Deterministic-First Clinical Concept Normalization in Brazilian Portuguese

Anonymous Authors[1]

Abstract

Mapping free-text clinical narratives to standardized ontologies is a critical bottleneck in healthcare NLP, yet remains largely unsolved for low-resource languages. To our knowledge, no automated system exists for SNOMED CT concept ID assignment in Brazilian Portuguese, limiting structured coding across one of the world's largest public health systems. We introduce SNOKind, a hybrid decision cascade that combines heuristic rules, embedding-based similarity, and LLM reasoning to map clinical entities extracted from Brazilian EHRs to SNOMED CT concepts. SNOKind is designed as a cost-aware pipeline: deterministic rules resolve the majority of cases cheaply and transparently, with stochastic methods reserved for genuinely ambiguous cases. This ordering strictly dominates LLM-first alternatives in cost and reliability while achieving better resolution coverage of 92%, compared to 57% for a standalone LLM baseline. Beyond performance, SNOKind demonstrates that orchestration over isolated model expressiveness is the key design principle for robust clinical NLP in constrained, real-world settings.

1. Introduction

Electronic Health Records (EHRs) contain rich clinical narratives critical for patient care, yet this information remains largely unstructured and difficult to use at scale. Standardized terminologies such as SNOMED CT®[1] transform free text into computable knowledge, enabling meaning-based retrieval, longitudinal tracking, epidemiological analysis, and reliable labels for downstream machine learning tasks. Despite its value, SNOMED CT concept ID assignment is costly and error-prone, especially in low-resource languages. This is particularly critical when mapping to concept identifiers (CID), which are numeric, language-agnostic, and enable interoperability and seamless information exchange across systems. Yet, no automated solution exists for PT-BR, limiting the use of structured data in one of the world's largest healthcare systems.

Prior approaches rely on supervised models or end-to-end LLMs, which struggle in data-scarce settings and introduce stochasticity. We propose SNOKind, a hybrid cascade for clinical SNOMED CT CID assignment in PT-BR that prioritizes deterministic methods (heuristics), followed by embedding similarity and LLM reasoning only when necessary. This design reduces cost and variability while preserving auditability, making structured SNOMED CT coding feasible in constrained settings such as the Brazilian SUS.

2. Related Work

Assigning standardized clinical codes to free text has been studied under a variety of paradigms. Rule-based systems were among the first to address this problem, relying on lexical matching and morphological normalization to link clinical mentions to SNOMED CT concepts, though these efforts were largely confined to English-language resources (Gaudet-Blavignac et al., 2021).

Neural approaches (e.g., Bi-GRU, BERT) achieved strong results on concept recognition benchmarks (Noori et al., 2025). More recent methods have explored direct SNOMED CT classification using transformer architectures trained on linked medical ontologies (Hristov et al., 2023), and two-stage entity linking pipelines that decouple candidate retrieval from concept matching (Kulyabin et al., 2024). Knowledge graph representations have also been leveraged to enrich concept embeddings, both for post-coordination (Castell-Díaz et al., 2023) and for contrastive learning over the Unified Medical Language System (UMLS) (Sakhovskiy et al., 2024).

LLM-based approaches have gained traction more recently. A scoping review across 37 studies found that the most common strategy is to incorporate SNOMED CT descrip-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]SNOMED CT® is a registered trademark of the International Health Terminology Standards Development Organization (IHTSDO). Used under license.

tions directly into model inputs, with concept normalization as the dominant downstream task (Chang & Sung, 2024). Performance gains are frequently reported, though the absence of standardized evaluation protocols makes cross-study comparison difficult. Beyond reproducibility concerns, LLM-based pipelines introduce computational overhead, non-deterministic outputs, and reduced auditability, factors that carry real consequences in clinical deployment.

Most of this work assumes English as the target language, for PT-BR, BioBERTpt demonstrated that domain-adaptive pretraining on clinical narratives improves NER performance over multilingual baselines (Schneider et al., 2020), and SemClinBr provided the field with a semantically annotated multi-specialty corpus that remains the primary benchmark for PT-BR clinical NLP (Oliveira et al., 2022). More recent work has examined resource-efficient LLM fine-tuning on Brazilian clinical data (de Souza Pinto et al., 2024). Oliveira and Rodrigues (de Oliveira & de Oliveira Rodrigues, 2026) derive a lightweight OWL ontology from SemClinBr annotations conceptually aligned with SNOMED CT, providing structured infrastructure for semantic interoperability, however, the work stops short of automated concept resolution and does not assign SNOMED CT CIDs. Most directly related is NormaTex-MapSNOMED (Araujo et al., 2026), which applies structured LLM prompting to map PT-BR clinical terms to SNOMED CT semantic categories. While promising, the approach assigns text-based category labels rather than concrete numerical SNOMED CT CIDs, limiting specificity, interoperability, and downstream usability. In contrast to language-agnostic CIDs, such labels are inherently language-dependent and less suitable for consistent information exchange across systems. Moreover, the method relies exclusively on LLM inference, leaving cost, variance, and auditability concerns unaddressed, while also underperforming overall. SNOKind addresses precisely these gaps: it resolves entities to SNOMED CT CIDs and reduces LLM exposure to 14% of cases through a deterministic-first cascade.

3. The SNOKind Framework

Figure 1 presents the proposed pipeline for mapping EHRs to SNOMED CT concepts. The proposed pipeline maps PT -BR clinical narratives from EHRs to SNOMED CT concepts through a multi-stage resolution process. All inputs are PT -BR clinical notes derived from real-world data from a Brazilian public hospital, de-identified and anonymized to ensure patient privacy and security.

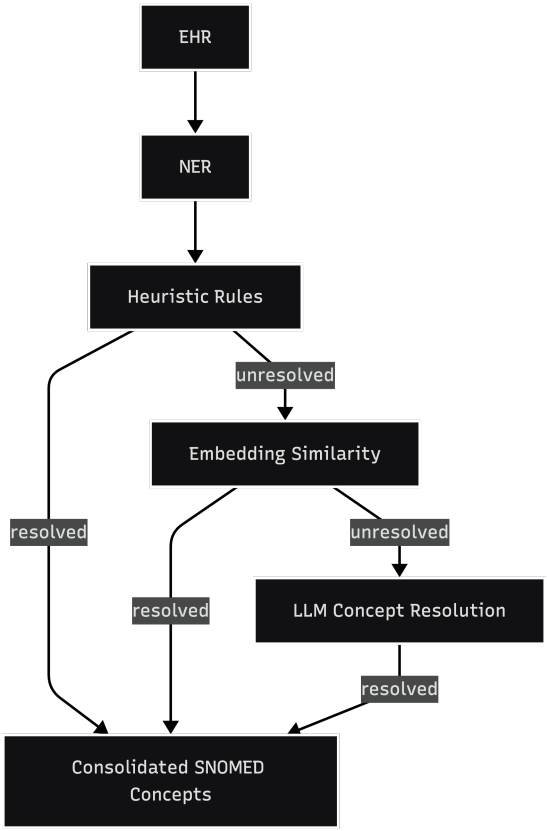

Figure 1. Overview of the SNOKind framework. The pipeline processes clinical text through multiple stages, producing structured clinical representations.

Electronic Health Records (EHR) Unstructured PT-BR clinical text containing symptoms, diagnoses, and procedures, often with noise, abbreviations, and terminological variability common in real-world settings.

Named Entity Recognition (NER) A LLM NER module identifies candidate clinical entities from the text. The model is adapted to Portuguese medical language and extracts spans corresponding to conditions, findings, and procedures.

SNOMED CT Vocabulary All resolution stages operate against a Portuguese SNOMED CT vocabulary, obtained by translating the official English concept descriptions using the GPT-4o model via OpenAI API. This translation step is applied once, prior to pipeline deployment, and serves as the shared reference for heuristic lookup, embedding indexing, and LLM-based disambiguation alike.

Heuristic Rules Extracted entities are first processed using deterministic rules. These include normalization, fuzzy matching, and direct lookup against the SNOMED CT vocabulary. Entities successfully matched at this stage are marked as resolved.

Embedding Similarity Unresolved entities are mapped using semantic similarity. Embeddings are computed for

Avaliamos paciente com derrame pericárdico (194970009) moderado em contexto de sepse (194394004), provável foco urinário (68566005), sem sinais ecocardiográficos de tamponamento (172185008). Encontra-se febril (386661006), taquicárdica (195069001) e com hipotensão (192835007) parcialmente responsiva à reposição volêmica (281800008), sem pulso paradoxal (52099006). Pericardiocentese (175181009) foi considerada, porém a deformidade torácica (301865001) acentuada inviabiliza o procedimento com segurança; indicada drenagem pericárdica (233189003) por janela subxifoide (224755008) em caráter prioritário. Há necessidade de hemodiálise (302497006) por falência de enxerto renal (105577008), cursando com anúria (213239005) e hipercalemia refratária (14140009).

Figure 2. Example clinical note with extracted SNOMED CT concept-ids. Concept-ids are highlighted in monospace for clarity.

both extracted entities and SNOMED CT terms, and perform nearest-neighbor search. If similarity exceeds a predefined threshold, the entity is resolved.

LLM Concept Resolution Remaining unresolved entities are passed to a LLM that uses full clinical context to disambiguate and map each entity to the most appropriate SNOMED CT concept, with outputs validated.

Consolidation All resolved concepts from the three stages are aggregated into a final SNOMED CT code set, enabling downstream applications such as classification algorithms, clinical database queries, and decision support systems, turning free-text narratives into machine-actionable data. Figure 2 demonstrates this on a real PT -BR EHR, with a full per-entity breakdown available in Table 3.

4. Experiments and Results

SNOKind is evaluated against a baseline to assess the impact of pipeline ordering on cost and reliability. For consistency, GPT-5 is used for NER and LLM-based resolution, while semantic retrieval uses `text-embedding-3-large`. This unified setup reduces confounding factors, ensuring differences stem from pipeline design rather than model variation, while supporting reproducibility and practical deployment. This homogeneous setup reduces confounding factors, ensuring differences arise from orchestration design rather than model variation.

SNOKind is model-agnostic: performance depends on the orchestration of NER, deterministic rules, embedding retrieval, and LLM reasoning, not on a specific model family. Alternative models (smaller, domain-specific or local) can be integrated at each stage.

Data Experiments use 100 ICU records from a Brazilian public hospital within the SUS network[2], yielding $2,092$ de-identified clinical entities as input.

Table 1. Overall performance comparison between LLM-only baseline and SNOKind pipeline

Method	Valid (%)	Unresolved (%)
LLM-only	57	43
SNOKind	92	8

Design Choices Conservative settings in early stages were adopted to prioritize precision over coverage. The heuristic layer uses restrictive rules to minimize false positives and provide a high-confidence deterministic foundation, at the cost of passing more cases downstream. For embeddings, a relatively strict similarity threshold (0.73) was selected, which improves match precision while deferring more ambiguous cases to the LLM. Overall, this reflects a deliberate trade-off: preserving high-confidence decisions early while shifting uncertainty to later, more expressive stages.

Baseline Our baseline is a standalone LLM without the NER preprocessing stage that SNOKind relies on. Without structured entity extraction, the LLM operates over raw clinical text and returns CID mappings directly.

SNOKind vs. Baseline SNOKind achieves 92% resolution over NER-extracted entities with valid SNOMED CT codes, compared to 57% for the LLM baseline, a gain of 36 percentage points. This improvement stems from two design decisions: (1) NER constrains the input to well-formed clinical entities before any matching occurs, and (2) the cascade prioritizes deterministic, auditable methods, reserving LLM inference for genuinely ambiguous cases.

Ordering Analysis To assess the role of orchestration, three pipeline orderings are compared: Rules→Emb→LLM (SNOKind), Emb→LLM→Rules, and LLM→Emb→Rules. As shown in Table 2, all configurations achieve similarly high resolution, with SNOKind slightly outperforming (92% vs. 91%), indicating that coverage is largely invariant to ordering. In contrast, cost and reliability vary substantially. This near-invariance across orderings arises from a set partition over an implicit error surface, where clinical entities are decomposed into disjoint regions of varying inferential difficulty, each handled by a distinct resolution mechanism.

As shown in Figure 3, placing either heuristics or embed-

[2]This study was conducted in compliance with Brazilian National Health Council Resolution CNS 466/12 and the General Data Protection Law (LGPD). The project was approved by the institutional Research Ethics Committee under opinion number [blinded for review] and CAAE [blinded for review], with a waiver of informed consent due to the use of a retrospective and de-identified dataset.

Table 2. Resolution performance across pipeline orderings. (A) Rules→Emb→LLM, (B) Emb→LLM→Rules, (C) LLM→Emb→Rules.

Metric	**(A)**	**(B)**	**(C)**
First stage (%)	36	66	64
Second stage (%)	35	25	28
Third stage (%)	21	0	0
Total (%)	**92**	**91**	**92**

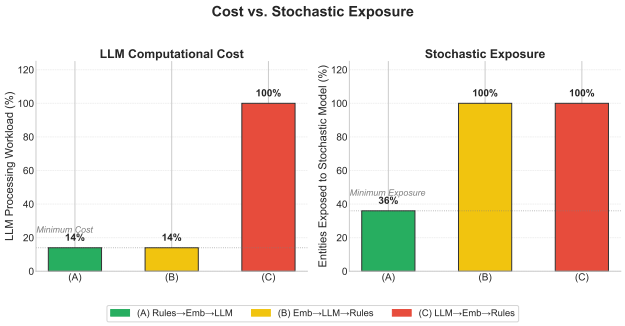

Figure 3. Computational cost and stochastic exposure across pipeline orderings. **Left:** Rules/embedding-first limit LLM usage, whereas LLM-first processes all inputs. **Right:** Only SNOKind pipeline preserves 36% of entities from stochastic processing.

dings first restricts LLM processing to only 14% of entities, whereas an LLM-first strategy imposes a 100% computational burden. Beyond cost, ordering also affects stochastic exposure: starting with deterministic rules reduces variance by resolving 36% of entities before any model inference, thereby eliminating hallucination risk for that subset.

Importantly, this invariance in final resolution is explained by complementary stage coverage: heuristic matches handle high-precision cases, embeddings recover semantically similar ones, and the LLM resolves the remaining hard cases via contextual reasoning. However, embedding-based matching introduces residual entropy due to its continuous similarity threshold, making it more susceptible to false positives on rare or morphologically variant clinical terms in Portuguese.

The heuristic layer thus acts as a *deterministic anchor*: its role is to reduce computational cost, variance, and auditability risk by filtering cases that do not require stochastic inference. Figure 3 highlights these two orthogonal effects. First, in terms of computational cost, both Rules-first and Embeddings-first pipelines reduce LLM load to 14% of the workload, whereas LLM-first exposes the entire pipeline to expensive inference. Second, in terms of stochastic exposure, only the Rules→Emb→LLM ordering ensures that a substantial fraction of entities (36%) is resolved deterministically before any probabilistic component is invoked, minimizing both variance. The other orderings expose the full input space to at least one stochastic stage, increasing sensitivity to model uncertainty regardless of cost efficiency.

4.1. Evaluation Under Label Scarcity

No benchmark dataset exists for SNOMED CT CID assignment in PT-BR, and SNOMED CT itself remains largely unadopted in Brazilian clinical practice. This absence extends to NER: no labeled corpus exists to assess whether extracted entities are genuinely SNOMED-relevant. We address this indirectly through resolution itself: every entity marked as *resolved* by SNOKind corresponds to a valid SNOMED CT code verified against the official terminology. Under this operationalization, resolution rate serves as a proxy for both CID assignment quality and NER relevance. Achieving 92% resolution across extracted entities suggests that the NER layer produces predominantly SNOMED-mappable clinical concepts, though we acknowledge this is an indirect signal rather than a gold-standard evaluation. A gold-standard annotated dataset would enable standard precision, recall, and F1 evaluation across pipeline stages — including assessment of the PT-BR SNOMED translation quality — and remains a critical contribution for the community. Furthermore, NER errors propagate silently through the pipeline and cannot be quantified without manual annotation, representing a fundamental limitation to be addressed in future work. Finally, SNOKind assumes every extracted entity has a valid SNOMED CT counterpart; NIL cases, ambiguous mappings, and concepts requiring post-coordination are outside the current scope. Additional detailed results are provided in the Appendix.

5. Conclusion

Clinical CID assignment demands robustness under noise and label scarcity, conditions that characterize PT-BR EHRs. SNOKind addresses this, a decision cascade that maps clinical narratives to SNOMED CT by combining heuristic rules, embedding similarity, and LLM reasoning, requiring no large-scale labeled data. Our key finding is that the heuristic→embedding→LLM ordering strictly dominates alternatives: it achieves equivalent resolution at lower cost and higher confidence by reserving expensive inference for genuinely ambiguous cases. Orchestration, not model expressiveness, is the primary driver of performance. To our knowledge, SNOKind is the first automated system for SNOMED CT CID assignment in PT-BR, offering a viable and auditable path for clinical NLP deployment in low-resource settings.

6. Impact Statement

This work introduces SNOKind, a cost-aware decision cascade for mapping PT-BR clinical text to SNOMED CT concepts. LLM usage in later stages may still introduce variability and limited explainability, hindering auditability in critical settings. Upstream biases may propagate and disproportionately affect certain patient groups. This work reports partial results of an ongoing research project.

References

Araujo, I., Moro, C., and Martinez, L. Normatex-mapsnomed: Bridging the gap between brazilian portuguese clinical narratives and snomed ct. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)-Vol. 1*, pp. 1085–1091, 2026.

Castell-Díaz, J., Miñarro-Giménez, J. A., and Martínez-Costa, C. Supporting snomed ct postcoordination with knowledge graph embeddings. *Journal of Biomedical Informatics*, 139:104297, 2023.

Chang, E. and Sung, S. Use of snomed ct in large language models: Scoping review. *JMIR Medical Informatics*, 12 (1):e62924, 2024.

de Oliveira, F. H. M. and de Oliveira Rodrigues, C. M. From annotated clinical narratives to ontology: Structuring brazilian portuguese clinical data. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)-Vol. 2*, pp. 128–134, 2026.

de Souza Pinto, J. G., Rodrigues de Freitas, A., Martins, A. C. G., Sawazaki, C. M. R., Vidal, C., and Silva e Oliveira, L. E. Developing resource-efficient clinical llms for brazilian portuguese. In *Brazilian Conference on Intelligent Systems*, pp. 46–60. Springer, 2024.

Gaudet-Blavignac, C., Foufi, V., Bjelogrlic, M., and Lovis, C. Use of the systematized nomenclature of medicine clinical terms (snomed ct) for processing free text in health care: Systematic scoping review. *J Med Internet Res*, 23(1):e24594, Jan 2021. ISSN 1438-8871. doi: 10.2196/24594. URL http://www.jmir.org/2021/1/e24594/.

Hristov, A., Ivanov, P., Aksenova, A., Asamov, T., Gyurov, P., Primov, T., and Boytcheva, S. Clinical text classification to snomed ct codes using transformers trained on linked open medical ontologies. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pp. 519–526, 2023.

Kulyabin, M., Sokolov, G., Galaida, A., Maier, A., and Arias-Vergara, T. Snobert: A benchmark for clinical notes entity linking in the snomed ct clinical terminology. In *International Conference on Pattern Recognition*, pp. 154–163. Springer, 2024.

Noori, A., Devkota, P., Mohanty, S., and Manda, P. Automated snomed ct concept annotation in clinical text using bi-gru neural networks, 2025. URL https://arxiv.org/abs/2508.02556.

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., and 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):13, 2022.

Sakhovskiy, A., Semenova, N., Kadurin, A., and Tutubalina, E. Biomedical entity representation with graph-augmented multi-objective transformer. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 4626–4643, 2024.

Schneider, E. T. R., de Souza, J. V. A., Knafou, J., e Oliveira, L. E. S., Copara, J., Gumiel, Y. B., de Oliveira, L. F. A., Paraiso, E. C., Teodoro, D., and 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, 2020.

A. SNOKind Extraction Example

Table 3 presents a complete SNOKind extraction over a single PT-BR EHR, illustrating the contribution of each pipeline stage. The heuristic rules layer resolves the majority of entities via direct lookup and fuzzy matching, covering well-formed clinical terms such as procedures and findings. The embedding similarity stage recovers semantically related concepts where surface form diverges from the SNOMED CT vocabulary, such as *tamponamento* and *instabilidade clínica*. Finally, the LLM stage handles the two remaining entities that required contextual reasoning to disambiguate. Together, the three stages resolve all 17 extracted entities to valid SNOMED CT concept IDs, with no unresolved cases.

Table 3. Example EHR and SNOMED CT concept extraction using the SNOKind

Extracted Term	**SNOMED CT Term**	**CID**	**Source**
derrame pericárdico	derrame pericárdico (distúrbio)	373945007	rules
taquicárdica	taquicardia (achado)	3424008	rules
hipotensão	hipotensão (transtorno)	155487000	rules
pulso paradoxal	pulso paradoxal (achado)	52099006	rules
pericardiocentese	pericardiocentese (procedimento)	309849004	rules
deformidade torácica	deformidade torácica (achado)	301865001	rules
drenagem pericárdica	drenagem do pericárdio (procedimento)	149186000	rules
hemodiálise	hemodiálise (procedimento)	302497006	rules
anúria	anúria (achado)	2472002	rules
tamponamento	tamponamento - ação (valor qualificativo)	257933004	similarity
falência de enxerto renal	falha de enxerto devido a necrose tubular aguda (achado)	73863006	similarity
hipercalemia refratária	hipercalemia crônica (distúrbio)	40777006	similarity
reposição volêmica	reposição de fluidos por via intravenosa (procedimento)	281800008	similarity
CTI pediátrico	cuidados pediátricos (regime/terapia)	700416004	similarity
instabilidade clínica	condição do paciente instável (achado)	162668006	similarity
sepse	Sepsis	91302008	llm
foco urinário	Focus of infection	44169009	llm

B. Qualitative Comparison

This section illustrates a qualitative comparison between SNOKind and LLM-only baseline on a single EHR excerpt. LLM-only approach operates directly over raw clinical text without NER preprocessing. Table 4 summarizes the outputs side by side, and the annotated excerpts below highlight the concept IDs inline.

SNOKind (all concepts valid)

> quadro de choque séptico (76571007) com foco inicialmente não definido, iniciada cobertura empírica com vancomicina (372735009), meropenem (387540000), polimixina b (372824005) e anidulafungina (422157006), com gentamicina (387321007) em período inicial. necessitou de noradrenalina em baixa dose por hipotensão (155487000) refratária ao volume, com posterior redução após estabilização.

GPT-5 LLM-only (three valid)

> quadro de choque séptico(91302008) com foco inicialmente não definido, iniciada cobertura empírica com vancomicina(372729009), meropenem(372687004), polimixina b(372749006) e anidulafungina(433724008), com gentamicina(372682005) em período inicial. necessitou de noradrenalina(372721003) em baixa dose por hipotensão(45007003) refratária ao volume, com posterior redução após estabilização.

As shown in Table 4, beyond producing invalid concept IDs, the LLM-only outputs exhibit two failure modes. First, annotation is incomplete: several entities remain unmapped, and seven are collapsed into two. Second, it introduces spurious spans absent from the NER-grounded input, a consequence of operating without structured entity extraction.

Table 4. Side-by-side concept ID outputs for the EHR excerpt above. SNOKind resolves all entities to valid SNOMED CT concept IDs via its deterministic-first cascade.

Extracted Term	SNOMED CT Term	CID	Model	Source
SNOKind				
choque séptico	choque séptico (distúrbio)	76571007	—	rules
vancomicina	vancomicina (substância)	372735009	—	rules
meropenem	meropeném (substância)	387540000	—	rules
polimixina b	polimixina b (substância)	372824005	—	rules
anidulafungina	anidulafungina (substância)	422157006	—	rules
gentamicina	gentamicina (substância)	387321007	—	rules
hipotensão	hipotensão (transtorno)	155487000	—	rules
LLM-only				
choque séptico	sepse (distúrbio)	91302008	GPT-5	llm
vancomicina	vancomicina (substância)	372729009	GPT-5	llm
meropenem	NA	372687004	GPT-5	llm
polimixina	NA	372749006	GPT-5	llm
anidulafungina	NA	433724008	GPT-5	llm
gentamicina	NA	372682005	GPT-5	llm
noradrenalina	NA	372721003	GPT-5	llm
hipotensão	pressão arterial baixa (distúrbio)	45007003	GPT-5	llm