Um Pipeline Baseado em LLMs para a Triagem Clínica e a Extração de Comorbidades Focado na Interoperabilidade da RNDS

Cristiano da Silveira Colombo[1], **Cauã Gomes Marvila, Lucas Mazioli, Beatriz Messias Correa Ruela, Alícia Maria Zanette Moreira, Isabella Vaillant, Alda Torres**

[1] Instituto Federal do Espírito Santo (IFES)
Laboratório de Inteligência Computacional na Saúde (LICS)
Cachoeiro de Itapemirim – ES – Brasil
cristianos@ifes.edu.br

Abstract. *Recent advances in Large Language Models (LLMs) have expanded the possibilities for processing clinical narratives and supporting healthcare workflows. However, most existing approaches address tasks such as triage, information extraction, or semantic standardization independently. This work proposes a pipeline based on LLMs to support clinical triage and structured extraction of comorbidities from textual medical reports. The architecture integrates automatic risk classification, identification of comorbid conditions, and mapping of extracted entities to ICD-10 codes using a retrieval-augmented mechanism. Experiments were conducted using clinical case narratives collected from biomedical literature. The results indicate that the proposed approach can transform unstructured clinical descriptions into structured data representations. By generating outputs compatible with standardized medical terminologies, the pipeline contributes to future interoperability with national health data infrastructures such as the Brazilian National Health Data Network (RNDS).*

Resumo. *Avanços recentes em modelos de linguagem de grande porte (LLMs) ampliaram as possibilidades de processamento de narrativas clínicas e apoio a processos assistenciais em saúde. Entretanto, a maioria das abordagens existentes trata tarefas como triagem clínica, extração de informações ou padronização semântica de forma isolada. Este trabalho propõe um pipeline baseado em LLMs para apoiar a triagem clínica e a extração estruturada de comorbidades a partir de relatos médicos textuais. A arquitetura integra classificação automática de risco, identificação de condições clínicas e mapeamento para códigos da Classificação Internacional de Doenças (CID-10) por meio de um mecanismo baseado em recuperação semântica. Experimentos foram conduzidos com narrativas clínicas provenientes da literatura biomédica. Os resultados indicam que a abordagem proposta é capaz de transformar descrições médicas não estruturadas em representações estruturadas de dados clínicos, favorecendo sua futura interoperabilidade com infraestruturas nacionais de dados em saúde, como a Rede Nacional de Dados em Saúde (RNDS).*

1. Introdução

A crescente digitalização dos serviços de saúde tem ampliado o volume de informações clínicas registradas em formato textual, como descrições de sintomas, relatos médicos e registros de atendimento. Embora esses dados contenham informações valiosas para suporte à decisão clínica, sua natureza não estruturada dificulta a integração com sistemas de informação em saúde e limita sua reutilização em processos assistenciais e analíticos.

Nesse contexto, modelos de linguagem de grande porte (*Large Language Models* - LLMs) têm demonstrado grande potencial para interpretar narrativas clínicas e extrair informações relevantes a partir de textos médicos. Estudos recentes indicam que esses modelos podem apoiar tarefas como classificação de risco, identificação de diagnósticos e extração de entidades clínicas em registros médicos eletrônicos.

Paralelamente, a interoperabilidade entre sistemas de informação em saúde tornou-se um requisito fundamental para a organização de ecossistemas digitais de dados clínicos. No Brasil, a Rede Nacional de Dados em Saúde (RNDS) [Ministério da Saúde 2023] busca viabilizar o compartilhamento seguro e padronizado de informações entre diferentes sistemas assistenciais. Para que isso seja possível, é necessário transformar informações clínicas originalmente registradas em linguagem natural em dados estruturados e codificados segundo terminologias padronizadas, como a Classificação Internacional de Doenças (CID-10).

Apesar dos avanços recentes no uso de LLMs na área médica, grande parte das pesquisas aborda tarefas como triagem clínica, extração de informações ou padronização terminológica de forma isolada. Ainda são escassos os trabalhos que integram essas etapas em um pipeline único capaz de transformar narrativas clínicas em dados estruturados interoperáveis.

Diante desse cenário, este trabalho propõe um pipeline baseado em modelos de linguagem de grande porte para apoiar processos de triagem clínica e extração estruturada de comorbidades a partir de relatos médicos textuais. A arquitetura proposta combina classificação automática de risco, identificação de comorbidades e mapeamento para códigos CID-10, produzindo saídas estruturadas compatíveis com infraestruturas nacionais de interoperabilidade em saúde.

As principais contribuições deste trabalho são:

- a proposição de um pipeline integrado baseado em LLMs para triagem clínica e extração de comorbidades a partir de narrativas médicas;
- a utilização de um mecanismo baseado em *Retrieval-Augmented Generation* (RAG) para mapear condições clínicas extraídas para códigos CID-10;
- a geração de dados estruturados semanticamente compatíveis com os princípios de interoperabilidade adotados pela RNDS.

2. Trabalhos Relacionados

Avanços recentes nas pesquisas sobre LLMs têm ampliado as possibilidades de aplicação do processamento de linguagem natural em saúde, especialmente em tarefas de raciocínio clínico, triagem e extração de informações a partir de registros textuais. Nesse contexto, Singhal et al. [Singhal et al. 2023] apresentaram o Med-PaLM, demonstrando que LLMs

podem alcançar desempenho competitivo em tarefas de raciocínio médico e apoio à decisão clínica.

Além do raciocínio clínico, a literatura recente também tem explorado o uso de LLMs para extração estruturada de informações clínicas. Guevara et al. [Guevara et al. 2024] mostraram o potencial desses modelos para identificar automaticamente informações relevantes em registros eletrônicos de saúde não estruturados, reforçando sua utilidade em tarefas de estruturação de dados clínicos. De modo semelhante, Martinez et al. [Martinez et al. 2024] investigaram pipelines de extração de entidades em textos médicos, evidenciando a relevância de abordagens automatizadas para transformar narrativas clínicas em dados estruturados.

No contexto específico da triagem, Xu et al. [Xu et al. 2025] avaliaram o desempenho de LLMs em tarefas de diagnóstico e classificação de risco a partir de vinhetas clínicas curtas, observando bom desempenho diagnóstico, mas também limitações relevantes na triagem, com tendência a *over-triage*. Em linha semelhante, Gaber et al. [Gaber et al. 2025] exploraram fluxos baseados em LLMs e recuperação de contexto para apoio à decisão clínica, mostrando que informações adicionais podem melhorar a classificação de urgência em cenários clínicos.

Abordagens baseadas em *Retrieval-Augmented Generation* (RAG) também têm sido investigadas para apoiar o raciocínio clínico e a interpretação de sintomas. Zhang et al. [Zhang et al. 2024] mostraram que a combinação entre recuperação semântica e modelos de linguagem pode melhorar a análise de casos em contextos de emergência. Já Singh et al. [Singh et al. 2026] investigaram o uso de LLMs para classificação de comorbidades em textos clínicos livres, apontando o potencial dessas arquiteturas para tarefas de mineração de informação médica.

Em conjunto, esses estudos evidenciam avanços importantes em triagem clínica, extração de informações e classificação de comorbidades com LLMs. Entretanto, tais tarefas são frequentemente tratadas de forma isolada. Assim, permanece uma lacuna quanto a arquiteturas capazes de integrar, em um único pipeline, a pré-classificação de risco, a extração estruturada de comorbidades e o mapeamento para terminologias padronizadas, com vistas à interoperabilidade em saúde. Este trabalho busca contribuir nesse ponto ao propor um pipeline unificado baseado em LLMs para triagem clínica, identificação de comorbidades e geração de dados semanticamente estruturados.

A arquitetura proposta neste trabalho fundamenta-se diretamente nos avanços reportados por esses estudos, buscando suprir a lacuna identificada quanto à integração das etapas em um pipeline unificado.

3. Arquitetura Proposta

A literatura recente evidencia avanços importantes no uso de LLMs para tarefas como triagem clínica, extração de informações médicas e classificação de diagnósticos. Entretanto, a maioria das abordagens trata essas tarefas de forma isolada, sem integrar os diferentes estágios necessários para transformar narrativas clínicas em dados estruturados e interoperáveis.

Este trabalho propõe um pipeline baseado em LLMs capaz de integrar três etapas fundamentais: (i) pré-classificação automática de risco clínico, (ii) identificação de

comorbidades a partir de narrativas médicas e (iii) mapeamento dessas condições para códigos padronizados da Classificação Internacional de Doenças (CID-10).

O objetivo deste trabalho não é propor um sistema clínico operacional pronto para implantação imediata, mas investigar a viabilidade de uma arquitetura baseada em LLMs para estruturar narrativas clínicas e apoiar etapas preliminares de triagem. Nesse sentido, o pipeline proposto deve ser interpretado como uma prova de conceito voltada à validação da arquitetura e de seu potencial para aplicações futuras em ambientes clínicos reais.

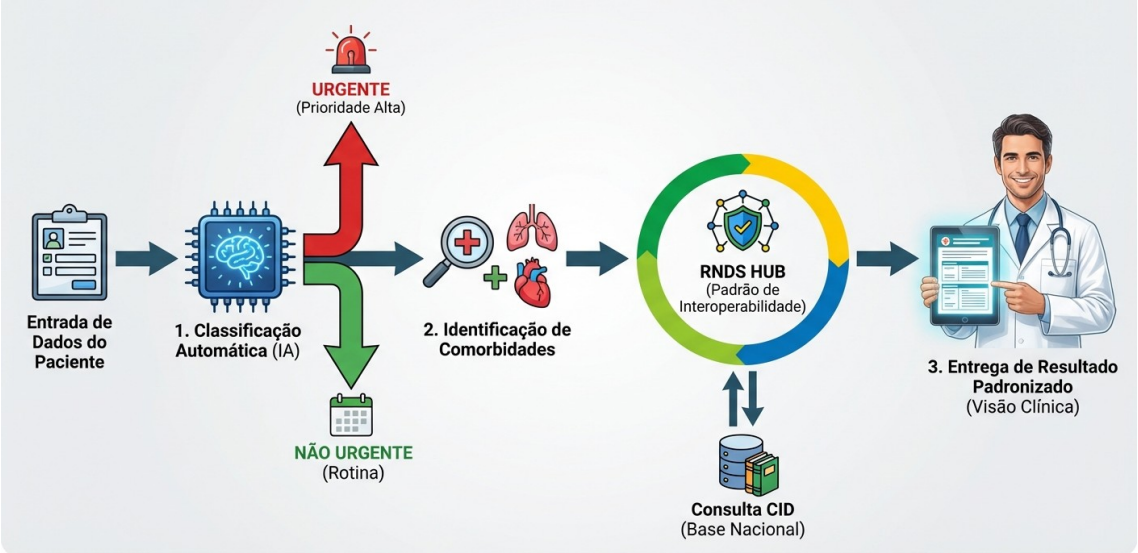

Figura 1. Arquitetura do pipeline proposto.

A Figura 1 apresenta a visão geral da arquitetura proposta. O fluxo inicia com o processamento de narrativas clínicas textuais, a partir das quais o sistema realiza uma classificação preliminar de risco e identifica possíveis comorbidades mencionadas no relato. Em seguida, as condições clínicas extraídas são mapeadas para códigos CID-10 por meio de um mecanismo baseado em recuperação semântica e modelos de linguagem.

Ao integrar essas etapas em um único fluxo, o pipeline permite transformar descrições clínicas não estruturadas em representações estruturadas de dados médicos, criando condições para sua futura integração com infraestruturas de interoperabilidade em saúde, como a RNDS.

4. Materiais e métodos

A construção do dataset, a integração das fontes de dados e o cálculo das métricas experimentais foram implementados em Python 3.12. Para o processamento e organização dos dados foi utilizada a biblioteca *pandas*, enquanto as métricas de avaliação foram calculadas com o auxílio do *scikit-learn*. Adicionalmente, foram empregadas as bibliotecas *sentence-transformers*, voltada à geração de embeddings semânticos, e *rouge-score*, utilizada para medir similaridade textual entre representações geradas pelos modelos, permitindo comparar diagnósticos e descrições clínicas extraídas a partir das narrativas médicas.

4.1. Construção e Validação do Dataset

Para a realização dos experimentos, foi construído um dataset unificado a partir de relatos de casos clínicos reais, coletados de duas bases de literatura biomédica de referência: o

PubMed [Canese and Weis 2013], principal repositório mundial de literatura biomédica, e o SciELO [Packer et al. 2018], plataforma de acesso aberto com ampla cobertura de periódicos científicos latino-americanos. O critério de seleção restringiu-se a *Clinical Case Reports* completos, resultando em 2.843 registros brutos, distribuídos da seguinte forma: 2.500 relatos provenientes do PubMed (88%, majoritariamente em inglês) e 343 relatos provenientes do SciELO (12%, em português). O arquivo bruto inicial continha três colunas: `url` (fonte única de identificação), `titulo` e `texto` (resumo médico original).

4.1.1. Estrutura de Extração via LLMs

Para transformar o texto não estruturado dos relatos clínicos em dados tabulares compatíveis com o padrão da RNDS, foram empregadas duas arquiteturas distintas de LLMs, atuando exclusivamente como extratores e estruturadores do conteúdo clínico original:

- Google Gemini 2.5 Flash (via API): modelo proprietário de alta escala, com foco em compreensão generalista e elevada sensibilidade semântica.
- Med-Gemma 1.5 4B-IT (execução local): modelo especializado no domínio médico, executado em ambiente local em uma workstation equipada com GPU NVIDIA RTX 5070 (12GB VRAM), garantindo privacidade e soberania dos dados clínicos processados nas etapas a ele atribuídas. Cabe destacar que, na configuração atual do pipeline, o Google Gemini 2.5 Flash é utilizado via API externa para determinados campos, o que implica que o pipeline completo, nesta fase experimental, não atende integralmente aos requisitos de conformidade com a LGPD e com os princípios da RNDS. A substituição do Gemini por um modelo local está prevista como etapa futura para viabilizar a conformidade plena.

O mecanismo RAG para mapeamento CID-10 utilizou o modelo MedEmbed (abhinand/MedEmbed-large-v0.1, 1024d), selecionado após avaliação comparativa com BioBERT-PT, PubMedBERT e abordagem híbrida *BiEncoder+CrossEncoder*. O banco de busca indexa cada código CID-10 com sua descrição completa como documento individual. Para cada comorbidade extraída, o sistema gera um *embedding* com prefixo de instrução e recupera o código mais similar por similaridade cosseno. O MedEmbed alcançou acurácia de 90,07% (IC 95%) na validação contra gabarito anotado manualmente.

Cada relato clínico foi processado por ambos os modelos, que extraíram e estruturaram os seguintes campos em formato JSON:

- `resumo_caso`: tradução e síntese dos sintomas em Português (PT-BR);
- `diagnostico_final`: identificação da patologia principal;
- `cid_10`: codificação internacional da doença, campo central para interoperabilidade com a RNDS;
- `comorbidades_identificadas`: lista de condições secundárias presentes no histórico clínico;
- `risco_estimado`: classificação de gravidade (Baixo, Médio, Alto, Emergência).

A fusão dos datasets seguiu uma estratégia de prioridade por campo, fundamentada em análise comparativa prévia entre os modelos. Os campos `resumo_caso` e `titulo_original` foram extraídos preferencialmente do Gemini, por sua maior fluência em português brasileiro. O campo `cid_10` e o campo `risco_estimado` foram atribuídos ao MedGemma, que apresentou maior precisão sintática e saídas mais padronizadas, respectivamente. Para o campo `diagnostico_final`, adotou-se estratégia de consenso por similaridade semântica: em caso de concordância, manteve-se a formulação do Gemini; em caso de divergência, ambas as formulações foram preservadas com identificação de origem e sinalizadas para revisão posterior.

4.1.2. Desafios de Alinhamento e Integração

Ao unificar os 2.843 registros foram identificados desafios críticos de alinhamento:

- Granularidade diagnóstica: o MedGemma tendeu a ser mais técnico e específico na codificação CID-10, enquanto o Gemini produziu descrições mais abrangentes;
- Barreira linguística: a validação cruzada enfrentou o desafio de comparar títulos em inglês com extrações em português;
- Baixa concordância exata: apenas 7,89% dos casos apresentaram concordância caractere-por-caractere no código CID-10, evidenciando que modelos distintos interpretam níveis de especificidade médica de formas diferentes, achado que motivou a adoção de métricas semânticas para validação.

4.1.3. Metodologia de Validação Estatística (*Silver Standard*)

Na ausência de especialistas humanos para revisão dos 2.843 casos, opção deliberada de escopo desta etapa, prevista como trabalho futuro, foi implementado um *framework* de validação multicamada para aferir o grau de confiabilidade do dataset, composto por três mecanismos complementares:

1. Validação sintática (Regex): verificação se o código CID-10 gerado segue a norma internacional (padrão Letra + 2 ou 3 dígitos numéricos), filtrando saídas malformadas antes de qualquer análise semântica.
2. Filtro NegEx (negação contextual): aplicação de expressões regulares baseadas no algoritmo NegEx para identificar e descartar condições mencionadas de forma negativa nas narrativas clínicas (e.g., “nega diabetes”, “sem histórico de hipertensão”), prevenindo a extração incorreta de comorbidades que o paciente não apresenta.
3. Triangulação semântica: cruzamento de radicais linguísticos entre o título original em inglês e o diagnóstico extraído em português, com dicionários expandidos por famílias de doenças (Cardio, Endo, Neuro, entre outros), para verificar coerência semântica entre as extrações dos dois modelos.

4.1.4. Resultados Comparativos e Seleção da Base de Avaliação

Os resultados da avaliação comparativa entre os dois modelos revelaram um *trade-off* estratégico relevante, sintetizado na Tabela 1.

Tabela 1. Métricas comparativas de extração entre Gemini e Med-Gemma. Fonte: os autores.

Métrica	Gemini (Generalista)	MedGemma (Especializado)
Precisão sintática CID-10	26,45%	78,29%
Recall	40,38%	12,96%
Completude de dados	57,82%	81,25%
Fidelidade (não-alucinação)	92,48%	93,33%

Os resultados evidenciam perfis complementares: o MedGemma apresenta superioridade na precisão sintática do CID-10 (78,29% vs. 26,45%) e na completude dos campos estruturados, sendo mais adequado para garantir interoperabilidade com a RNDS. O Gemini, por sua vez, apresenta recall significativamente superior (40,38% vs. 12,96%), demonstrando maior sensibilidade na detecção de comorbidades mencionadas de forma sutil nas narrativas clínicas.

A priorização do recall neste contexto é clinicamente justificada: em sistemas de triagem, o custo de um falso negativo, ignorar uma comorbidade real, é substancialmente mais grave do que o de um falso positivo, que seria revisado pelo profissional de saúde. Essa complementaridade fundamenta a estratégia de *ensemble* adotada no pipeline: unir a sensibilidade do Gemini ao rigor técnico do MedGemma para maximizar o F1-score do sistema integrado.

A partir da aplicação do *framework* de validação multicamada, foi identificado um subconjunto de 789 casos (28,17%) com validação tripla simultânea (consenso entre modelos + coerência com título + CID-10 sintaticamente válido), constituindo a base mais confiável do dataset para avaliação do pipeline proposto. Desse subconjunto, 429 casos apresentaram disponibilidade simultânea de todos os campos de referência necessários para os experimentos, incluindo `risco_estimado` e `diagnostico_final` provenientes de ambos os modelos extratores, sendo utilizados integralmente (sem amostragem adicional) como base de avaliação nas Fases 1 e 2 deste trabalho. A validação por especialistas humanos sobre o subconjunto de 789 casos está prevista como etapa subsequente desta pesquisa.

Tabela 2. Desempenho dos modelos na extração de comorbidades.

Modelo	Cos. Semântico	Taxa Detecção	Média Comorbidades
Claude Sonnet 4	**0,428**	**1,000**	**2,8**
MedGemma 1.5 4B	0,356	**1,000**	2,5
Sabiazinho-4	0,355	**1,000**	2,1
Sabiá-4	0,350	**1,000**	2,1
Gemini 2.5 Flash	0,345	**1,000**	1,8

Os resultados do Experimento 2, apresentados na Tabela 2, revelam um padrão distinto em relação ao Experimento 1: todos os modelos detectaram comorbidades em 100% dos casos avaliados (Taxa de Detecção=1,000), demonstrando robustez uniforme na tarefa de identificação. A diferenciação entre os modelos se manifesta na qualidade semântica das extrações, medida pela similaridade cosseno entre os termos extraídos e o diagnóstico de referência.

O Claude Sonnet 4 obteve o melhor desempenho semântico (Cos. Semântico=0,428; Média=2,8 comorbidades por caso), resultado que contrasta com sua posição intermediária no Experimento 1. Esse comportamento sugere que a capacidade de raciocínio contextual e fluência linguística do modelo se traduz em extrações semanticamente mais precisas em tarefas abertas, mesmo que seu desempenho em classificação categórica seja inferior ao de modelos especializados. A análise conjunta dos dois experimentos aponta para uma complementaridade entre os modelos: o MedGemma 1.5 4B, melhor classificador de risco no Experimento 1 (Acurácia=0,476; Kappa=0,248), ocupa a segunda posição no Experimento 2 (Cos.=0,356), enquanto o Claude lidera a extração semântica mas fica em quarto lugar na classificação. Essa inversão de ranking fundamenta a adoção de uma estratégia de *ensemble* no pipeline TRIA, combinando MedGemma para triagem de risco e Claude Sonnet 4 para extração de comorbidades, de forma a maximizar o desempenho em ambas as tarefas.

O Gemini 2.5 Flash registrou o menor desempenho semântico (Cos.=0,345) e a menor média de comorbidades extraídas por caso (1,8), resultado que reforça a ausência de viés favorável ao modelo gerador do dataset: mesmo tendo sido responsável pela extração dos diagnósticos de referência utilizados na avaliação, o Gemini 2.5 Flash não obteve vantagem sobre os demais modelos, sendo, inclusive, o menos preciso semanticamente no Experimento 2. Esse achado fortalece a validade do *silver standard* como referência de avaliação independente.

Por fim, os modelos Sabiá-4 e Sabiazinho-4 apresentaram desempenhos próximos entre si (Cos.=0,350 e 0,355, respectivamente), com média de 2,1 comorbidades por caso. Considerando que ambos são modelos de propósito geral otimizados para o português brasileiro, o resultado é competitivo, especialmente para o Sabiazinho-4, que possui menor porte e maior velocidade de inferência, podendo ser viável em cenários de triagem com restrições computacionais.

A alta sensibilidade aliada à baixa precisão na classe Alto motivou uma análise exploratória dos casos de falha. A inspeção de falsos positivos sugere que os modelos tendem a superestimar a gravidade em narrativas com terminologia clínica densa, comportamento possivelmente amplificado pela prevalência de casos graves no dataset. Investigações futuras incluirão análise sistemática desses padrões para identificar construções linguísticas que contribuam para o *over-triage*.

5. Resultados e Discussão

Para avaliar o pipeline proposto, foram conduzidos dois experimentos distintos. O Experimento 1 avaliou o desempenho dos modelos na tarefa de classificação automática de risco clínico, utilizando as 429 narrativas do subconjunto validado como base de avaliação, com as classes de risco extraídas pelo pipeline como referência. O Experimento 2 avaliou a extração de comorbidades, avaliando a qualidade semântica dos termos extraídos por cada modelo em relação ao diagnóstico de referência, por meio de similaridade cosseno entre *embeddings*.

Os resultados do Experimento 1, sintetizados nas Tabelas 3 e 4, revelam um padrão consistente entre os cinco modelos avaliados: desempenho global moderado, com acurácias entre 34,7% e 47,6% e coeficientes Kappa entre 0,115 e 0,248, situando todos os modelos na faixa de concordância *fraca* segundo a escala de Landis e

Koch [Landis and Koch 1977]. Esse resultado é esperado dado o caráter desafiador da tarefa, a classificação de risco clínico a partir de narrativas textuais sem acesso a sinais vitais ou dados estruturados do paciente, e está alinhado com achados reportados na literatura [Xu et al. 2025, Gaber et al. 2025].

O MedGemma 1.5 4B obteve o melhor desempenho em todas as métricas globais (Acurácia=0,476; Kappa=0,248; F1-macro=0,442; F1-weighted=0,542), resultado atribuível ao seu treinamento especializado no domínio médico. Seu diferencial mais expressivo reside na classe *Médio* (F1=0,475; P=0,649), a mais difícil do conjunto, onde os demais modelos apresentaram desempenho substancialmente inferior. A capacidade de distinguir casos de urgência intermediária é clinicamente relevante, pois erros nessa classe impactam diretamente a priorização do fluxo hospitalar.

O Sabiá-4 ocupou a segunda posição (Acurácia=0,444; F1-weighted=0,500), resultado expressivo considerando que se trata de um modelo de propósito geral otimizado para o português brasileiro, avaliado sobre um dataset majoritariamente traduzido do inglês. O modelo apresentou a maior precisão na classe *Baixo* entre todos os modelos avaliados (P=0,841), indicando alta confiabilidade ao classificar casos de menor urgência, comportamento relevante para reduzir sobrecarga desnecessária em serviços de urgência do SUS.

O Gemini 2.5 Flash ficou em terceiro lugar (Acurácia=0,420; Kappa=0,169), resultado que merece atenção especial do ponto de vista metodológico: trata-se do mesmo modelo utilizado na construção do *silver standard* do dataset. O fato de não ter se destacado entre os demais avaliados, e de ter ficado abaixo de modelos menores como o MedGemma 1.5 4B e o Sabiá-4, constitui evidência relevante de que o processo de anotação automática não introduziu viés sistemático favorável ao modelo gerador, reforçando a validade do dataset como referência de avaliação independente [Gaber et al. 2025]. Adicionalmente, o Gemini 2.5 Flash apresentou o maior *recall* na classe *Alto* entre todos os modelos (R=0,833), comportamento que, aliado à sua precisão modesta nessa classe (P=0,149), segue o padrão de *over-triage* descrito adiante.

O Claude Sonnet 4 apresentou o maior *recall* na classe *Baixo* entre todos os modelos (R=0,596; F1=0,670), além de destacar-se pelo *recall* de 0,800 na classe *Alto*. Embora sua precisão nessa classe seja baixa (P=0,127), esse comportamento é clinicamente defensável em sistemas de triagem preliminar, onde o custo de um falso negativo, deixar de identificar um caso urgente, é substancialmente maior que o de um falso positivo.

O Sabiazinho-4, modelo de menor porte e maior velocidade da família Sabiá, obteve o menor desempenho geral (Acurácia=0,347; F1-macro=0,337), como esperado para sua escala. Ainda assim, registrou o maior *recall* na classe *Alto* junto ao Gemini 2.5 Flash (R=0,833), sugerindo utilidade como camada inicial de triagem rápida para identificação de casos críticos, com posterior refinamento por modelos de maior capacidade.

Um achado transversal de especial relevância é a tendência de *over-triage* na classe *Alto*, observada em todos os cinco modelos: os valores de *recall* são consistentemente elevados (0,700–0,833) enquanto as precisões permanecem baixas (0,120–0,149). Esse padrão não é específico de nenhuma arquitetura, família ou origem linguística dos modelos avaliados, está presente tanto em modelos especializados (MedGemma) quanto em modelos de propósito geral (Sabiá-4, Claude Sonnet 4, Sabiazinho-4, Ge-

mini 2.5 Flash), indicando um comportamento sistêmico dos LLMs frente à ambiguidade clínica, coerente com protocolos de triagem que priorizam a segurança do paciente [Xu et al. 2025]. A distinção entre as classes *Alto* e *Médio* representa o principal desafio do pipeline, motivando a investigação de estratégias de *ensemble* como trabalho futuro, combinando a sensibilidade dos modelos na detecção de casos críticos com a precisão do MedGemma na classificação intermediária..

Tabela 3. Desempenho dos modelos na classificação automática de risco clínico.

Modelo	Acurácia	Kappa	F1-macro	F1-weighted
MedGemma 1.5 4B	**0,476**	**0,248**	**0,442**	**0,542**
Sabiá-4	0,444	0,198	0,418	0,500
Gemini 2.5 Flash	0,420	0,169	0,394	0,479
Claude Sonnet 4	0,408	0,166	0,363	0,444
Sabiazinho-4	0,347	0,115	0,337	0,404

Tabela 4. Precisão (P), Revocação (R) e F1-score por classe de risco.

Modelo	Classe	P	R	F1
MedGemma 1.5 4B	Alto	0,123	0,700	0,209
	Médio	0,649	0,375	**0,475**
	Baixo	0,745	0,563	0,642
Sabiá-4	Alto	0,146	0,793	0,247
	Médio	0,401	0,329	0,362
	Baixo	**0,841**	0,521	0,643
Gemini 2.5 Flash	Alto	0,149	0,833	0,253
	Médio	0,374	0,256	0,304
	Baixo	0,767	0,526	0,624
Claude Sonnet 4	Alto	0,127	**0,800**	0,219
	Médio	0,324	0,143	0,198
	Baixo	0,765	**0,596**	**0,670**
Sabiazinho-4	Alto	0,120	**0,833**	0,210
	Médio	0,297	0,196	0,237
	Baixo	0,827	0,427	0,563

5.1. Seleção e Análise de Desempenho de LLMs na Pré-classificação de Risco

A classificação de risco é uma das etapas mais cruciais em triagens clínicas, definida como um processo dinâmico de identificação e distribuição de usuários para o serviço ou ambiente de cuidado mais adequado em tempo oportuno [Sacoman et al. 2019]. Ela desempenha papel central na priorização de casos e na organização do fluxo hospitalar, contexto em que o Brasil enfrenta desafios significativos de superlotação nas Unidades de Pronto-Atendimento (UPAs) [Bittencourt and Hortale 2009].

A superlotação implica entraves como aumento do tempo de espera e sobrecarga do corpo clínico, comprometendo o desempenho assistencial e expondo o paciente a erros na classificação de risco [Bittencourt and Hortale 2009]. Um dos obstáculos que agravam

esse cenário é o atendimento por ordem de chegada antes da triagem clínica. A ineficiência nesse fluxo inicial impede que os casos mais urgentes sejam priorizados e que a equipe médica possa agir de forma estratégica para tornar o fluxo mais otimizado. Um dos principais obstáculos é o atendimento por ordem de chegada antes da triagem, impedindo a priorização dos casos mais urgentes. Para mitigar esse problema, propõe-se o sistema `TRIA (Triagem Rápida Inteligente e Automatizada)`, baseado em LLMs orientadas por protocolos reconhecidos como o de Manchester, responsável por receber os sintomas do paciente, gerar uma classificação preliminar e ordenar a fila em tempo real.

Por fim cabe ressaltar que TRIA é uma ferramenta de apoio à decisão clínica, não um diagnóstico ou função determinística, voltada para serviços de urgência do SUS. Para os experimentos, foram selecionadas duas LLMs com perfis complementares, uma clínica e uma geral, conforme critérios de desempenho e viabilidade operacional.

5.2. Ameaças à validade

Esta pesquisa apresenta limitações que devem ser consideradas na interpretação dos resultados. A avaliação foi conduzida com base em um padrão de referência automatizado, sem validação por especialistas clínicos humanos, etapa prevista como trabalho futuro. O dataset, construído a partir de literatura biomédica, pode não refletir integralmente a variabilidade linguística de prontuários reais. Adicionalmente, a triagem foi modelada como classificação textual, sem acesso a sinais vitais ou histórico clínico completo, o que limita a comparação direta com sistemas de triagem reais. Por fim, diferentes modelos, versões ou estratégias de prompt podem produzir variações nos resultados.

6. Conclusão

Este trabalho apresentou um pipeline baseado em LLMs para triagem clínica e a extração estruturada de comorbidades a partir de narrativas médicas, integrando classificação automática de risco, identificação de comorbidades e mapeamento para códigos CID-10 compatível com a RNDS. Os experimentos demonstram o potencial de LLMs para triagem e mineração de informações clínicas a partir de registros textuais, combinando recuperação semântica e validação hierárquica de códigos CID-10 para garantir consistência na estruturação dos dados.

Como principal contribuição, o pipeline integra triagem clínica, extração de comorbidades e interoperabilidade com padrões nacionais de saúde digital, evidenciando o potencial de LLMs para apoiar fluxos clínicos e a padronização de dados hospitalares. Entre as limitações, destacam-se a ausência de validação por especialistas clínicos e a dependência de dados provenientes de literatura médica. A avaliação por profissionais de saúde é etapa indispensável antes de qualquer aplicação clínica real. Trabalhos futuros incluem validação com dados reais e expansão do pipeline para padrões internacionais como HL7 FHIR.

7. Declaração do Uso de ferramentas IA

Os autores utilizaram o ChatGPT e o Gemini para auxílio na revisão gramatical, estruturação estilística e organização textual deste artigo[1]. O conteúdo técnico, a

[1]Código de Conduta para Autores em Publicações da Sociedade Brasileira de Computação (SBC). Disponível em: `https://sol.sbc.org.br/index.php/indice/conduta`

condução dos experimentos, a coleta de dados e a interpretação dos resultados foram realizados e verificados integralmente pelos autores humanos, que assumem total responsabilidade pela veracidade e integridade das informações apresentadas.

Referências

Bittencourt, R. J. and Hortale, V. A. (2009). Intervenções para solucionar a superlotação nos serviços de emergência hospitalar: uma revisão sistemática. *Cadernos de Saúde Pública*, 25(7):1439–1454.

Canese, K. and Weis, S. (2013). Pubmed: the bibliographic database. *The NCBI Handbook*.

Gaber, F., Shaik, M., Franke, V., and Akalin, A. (2025). Evaluating large language model workflows in clinical decision support: Referral, triage, and diagnosis. *Preprint*. Berlin Institute for Medical Systems Biology (BIMSB), Max Delbrück Center for Molecular Medicine.

Guevara, E., Chen, Y., et al. (2024). Extracting social determinants of health from electronic health records using large language models. *npj Digital Medicine*.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Martinez, D., Silva, R., and Thompson, P. (2024). An entity extraction pipeline for medical text records. *Journal of Medical Internet Research*, 26.

Ministério da Saúde (2023). Rede nacional de dados em saúde (rnds). Acessado em: 9 mar. 2026.

Packer, A. L., Cop, N., Luccisano, A., Ramalho, A., and Spinak, E. (2018). Scielo—20 years of open access: an analytic study of open access and scholarly communication. *Learned Publishing*, 31(3):205–213.

Sacoman, T. M., Beltrammi, D. G. M., Andreazza, R., Cecílio, L. C. d. O., and Chioro dos Reis, A. A. (2019). Implantação do sistema de classificação de risco manchester em uma rede municipal de urgência. *Saúde em Debate*, 43(121):354–367.

Singh, R., Patel, A., and Kumar, S. (2026). Comorbidity classification from clinical free-text using large language models. *Journal of Medical Systems*.

Singhal, K., Azizi, S., Tu, T., et al. (2023). Towards expert-level medical question answering with large language models. *Nature*.

Xu, L., Zhao, W., and Huang, X. (2025). Diagnosis and triage performance of contemporary large language models on short clinical vignettes. *Journal of Medical Systems*, 49(141).

Zhang, Y., Li, H., Chen, X., and Wang, J. (2024). Multi-evidence clinical reasoning with retrieval-augmented generation for emergency triage. *Artificial Intelligence in Medicine*.