Journal of Health Informatics

J Health Inform. 2026;18:e1781
DOI: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1781

ARTIGO COMPLETO
CBIS 2026

# Clinical named entity recognition in the Portuguese language: a benchmark of modern BERT models and LLMs

Reconhecimento de entidades nomeadas clínicas em língua portuguesa: comparação de modelos BERT modernos e LLMs

Vinicius Anjos de Almeida[1] ✉, Sandro Saorin da Silva[1], Nicolas Henrique Borges[1], Leonardo Vicenzi[1], Helena Kociolek[2], Sarah Miriã de Castro Rocha[2], Frederico Nassif Gomes[2], Julia Cristina Ferreira[2], Oge Marques[3], Lucas Emanuel Silva e Oliveira[1]

ORCID: 0009-0001-1273-586X · 0009-0007-8836-1212 · 0000-0003-2529-3869 · 0009-0001-6273-8065 · 0009-0004-2593-8262 · 0009-0009-3726-4572 · 0009-0009-4597-4136 · 0009-0000-3375-4595 · 0000-0003-4321-2719 · 0000-0003-1811-5087

[1] Spesia, Curitiba (PR), Brazil.
[2] Pontifícia Universidade Católica do Paraná, Curitiba (PR), Brazil.
[3] Florida Atlantic University, Boca Raton (Florida), United States of America.

ARTICLE INFORMATION

TYPE
Artigo completo · CBIS 2026

IDENTIFIER
e1781

DOI
10.59681/2175-4411.v18.iEspecial.2026.1781

RECEIVED
17 May 2026

ACCEPTED
20 July 2026

REVIEW PROCESS
Double Blind Review

HANDLING EDITOR
Juliano de Souza Gaspar

ACCESS
Open, CC BY 4.0

CORRESPONDING AUTHOR
Vinicius Anjos de Almeida
vinicius.almeida@spesia.com.br

LICENSE

This is an open access article published under the Creative Commons CC BY 4.0 license, which permits use, distribution, and reproduction in any medium, provided the original work is properly cited.

CC BY 4.0

## ABSTRACT

**Objective:** To evaluate BERT models and large language models (LLMs) for named entity recognition (NER) in Portuguese clinical text and assess strategies to address multilabel imbalance. **Method:** We compared BioBERTpt, BERTimbau, ModernBERT, and mmBERT with LLMs (GPT-5 and Gemini-2.5) using the SemClinBr corpus and a private breast cancer dataset. Models were trained under standardized conditions and evaluated using precision, recall, and F1-score. Iterative stratification, weighted loss, and oversampling were explored. **Results:** mmBERT-base achieved the best performance (micro F1 = 0.76), outperforming all other models. Iterative stratification improved class balance and overall results. **Conclusion:** Multilingual BERT models, particularly mmBERT, are effective for Portuguese clinical NER and can run locally with limited resources. Data balancing strategies further enhance performance.

**Keywords:** machine learning; natural language processing; neural networks, computer.

## RESUMO

**Objetivo:** Avaliar modelos BERT e modelos de linguagem de grande porte (LLMs) para reconhecimento de entidades nomeadas (NER) em textos clínicos em português e investigar estratégias para lidar com desbalanceamento *multilabel*. **Método:** Comparamos BioBERTpt, BERTimbau, ModernBERT e mmBERT com LLMs (GPT-5 e Gemini-2.5) utilizando o corpus público SemClinBr e um conjunto privado de câncer de mama. Os modelos foram treinados sob condições padronizadas e avaliados por precisão, *recall* e *F1-score*. Testamos estratificação iterativa, perdas ponderadas e *oversampling*. **Resultados:** O mmBERT-base apresentou o melhor desempenho (micro F1 = 0,76), superando os demais modelos. A estratificação iterativa melhorou o balanceamento e o desempenho global. **Conclusão:** Modelos BERT multilíngues, especialmente o mmBERT, são eficazes para NER clínico em português, com viabilidade de execução local. Estratégias de balanceamento de dados aumentam o desempenho.

**Descritores:** aprendizado de máquina; processamento de linguagem natural; redes neurais de computação.

CBIS 2026 XXI CONGRESSO BRASILEIRO DE INFORMÁTICA EM SAÚDE

Paper presented at the 21st Brazilian Congress of Health Informatics (CBIS 2026), held from September 23 to 25, 2026, in Brasília (DF), Brazil, and published in the CBIS 2026 section of the Journal of Health Informatics.

**HOW TO CITE THIS ARTICLE:** de Almeida VA, da Silva SS, Borges NH, Vicenzi L, Kociolek H, de Castro Rocha SM, et al. Clinical named entity recognition in the Portuguese language: a benchmark of modern BERT models and LLMs. J Health Inform 2026;18:e1781. doi: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1781

## 1 INTRODUCTION

Unstructured text data is ubiquitous in healthcare [1,2]. Patient complaints, clinical findings, relevant laboratory results, care plans, and the healthcare professional's rationale for their decisions are essential information often lost in electronic health record notes. Natural language processing (NLP) techniques, particularly with the current advancements, show potential for healthcare institutions to leverage text data for management, research, and quality of care improvement [3].

Recent advancements in language modeling, especially with large language models (LLMs), have motivated research in several NLP tasks, such as named-entity extraction [4,5], information retrieval [6], and question answering [7]. This momentum has fueled the development of specialized encoders tailored to specific languages and domains.

In the Portuguese context, models based on Bidirectional Encoder Representations from Transformers (BERT) have been adapted for complex legislative jargon [8], applied to financial named entity recognition (NER) [8,9], and subjected to comparative evaluations across general and specific domains [10].

With respect to NER in the clinical context, modern BERT models have emerged; however, they have not yet been evaluated on clinical datasets in the Portuguese language. The mmBERT models [11] in particular demonstrated improved multilingual capabilities that may benefit Portuguese clinical NLP tasks.

Finally, the results found in benchmarks using public datasets do not always translate to private data [12,13]. The reproducibility of results in private data is essential for technology adoption, especially in healthcare.

In this scenario, the objectives of this study are twofold: to evaluate the performance of BERT models and LLMs in clinical NER tasks using both public and private datasets, and to explore strategies for handling multilabel token classification datasets with severely imbalanced classes.

## 2 METHODS

### 2.1 Datasets

Two datasets were used in this benchmark: the SemClinBr dataset and a breast cancer clinical notes dataset.

**SemClinBr.** The SemClinBr [14] corpus is a publicly available, semantically annotated dataset of clinical notes from more than twelve medical specialties, encompassing both inpatient and outpatient scenarios. It contains 1,000 entries, including discharge summaries, admission notes, nursing notes, and ambulatory notes. The annotation protocol for NER tasks included 15 Unified Medical Language System (UMLS) [15] semantic groups, 12 of which contained at least 30 entities in the dataset. The quantity of clinical records containing each semantic group, together with examples, is available in **Table 1**.

**Table 1.** Number of records per entity in the SemClinBr dataset and entity examples. The UMLS semantic groups "Genes & Molecular Sequences", "Occupations", and "Geographic Areas" were not included in this study due to the small number of examples in the dataset.

| Entity | Number of records | Examples |
|---|---|---|
| Procedures | 984 | *"Medicada", "soroterapia", "mudança de decúbito"* |
| Disorders | 979 | *"Sonolenta", "dependente", "diabetes"* |
| Concepts & Ideas | 887 | *"2 dias", "2 anos"* |
| Living Beings | 715 | *"Filho", "neto"* |
| Anatomy | 705 | *"Calcâneo", "pescoço", "perna E"* |
| Chemicals & Drugs | 601 | *"Tazocin", "ATB", "cálcio"* |
| Physiology | 566 | *"Evacuação", "diurese"* |
| Phenomena | 492 | *"Dextro de 88mg/dl"* |
| Devices | 416 | *"Acesso venoso periférico"* |
| Organizations | 348 | *"Setor de radiografia", "hospital"* |
| Objects | 122 | *"Leito", "óculos", "próteses dentárias"* |
| Activities & Behaviors | 36 | *"Fuma", "empregada doméstica", "sedentarismo"* |

**Source:** Prepared by the authors (2026).

**Table 2.** Number of records per entity in the breast cancer clinical notes dataset and entity examples. ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; PR, progesterone receptor; BRCA, breast cancer susceptibility gene; FISH, fluorescence in situ hybridization; CISH, chromogenic in situ hybridization; SISH, silver in situ hybridization.

| Entity | Number of records | Examples |
|---|---|---|
| ER | 346 | *"HQ: RE-/RP-/Her2 3+/ gata3 +++"* |
| HER2 | 343 | *"HQ: RE-/RP-/Her2 3+/ gata3 +++"* |
| Histological type | 336 | *"Adenocarcinoma pouco diferenciado associado a..."* |
| PR | 335 | *"HQ: RE-/RP-/Her2 3+/ gata3 +++"* |
| Surgery | 296 | *"Mastectomia mama dir + EA em 2011"* |
| Menopause status | 212 | *"Paciente de 51 anos, pré-menopausa"* |
| Metastasis location | 117 | *"Ca de mama progressão de doença para ossos"* |
| Distant metastasis | 72 | *"nódulos peritoneais e lesão na bacia metastástica"* |
| BRCA | 62 | *"Análises genéticas: BRCA 1 e 2 não mutados"* |
| FISH/CISH/SISH | 32 | *"FISH negativo"* |

**Source:** Prepared by the authors (2026).

**Breast cancer clinical notes.** A restricted-access dataset of clinical notes was organized with a sample from a private oncological healthcare center in the southern region of Brazil. Five hundred ambulatory visit notes from breast cancer patients were sampled from the center's database.

A team of five medical students, coordinated by one medical doctor, was responsible for annotating entities and resolving disagreements until consensus was reached. The annotation team held regular meetings to address doubts and inconsistencies in the annotation process.

Ten entities were annotated based on their relevance to the care process of breast cancer patients: estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), histological type, surgical procedure, menopause status, presence of metastasis, metastasis location, breast cancer susceptibility gene (BRCA), and results of FISH (fluorescence in situ hybridization), CISH (chromogenic in situ hybridization), or SISH (silver in situ hybridization) tests. The quantity of clinical records containing each entity and an annotated example of each one is available in **Table 2**.

## 2.2 Metrics

Models were evaluated using the F1-score, precision, and recall, aggregated across all labels with both micro and macro averages. These metrics were computed using the scikit-learn Python library [16].

In order to compute these metrics, a probability threshold must be defined for each model. The threshold that maximized each model's F1-score was chosen for the sake of comparison.

In addition, precision-recall curves were plotted for each model and label to better describe the precision-recall trade-off.

## 2.3 Base models

BERT [17] is a transformer-based neural network architecture for language modeling. It has demonstrated state-of-the-art performance in many NLP tasks, including NER in the clinical domain in multiple languages [18-20].

**BioBERTpt**. The BioBERTpt models [18] were built on top of the original multilingual BERT models from Google [17]. Three versions were released: *biobertpt-clin*, trained on 2 million Brazilian Portuguese clinical notes from Brazilian hospitals; *biobertpt-bio*, trained on 16.4 million words from scientific papers from PubMed and SciELO; and *biobertpt-all*, trained with a combination of both corpora. To the best of the authors' knowledge, it holds state-of-the-art performance on the SemClinBr dataset.

**ModernBERT**. The ModernBERT models [21] incorporate many recent optimizations of language modeling in an encoder-only architecture. These optimizations include both architectural improvements (e.g., disabling bias terms in linear layers, adopting rotary positional embeddings, and using pre-normalization) and efficiency improvements (e.g., alternating attention, unpadding during training and inference, and flash attention). This strategy demonstrated significant improvements in the context window, natural language understanding, and information retrieval tasks.

**mmBERT**. The mmBERT models [11] were built with an architecture inspired by ModernBERT [21] and a novel pre-training protocol aimed at improving multilingual capabilities. The pre-training involved three phases in which low-resource language data was gradually included, and the ratio between high-, mid-, and low-resource languages was dynamically adjusted. The authors demonstrated improved performance in NLP

tasks when compared to similar multilingual models, including retrieval, reranking, and summarization.

**BERTimbau**. The BERTimbau models [22] build upon the same architecture as the original BERT models [17]. They were trained on 2.68 billion tokens from the brWaC corpus, the largest Portuguese open corpus up to that time. They also demonstrated strong results in Portuguese NER and text similarity tasks.

This study also compared the performance of BERT models to LLMs prompted with a few-shot strategy. The prompt template is available in **Table S.1** in the supplementary material. The models included in this baseline were: *gemini-2.5-flash*, *gemini-2.5-flash-lite*, *gemini-2.5-pro*, *gpt-4.1*, *gpt-5-mini*, *gpt-5*, and *gpt-5-nano*. All *gpt-5* models were prompted with the minimal reasoning effort option to control costs and reduce inference time. Only the *gpt-5* model was tested on "low", "medium", and "high" reasoning efforts. LLMs were evaluated using only the SemClinBr dataset to preserve the privacy of the restricted-access breast cancer dataset. Details on the training procedure and the hardware used are provided in **sections S.1 and S.2** in the supplementary material. The source code used to run the experiments and data analysis is available at: https://github.com/GRUPOMED4U/clinical_ner_benchmark_paper.

## 3 RESULTS

### 3.1 NER benchmark

All BERT models were trained and evaluated on both datasets. The metrics on the test split are presented in **Table 3**. The F1-scores reported are the maximum achievable F1-score computed with a shared probability threshold across all classes.

The *mmBERT* models achieved the best results in both datasets, and *mmBERT-base* established a new state of the art on the SemClinBr dataset (micro F1-score = 0.7646). BERT models, in general, have demonstrated better performance when compared to leading LLMs in the industry. They also had faster inference and ran locally on domestic hardware, enabling experiments with no privacy risks.

The experiments involving increased reasoning effort using the *gpt-5* model demonstrate a modest gain in performance that was not able to surpass every BERT model. The increase in performance was followed by a substantial increase in cost and inference time, as shown in **Table S.2** in the supplementary material. The increased resource requirements may be prohibitive, depending on the setting and the application.

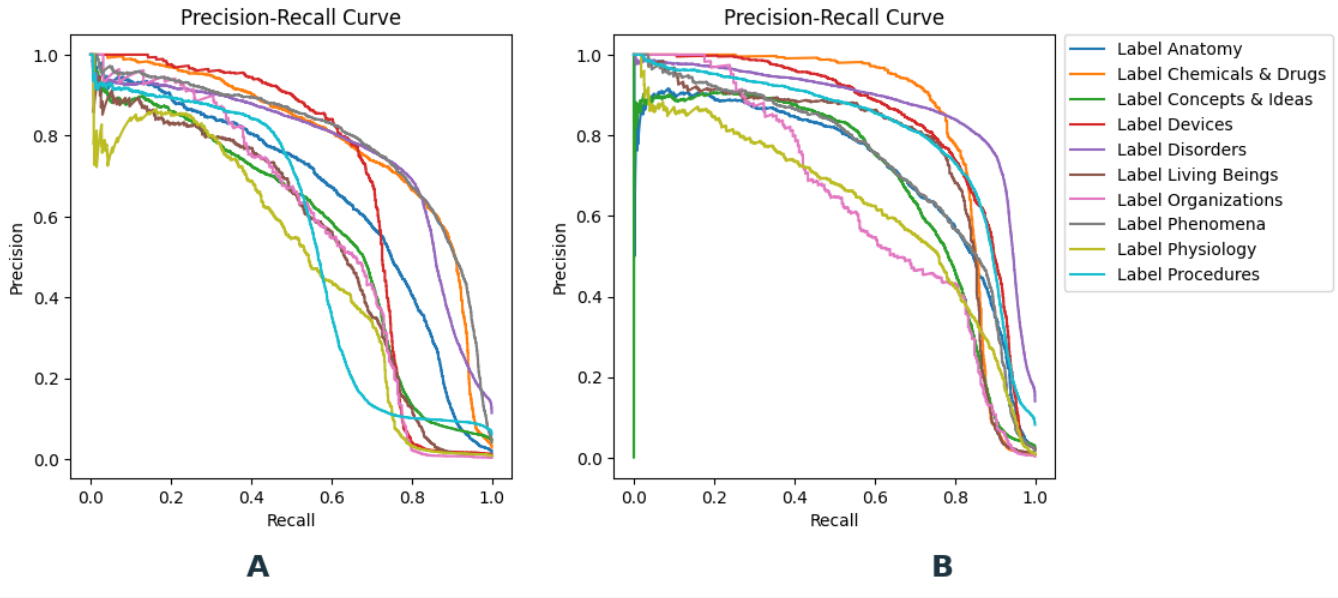
### 3.2 Dealing with imbalanced classes

Two data splitting strategies were compared: simple random split and iterative stratification. The impact of these methods was measured using the best-performing BERT method in this study: *mmBERT-base*.

In the SemClinBr dataset with the inside-outside (IO) data annotation scheme, *mmBERT-base* achieved a maximum micro F1-score of 0.6904 (±0.0287) with a simple random split and 0.7646 (±0.0094) with iterative stratification. Detailed metrics are presented in **Table S.3** in the supplementary material. The precision-recall curves are presented in **Figure 1**.

Tests using weighted loss and oversampling to deal with imbalanced classes were also explored. Due to the high frequency of certain classes, some computed weights were very small, limiting the model's capacity to learn those classes. In these cases, defining the minimum weight value as 1.0 was also explored.

**Figure 1.** A: precision-recall curve using random data split. B: precision-recall curve using iterative stratification. The iterative stratification results in curves that better approximate the top-right corner of the plot, demonstrating better precision-recall trade-off.




**Source:** Prepared by the authors (2026).

**Table 3.** Benchmark metrics for each model (BERT and LLMs) and each dataset. With respect to BERT models, the reported results reflect the use of a weighted loss function. Iterative stratification was used to split the data. Results may vary due to stochastic processes on training, such as in weight initialization and token sampling in the case of LLMs. Specific LLM versions: gpt-5 refers to gpt-5-2025-08-07; gpt-5-nano, gpt-5-nano-2025-08-07; gpt-5-mini, gpt-5-mini-2025-08-07; gpt-4.1, gpt-4.1-2025-04-14. Gemini models are stable models and have the same name in the API.

| Model | SemClinBr | | | | Breast cancer dataset | | | |
|---|---|---|---|---|---|---|---|---|
| | Macro F1-score | | Micro F1-score | | Macro F1-score | | Micro F1-score | |
| | IO | BIO | IO | BIO | IO | BIO | IO | BIO |
| BioBERTpt | | | | | | | | |
| biobertpt-all | 0.6443 | 0.6207 | 0.6830 | 0.6573 | 0.7473 | 0.3671 | 0.7602 | 0.6124 |
| biobertpt-clin | 0.6489 | 0.6470 | 0.6981 | 0.6959 | 0.7527 | 0.3982 | 0.7728 | 0.6508 |
| biobertpt-bio | 0.5839 | 0.5543 | 0.6548 | 0.6208 | 0.7268 | 0.3741 | 0.7458 | 0.6319 |
| mmBERT | | | | | | | | |
| mmBERT-base | 0.7139 | 0.4231 | 0.7646 | 0.7444 | 0.7445 | 0.4945 | 0.7723 | 0.6598 |
| mmBERT-small | 0.6667 | 0.3968 | 0.7264 | 0.7055 | 0.7212 | 0.5398 | 0.7558 | 0.6829 |
| ModernBERT | | | | | | | | |
| ModernBERT-base | 0.5475 | 0.2277 | 0.6494 | 0.5654 | 0.6987 | 0.3814 | 0.7259 | 0.6796 |
| ModernBERT-large | 0.5895 | 0.3258 | 0.6987 | 0.6851 | 0.7104 | 0.5056 | 0.7488 | 0.6648 |
| BERTimbau | | | | | | | | |
| bert-base-portuguese-cased | 0.5839 | 0.5000 | 0.6592 | 0.6094 | 0.0055 | 0.1220 | 0.3182 | 0.1218 |
| bert-large-portuguese-cased | 0.5769 | 0.6264 | 0.6421 | 0.6474 | 0.0130 | 0.0038 | 0.0130 | 0.0315 |
| OpenAI* | | | | | | | | |
| gpt-4.1 | 0.4795 | | 0.5328 | | | | | |
| gpt-5-nano minimal | 0.1870 | | 0.2507 | | | | | |
| gpt-5-mini minimal | 0.4489 | | 0.4746 | | | | | |
| gpt-5 minimal | 0.5878 | | 0.6604 | | | | | |
| gpt-5 low | 0.6311 | | 0.6828 | | | | | |
| gpt-5 medium | 0.6338 | | 0.6907 | | | | | |
| gpt-5 high | ** | | ** | | | | | |
| Gemini | | | | | | | | |
| gemini-2.5-flash-lite | 0.4849 | | 0.5781 | | | | | |
| gemini-2.5-flash | 0.5965 | | 0.6528 | | | | | |
| gemini-2.5-pro | 0.5965 | | 0.6528 | | | | | |

* "minimal", "low", "medium", and "high" refer to the reasoning effort. ** Inference with gpt-5 with high reasoning effort was aborted due to very long inference time and higher cost, even with optimized prompt.

**Source:** Prepared by the authors (2026).

The results show that the best approach for dealing with class imbalance differs among datasets. In the SemClinBr, maximum performance was achieved through the iterative stratification of data alone. In the breast cancer dataset, it was the combination of iterative stratification with a weighted loss, with the minimum weight fixed at 1.0.

## 4 DISCUSSION

### 4.1 Key findings

This study evaluated the performance of modern BERT models and LLMs on two clinical NER datasets: one public and the other private. BERT models, especially mmBERT [11] followed by BioBERTpt [18], demonstrated the best performance in the NER task across both datasets and achieved an F1-score of up to 0.7646.

Several characteristics of mmBERT may have contributed to these results [11]. It was trained on about 3 trillion tokens, with hundreds of different languages represented. BioBERTpt [18], on the other hand, was fine-tuned on fewer than 50 million words in the Portuguese language. mmBERT incorporates language modeling optimizations from ModernBERT [21] and dynamically includes low-resource languages as the training phases progress, allowing for faster learning and better representation of these languages, even though less data is available for them.

It is expected that additional in-domain pre-training improves model performance, as seen with BioBERTpt

in the clinical context [18]. Future work should explore fine-tuning mmBERT on biomedical data in the Portuguese language.

The possibility of data contamination cannot be excluded as a cause of the difference in performance between BioBERTpt and mmBERT in the SemClinBr dataset. mmBERT was trained on 3 trillion tokens and was published after the SemClinBr dataset was made public. The smaller difference in performance between both models in a private clinical dataset raises that concern even further. It was beyond the scope of this study to thoroughly investigate data contamination in the pre-training data of the evaluated BERT models, but this matter should be addressed in future studies.

The difference in performance between the IO and BIO annotation schemes, particularly in the macro F1-score, was greater in the mmBERT and ModernBERT models compared to BioBERTpt and BERTimbau. Since these results were obtained from the same datasets, it is reasonable to hypothesize that there are architectural aspects influencing how these models scale in larger label spaces. The beginning-inside-outside (BIO) annotation scheme doubles the number of targeted classes by adding the beginning (B) tag to each class. Future work should explore these architectural differences and strategies to mitigate diminishing performance on less frequent classes.

### 4.2 Strengths and limitations

This study has several strengths. Our results established a new state of the art for the SemClinBr dataset and demonstrated that these results can also be reproduced with private data. The private breast cancer dataset was annotated by a mixed team of clinicians and students, with disagreements resolved by consensus. This reflects a realistic, high-quality annotation pipeline that could be reproduced in diverse healthcare settings. With only 500 clinical notes, the authors were able to train a well-performing NER model for oncological entities relevant to breast cancer care, demonstrating its feasibility. This study also systematically tested and quantified the effects of data iterative stratification, loss weighting, and data oversampling on multilabel imbalanced clinical data. The results pave the way for broader exploration of these techniques in clinical NLP tasks. With reproducibility as a cornerstone, all source code has been made available to the research community.

This study also has many limitations. Exploring LLMs' reasoning capabilities on NER tasks was beyond its scope. Time, cost, and data privacy concerns were among the reasons not to explore them further. No LLM fine-tuning was performed. The authors believe that small language models fine-tuned with biomedical data in the Portuguese language or prompt optimization strategies, such as GEPA [23], may be a good option for clinical NER [24]. These strategies should be explored in future studies.

BERT models were not fine-tuned on additional biomedical data, which could influence the results. Future studies should investigate the impact of fine-tuning models on private health data on the performance of BERT models in public and private NER datasets. Although both datasets cover a broad range of topics and medical specialties, the complexity of healthcare is far greater than what this study was able to address. For this reason, the results may not apply to different medical specialties or to data from different regions. Finally, this study did not include hypothesis testing to rigorously compare the different techniques for handling class imbalance, which limits the generalizability of the findings.

## 5 CONCLUSION

This study benchmarked modern BERT models and LLMs for clinical NER in the Portuguese language, using both public and private datasets. Among all tested models, mmBERT achieved the best performance, establishing a new state of the art on the SemClinBr dataset and demonstrating strong generalization to a private breast-cancer corpus. The findings highlight the robustness of encoder-only architectures for clinical NER, especially when trained with multilingual strategies. This work also contributes by systematically exploring strategies to handle multilabel class imbalance, such as iterative stratification, weighted loss, and oversampling. The results show that appropriate data-splitting strategies can substantially improve performance and that the use of oversampling and weighted loss may or may not help, depending on the data. The tests also showed that using weighted loss with a fixed minimum weight of 1.0 for frequent classes preserves performance in those classes while also enhancing it in the less frequent ones.

This study reinforces the feasibility of developing high-quality Portuguese clinical NER systems within real-world healthcare institutions, even with modest annotation efforts. Future research should explore domain-specific pre-training of multilingual models like mmBERT on Portuguese biomedical corpora, generalization across institutions and medical specialties, and the use of refined small language models on Portuguese clinical NER data.

## SUPPLEMENTARY MATERIAL

Link: https://github.com/GRUPOMED4U/clinical_ner_benchmark_paper/blob/main/assets/CBIS%202026%20submission%20-%20Supplementary%20material.pdf

## 6 ETHICAL STATEMENT

This study was approved by the Ethics Committee of the Hospital de Clínicas da Universidade Federal do Paraná (HCUFPR) under protocol number 90422925.6.0000.0096. All procedures were conducted in accordance with the principles of the 1975 Declaration of Helsinki, revised in 2013. The research was based exclusively on the analysis of previously collected data, with no direct patient involvement, clinical interventions, or contact with participants. All information was handled in accordance with ethical and confidentiality standards as well as with the LGPD (Lei Geral de Proteção de Dados Pessoais, Lei nº 13.709/2018).

**Author contributions.** Vinicius Anjos de Almeida: Conceptualization, Formal analysis, Investigation, Methodology, Resources, Software, Supervision, Data Curation, Validation, Visualization, Writing - original draft, Writing - review and editing. Sandro Saorin da Silva: Software. Nicolas Henrique Borges: Data Curation, Writing - review and editing. Leonardo Vicenzi: Writing - review and editing. Helena Kociolek, Sarah Miriã de Castro Rocha, Frederico Nassif Gomes, Julia Cristina Ferreira: Data Curation. Oge Marques, Lucas Emanuel Silva e Oliveira: Writing - review and editing. All authors approved the final manuscript.

**Acknowledgments.** We thank Josimar Chire for his contribution to the oversampling algorithms tested. We acknowledge his efforts in this section due to the journal's limit on the number of authors.

**Funding.** The authors did not declare any source of funding.

**Data and materials availability.** The authors did not provide a data and materials availability statement.

**Conflicts of interest.** The authors did not declare any conflicts of interest.

**Use of artificial intelligence.** The authors declare that language models were used solely for grammar and spelling revision and for programming support.

## REFERENCES

1. Sedlakova J, Daniore P, Horn Wintsch A, Wolf M, Stanikic M, Haag C, et al. Challenges and best practices for digital unstructured data enrichment in health research: A systematic narrative review. PLOS Digit Health. 2023;2(10):e0000347. doi: https://doi.org/10.1371/journal.pdig.0000347
2. Nesca M, Katz A, Leung CK, Lix LM. A scoping review of preprocessing methods for unstructured text data to assess data quality. Int J Popul Data Sci. 2022;7(1):1757. doi: https://doi.org/10.23889/ijpds.v7i1.1757
3. Alafari F, Driss M, Cherif A. Advances in natural language processing for healthcare: A comprehensive review of techniques, applications, and future directions. Comput Sci Rev. 2025;56:100725. doi: https://doi.org/10.1016/j.cosrev.2025.100725
4. Kocaman V, Talby D. Accurate Clinical and Biomedical Named Entity Recognition at Scale. Softw Impacts. 2022;13:100373. doi: https://doi.org/10.1016/j.simpa.2022.100373
5. Monajatipoor M, Yang J, Stremmel J, Emami M, Mohaghegh F, Rouhsedaghat M, et al. LLMs in Biomedicine: A study on clinical Named Entity Recognition. arXiv [Preprint]. 2024. doi: https://doi.org/10.48550/arXiv.2404.07376
6. Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. J Am Med Inform Assoc. 2025;32(4):605-15. doi: https://doi.org/10.1093/jamia/ocaf008
7. Kell G, Roberts A, Umansky S, Qian L, Ferrari D, Soboczenski F, et al. Question answering systems for health professionals at the point of care-a systematic review. J Am Med Inform Assoc. 2024;31(4):1009-24. doi: https://doi.org/10.1093/jamia/ocae015
8. Nunes RO, Balreira DG, Spritzer AS, Freitas CMDS. A Named Entity Recognition Approach for Portuguese Legislative Texts Using Self-Learning. In: Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR 2024), Vol. 1; 2024 Mar; Santiago de Compostela, Spain. Stroudsburg (PA): Association for Computational Linguistics; 2024. p. 290-300. [cited 2026 Oct 13]. Available from: https://aclanthology.org/2024.propor-1.30
9. Abilio R, Coelho GP, da Silva AEA. Evaluating Named Entity Recognition: A comparative analysis of mono- and multilingual transformer models on a novel Brazilian corporate earnings call transcripts dataset. Appl Soft Comput. 2024;166:112158. doi: https://doi.org/10.1016/j.asoc.2024.112158
10. Monteiro M, Zanchettin C. Optimization Strategies for BERT-Based Named Entity Recognition. In: Anais da XII Brazilian Conference on Intelligent Systems (BRACIS); 2023; Belo Horizonte, Brazil. Porto Alegre: Sociedade Brasileira de Computação; 2023. p. 80-94. [cited 2026 Oct 13]. Available from: https://sol.sbc.org.br/index.php/bracis/article/view/28408
11. Marone M, Weller O, Fleshman W, Yang E, Lawrie D, Van Durme B. mmBERT: A Modern Multilingual Encoder with Annealed Language Learning. arXiv [Preprint]. 2025. doi: https://doi.org/10.48550/arXiv.2509.06888
12. Deng C, Zhao Y, Tang X, Gerstein M, Cohan A. Investigating Data Contamination in Modern Benchmarks for Large Language Models. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers); 2024 Jun; Mexico City, Mexico. Stroudsburg (PA): Association for Computational Linguistics; 2024. p. 8706-19. doi: https://doi.org/10.18653/v1/2024.naacl-long.482

13. Dong Y, Jiang X, Liu H, Jin Z, Gu B, Yang M, et al. Generalization or Memorization: Data Contamination and Trustworthy Evaluation for Large Language Models. In: Findings of the Association for Computational Linguistics: ACL 2024; 2024 Aug; Bangkok, Thailand. Stroudsburg (PA): Association for Computational Linguistics; 2024. p. 12039-50. doi: https://doi.org/10.18653/v1/2024.findings-acl.716

14. Oliveira LESE, Peters AC, da Silva AMP, Gebeluca CP, Gumiel YB, Cintho LMM, et al. SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks. J Biomed Semantics. 2022;13(1):13. doi: https://doi.org/10.1186/s13326-022-00269-1

15. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993;32(4):281-91. doi: https://doi.org/10.1055/s-0038-1634945

16. scikit-learn developers. sklearn.metrics: scikit-learn 1.9.1 documentation [Internet]. 2026 [cited 2026 Oct 13]. Available from: https://scikit-learn.org/stable/api/sklearn.metrics.html

17. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); 2019 Jun; Minneapolis, Minnesota. Stroudsburg (PA): Association for Computational Linguistics; 2019. p. 4171-86. doi: https://doi.org/10.18653/v1/N19-1423

18. Schneider ETR, de Souza JVA, Knafou J, Oliveira LESE, Copara J, Gumiel YB, et al. BioBERTpt - A Portuguese Neural Language Model for Clinical Named Entity Recognition. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop; 2020 Nov; Online. Stroudsburg (PA): Association for Computational Linguistics; 2020. p. 65-72. doi: https://doi.org/10.18653/v1/2020.clinicalnlp-1.7

19. Danu MD, Marica G, Suciu C, Itu LM, Farri O. Multilingual Clinical NER for Diseases and Medications Recognition in Cardiology Texts using BERT Embeddings. arXiv [Preprint]. 2025. doi: https://doi.org/10.48550/arXiv.2510.17437

20. Tiwari SS, Goyal HR, Adnan MM, Shrivastava A, Yadav K, Rao ALN. Clinical Text Analysis with Natural Language Processing: A BERT-based Approach. In: 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE); 2024 May 9; Gautam Buddha Nagar, India. [place unknown]: IEEE; 2024. p. 1382-8. doi: https://doi.org/10.1109/IC3SE62002.2024.10593504

21. Warner B, Chaffin A, Clavié B, Weller O, Hallström O, Taghadouini S, et al. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2025 Jul; Vienna, Austria. Stroudsburg (PA): Association for Computational Linguistics; 2025. p. 2526-47. doi: https://doi.org/10.18653/v1/2025.acl-long.127

22. Souza F, Nogueira R, Lotufo R. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In: Cerri R, Prati RC, editors. Intelligent Systems. BRACIS 2020. Lecture Notes in Computer Science, vol. 12319. Cham: Springer; 2020. p. 403-17. doi: https://doi.org/10.1007/978-3-030-61377-8_28

23. Agrawal LA, Tan S, Soylu D, Ziems N, Khare R, Opsahl-Ong K, et al. GEPA: Reflective Prompt Evolution Can Outperform Reinforcement Learning. arXiv [Preprint]. 2026. doi: https://doi.org/10.48550/arXiv.2507.19457

24. Belcak P, Heinrich G, Diao S, Fu Y, Dong X, Muralidharan S, et al. Small Language Models are the Future of Agentic AI. arXiv [Preprint]. 2025. doi: https://doi.org/10.48550/arXiv.2506.02153

## EDITORIAL NOTES

The Journal of Health Informatics is a scientific journal maintained by the Brazilian Society of Health Informatics (Sociedade Brasileira de Informática em Saúde), a non-profit scientific association dedicated to the development of Health Informatics and Digital Health in Brazil. The editorial process is conducted with scientific independence, free from institutional, commercial, or sponsor interference, and follows international best practices in publication ethics.

This article was double-blind peer reviewed by specialists in the field. The editorial team used artificial intelligence tools, under human supervision, in the initial screening, in checking the text, references and numbers and, when needed, in redrawing figures, charts and tables, which the authors approved in the proof. The tools and safeguards are described in the journal's Policy on the use of artificial intelligence. Editorial decisions were made by people.

SBIS Sociedade Brasileira de Informática em Saúde