NormaTex-MapSNOMED: Bridging the Gap Between Brazilian Portuguese Clinical Narratives and SNOMED CT

Isabela Araujo and Claudia Moro and Layslla Martinez
Pontifícia Universidade Católica do Paraná
Curitiba, Brazil
isabela.fontes@pucpr.edu.br, c.moro@pucpr.br, layslla.troya@pucpr.edu.br

Abstract

Clinical narratives written in free text contain valuable information for patient care. However, their unstructured nature and linguistic variability pose significant challenges for automatic processing and interoperability. In particular, mapping clinical terms to standardized terminologies such as SNOMED Clinical Terms (SNOMED CT) remains difficult for languages other than English, including Brazilian Portuguese. This paper presents NormaTex-MapSNOMED, a proposed component of the NormaTex framework that focuses on mapping clinical terms to predefined categories aligned with SNOMED CT. Given previously extracted terms, the method leverages large language models (LLMs) guided by a structured prompt to assign terms to target categories. Experiments were conducted on Portuguese-language clinical narratives and evaluated using three complementary strategies: lexical similarity based on Levenshtein distance, contextual similarity using a BERT-based model, and semantic validation using LLMs. The results indicate that LLM-based evaluation consistently outperforms lexical and contextual baselines across different models, with higher precision observed for disease-related terms compared to symptom-related expressions. These findings indicate that LLMs are a promising approach for semantic mapping of clinical terms in Brazilian Portuguese and can support clinical term normalization and interoperability with standardized terminologies.

1 Introduction

Clinical narratives written in free text contain valuable information for patient care; however, their unstructured nature, frequent use of medical jargon, abbreviations, and lexical variability pose substantial challenges for automatic processing. Despite these challenges, such narratives convey clinically relevant information, including diagnoses, signs and symptoms, procedures, and exam results, which are essential for clinical decision-making and continuity of care Bhagat et al. (2024).

Structuring information extracted from clinical text facilitates interoperability across healthcare systems and supports downstream applications such as epidemiological analysis and clinical decision support Nan and Xu (2023). To ensure semantic consistency and effective data exchange, the adoption of health information standards and terminologies, such as HL7 (Health Level Seven International, 2024b), FHIR (Health Level Seven International, 2024a), and SNOMED CT (NHS Digital, 2024), is required.

SNOMED CT provides comprehensive clinical coverage and formal semantic representation. Nevertheless, mapping free-text clinical narratives to this terminology remains challenging, particularly for languages other than English. In Brazilian Portuguese, this task is further constrained by linguistic variability and the scarcity of specialized normalization approaches.

NormaTex is an ongoing research effort focused on the normalization of clinical terms extracted from Brazilian Portuguese clinical narratives using SNOMED CT. The framework is being developed to normalize clinical terms into SNOMED CT terminology. The method encompasses term extraction, normalization, which includes mapping and term codification, and evaluation. NormaTex has not yet been described in previous publications and is still in the development stage at the time of writing. This paper concentrates on NormaTex-MapSNOMED, a proposed method, part of the NormaTex framework that corresponds with the term mapping stage of normalization.

Using terms that have been previously extracted, NormaTex-MapSNOMED assigns clinical terms to predetermined categories that are in accordance with SNOMED CT. The study uses large language models (LLMs) (Hu et al., 2024) to map terms into predefined categories and evaluates the results

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 1085–1091

using three different strategies: the Levenshtein-distance-based, BERT-based, and LLM-based approaches.

2 Related Work

The interest in applying LLM solutions to clinical text processing has grown over the years, particularly for information extraction and semantic normalization tasks Bhagat et al. (2024). Recent studies have indicated that LLMs can effectively identify clinical entities and symptoms from free-text narratives, even in the presence of high linguistic variability and domain-specific terminology. For example, Bai et al. (2025) demonstrated that open LLMs achieve robust performance in identifying genitourinary symptoms from clinical notes, while Khan (2024) showed that generative models can extract normalized symptom terms without requiring extensive supervised training. The ability of LLMs to interpret medical narratives across multiple languages is further highlighted by studies such as Menezes et al. (2025).

Comparative analyses between classical encoder-based architectures and generative models consistently indicate advantages for LLMs in clinical NLP tasks. In oncology-related entity recognition, generative models were shown to outperform BERT-based approaches, particularly under heterogeneous textual conditions Arzideh et al. (2025). Similarly, Zhang et al. (2025) introduced a zero-shot generative approach for clinical information extraction, demonstrating that LLMs can perform complex tasks without task-specific fine-tuning. Together, these findings suggest that semantic reasoning capabilities are crucial for handling ambiguity and implicit clinical meaning.

Despite these advances, mapping clinical terms to standardized terminologies such as SNOMED CT remains a challenging problem. Prior work has explored embedding-based and hybrid approaches for concept normalization Abdulnazar et al. (2023); Xu and Miller (2022), as well as heuristic and rule-based strategies applied to clinical narratives Silva et al. (2020). However, the majority of these studies focus on English-language corpora, which limits the development of other language applications, such as Brazilian Portuguese.

In this context, research on mapping Brazilian Portuguese clinical terms to SNOMED CT remains scarce (e.g., Oliveira et al. (2022) for the SemClinBr corpus, Hasan et al. (2015) for ontology-driven search in Brazilian Portuguese clinical notes). NormaTex-MapSNOMED addresses this gap by using LLMs to map terms from clinical narratives, combined with a comparative evaluation against lexical and contextual similarity baselines.

3 Method

NormaTex-MapSNOMED is a method designed to map previously extracted clinical terms into predefined semantic categories aligned with SNOMED CT. It represents the normalization component of the broader NormaTex pipeline. In the complete pipeline, clinical narratives first undergo information extraction to identify entity spans, followed by a normalization stage that performs semantic category assignment and terminology codification. The present study focuses exclusively on the normalization stage. Therefore, all experiments assume that clinical entities have already been correctly extracted and provided as input to the mapping module (Figure 1). Accordingly, the experiments were conducted under the assumption that clinical terms had already been extracted. NormaTex-MapSNOMED applies LLMs to map these extracted terms into predefined categories using a structured prompt Sivarajkumar et al. (2024), followed by an evaluation of the mapping outcomes. Errors originating from the entity extraction stage, such as missed entities, incorrect span boundaries, or spurious detections, may propagate to the mapping stage because the LLM receives only the spans produced by the upstream component. Consequently, the evaluation reported in this study isolates the mapping performance and does not account for extraction errors. A systematic assessment of robustness to such upstream noise is left for future work. Two annotated corpora were used in the experiments: SemClinBr (Oliveira et al., 2022), a corpus of 1,000 Brazilian Portuguese clinical narratives annotated with 89 semantic categories, and TempClinBr, a curated subset of SemClinBr containing 126 cardiology narratives. TempClinBr served as the gold standard dataset for method development and evaluation. SemClinBr provides fine-grained annotations that distinguish between Disease or Syndrome and Sign or Symptom. In contrast, TempClinBr adopts a broader umbrella category (Problem) that encompasses both concepts. For the purposes of this study, mapping decisions were guided by explicit category definitions derived from the literature (e.g., (Carneiro, 2017)) to en-

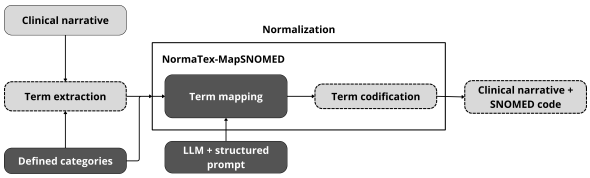

Figure 1: Overview of the NormaTex framework highlighting the term mapping stage, NormaTex-MapSNOMED. This work focuses solely on the NormaTex-MapSNOMED component (term mapping). SNOMED CT codification is future work.

sure consistent interpretation during prompt-based classification.

3.1 Prompt

Prompt construction determines the execution of the method by guiding the LLM through a structured sequence of instructions. During prompt development, instructions were decomposed into sequential steps to guide the reasoning process of the model, following chain-of-thought prompting principles (Liu et al., 2021). Few-shot learning was also applied by including representative examples to clarify expected behavior and improve mapping precision. The prompt was refined iteratively (Zaghir et al., 2024) to minimize hallucinations and reduce ambiguity during term mapping. Instructions were simplified by avoiding distinctions between simple and compound terms, excluding explicit negation handling at this stage, and explicitly ignoring certain categories, such as exam findings and test results. The use of structured markers inspired by known languages (e.g., HTML) further improved instruction clarity and task comprehension.

3.2 Language Models

Four large language models were evaluated to assess the generalizability of the proposed approach across different deployment conditions. LLaMA 3 (8B, Meta) (Dubey et al., 2024) was executed locally using the Ollama framework, representing an open-weight model. Gemini 2.5 Pro (Google DeepMind, 2025) was accessed through an academic license. GPT-3.5-Turbo and GPT-4.1 (2025-04-14) (OpenAI, 2023) were accessed via the OpenAI API. This configuration enabled comparison across locally deployable models, academically licensed models, and commercial API-based models within the same experimental framework.

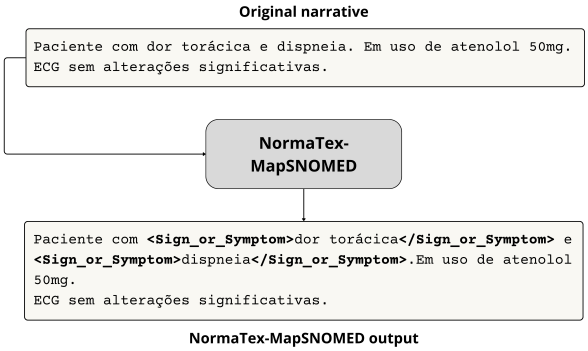

Figure 2: Example of a Brazilian Portuguese clinical narrative before and after processing with NormaTex-MapSNOMED, illustrating inline annotation of mapped clinical terms while non-target concepts are ignored.

4 Results and Evaluation

This section presents the results of NormaTex-MapSNOMED, considering both the finalized prompt used for term mapping and the observed performance across different LLMs and semantic categories.

4.1 Final Prompt

The final prompt is a concrete outcome of the NormaTex-MapSNOMED method and encapsulates the rules required for semantic term mapping. It defines constraints governing the application of the target categories, Disease or Syndrome and Sign or Symptom.

The prompt also specifies non-target concepts to be ignored, including medications, dosage information, and complementary exam findings, ensuring that only clinically relevant entities are considered. To ensure consistent behavior, models are required to return only the original clinical narrative with inline annotations reflecting the mapped terms. An example prompt structure illustrating the instructions and format used for term mapping is provided in Appendix A.

4.2 Evaluation Setup and Methodology

Experiments were conducted using ten clinical narratives from the TempClinBr corpus. The choice of ten narratives allowed for full manual validation of all mapping outcomes and a controlled comparison across models and evaluation strategies. The small sample size limits the generalizability of the results and motivates evaluation on a larger set in future work. Narrative length ranged from 33 to 304 tokens, with an average of approximately 140 tokens. The gold standard comprises 111 clin-

ical terms (spans) across these narratives, corresponding to the Problema category in TempClinBr (which subsumes Disease or Syndrome and Sign or Symptom). Performance was assessed using three complementary evaluation approaches: Levenshtein distance (Levenshtein, 1966) as a classical lexical similarity baseline, a BERT-based model (Devlin et al., 2019) capturing bidirectional contextual representations, and an LLM-based semantic validation strategy using a simple prompt. In the LLM-based approach, the same model used for term mapping was also employed as the evaluator. Mapping and evaluation were executed as independent, sequential steps: mapping outputs were first saved, and evaluation was performed in a separate execution with no shared context or cached state, thereby avoiding circular bias. In addition to the automated evaluation strategies, a manual review was conducted to assess each mapping outcome. For every term classified as correct, partially correct, or incorrect by the automated methods, the reviewer independently verified whether the classification was semantically coherent, recording agreement or disagreement along with qualitative observations. This human validation layer served as an additional safeguard against potential evaluation bias. Method outputs were classified into three categories: correct, partially correct and incorrect. Correct outcomes correspond to exact matches or clear semantic equivalence. Partially correct outcomes reflect semantic or conceptual relatedness without full equivalence, while incorrect outcomes indicate semantic mismatch or lack of correspondence. During experimentation, the API usage limit associated with the academic license was exceeded, preventing the application of the LLM-based evaluation strategy to Gemini 2.5 Pro. As a result, this model was evaluated exclusively using BERT-based similarity and Levenshtein distance.

4.3 Analysis by LLMs

Table 1 reports, for each model and evaluation strategy, the counts and proportions of mappings classified as correct (VP), partially correct (VPP), and incorrect (FP) by manual validation. The total number of mapped terms varies by model because each LLM produces a different set of spans. Although the same LLM was used for both mapping and evaluation, the two stages were strictly decoupled: each execution operated without access to prior outputs or session memory, ensuring that

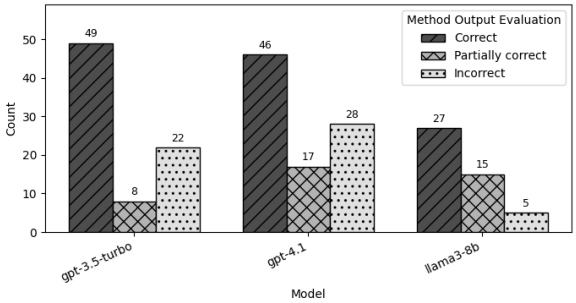

Figure 3: Distribution of correct, partially correct, and incorrect mappings across models when considering LLM-based evaluation.

the evaluation was not influenced by the mapping process. Across the models and strategies where LLM-based evaluation was applied, it consistently identified a larger share of mappings as correct than did BERT-based or Levenshtein-based assessment for the same model. Figure 3 summarizes the distribution of correct, partially correct, and incorrect mappings across models when considering LLM-based evaluation.

Levenshtein distance was particularly sensitive to surface-level variation, including spelling differences and abbreviations, leading to more incorrect outcomes. The BERT-based approach improved robustness by capturing contextual similarity but remained limited when handling domain-specific terminology and implicit semantic relations. In contrast, LLM-based evaluation proved more effective in validating semantic equivalence beyond lexical overlap.

Gemini 2.5 Pro was evaluated only with BERT and Levenshtein. LLM-based evaluation was not run for this model due to API constraints. Under the BERT-based strategy, GPT-4.1 attains the highest proportion of correct mappings (63.7%), followed by LLaMA-3-8B (53.2%) and Gemini-2.5-Pro (47.4%). Under Levenshtein, Gemini-2.5-Pro shows the highest proportion of incorrect outcomes (59.6%), consistent with greater sensitivity to surface variation. LLM-based evaluation yields the highest % correct for GPT-3.5-Turbo (62.0%) and LLaMA-3-8B (57.4%) in this setup. The proportions in Table 1 correspond to precision at the span level (i.e., the fraction of each model's mapped terms that were judged correct, partially correct, or incorrect). Recall relative to the gold set was not computed, as the number of mapped spans differs per model.

Model	Method	VP	VPP	FP	Total	% Correct	% Partial	% Incorrect
Gemini-2.5-Pro	BERT	27	25	5	57	47.4	43.9	8.8
Gemini-2.5-Pro	Levenshtein	17	6	34	57	29.8	10.5	59.6
GPT-3.5-Turbo	BERT	36	24	19	79	45.6	30.4	24.1
GPT-3.5-Turbo	Levenshtein	35	15	29	79	44.3	19.0	36.7
GPT-3.5-Turbo	LLM	49	8	22	79	62.0	10.1	27.8
GPT-4.1	BERT	58	29	4	91	63.7	31.9	4.4
GPT-4.1	Levenshtein	46	36	9	91	50.5	39.6	9.9
GPT-4.1	LLM	46	17	28	91	50.5	18.7	30.8
LLaMA-3-8B	BERT	25	13	9	47	53.2	27.7	19.1
LLaMA-3-8B	Levenshtein	23	11	13	47	48.9	23.4	27.7
LLaMA-3-8B	LLM	27	15	5	47	57.4	31.9	10.6

Table 1: Counts and proportions of correct (VP), partially correct (VPP), and incorrect (FP) mappings per model and evaluation strategy (manual validation).

4.4 Analysis by Category

Disease or Syndrome consistently exhibited higher precision than Sign or Symptom across all models and evaluation approaches. This result is expected, as disease-related terms tend to be more standardized, whereas symptom descriptions in clinical narratives display greater lexical variability and semantic ambiguity.

These findings underscore the increased difficulty of mapping symptom-related expressions and highlight the importance of clearly defined category boundaries and decision rules during prompt development.

5 Conclusion

This paper presented NormaTex-MapSNOMED, a normalization component of the NormaTex framework designed to map terms from Brazilian Portuguese clinical narratives to SNOMED CT. The method leverages structured prompts and well-defined semantic categories to address challenges commonly found in unstructured clinical text, such as lexical variation, abbreviations, and lack of standardization.

Experimental results indicate that semantic validation using LLMs surpasses classical lexical and contextual similarity approaches, particularly for disease-related terms. Although symptom mapping remains more challenging due to inherent linguistic variability, the method demonstrates consistent and robust performance across different models.

Overall, the findings suggest that the use of LLMs for mapping terms from Brazilian Portuguese clinical narratives is a promising direction and has the potential to support related tasks, including clinical term normalization and interoperability with standardized terminologies such as SNOMED CT.

Limitations

NormaTex-MapSNOMED has several limitations. Negation is not explicitly modeled in the current prompt, which may affect the accurate representation of certain clinical expressions in SNOMED CT. Additionally, the approach is currently limited to a small set of semantic categories, constraining its applicability to more complex clinical scenarios. This study does not include comparisons with concept-normalization baselines such as SapBERT or rule- and dictionary-based methods for Portuguese. Such comparisons are left for future work.

The choice to use generative LLMs for category assignment was motivated by their semantic flexibility (Leiser et al., 2025) and the scarcity of existing resources for mapping clinical terms to standardized terminologies in Brazilian Portuguese (Hasan et al., 2015). A systematic comparison with traditional NER and entity-linking pipelines is planned as a next step to better quantify the relative gains of the LLM-based approach.

Evaluation results are also influenced by the quality of the SemClinBr annotations. During analysis, inconsistencies such as missing or imprecise category assignments were identified in the reference corpus. In some cases, NormaTex-MapSNOMED produced semantically coherent mappings that diverged from the gold-standard annotations, directly affecting evaluation outcomes. Finally, the experimental evaluation relied on a limited number of

clinical narratives, and one model could not be evaluated using the LLM-based strategy due to API usage constraints. These factors limit the generalizability of the results and motivate future work involving larger datasets, additional categories, explicit negation handling, and more comprehensive evaluations. Furthermore, this work does not perform actual concept-ID coding to SNOMED CT.

The contribution is limited to a preliminary categorization step that aligns with SNOMED CT categories and prepares for later codification. The limitations above suggest concrete directions for future work: prompt redesign strategies (e.g., to handle negation explicitly and to refine category boundaries), integration with rule-based clinical NLP components for pre-filtering or post-validation of LLM outputs, and extension of the set of target categories to cover a broader range of clinical entities.

References

Akhila Abdulnazar, Markus Kreuzthaler, Roland Roller, and Stefan Schulz. 2023. Sapbert-based medical concept normalization using snomed ct. In *Studies in Health Technology and Informatics*, volume 302, pages 825–826.

Kamyar Arzideh, Henning Schäfer, and 1 others. 2025. From bert to generative ai: Comparing encoder-only vs. large language models in a cohort of lung cancer patients for named entity recognition in unstructured medical reports. *Computers in Biology and Medicine*, page 110665.

Yunbing Bai, Wanting Cui, and Joseph Finkelstein. 2025. Performance of open-source large language models to extract symptoms from clinical notes. *Studies in Health Technology and Informatics*, 329:663–667.

Navya Bhagat, Olivia Mackey, and Adam Wilcox. 2024. Large language models for efficient medical information extraction. *AMIA Summits on Translational Science Proceedings*, 2024:509.

R. Carneiro. 2017. Sintomas e sinais: uma abordagem científica do exame clínico. *Revista Portuguesa de Medicina Geral e Familiar*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Google DeepMind. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261.

S. A. Hasan and 1 others. 2015. Ontology-driven semantic search for brazilian portuguese clinical notes. In *MEDINFO 2015: eHealth-enabled Health – Proceedings of the 15th World Congress on Health and Biomedical Informatics*, volume 216 of *Studies in Health Technology and Informatics*, São Paulo, Brazil. IOS Press.

Health Level Seven International. 2024a. Fhir: Fast healthcare interoperability resources. Accessed: Sep 30, 2024.

Health Level Seven International. 2024b. Hl7 international. Accessed: Sep 30, 2024.

Yan Hu and 1 others. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820.

Afia Z. Khan. 2024. Extraction of normalized symptom mentions from clinical narratives using large language models. *Journal of the American Medical Informatics Association*.

Florian Leiser, Richard Guse, and Ali Sunyaev. 2025. Large language model architectures in health care: Scoping review of research perspectives. *Journal of Medical Internet Research*.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Pengfei Liu and 1 others. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Maria Clara Saad Menezes, Alexander F. Hoffmann, Amelia L. M. Tan, and 1 others. 2025. The potential of generative pre-trained transformer 4 (gpt-4) to analyse medical notes in three different languages: a retrospective model-evaluation study. *The Lancet Digital Health*, 7(1):e35–e43.

Jingwen Nan and Li-Qun Xu. 2023. Designing interoperable health care services based on fast healthcare interoperability resources: literature review. *JMIR Medical Informatics*, 11(1):e44842.

NHS Digital. 2024. Snomed ct browser. Accessed: Jul 21, 2024.

João Oliveira and 1 others. 2022. Semclinbr - a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical nlp tasks. *Journal of Biomedical Semantics*.

OpenAI. 2023. GPT-4 technical report. arXiv preprint arXiv:2303.08774.

João Figueira Silva, Rui Antunes, João Rafael Almeida, and Sérgio Matos. 2020. Clinical concept normalization on medical records using word embeddings and heuristics. In *Studies in Health Technology and Informatics*, volume 270, pages 93–97.

Sonish Sivarajkumar and 1 others. 2024. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing. *JMIR Medical Informatics*.

Dongfang Xu and Timothy Miller. 2022. A simple neural vector space model for medical concept normalization using concept embeddings. *Journal of Biomedical Informatics*, 130:104080.

A. Zaghir and 1 others. 2024. Prompt engineering paradigms for medical applications: Scoping review. *JMIR*.

Kai Zhang, Tongtong Huang, Bradley A. Malin, and 1 others. 2025. Introducing mcodegpt as a zero-shot information extraction from clinical free text data tool for cancer research. *Communications Medicine*.

A Prompt structure for NormaTex-MapSNOMED

This appendix illustrates the structure of the prompt used by NormaTex-MapSNOMED for mapping clinical terms to semantic categories. The prompt targets Brazilian Portuguese clinical narratives; the version used in the experiments may include additional few-shot examples and refinements.

Instruction summary. The prompt specifies: (1) **Task**: given a clinical narrative in Brazilian Portuguese, identify each clinical term that refers to a disease/syndrome or a sign/symptom and assign it to exactly one of: **Disease or Syndrome** (conditions, diagnoses, pathologies), **Sign or Symptom** (signs, symptoms, complaints, clinical findings that are not diagnoses), or **Concepts to ignore** (medications, dosage information, complementary exam results such as lab values or imaging findings, and other non-target entities); (2) **Output format**: return only the original narrative with each mapped term wrapped in inline tags indicating its category (e.g., `<category>term</category>` or equivalent); (3) **Constraints**: no explanations, only the annotated narrative; each term receives exactly one category.

Example.

Input:

Paciente refere dor torácica e dispneia. Hipertensão em uso de losartana. ECG sem alterações.

Output:

Paciente refere `<Sign or Symptom>dor torácica</Sign or Symptom>` e `<Sign or Symptom>dispneia</Sign or Symptom>`. `<Disease or Syndrome>Hipertensão</Disease or Syndrome>` em uso de losartana. ECG sem alterações.

In this example, “Hipertensão” is tagged as Disease or Syndrome; “losartana” and “ECG sem alterações” are left unannotated (medication and exam finding). Figure 4 illustrates another narrative before and after processing.

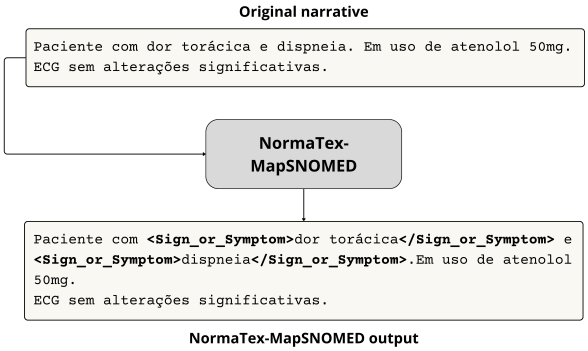

Figure 4: Example of a Brazilian Portuguese clinical narrative before and after processing with NormaTex-MapSNOMED: inline annotation of mapped terms; non-target concepts ignored.