Expansão Adaptativa de Vocabulário Clínico com LLMs Biomédicos em Registros Eletrônicos de Saúde

Nadine Anderle[1], **Dalvan Griebler**[1]

[1]Escola Politécnica, Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Porto Alegre – RS – Brasil

nadine.anderle@edu.pucrs.br, dalvan.griebler@pucrs.br

Abstract. *Biomedical ontologies often fail to capture the lexical variability present in real clinical narratives, limiting Natural Language Processing (NLP) applications over electronic health records. This work proposes a weakly supervised pipeline for adaptive clinical vocabulary expansion using ICD codes from MIMIC-IV v3.1. From 842 normalized root codes, the BioMistral-7B model generated 18,017 candidate terms. After semantic validation using SapBERT embeddings ($\theta = 0.60$), 4,094 terms (22.7%) were retained, representing the highest lexical acceptance rate among the evaluated configurations. In a downstream task of disease mention detection in clinical text (627 diseases), the configuration $\theta = 0.50$ achieved the best empirical performance, increasing macro recall from 4.8% to 21.4% and macro F1 from 2.0% to 5.2%. The results indicate that combining LLM-based lexical generation with embedding-based semantic validation enables scalable expansion of clinical vocabularies, improving diagnostic coverage in clinical text mining tasks.*

Resumo. *Ontologias biomédicas frequentemente não captam a variabilidade lexical presente em textos clínicos reais, limitando aplicações de Processamento de Linguagem Natural (PLN) em prontuários eletrônicos. Este trabalho propõe um pipeline de supervisão fraca para expansão adaptativa de vocabulário clínico utilizando códigos ICD do MIMIC-IV v3.1. A partir de 842 root codes normalizados, o modelo BioMistral-7B gerou 18.017 termos candidatos. Após validação semântica com embeddings SapBERT ($\theta = 0{,}60$), 4.094 termos (22,7%) foram aceitos, representando a maior taxa de aprovação lexical entre as configurações avaliadas. Na tarefa downstream de detecção de menções de doenças (627 doenças), a configuração $\theta = 0{,}50$ apresentou o melhor desempenho, elevando o recall macro de 4,8% para 21,4% e o F1 macro de 2,0% para 5,2%. Os resultados indicam que a combinação de geração lexical por LLM com validação semântica baseada em embeddings permite expandir vocabulários clínicos de forma escalável, ampliando a cobertura diagnóstica em tarefas de mineração de texto clínico.*

Palavras-chave: *Processamento de Linguagem Natural, Modelos de Linguagem, Vocabulário Clínico, Supervisão Fraca, MIMIC-IV*

1. Introdução

Os Registros Eletrônicos de Saúde (RES) constituem uma das principais infraestruturas informacionais para suporte à decisão clínica e pesquisa biomédica. Entretanto, parcela expressiva das informações permanece registrada em texto livre, o que dificulta a

interoperabilidade e a análise automatizada. Esse cenário impulsiona o uso de Processamento de Linguagem Natural (PLN) para estruturar conteúdo clínico não estruturado [Agrawal et al. 2022]. Ainda assim, a extração automática enfrenta um desafio estrutural: a elevada variabilidade lexical da linguagem clínica.

Na prática assistencial, diagnósticos são frequentemente descritos por abreviações, variações ortográficas e formulações não canônicas, muitas vezes distintas das representações padronizadas em ontologias biomédicas como UMLS [Bodenreider 2004]. Essa discrepância reduz a cobertura de tarefas como reconhecimento de entidades nomeadas e normalização semântica [Chen et al. 2025], fragmentando a representação conceitual de diagnósticos.

Diante desse contexto, investigamos a seguinte questão de pesquisa: *como realizar a expansão lexical automática de diagnósticos clínicos, sob supervisão fraca, a partir de um termo âncora validado, preservando equivalência semântica entre as variações geradas e o conceito original?*. Partimos da hipótese de que modelos de linguagem biomédicos podem gerar variações lexicais semanticamente equivalentes para diagnósticos âncora, e que a aplicação de validação semântica vetorial permite controlar explicitamente o compromisso entre cobertura lexical e fidelidade conceitual.

Grandes Modelos de Linguagem (LLMs) têm sido explorados para ampliar a cobertura lexical devido à sua capacidade generativa. Contudo, a geração livre pode introduzir inconsistências semânticas, exigindo mecanismos externos de validação antes do uso operacional [Chen et al. 2025]. A literatura carece de abordagens que não apenas vinculem entidades a catálogos fechados, mas que proponham explicitamente a expansão controlada e validada do vocabulário clínico para uso em sistemas downstream.

Neste trabalho, propomos um pipeline de expansão lexical guiada ancorado nos diagnósticos estruturados do MIMIC-IV [Johnson et al. 2023]. Os códigos ICD são normalizados, submetidos à geração lexical sintética via BioMistral-7B [Labrak et al. 2024] e rigorosamente filtrados por similaridade semântica utilizando *embeddings* biomédicos SapBERT [Liu et al. 2021]. O método é avaliado tanto intrinsecamente, na consistência do vocabulário produzido, quanto extrinsecamente em uma tarefa *downstream* de detecção de doenças em texto clínico.

O restante do artigo está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados; a Seção 3 descreve o método proposto; a Seção 4 apresenta os experimentos e resultados; a Seção 5 discute os resultados e limitações; e a Seção 6 sintetiza as contribuições.

2. Trabalhos Relacionados

Na literatura de PLN clínico, a extração de informação e a estruturação de textos livres para tarefas *downstream* baseiam-se frequentemente em vocabulários estáticos e ontologias, como o UMLS, combinadas a modelos supervisionados especializados, como BioBERT [Bodenreider 2004, Lee et al. 2020]. Embora eficazes para normalizar conceitos catalogados, essas abordagens assumem uma estabilidade terminológica que frequentemente não se sustenta em textos clínicos reais, marcados por abreviações e variações linguísticas, resultando em perda de cobertura lexical [Fries et al. 2021]. Mesmo soluções baseadas em aprendizado de representações robustas, como o MedCAT

[Kraljevic et al. 2021], permanecem centradas na identificação de entidades de ontologias consolidadas.

Para reduzir a dependência de anotação manual em larga escala, o paradigma de supervisão fraca (*weak supervision*) tem sido explorado. Abordagens como *Data Programming* utilizam funções heurísticas e regras de rotulação para gerar conjuntos de treinamento probabilísticos de forma programática, permitindo treinar modelos discriminativos sem dados anotados manualmente [Ratner et al. 2017]. No domínio biomédico, as estratégias de supervisão fraca são aplicadas à extração de entidades clínicas, mas geralmente focam em alinhar textos a ontologias já existentes [Fries et al. 2021]. Contudo, essas aplicações operam sobre espaços conceituais fixos, limitando a expansão sistemática do repertório lexical para termos ausentes em dicionários predefinidos.

Com a ascensão dos Grandes Modelos de Linguagem (LLMs), possibilitou-se extrair e gerar variações semânticas a partir de instruções controladas, demonstrando forte capacidade de generalização no domínio biomédico [Yang et al. 2024]. Entretanto, a geração livre introduz desafios severos de controle, como sensibilidade ao *prompt* e o risco crítico de alucinações, situações em que o modelo produz saídas plausíveis, porém incorretas, incompletas ou inconsistentes com a realidade clínica [Labbé et al. 2023, Chen et al. 2025]. Por essa razão, mecanismos externos de restrição e validação tornam-se componentes essenciais para uso confiável desses modelos [Agrawal et al. 2022].

Abordagens recentes em *biomedical entity linking* mitigam esses riscos ao combinar LLMs a estratégias de decodificação restritiva e desambiguação [Ye and Mitchell 2025]. Ainda assim, esses métodos assumem um espaço de busca previamente definido e dependem estritamente dos candidatos gerados, dificultando a incorporação de novas variantes linguísticas [Ye and Mitchell 2025]. De forma semelhante, *benchmarks* contemporâneos de codificação clínica concentram-se na atribuição automática de códigos ICD a partir de notas textuais, o que caracteriza essencialmente tarefas de classificação supervisionada, e não de expansão do repertório lexical [Edin et al. 2023].

Diante desse cenário, observam-se duas lacunas principais: (i) a ausência de abordagens que explorem supervisão fraca para expansão lexical generativa com controle semântico explícito; e (ii) a limitada investigação sobre a estabilidade estrutural e semântica da saída de LLMs biomédicos sob restrições formais de geração e validação vetorial. O presente trabalho posiciona-se nesse contexto ao integrar geração estruturada com BioMistral-7B [Labrak et al. 2024], avaliação complementar com LLaMA2-MedTuned [Rohanian et al. 2024] e validação semântica automática baseada em embeddings SapBERT [Liu et al. 2021], compondo um pipeline controlado de expansão adaptativa de vocabulário clínico.

3. Método de Expansão Adaptativa do Vocabulário Clínico

Esta seção descreve o método proposto para expansão adaptativa de vocabulário clínico sob supervisão fraca. O processo utiliza diagnósticos estruturados como âncoras conceituais para gerar variações textuais semanticamente equivalentes, dispensando anotação manual. O pipeline é organizado em três etapas: (i) preparação e normalização do espaço diagnóstico, (ii) expansão lexical generativa com validação semântica e (iii) aplicação do vocabulário resultante em uma tarefa *downstream* de detecção de menções de doenças. A

Figura 1 apresenta uma visão geral do fluxo completo do método.

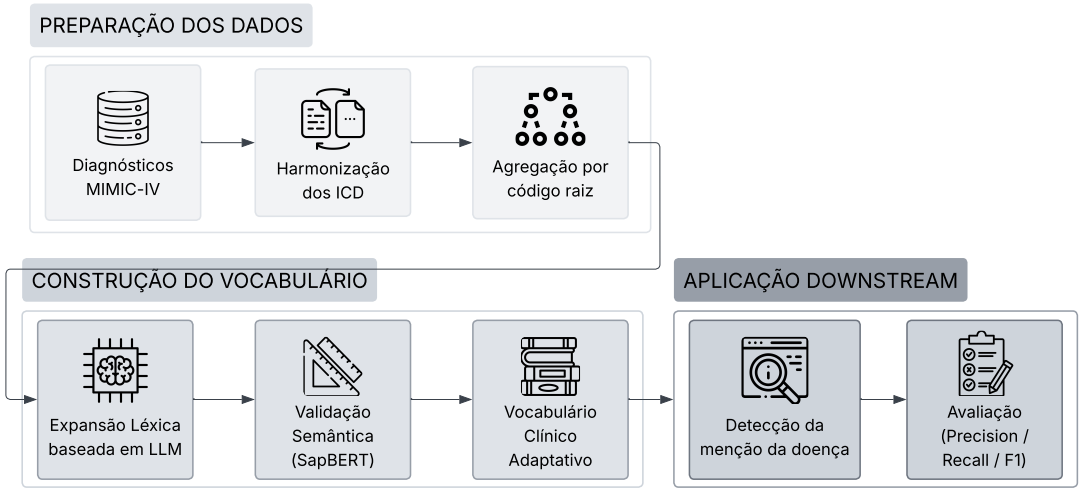

Figura 1. Pipeline completo do método proposto, organizado em três fases

3.1. Dados e Preparação do Espaço Diagnóstico

Utilizamos o MIMIC-IV v3.1 [Johnson et al. 2023], base pública de registros eletrônicos de saúde contendo dados estruturados e narrativas clínicas do Beth Israel Deaconess Medical Center. A extração concentrou-se nos diagnósticos clínicos estruturados e na tabela de referência que associa cada código ICD à sua descrição textual. A adoção do ICD como vocabulário de referência decorre por ser o padrão nativo do MIMIC-IV, dispensando mapeamento ontológico intermediário.

Como o MIMIC-IV combina versões ICD-9-CM e ICD-10-CM, aplicamos uma adaptação da estratégia de conversão e agregação proposta por Gupta et al. [Gupta et al. 2023]. Inicialmente, priorizamos os 4.683 diagnósticos mais frequentes para mitigar efeitos de cauda longa [Edin et al. 2023]. Em seguida, realizamos a harmonização semântica por conversão para ICD-10-CM.

Os diagnósticos harmonizados foram agregados por código raiz (*root code*), colapsando subcategorias em categorias de três caracteres [Gupta et al. 2023, Edin et al. 2023]. Esse procedimento resultou em 842 *root codes*, preservando as 5.753.571 ocorrências originais e reduzindo aproximadamente 82% do espaço diagnóstico. Esses códigos constituem as âncoras conceituais utilizadas na etapa de expansão lexical e, adicionalmente, servem como rótulos fracos na avaliação extrínseca: uma nota de alta é instância positiva de um *root code* quando esse código está presente nos diagnósticos estruturados do respectivo internamento.

3.2. Expansão Lexical e Validação Semântica

Cada título diagnóstico associado aos *root codes* é submetido à expansão lexical generativa utilizando o modelo biomédico BioMistral-7B [Labrak et al. 2024]. O modelo é instruído a produzir variações semanticamente equivalentes ao diagnóstico âncora, incluindo sinônimos clínicos, variações morfológicas e abreviações.

Os termos candidatos são posteriormente submetidos a validação semântica automática por meio de *embeddings* SapBERT [Liu et al. 2021]. Um termo é aceito quando a similaridade de cosseno com o diagnóstico âncora supera o limiar $\theta = 0.60$, convertendo a geração probabilística em um processo controlado por critério quantitativo de equivalência semântica.

O limiar $\theta = 0{,}60$ foi definido a partir da distribuição de similaridades: valores inferiores a 0,55 capturavam termos semanticamente adjacentes, enquanto valores superiores a 0,70 eliminavam variantes legítimas. Assim, $\theta = 0{,}60$ foi adotado como referência para a análise intrínseca, equilibrando restrição semântica e cobertura lexical.

Para análise de comparação experimental, o pipeline também foi executado com o modelo LLaMA2-MedTuned-7B [Rohanian et al. 2024], mantendo o mesmo protocolo de geração e validação semântica.

Sob a perspectiva metodológica, o processo caracteriza-se como supervisão fraca (*weak supervision*), combinando: (i) a ontologia ICD como mecanismo de *distant supervision*, (ii) um modelo generativo para produção de hipóteses lexicais e (iii) um modelo de *embeddings* biomédicos como filtro semântico determinístico.

3.3. Aplicação Downstream

O vocabulário é aplicado em uma tarefa de detecção de menções de doenças em notas de alta do MIMIC-IV-Note v2.2 [Johnson et al. 2023], um corpus de narrativas clínicas em inglês do Beth Israel Deaconess Medical Center. Os rótulos de referência correspondem aos *root codes* ICD presentes nos diagnósticos de cada internamento, conforme o mecanismo de supervisão fraca descrito na Seção 3.1.

Nesse contexto, abordagens baseadas em terminologias e *gazetteers* surgem como alternativas utilizadas em PLN clínico em cenários com escassez de dados anotados, devido à simplicidade operacional e transparência interpretativa [Kraljevic et al. 2021, Fries et al. 2021].

Avaliamos quatro condições: `baseline` (título normalizado por código) e `expanded@0.5`, `expanded@0.6` e `expanded@0.7`, correspondentes aos vocabulários expandidos validados em cada limiar. As condições são comparadas como *gazetteers* independentes por código, e a detecção é realizada por correspondência lexical exata da palavra inteira entre termos e textos clínicos.

A eficácia é medida por Precisão, *Recall* e F1 em nível de doença, com agregação macro. Como a avaliação é totalmente automática e não envolve curadoria humana, os resultados devem ser interpretados como evidência inicial de impacto e não como validação clínica definitiva.

4. Experimentos e Resultados

Esta seção avalia o vocabulário clínico expandido produzido pelo pipeline em duas frentes complementares: (i) avaliação intrínseca das etapas de geração e validação semântica e (ii) avaliação extrínseca em uma tarefa *downstream* de detecção de doenças em texto clínico.

Como verificação preliminar do fluxo de geração estruturada, conduzimos um ensaio exploratório com BioGPT [Luo et al. 2022]. Em um subconjunto de 10 *root codes*,

o modelo gerou 192 candidatos lexicais, dos quais apenas 2 foram aceitos após validação semântica (1,04%). O resultado sugere baixa aderência do modelo às restrições estruturais do *prompt*, contrastando com modelos *instruction-tuned*, que apresentam maior consistência de saída [Labrak et al. 2024, Chen et al. 2025]. Esse experimento foi utilizado apenas para validar o fluxo experimental; os experimentos principais foram conduzidos com o BioMistral-7B.

4.1. Geração e Validação Semântica

Na etapa de geração do vocabulário, o BioMistral-7B produziu saída válida para 684 dos 842 *root codes* (81,2%), gerando 18.017 termos candidatos. Após aplicação do filtro semântico com SapBERT em $\theta = 0{,}60$, 4.094 termos foram aprovados (22,7%), enquanto 13.923 foram rejeitados.

A Tabela 1 apresenta o funil de cobertura em nível de código ao longo do pipeline. A primeira linha representa o universo inicial de diagnósticos estruturados utilizados como âncoras conceituais. A segunda linha indica a proporção de códigos para os quais o modelo conseguiu produzir saída estruturada válida. A terceira linha reflete o subconjunto final de códigos que mantiveram ao menos um termo após a validação semântica baseada em *embeddings*. Dessa forma, a diferença entre as duas últimas etapas quantifica o impacto do filtro semântico na eliminação de candidatos lexicalmente plausíveis, porém semanticamente distantes do diagnóstico original.

Tabela 1. Funil de cobertura por código ao longo do pipeline.

Etapa	Códigos	% sobre 842
Root codes normalizados	842	100,0%
Com geração válida	684	81,2%
Com $\geq$1 termo aceito	498	59,1%

Para analisar a sensibilidade do filtro semântico, mantivemos fixo o conjunto de 18.017 candidatos e variamos apenas o limiar θ. Como mostrado na Tabela 2, observa-se redução monotônica na taxa de aceitação à medida que o critério de similaridade se torna mais restritivo.

Tabela 2. Sensibilidade do SapBERT sobre 18.017 candidatos.

Threshold	Aceitos	Rejeitados	Taxa de Aceitação
0,50	8.666	9.351	48,1%
0,60	4.094	13.923	22,7%
0,70	1.132	16.885	6,3%

A Tabela 2 deve ser interpretada como uma análise isolada do comportamento do filtro semântico. Cada linha corresponde à aplicação do mesmo conjunto de candidatos sob diferentes limiares de similaridade, de modo que as diferenças observadas refletem exclusivamente o efeito do critério de validação. Observa-se que limiares mais permissivos ampliam substancialmente a cobertura lexical, enquanto limiares mais restritivos reduzem o vocabulário aprovado. Esse comportamento evidencia o papel do parâmetro θ como mecanismo de controle do *trade-off* entre cobertura e precisão semântica.

No plano intrínseco, $\theta = 0{,}60$ apresentou um compromisso intermediário entre cobertura lexical e restrição semântica, reduzindo substancialmente o ruído gerado sem colapsar o vocabulário resultante. Esse resultado deve ser interpretado no contexto da avaliação intrínseca, não implicando necessariamente melhor desempenho em tarefas *downstream*.

4.2. Execução Complementar com LLaMA

Para avaliar a sensibilidade do pipeline ao modelo gerador, realizamos uma execução complementar com o Llama2-MedTuned-7b [Rohanian et al. 2024], mantendo o mesmo *prompt* e o protocolo de validação. O modelo produziu saída válida para 348 dos 842 *root codes* (41,3%), aproximadamente metade da cobertura obtida com o BioMistral-7B. Após validação semântica, 287 códigos (34,1%) mantiveram ao menos um termo aceito, totalizando 7.701 candidatos gerados.

Tabela 3. Funil de cobertura por código para o LLaMA2-MedTuned.

Etapa	Códigos	% sobre 842
Root codes normalizados	842	100,0%
Com geração válida	348	41,3%
Com $\geq$1 termo aceito	287	34,1%

A análise de sensibilidade apresentada na Tabela 4 indica comportamento semelhante do filtro semântico. Em $\theta = 0{,}60$, 1.792 termos foram aprovados (23,2%), proporção próxima à observada com BioMistral-7B, embora aplicada a um conjunto bruto substancialmente menor.

Tabela 4. Sensibilidade do SapBERT sobre 7.701 candidatos (LLaMA).

Threshold	Aceitos	Rejeitados	Taxa
0,50	3.306	4.395	42,9%
0,60	1.792	5.909	23,2%
0,70	604	7.097	7,8%

Esse contraste sugere que o principal gargalo do LLaMA2-MedTuned-7b está na etapa de geração estrutural, enquanto o comportamento seletivo observado após a validação decorre principalmente do critério semântico adotado.

4.3. Composição do Vocabulário Final

O vocabulário final gerado com BioMistral contém 4.094 termos distribuídos em 498 códigos diagnósticos, correspondentes ao subconjunto de *root codes* que mantiveram ao menos um candidato após a etapa de validação semântica. Em média, cada diagnóstico recebeu 8,2 variações lexicais aprovadas.

A distribuição tipológica é apresentada na Tabela 5.

Sinônimos e variações representam 87,3% do vocabulário validado, indicando predominância de reformulações lexicais próximas ao diagnóstico original. Em contraste, a menor proporção de abreviações reflete a maior ambiguidade semântica de siglas clínicas quando avaliadas fora de contexto [Ye and Mitchell 2025].

Tabela 5. Distribuição por categoria no vocabulário final.

Categoria	Termos	% sob 4.094
Sinônimos	1.890	46,2%
Variações	1.686	41,1%
Abreviações	518	12,7%

4.4. Avaliação Downstream de Detecção

Para avaliar o impacto prático do vocabulário expandido, realizamos uma avaliação extrínseca nas notas de alta do MIMIC-IV-Note v2.2 [Johnson et al. 2023], usando os códigos ICD do internamento como rótulos fracos de referência para detecção de menções de doenças.

Foram consideradas 633 doenças derivadas da união dos vocabulários gerados nas configurações `expanded@0.5`, `expanded@0.6` e `expanded@0.7`. Destas, 627 apresentaram instâncias suficientes para cálculo das métricas. O desempenho foi comparado a um cenário *baseline* baseado apenas nos termos diagnósticos originais.

Tabela 6. Resultados downstream de detecção de doenças

Condição	Precision (macro)	*Recall* (macro)	F1 (macro)	TP total	ΔRecall (pp)	ΔF1 (pp)
baseline	3,9%	4,8%	2,0%	2.966	—	—
expanded@0.5	6,3%	21,4%	5,2%	13.171	+16,6	+3,2
expanded@0.6	5,1%	12,5%	3,8%	7.678	+7,7	+1,8
expanded@0.7	2,1%	3,5%	1,4%	2.112	-1,3	-0,6

A Tabela 6, a configuração `expanded@0.5` apresentou o maior ganho de cobertura, elevando o *recall* macro(+16,6 pp) e aumentando o número total de verdadeiros positivos de 2.966 para 13.171. Em contraste, a configuração `expanded@0.7` apresentou queda de desempenho em relação ao *baseline*, indicando que limiares mais permissivos favorecem cobertura lexical em tarefas de detecção, desta forma códigos para os quais nenhum candidato atingiu $\theta = 0{,}7$ permanecem sem representação no *gazetteer*.

Embora os valores absolutos de precisão e F1 permaneçam modestos, esse comportamento é consistente com abordagens baseadas em *gazetteers*, nas quais o objetivo principal é ampliar a cobertura lexical inicial de menções clínicas. A análise por doença apresenta tendência semelhante: `expanded@0.5` aumentou o *recall* em 325 das 627 doenças avaliadas (51,8%) e reduziu essa métrica em 44 (7,0%), enquanto o *F1-score* aumentou em 304 doenças (48,5%).

5. Discussão

Os resultados evidenciam dois regimes operacionais do pipeline: limiares mais restritivos na validação semântica baseada em embeddings SapBERT favorecem maior fidelidade conceitual, enquanto limiares mais permissivos ampliam a cobertura lexical em texto livre. Esse comportamento reforça o papel do parâmetro θ como mecanismo explícito de controle do trade-off entre precisão semântica e recall.

Na avaliação intrínseca, a etapa generativa apresentou elevada produtividade, mas também ruído semântico substancial. Dos 18.017 candidatos produzidos pelo BioMistral-7B, apenas 4.094 foram aceitos após validação semântica com $\theta = 0{,}60$ (22,7%). Esse

comportamento é consistente com evidências de que LLMs biomédicos em regime *zero-shot* frequentemente produzem variações lexicalmente plausíveis, porém semanticamente imprecisas [Chen et al. 2025]. Nesse contexto, o filtro vetorial desempenha papel essencial como mecanismo de controle de qualidade da expansão lexical.

A análise de sensibilidade mostrou que o limiar de similaridade atua como mecanismo explícito de controle entre cobertura lexical e fidelidade conceitual. Na avaliação intrínseca, $\theta = 0{,}60$ apresentou melhor equilíbrio semântico, preservando variações plausíveis sem introduzir expansão excessivamente ruidosa. Contudo, como a similaridade vetorial não garante equivalência conceitual estrita, limiares mais permissivos podem capturar variantes textuais adicionais presentes em narrativas clínicas reais.

Essa diferença torna-se evidente na avaliação *downstream*. Na tarefa de detecção de doenças, o vocabulário validado com $\theta = 0{,}50$ apresentou melhor desempenho, elevando o *recall* macro de 4,8% para 21,4% e o *F1-score* macro de 2,0% para 5,2%. Em termos práticos, isso sugere dois regimes de uso do pipeline: limiares mais altos favorecem maior fidelidade conceitual em tarefas de normalização terminológica, enquanto limiares mais baixos ampliam a recuperação de menções em texto livre, funcionando como etapa inicial de detecção em pipelines de extração de informação clínica.

Do ponto de vista metodológico, o pipeline situa-se entre abordagens baseadas em terminologias estáticas, que oferecem alta consistência conceitual porém baixa cobertura lexical, e estratégias de geração livre com LLMs, que ampliam a variabilidade linguística ao custo de maior ruído semântico. A combinação de geração aberta com validação vetorial independente permite equilibrar criatividade lexical e fidelidade conceitual, contribuindo para reduzir o descompasso entre ontologias biomédicas padronizadas e a linguagem observada em narrativas clínicas.

5.1. Limitações

O estudo apresenta limitações relevantes. A validação semântica baseia-se exclusivamente na similaridade de cosseno entre *embeddings* SapBERT de termos isolados.

Além disso, a avaliação *downstream* foi totalmente automática e sem curadoria humana. Parte dos erros pode decorrer de imperfeições no pareamento entre códigos estruturados e menções textuais, especialmente em notas com múltiplos diagnósticos [Oliveira et al. 2022]. Além disso, diagnósticos estruturados podem não refletir integralmente o conteúdo narrativo das notas: as detecções de doenças mencionadas mas não codificadas no internamento são tratadas como falsos positivos pelo protocolo automático, podendo subestimar a precisão real do vocabulário.

Os experimentos foram conduzidos no MIMIC-IV v3.1 (dados diagnósticos estruturados) e MIMIC-IV-Note v2.2 (notas de alta), ambos em inglês e provenientes de uma única instituição, o que limita a validade externa [Gupta et al. 2023]. A comparação entre modelos também possui ressalvas, pois apenas uma execução complementar com o LLaMA2-MedTuned-7b foi realizada, restringindo o *benchmarking* frente a outros Grandes Modelos de Linguagem [Chen et al. 2025].

Por fim, a priorização dos diagnósticos mais frequentes reduz a esparsidade, mas pode introduzir viés de cobertura, possivelmente superestimando o desempenho do modelo em cenários de doenças raras.

6. Conclusão

Este trabalho investigou a viabilidade de automatizar a construção de recursos léxicos clínicos sob regime de supervisão fraca, buscando mitigar o gargalo da anotação manual e a dependência de corpora extensivamente rotulados. Foi proposto um pipeline que integra geração lexical estruturada via modelo biomédico instruído (BioMistral-7B) e validação semântica independente baseada em embeddings contrastivos (SapBERT), utilizando exclusivamente códigos ICD estruturados como âncora conceitual e operando em ambiente local, em conformidade com requisitos de privacidade.

Os resultados evidenciam que a capacidade generativa do modelo é elevada, porém intrinsecamente ruidosa. A aplicação de um filtro vetorial com limiar $\theta = 0{,}60$ reduziu o conjunto bruto para uma taxa de aceitação de 22,7%, demonstrando que o controle semântico não constitui etapa acessória, mas condição necessária para tornar a expansão lexical operacionalmente confiável. A predominância de sinônimos e variações (87,3%) no vocabulário final sugere preservação de equivalência conceitual, enquanto a menor taxa de aceitação de abreviações (12,7%) confirma a limitação de validações baseadas exclusivamente em similaridade vetorial para termos polissêmicos descontextualizados [Chen et al. 2025].

Na avaliação *downstream* automática, considerando 627 doenças avaliadas, o *gazetteer* `expanded@0.5` apresentou melhora substancial em relação ao *baseline*. O *recall* macro aumentou de 4,8% para 21,4%, enquanto o F1 macro passou de 2,0% para 5,2%. Isso corresponde a ganhos de +16,6 e +3,2 pontos percentuais, respectivamente, indicando aumento consistente da cobertura diagnóstica em texto clínico. Esse resultado indica que limiares mais permissivos podem favorecer a recuperação de menções em texto livre, nas quais variações ortográficas, abreviações e formulações não canônicas são frequentes.

A execução complementar com o LLaMA2-MedTuned-7B, sob protocolo idêntico de geração estruturada e validação semântica, apresentou diferenças substanciais na taxa de geração estrutural válida, porém manteve proporção relativa de aceitação semelhante após o filtro vetorial. Esse comportamento sugere que o controle semântico mantém padrão estável entre modelos, enquanto a cobertura final do pipeline permanece dependente da robustez do modelo gerador.

Diferentemente de abordagens baseadas exclusivamente em ontologias estáticas, que restringem a descoberta de variantes locais, ou em geração irrestrita, que maximiza cobertura à custa de consistência, o pipeline demonstra empiricamente que é possível equilibrar criatividade lexical e fidelidade conceitual por meio de validação independente. Em termos práticos, esse equilíbrio pode ser ajustado pelo limiar de similaridade: valores mais restritivos, como $\theta = 0{,}60$, tendem a preservar maior fidelidade conceitual e são mais adequados para tarefas de normalização terminológica ou integração com ontologias clínicas, enquanto limiares mais permissivos, como $\theta = 0{,}50$, ampliam a cobertura lexical e favorecem a detecção inicial de menções em texto livre.

Como trabalhos futuros, propõe-se: (i) aprofundar a avaliação *downstream* com anotação manual de menções para estimar desempenho real em tarefas de NER no MIMIC-IV Note e em corpora clínicos em português, como o SemClinBr [Oliveira et al. 2022]; (ii) investigar estratégias de validação contextual, incluindo abor-

dagens de Geração Aumentada por Recuperação (RAG); (iii) conduzir análises de erro por doença e por perfil de nota para calibrar o compromisso entre cobertura e precisão do vocabulário expandido; (iv) avaliar formalmente o enriquecimento multi-rótulo de atendimentos a partir das detecções textuais, com protocolo de validação clínica para evitar propagação de ruído. e (v) explorar a SNOMED CT como âncora ontológica de maior granularidade semântica.

Em síntese, os achados indicam que a expansão lexical clínica orientada por LLMs é tecnicamente viável sob supervisão fraca, desde que acompanhada por mecanismos explícitos e ajustáveis de controle semântico. O estudo contribui ao demonstrar empiricamente o compromisso entre geração aberta e equivalência conceitual mensurável, oferecendo um caminho escalável e metodologicamente controlado para enriquecimento terminológico em registros eletrônicos de saúde.

Agradecimentos

Este trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Código de Financiamento 001, FAPERGS (Nº 24/2551-0001400-4), e CNPq (Nº311012/2025-6).

Referências

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., and Sontag, D. (2022). Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(90001):267D–270.

Chen, Q., Hu, Y., Peng, X., Xie, Q., Jin, Q., Gilson, A., Singer, M. B., Ai, X., Lai, P.-T., Wang, Z., Keloth, V. K., Raja, K., Huang, J., He, H., Lin, F., Du, J., Zhang, R., Zheng, W. J., Adelman, R. A., Lu, Z., and Xu, H. (2025). Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature Communications*, 16(1):3280.

Edin, J., Junge, A., Havtorn, J. D., Borgholt, L., Maistro, M., Ruotsalo, T., and Maaløe, L. (2023). Automated Medical Coding on MIMIC-III and MIMIC-IV: A Critical Review and Replicability Study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2572–2582. arXiv:2304.10909 [cs].

Fries, J. A., Steinberg, E., Khattar, S., Fleming, S. L., Posada, J., Callahan, A., and Shah, N. H. (2021). Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nature Communications*, 12(1):2017.

Gupta, M., Gallamoza, B., Cutrona, N., Dhakal, P., Poulain, R., and Beheshti, R. (2023). An Extensive Data Processing Pipeline for MIMIC-IV.

Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L.-w. H., Celi, L. A., and Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1).

Kraljevic, Z., Searle, T., Shek, A., Roguski, L., Noor, K., Bean, D., Mascio, A., Zhu, L., Folarin, A. A., Roberts, A., Bendayan, R., Richardson, M. P., Stewart, R., Shah, A. D., Wong, W. K., Ibrahim, Z., Teo, J. T., and Dobson, R. J. (2021). Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. *Artificial Intelligence in Medicine*, 117:102083.

Labbé, T., Castel, P., Sanner, J.-M., and Saleh, M. (2023). ChatGPT for phenotypes extraction: one model to rule them all? In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4, Sydney, Australia. IEEE.

Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., and Dufour, R. (2024). BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. arXiv:2402.10373 [cs].

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Liu, F., Shareghi, E., Meng, Z., Basaldella, M., and Collier, N. (2021). Self-Alignment Pretraining for Biomedical Entity Representations. arXiv:2010.11784 [cs].

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., and Liu, T.-Y. (2022). BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining. *Briefings in Bioinformatics*, 23(6):bbac409. arXiv:2210.10341 [cs].

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. (2022). SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks. *Journal of Biomedical Semantics*, 13(1):13.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., and Ré, C. (2017). Snorkel: Rapid Training Data Creation with Weak Supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282. arXiv:1711.10160 [cs].

Rohanian, O., Nouriborji, M., Kouchaki, S., Nooralahzadeh, F., Clifton, L., and Clifton, D. A. (2024). Exploring the effectiveness of instruction tuning in biomedical language processing. *Artificial Intelligence in Medicine*, 158:103007.

Yang, J., Liu, C., Deng, W., Wu, D., Weng, C., Zhou, Y., and Wang, K. (2024). Enhancing phenotype recognition in clinical notes using large language models: PhenoBCBERT and PhenoGPT. *Patterns*, 5(1):100887.

Ye, C. and Mitchell, C. S. (2025). LLM as entity disambiguator for biomedical entity-linking. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 301–312, Vienna, Austria. Association for Computational Linguistics.