underscoring their superior coherence.

4.2 Error Analysis

In this section, we perform an error analysis in the task of generating clinical texts, both from the models we trained and the original baseline models. Table 4 presents some examples of generated clinical texts that obtained low scores for one or more evaluation criteria.

All the evaluated models have demonstrated a behavior in which they repeat certain words or sequences multiple times in succession, like on error #2. This issue can probably be solved by adjusting some inference parameters. For example, when applying the Temperature = 0.1 and Repetition Penalty = 1.2, the output for error #2 would be "por sensação de pressão no peito. A doença foi diagnosticada há mais de um ano atrás, mas o paciente não tomou nenhuma medicação para tratamento. O exame clínico revelou: pressão arterial sistólica de 140 mmHg; frequência cardíaca regular de 76 bpm; pulso normal; auscultação cardiovascular sem alterações. . . ".

Although clinical texts are known for not having a standard formal structure, as each physician or institution may establish a different format, we expect a clinical note to contain at least a partial description of the patient's condition or visit, either in narrative or semi-structured form. Therefore, some notes generated contained an output more similar to a test question or teaching documentation than a clinical narrative (as shown in error #1). This was a more prevalent behavior of the Sabia-7B model, most likely due to its training data related to teaching exams. Regarding the problems in the spelling metric, most of the issues came from generating an output in English instead of Portuguese (like on error #3), being a more recurrent problem in the Mistral-7B-v0.2 model.

As for the clinical coherence dimension, the most complex for models to achieve maximum scores, most of the problems are associated with the use of medical terms that would make sense to be in that text section, but not for the patient's reported condition. For example, error #4 shows a text that suggests the use of atenolol, a drug widely used to treat hypertension, for the treatment of dyslipidemia. This is an expected problem, given that the standard behavior of models based on vector representations interprets all terms related to drugs as similar, since they are always used in similar contexts.We also had clinical coherence problems, in which the generated output uses keywords related to the patient's context, but the construction of sentences with these words does not make sense, as in error #5, where the word "dialysis" makes sense, since the patient has chronic kidney disease, but the presented sentence does not make sense from a clinical point of view.

Table 4. Examples of poorly generated clinical notes

#	Model	Input context	Output
1	Sabia-7B	Paciente com histórico de hipertensão tabagismo cardiopatia isquêmica com infarto agudo do miocárdio em 2000 que já havia recebido implante de stent na artéria descendente anterior proximal e média (ACTP primária em 2000) retorna	para avaliação de dor precordial em repouso. Qual a conduta a ser tomada? a) Realizar cateterismo cardíaco de urgência. b) Realizar cateterismo cardíaco eletivo. c) Realizar estudo hemodinâmico. d) Realizar estudo eletrofisiológico. e) Realizar estudo ecocardiográfico. Resolver Comentários (0)
2	Mistral-7B-v0.2	# HAS # IRC em HD desde jan/05 # prostatectomia em 1993 # RTU em 2003 # úlcera péptica em 2003 # FA com MP (Bradarritmia sintomática) desde 1998 # CRM em 2002 # angina instável e CAT em 2003 # fratura colo fêmur em 2005 Chega à EME-HCPA queixando-se de dor precordial de forte intensidade próxima à região da fúrcula esternal acompanhada	de dor no peito dor no esterno dor no epigástrico dor no abdômen dor no costela dor no flanco dor no quadril dor no perineu dor no pé dor no braço dor no pulso dor no dedo dor no joelho dor no tornozelo dor no pé dor no tornozelo dor no joelho dor no tornozelo...
3	Mistral-7B-v0.2	Diluição: Extraia 40ml do medicamento e misture com 80ml de Solução de Cloreto de Sódio a 0,9% em um frasco de 100ml. Administre 120ml por infusão intravenosa. Estabilidade:	24 horas a 25°C. # Description # What is this medicine? DILTIAZEM (dil TYE a zem) is a calcium
4	Clinical-BR-Mistral-7B-v0.2	paciente expressa preocupações com a possibilidade de complicações metabólicas devido à SOP como diabetes tipo 2 e dislipidemia. Os exames de glicose em jejum e perfil lipídico apresentam alterações indicativas de resistência à insulina e dislipidemia. O tratamento proposto envolve a abordagem multifacetada da SOP.	Para o tratamento da dislipidemia foi iniciado a atenolol 100 mg/dia e a simvastatina 20 mg/dia. Para o tratamento da resistência à insulina foi iniciado a metformina 500 mg/dia.
5	Clinical-BR-LlaMA-2-7B	Paciente com histórico de doença renal crônica (DRC) de estágio 3 diagnosticado há três anos durante um exame de rotina. Apresenta proteinúria persistente com níveis de creatinina sérica elevados e clearance de creatinina estimado em 45 ml/min/$1.73m^2$.	Apresenta história de frequência de uso de frascos de diálise com uso de frascos de diálise com alta e baixa qualidade...

5 DISCUSSION AND FUTURE WORK

In this study, we developed and evaluated two medical large language models, Clinical-BR-LlaMA-2-7B and Clinical-BR-Mistral-7B-v0.2, tailored for pt-BR. Our findings showed significant improvements in generating synthetic clinical text, particularly in structure, orthographic accuracy, and clinical coherence.

These improvements are due to targeted pre-training and fine-tuning on clinical narratives using Low-Rank Adaptation (LoRA) for efficient adaptation. The Clinical-BR-LlaMA-2-7B and Clinical-BR-Mistral-7B-v0.2 models produced outputs that adhered to clinical document norms, maintained high spelling accuracy, and generated clinically coherent content.

Clinical-BR-Mistral-7B-v0.2 handled long contexts better due to its larger context window, a critical attribute for long clinical documents. It is worth noting that the Mistral-7B-v0.2 model performed worse than LlaMA-2-7B in all evaluation criteria, maybe due to the low volume of pt-BR data in its training (this information was not made available by the team that developed the model). However, after our training on clinical data, the Clinical-BR-Mistral-7B-v0.2 model performed better or similarly to the Clinical-BR-LlaMA-2-7B model. This may indicate that architectural details of the Mistral-based models better handle adjustments to new contexts in a low volume of data scenario and with the use of PEFT techniques.

Our work focuses on medical LLMs for pt-BR, addressing unique challenges compared to models trained in other languages. This distinction is critical as pt-BR linguistic and clinical terminologies differ significantly, affecting model generalization.

BioMistral and MediTron models, trained on extensive medical corpora like PubMed Central, show performance gains in English medical QA tasks. However, our pt-BR models face fewer resources and less standardized evaluations, complicating direct performance comparisons. GatorTronGPT and OpenBioLLM-70B handle extensive clinical data in English, showcasing improvements in biomedical NLP tasks. In contrast, our approach optimizes limited resources while maintaining high performance in pt-BR clinical narratives, emphasizing computational efficiency with techniques like LoRA. This aspect is crucial for making advanced LLMs accessible to research communities with limited computational infrastructure.

Evaluation protocols often involve well-defined benchmarks in English. BioMistral's evaluation includes multilingual medical QA tasks, reflecting its capabilities. Our evaluation metrics are tailored to pt-BR, ensuring relevance but limiting direct comparisons with English-centric models.

As another limitation in the evaluation protocol of our project, we can mention that the use of medical students may not have been an optimal solution, especially when evaluating the clinical cohesion of the notes, as students have limited expertise regarding some conditions and treatments. Moreover, the length of the clinical notes, a critical factor in generating synthetic health data, was not included in our structure or cohesion assessment. Our primary aim was to identify the impact of pt-BR clinical training data on model performance, even in small volumes, compared to existing models. Consequently, our models are not recommended for generating fully coherent synthetic data. Despite high scores, we did not consider the note length or completeness, leading models to sometimes generate only a few words post-input. Additionally, using other clinical notes as input references may complicate synthetic data generation in environments with limited real clinical data.

While our models show promising results in pt-BR, future work should aim to fine-tune and evaluate the models in clinical downstream tasks, and align evaluation protocols more closely with international benchmarks to facilitate better comparisons. In addition to also building more specific benchmarks for the context of clinical texts, which have a very peculiar format and structure, differing greatly from scientific articles or medical licensing exams. Furthermore, expanding our dataset and incorporating more diverse clinical narratives can enhance model robustness and generalization.

Training a model with medical records data tends to perform better in tasks related to data extraction, interpretation, and generation in this context. However, we understand that several tasks in the medical field involve medical reasoning, and much of the knowledge needed for the model to perform this type of inference is found in biomedical texts and specific question-answering datasets. In this context, we intend to expand our models to the biomedical context as well, so that they are able to deal well with all types of medical data.

By addressing these differences, we highlight the unique contributions of our work and set the stage for future advancements in developing and evaluating medical LLMs across diverse languages and resource settings.

6 CONCLUSION

This study focused on developing and evaluating two medical large language models, Clinical-BR-LlaMA-2-7B and Clinical-BR-Mistral-7B-v0.2, made for medical Brazilian Portuguese context. These models demonstrated significant improvements over baseline models in generating synthetic clinical text, particularly in terms of Authenticity of Format and Structure, Spelling Accuracy, and Clinical Coherence, largely due to the efficient use of the LoRA technique and continued pretraining on a dataset of clinical notes.

The models' performance underscores the potential for integrating resource-efficient medical LLMs into clinical practice, facilitating the generation of clinical notes and supporting medical research in Portuguese-speaking contexts. However, limitations such as the subjectivity of the evaluation protocol, the expertise of the evaluators, and the scope of the dataset highlight areas for further enhancement.

Future research should aim to perform the models fine-tuning for specific tasks, align evaluation protocols with international benchmarks, build specific benchmarks for clinical texts, and expand datasets to include diverse clinical narratives and biomedical texts. These efforts will help improve model robustness and generalization, enabling the models to perform better in various medical tasks.

Our work highlights the feasibility of creating resource-efficient medical LLMs for pt-BR, paving the way for their broader adoption in healthcare. Continued improvements and expansions will contribute to more accurate and efficient healthcare solutions, benefiting both patients and medical professionals.

Acknowledgments. We would like to thank Comsentimento and the Google for Startups Cloud Program for providing infrastructure and support to develop this project. We acknowledge the use of Generative AI tools, particularly ChatGPT, in the preparation of this article. ChatGPT was employed for tasks such as text revision, paragraph summarization, and enhancing the overall clarity and coherence of the manuscript.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Almeida, T.S., Abonizio, H., Nogueira, R., Pires, R.: Sabiá-2: A new generation of portuguese large language models (2024)
2. Ankit Pal, M.S.: Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B (2024)
3. Brown, T.B., Mann, B., et al.: Language models are few-shot learners (2020)
4. Chen, Z., Cano, A.H., et al.: Meditron-70b: Scaling medical pretraining for large language models (2023)
5. Clusmann, J., Kolbinger, F.R., Muti, H.S., Carrero, Z.I., Eckardt, J.N., Laleh, N.G., Löffler, C.M.L., Schwarzkopf, S.C., Unger, M., Veldhuizen, G.P., Wagner, S.J., Kather, J.N.: The future landscape of large language models in medicine. Communications Medicine **3**(1) (Oct 2023)
6. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale (2020)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
8. Dias, H., Ulbrich, A.H.D.P.d.: Brateca (brazilian tertiary care dataset): a clinical information dataset for the portuguese language (2022), https://physionet.org/content/brateca/1.1/
9. Garcia, G.L., Paiola, P.H., Morelli, L.H., Candido, G., Júnior, A.C., Jodas, D.S., Afonso, L.C.S., Guilherme, I.R., Penteado, B.E., Papa, J.P.: Introducing bode: A fine-tuned large language model for portuguese prompt-based task (2024)
10. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (2000), [Online]
11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021)
12. Huang, K., Altosaar, J., Ranganath, R.: Clinicalbert: Modeling clinical notes and predicting hospital readmission (2020)
13. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023)
14. Jin, D., Pan, E., Oufattole, N., Weng, W.H., Fang, H., Szolovits, P.: What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences **11**(14), 6421 (Jul 2021)
15. Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.A., Rouvier, M., Dufour, R.: Biomistral: A collection of open-source pretrained large language models for medical domains (2024)
16. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (Sep 2019)
17. Lopes, F., Teixeira, C., Gonçalo Oliveira, H.: Contributions to clinical named entity recognition in Portuguese. In: Proceedings of the 18th BioNLP Workshop and Shared Task. pp. 223–233. Association for Computational Linguistics, Florence, Italy (Aug 2019)

18. Oliveira, L.E.S.e., Peters, A.C., da Silva, A.M.P., Gebeluca, C.P., Gumiel, Y.B., Cintho, L.M.M., Carvalho, D.R., Al Hasan, S., Moro, C.M.C.: Semclinbr - a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical nlp tasks. Journal of Biomedical Semantics **13**(1) (May 2022)
19. Peng, C., Yang, X., Chen, A., Smith, K.E., PourNejatian, N., Costa, A.B., Martin, C., Flores, M.G., Zhang, Y., Magoc, T., Lipori, G., Mitchell, D.A., Ospina, N.S., Ahmed, M.M., Hogan, W.R., Shenkman, E.A., Guo, Y., Bian, J., Wu, Y.: A study of generative large language model for medical research and healthcare. npj Digital Medicine **6**(1) (Nov 2023)
20. Pires, R., Abonizio, H., Almeida, T.S., Nogueira, R.: Sabiá: Portuguese Large Language Models, p. 226–240. Springer Nature Switzerland (2023)
21. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D.: Med-bert: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction (2020)
22. Schneider, E.T.R., Gumiel, Y.B., de Souza, J.V.A., Mie Mukai, L., Emanuel Silva e Oliveira, L., de Sa Rebelo, M., Antonio Gutierrez, M., Eduardo Krieger, J., Teodoro, D., Moro, C., Paraiso, E.C.: Cardiobertpt: Transformer-based models for cardiology language representation in portuguese. In: 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS). pp. 378–381 (2023)
23. Schneider, E.T.R., de Souza, J.V.A., Knafou, J., Oliveira, L.E.S.e., Copara, J., Gumiel, Y.B., Oliveira, L.F.A.d., Paraiso, E.C., Teodoro, D., Barra, C.M.C.M.: BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop. pp. 65–72. Association for Computational Linguistics, Online (Nov 2020)
24. Schneider, E.T.R., de Souza, J.V.A., Gumiel, Y.B., Moro, C., Paraiso, E.C.: A gpt-2 language model for biomedical texts in portuguese. In: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS). pp. 474–479 (2021)
25. Touvron, H., Martin, L., et al.: Llama 2: Open foundation and fine-tuned chat models (2023)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2023)
27. Wu, S., Dredze, M.: Beto, bentz, becas: The surprising cross-lingual effectiveness of bert (2019)
28. Yu, H., Fan, L., Li, L., Zhou, J., Ma, Z., Xian, L., Hua, W., He, S., Jin, M., Zhang, Y., Gandhi, A., Ma, X.: Large language models in biomedical and health informatics: A bibliometric review (2024)
29. Zheng, Y., Gan, W., Chen, Z., Qi, Z., Liang, Q., Yu, P.S.: Large language models for medicine: A survey (2024)
30. Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z., Ma, Y.: Llamafactory: Unified efficient fine-tuning of 100+ language models (2024)