UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
CURSO DE CIÊNCIA DA COMPUTAÇÃO

Pedro Nack Martins

Extração de Informação de Notas Fiscais apoiada em Conhecimento

Florianópolis
2025

Pedro Nack Martins

Extração de Informação de Notas Fiscais apoiada em Conhecimento

Trabalho de Conclusão de Curso submetido ao Curso de Ciência da Computação do Departamento de Informática e Estatística da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.
Orientador: Prof. Renato Fileto, Dr.

Florianópolis
2025

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Martins, Pedro Nack
Extração de Informação de Notas Fiscais apoiada em Conhecimento / Pedro Nack Martins ; orientador, Renato Fileto, 2025.
39 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Ciências da Computação, Florianópolis, 2025.

Inclui referências.

1. Ciências da Computação. 2. Processamento de Linguagem Natural. 3. Grandes Modelos de Linguagem. 4. Extração de Informação. 5. Ligação de Entidades. I. Fileto, Renato. II. Universidade Federal de Santa Catarina. Graduação em Ciências da Computação. III. Título.

Pedro Nack Martins

Extração de Informação de Notas Fiscais apoiada em Conhecimento

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do título de Bacharel em Ciência da Computação e aprovado em sua forma final pelo Curso de Ciência da Computação.

Florianópolis, 10 de Novembro de 2025.

Prof. Álvaro Junio Pereira Franco, Dr.
Coordenador do Curso

Banca Examinadora:

Prof. Renato Fileto, Dr.
Orientador
Universidade Federal de Santa Catarina

Prof. Carina Friedrich Dorneles, Dra.
Avaliadora
Universidade Federal de Santa Catarina

Prof. Osmar de Oliveira Braz Junior, Dr.
Avaliador
Universidade Federal de Santa Catarina

Florianópolis, 2025.

Dedico este trabalho a meus pais e amigos.

AGRADECIMENTOS

Agradeço aos meus pais, que apoiaram e incentivaram meus estudos desde pequeno, e aos meus amigos que me acompanharam durante toda a jornada acadêmica.

RESUMO

A descrição de produto em uma nota fiscal é um campo de texto não estruturado e não padronizado, dificultando a identifição do produto e suas caracterísicass, e consequentemente a análise e a comparação de informações como preço dos mesmos produtos em diferentes notas fiscais. A extração da informação contida nessas descrições e a identificação da correspondência do produto descrito com produtos cadastrados em bases de dados oficiais promove a acessibilidade e possibilita a exploração dessa informação. A presente monografia explora o uso de técnicas de Processamento de Linguagem Natural (PLN), desde expressões regulares até Grandes Modelos de Linguagem (LLMs), para a extração da informação e ligação de entidades de notas fiscais eletrônicas (NF-es), com informações armazenadas em um Grafo de Conhecimento (KG). O estudo será conduzido inicialmente no domínio de medicamentos, para o qual foi construído KG com dados sobre medicamentos aprovados pela Anvisa, seus princípios ativos, dosagens, apresentações e outras informações.

Palavras-chave: Grandes Modelos de Linguagem. Inteligência Artificial. Extração de informação. Ligação de entidades.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Na era contemporânea, o vasto volume de dados disponíveis em formato digital representa um potencial imensurável para análises estatísticas e investigações. Contudo, o crescimento acelerado da quantidade e da variedade desses dados, além de sua complexidade, traz grandes desafios. Como integrar e processar eficientemente uma quantidade gigantesca de dados, frequentemente semi-desestruturados ou não estruturados como textos livres como é o caso da descrição de produto em uma nota fiscal eletrônica (NFe) para extrair análises valiosas e significativas?

Nesse cenário, a extração de informação de textos e sua integração com bases de conhecimento pré-existentes se apresenta como uma parte crucial do processamento e análise de dados. No contexto de notas fiscais, vendedores costumam preencher as descrições textuais de produtos em itens de NF-es sem padronização, gerando problemas no processo de estudo desses dados em grandes quantidades, e dificultando a união desses dados com informações de outras fontes.

A utilização de grandes modelos de linguagem (do inglês *Large Language Models* - LLMs) é uma abordagem recente, porém muito poderosa, para realizar tarefas de Processamento de Linguagem Natural. Entretanto, como apontado por (Pan et al., 2024), estes modelos apresentam algumas dificuldades, como a imprecisão dentro de domínios específicos, fuga de resposta e resultados errôneos sem fundamentação (alucinações).

Este projeto tem como objetivo utilizar LLMs para desambiguar e ligar entidades pertencentes a diferentes fontes de informação, identificando pares de itens correspondentes e permitindo a união destes dados desestruturados, gerando assim um aumento no potencial analítico desta informação. Ao final, espera-se que esse trabalho demonstre o potencial desta ferramenta em processos de integração e análise de dados textuais em grande escala, contribuindo para o avanço de soluções baseadas em inteligência artificial no campo de medicamentos.

1.1 DESCRIÇÃO DO PROBLEMA

Atualmente, as descrições de medicamentos em notas fiscais eletrônicas descrevem os produtos de forma não estruturada, com todas as suas características sendo apresentadas em uma única string, e sem um valor que identifique o produto descrito em uma base de dados de medicamentos registrados pelos órgãos responsáveis. Essa representação limita a organização e a execução de pesquisas com estes dados, que não são adequados para uso em bancos de dados relacionais devido à delimitação indeterminada entre os atributos dos registros, e dificulta a união dessa informação com dados oriundos de outras fontes.

Este trabalho propõe uma avaliação do uso de técnicas de PLN apoiadas em

bases de conhecimento para as funções de Reconhecimento de Entidades Nomeadas (do inglês, *Named Entity Recognition* - NER) e Ligação de Entidades (do inglês, *Entity Linking* - EL), em uma tentativa de ligar descrições presentes em notas fiscais com medicamentos registrados pela Anvisa e CMED, possibilitando assim a identificação precisa do medicamento correspondente a uma nota fiscal e suas características.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Realizar uma exploração e análise do uso de grandes modelos de linguagem para a ligação de entidades em uma base de conhecimento, dirigindo um estudo no campo de medicamentos.

1.2.2 Objetivos Específicos

1. Analisar o estado da arte em LLMs e ligação de entidades;
2. Identificar e minimizar o conjunto de registros candidatos à desambiguação;
3. Selecionar LLMs atuais e que possam ser executados localmente para serem usados nos experimentos;
4. Implementar a ligação de entidades com modelos de linguagem;
5. Avaliar o desempenho das estratégias empregadas;
6. Publicar os resultados do trabalho na forma de contribuições em artigos.

1.3 METODOLOGIA

Este trabalho foi desenvolvido como parte do Projeto Céos, parceria entre UFSC e MPSC. Os dados de notas fiscais foram fornecidos pelo MPSC, originados de licitações públicas, e foram selecionados aqueles que pertencem a medicamentos. Os dados sobre medicamentos foram adquiridos através de APIs da Anvisa e CMED.

Foi executada uma análise dos dados utilizando a linguagem Python para observação das características dos dados. A partir disso, são propostas soluções para os problemas encontrados, utilizando uma abordagem mista de técnicas de Processamento de Linguagem Natural tradicionais (expressões regulares) e modernas (grandes modelos de linguagem). Com isso, são executados experimentos sobre os dados e expostos os resultados obtidos.

1.4 ESTRUTURA DO TRABALHO

O restante deste trabalho está organizado da seguinte maneira. O Capítulo 2 esclarece as tecnologias utilizadas para o desenvolvimento da tese. O Capítulo 3 apre-

senta e compara trabalhos correlatos a este. O Capítulo 4 explica o processo proposto para a solução do problema. O Capítulo 5 define as variáveis que serão utilizadas para a realização dos experimentos. O Capítulo 6 descreve a realização dos experimentos propostos por esse trabalho e a avaliação dos resultados obtidos.

2 FUNDAMENTOS

Este capítulo apresenta os conceitos fundamentais ao entendimento do trabalho, bem como descreve as técnicas e ferramentas utilizadas para o desenvolvimento da ideia e estrutura proposta.

2.1 PROCESSAMENTO DE LINGUAGEM NATURAL

O Processamento de Linguagem Natural (PLN) é uma área da Inteligência Artificial que estuda a capacidade de um computador de interpretar texto e fala de maneira análoga à humana (IBM, 2025). O objetivo destes estudos é possibilitar que computadores possam realizar atividades de extração, análise e categorização de informação, bem como geração de texto em linguagem humana.

A área de PLN abrange uma grande variedade de processos, como extração de informação, tradução, reconhecimento de fala, etc.

Para que um texto possa ser mais facilmente analisado por máquina, é conveniente que seja feita uma padronização da informação. O processo de tokenização, portanto, descreve a etapa de divisão de um texto em partes menores denominadas *tokens*, que podem ser pequenas frases, palavras, subpalavras ou até mesmo letras (Jurafsky; Martin, 2025). Cada token agrega um valor semântico à sentença, que é utilizado pelo computador para interpretar o sentido presente na informação.

2.2 EXTRAÇÃO DE INFORMAÇÃO

A tokenização é base para diversas tarefas de PLN, incluindo tarefas de extração de informação. Extração de Informação (do inglês, Information Extraction - IE) se refere ao processo de extrair e estruturar dados de forma automatizada a partir de fontes não estruturadas (Jurafsky; Martin, 2025). Esse processo transforma páginas web, documentos ou outras representações de informação sem estrutura fixa em fontes valiosas de dados, que podem ser organizados para realização de consultas e análises.

Ao estruturar a informação, é facilitada a execução de buscas e análises sobre esses dados. Com isso, percebe-se a importância da IE, dado que a internet abriga inúmeras páginas web com inestimável volume de informação representada em linguagem humana.

Este trabalho lidará com dados textuais, em linguagem natural, presentes em descrições de notas fiscais. Como exemplo, temos a seguinte descrição de nota fiscal de um medicamento, para observação de como as informações podem ser apresentadas:

ARTRINID 100MG - CETOPROFENO-PO SOL INJ IV-50FA-UNIAO QUIMICA(POS) - LOTE:1806802 - Val:29/02/2020

As tarefas de extração de informação de texto empregadas neste trabalho e as características da descrição apresentada são exploradas nas seções subsequentes.

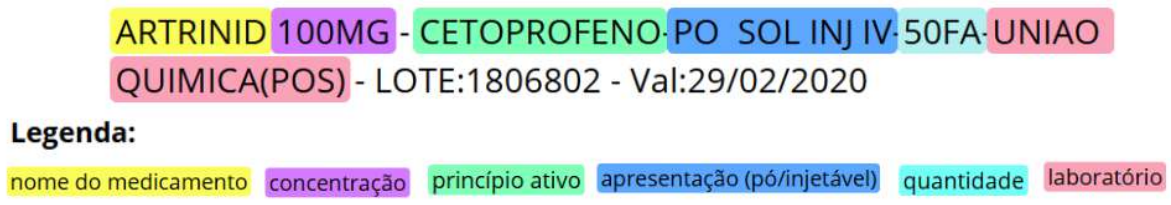
2.2.1 Reconhecimento de Entidades Nomeadas

Entidades nomeadas são qualquer objeto a qual se pode atribuir um nome próprio: pessoas, cidades, organizações, etc. O reconhecimento e categorização dessas entidades é importante para identificação de relacionamentos entre entidades (Ma; Hiraoka; Okazaki, 2022), solucionando problemas de desambiguação ou análises de sentimento.

O conceito de Entidade Nomeada, contudo, é comumente estendido para incluir objetos que podem não ser necessariamente entidades, como datas, porcentagens, quantidades de dinheiro, etc. (Jurafsky; Martin, 2025).

A Figura 1 apresenta um exemplo de extração de informações na descrição do medicamento apresentado na descrição apresentada anteriormente, categorizando os diferentes dados presentes. Na legenda, temos os diferentes tipos de entidades mais comumente presentes no conjunto de dados, como nome do medicamento, concentração e quantidade.

Figura 1 – Exemplo de NER em Descrição de Medicamento



Fonte: Elaborada pelo autor.

2.2.2 Desambiguação de Entidades

Em muitos casos, ao reconhecer as entidades de um texto, se torna interessante a incorporação de informação externa, por exemplo, presente em uma base de dados, para acrescentar semântica ao contexto de uma sentença. A recuperação desta informação pode se tornar difícil em cenários onde existe ambiguidade, como é o caso de palavras homônimas, ou contextos em que instâncias possuem definições muito similares, como na área biomédica (Yuan, H.; Yuan, Z.; Yu, 2022). Para isto, muitas dessas entidades precisam passar por um processo de desambiguação, orientado a definir a qual exato objeto de uma base de conhecimento aquela entidade se refere.

Ao efetuar uma busca por similaridade léxica utilizando a descrição demonstrada na Figura 1, com base no nome do medicamento e princípio ativo, em uma base de conhecimento, podemos alcançar os registros apresentados na Figura 2 como candidatos ao processo de desambiguação.

Figura 2 – Resultado da pesquisa pelo medicamento descrito

nome do medicamento	princípio ativo	laboratório	apresentação
CETOPROFENO	CETOPROFENO	CRISTALIA PRODUTOS QUIMICOS FARMACEUTICOS LTDA.	100 MG PO LIOF P/ SOL INJ CX 50 FA VD TRANS
ARTRINID	CETOPROFENO	UNIAO QUIMICA FARMACEUTICA NACIONAL S/A	50 MG/ML SOL INJ IM CT 50 AMP VD AMB X 2 ML
ARTRINID	CETOPROFENO	UNIAO QUIMICA FARMACEUTICA NACIONAL S/A,	100 MG PO LIOF SOL INJ IV CT 50 FA VD TRANS

Fonte: Elaborada pelo autor.

Ao analisar os candidatos, observa-se que o primeiro possui o mesmo princípio ativo do medicamento original, porém possui nome incorreto e laboratório diferente de União Química, que era previsto na descrição. Os medicamentos das linhas 2 e 3, contudo, apresentam nome de medicamento e laboratório corretos. Ao analisar a apresentação de cada um, nota-se que o medicamento 2 apresenta concentração de 50MG/ML, enquanto o medicamento 3 apresenta 100MG, correspondendo ao medicamento original. Ao final do processo, define-se então que o medicamento descrito trata-se do medicamento número 3.

2.2.3 Métodos para Extração de Informação

Uma das formas mais tradicionais e eficientes para extração de informação é através de expressões regulares (regex). Expressões regulares são definições de padrões para uma string que deve ser procurada dentro de um determinado texto. Por exemplo, se o objetivo fosse encontrar dentro da descrição de um produto o código de um item, poderia ser construída a expressão regular: "Código: [0-9]*".

A substring "Código: "define exatamente o que deve ser encontrado no texto para que haja correspondência com o que se procura. O trecho [0-9] define um intervalo, ou seja, qualquer valor numérico entre 0 e 9 seria aceito pela expressão regular, enquanto o símbolo de asterisco (*) define que é necessária a presença de pelo menos um valor do intervalo, mas que também deve ser aceita uma sequência destes números (ex: 25835).

Essa abordagem, portanto, se torna bastante altamente limitada quando o corpo de texto em que se busca a informação não apresenta padrões facilmente identificáveis. Neste cenário, o uso de LLMs, modelos de Machine Learning treinados em um vasto volume dados, representam alternativas mais flexíveis ao se desprender da rigidez das expressões regulares, utilizando técnicas de PLN para compreender contexto e variações linguísticas com precisão.

2.3 GRANDES MODELOS DE LINGUAGEM (LLM)

A ascensão dos Grandes Modelos de Linguagem (LLMs) representa um grande marco no avanço das áreas de PLN e IE. Estes modelos são sustentados pela tecnologia de Deep Learning, em que são construídas redes neurais artificiais, com o objetivo de detectar padrões em um conjunto de dados de entrada através de múltiplas camadas de neurônios (Sarker, 2021). Essa estrutura permite que máquinas identifiquem padrões com alta efetividade, sendo capazes de prever tokens futuros e sequências de palavras com precisão.

Os LLMs, modelos que estudam enormes quantidades de informação através de texto em uma etapa denominada pré-treinamento, demonstram ser uma solução para tarefas de interpretação e geração textual, com avançada capacidade de análise semântica e respondimento de perguntas (Jurafsky; Martin, 2025).

Esses avanços na comunicação em linguagem natural entre humano e máquina são visíveis através do recente sucesso de ferramentas como ChatGPT (OpenAI, 2022), LLMs orientados a conversação. Ao prestarem o papel de assistente virtual para múltiplas diferentes tarefas com significativa precisão, estes modelos se tornaram parte do cotidiano da população, e a expansão do tamanho destes modelos promete capacidade cada vez maior na resolução de problemas de PLN (Zhao et al., 2025).

2.4 GERAÇÃO AUMENTADA VIA RECUPERAÇÃO

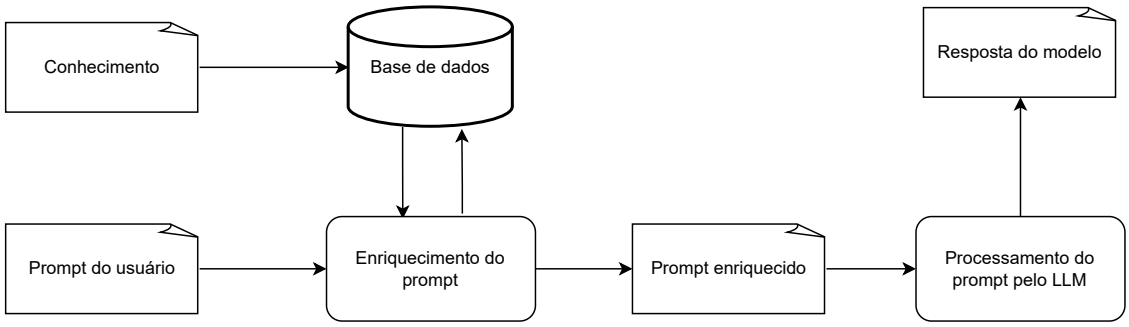
Como constatado, LLMs são eficientes para resolver uma vasta multitude de tarefas ao serem treinados com as enormes quantidades de dados presentes na internet, com conteúdo que representa diversas categorias de informação diferentes. Contudo, para a resolução de tarefas cujo contexto demonstra ser específico ou menos presente na etapa de treinamento, esses LLMs com propósito geral podem apresentar resultados menos consistentes (Kandpal et al., 2023).

Neste âmbito, a tecnologia de Geração Aumentava via Recuperação (do inglês, Retrieval Augmented Generation - RAG) se apresenta como uma ferramenta para contornar estas inconsistências, e é definida pela recuperação de informação relevante à tarefa a ser tratada em uma base de dados, somando-a ao contexto. Esse contexto enriquece o prompt oferecido ao modelo, que utiliza esse conhecimento para gerar respostas mais precisas (Song et al., 2025). A Figura 3 representa uma simples execução de RAG.

O uso de RAG, portanto, se torna bastante eficiente para tratar tarefas pertencentes a domínios específicos, como da medicina ou direito, ao incorporar informações relevantes e confiáveis sobre estes assuntos, oriundas de bases vetoriais ou grafos de conhecimento (Beckhauser; Fileto, 2024), para melhorar os resultados.

Para aplicações no campo farmacêutico, RAG se torna uma ferramenta bas-

Figura 3 – Geração Aumentada via Recuperação (RAG)



Fonte: Elaborada pelo autor.

tante valiosa, ao conceder contexto adicional para auxiliar na resolução de entidades, considerando a alta variabilidade de medicamentos, concentrações e dosagens. Por exemplo, quando lidando com dois medicamentos com nomes diferentes - mas baseados em um mesmo princípio ativo - uma base de fatos auxilia na determinação da similaridade entre estes medicamentos em determinados contextos.

3 TRABALHOS RELACIONADOS

As melhores metodologias para realizar a tarefa de Ligação de Entidades ainda são alvo de muito debate. O trabalho de (Xiao et al., 2023) explora o potencial de LLMs como poderosos ligadores de entidades, demonstrando sua efetividade ao empregar esses modelos combinados com uma base de conhecimento para a execução da tarefa de desambiguação.

(Liu et al., 2024) destaca a diminuição da precisão das respostas em tarefas de Entity Linking ao apresentar a LLMs quantidades muito grandes de registros candidatos. O trabalho de (Choudhary et al., 2021) destaca a possibilidade do uso de KGs para amenizar estes problemas, colaborando para a filtragem de entidades significantes ao aplicar funções de similaridade e definir um ranking de relevância entre as possíveis escolhas.

(Liu et al., 2024) também propõe uma abordagem de EL estruturada em três etapas: (i) sumarização e filtragem da entrada, para reduzir a quantidade de candidatos, (ii) uso de informações do contexto e conhecimento prévio para executar a desambiguação, e (iii) uma redução no número de alucinações ao implementar um algoritmo de consistência.

(Ding et al., 2024) explora as melhores estratégias de prompting para tarefas de entity linking utilizando LLMs. São utilizados prompts com múltipla escolha para seleção da entidade mais adequada.

Um exemplo próximo ao problema de Entity Linking no campo farmacêutico é o trabalho de (Xu; Chen; Hu, 2023), que realça as dificuldades da realização de tarefas de NLP no campo da saúde. As principais barreiras apontadas são as altas variabilidade de nomenclaturas, com alta quantidade de palavras sinônimas e variações morfológicas, adicionando bastante complexidade à tarefa.

O uso de RAG se destaca entre uma das soluções para melhoria de resultados para tarefas de NLP. O trabalho de (Beckhauser; Fileto, 2024) foca em apresentar o aumento da precisão nas respostas de LLMs ao integrá-los a grafos de conhecimento. Seus resultados demonstram melhorias de até mais de 40% de precisão, e mostram que modelos de linguagem menores, quando integrados com KGs, podem alcançar resultados melhores que modelos significativamente maiores.

Tabela 1 – Comparação entre trabalhos correlatos em Entity Linking com LLMs

Trabalho	**Ano**	**Estratégia principal**	**Uso de RAG**	**Notas fiscais**	**Medicamentos**
Xiao et al. (INSGENEL) (Xiao et al., 2023)	2023	LLMs instruídos combinados com recuperadores (retrievers).	Sim	Não	Não
Liu et al. (OneNet) (Liu et al., 2024)	2024	Pipeline em 3 etapas: filtragem, uso de contexto, consistência.	Parcial	Não	Não
Choudhary et al. (Choudhary et al., 2021)	2021	Knowledge Graphs e embeddings para ranking de entidades.	Não	Não	Não
Ding et al. (EntGPT) (Ding et al., 2024)	2024	Estratégias de prompting (múltipla escolha).	Não	Não	Não
Zhang et al. (Biomedical EL) (Xu; Chen; Hu, 2023)	2023	Interação entre entidades para EL biomédico.	Não	Não	Sim
Almeida et al. (Beckhauser; Fileto, 2024)	2023	Integração de LLMs com KGs e RL.	Sim	Não	Não

4 PROCESSO PROPOSTO

Esse capítulo descreve o processo proposto para a tarefa de extração de informação e ligação de entidades de notas fiscais de medicamentos aliada à conhecimento. As duas tarefas compreendidas neste trabalho são a Filtragem da Lista de Candidatos, que visa minimizar a quantidade de registros oferecidos aos LLMs para o processo de desambiguação, e a Desambiguação de Entidades Nomeadas, que tem como objetivo definir qual medicamento cadastrado em bases de dados oficiais corresponde ao produto descrito em uma nota fiscal.

A abordagem proposta é centrada em diferentes técnicas de PLN. Para a etapa de Filtragem, foram desenvolvidas expressões regulares baseadas nos padrões de representação encontrados nos dados, com o objetivo de capturar informação que pudesse diminuir o número de candidatos com um baixo custo computacional. Na etapa de Desambiguação, para lidar com a complexidade e despadronização da informação, são utilizados LLMs pequenos, que possam ser executados localmente nos servidores da UFSC, devido à privacidade dos dados.

A figura 4 apresenta o processo desenvolvido, desde a definição da origem e geração dos dados utilizados nos experimentos deste trabalho até a etapa final da resolução do problema apresentado. As etapas do processo em cor branca representam tarefas previamente desenvolvidas por outros estudantes do projeto, enquanto em verde temos as etapas efetivamente realizadas e analisadas neste trabalho.

A **etapa 1** corresponde à seleção dos dados de notas fiscais a serem utilizados nos experimentos. São relevantes para este trabalho aquelas que correspondem a medicamentos, que constituem o foco principal desta pesquisa.

A **etapa 2** é a indexação das descrições das notas fiscais em uma base de dados vetorial. O objetivo dessa indexação é viabilizar operações mais eficientes sobre dados desestruturados, armazenados em formato textual. A representação vetorial das descrições permite que métodos de busca léxica sejam aplicados posteriormente, facilitando a recuperação e comparação de informações de maneira mais precisa.

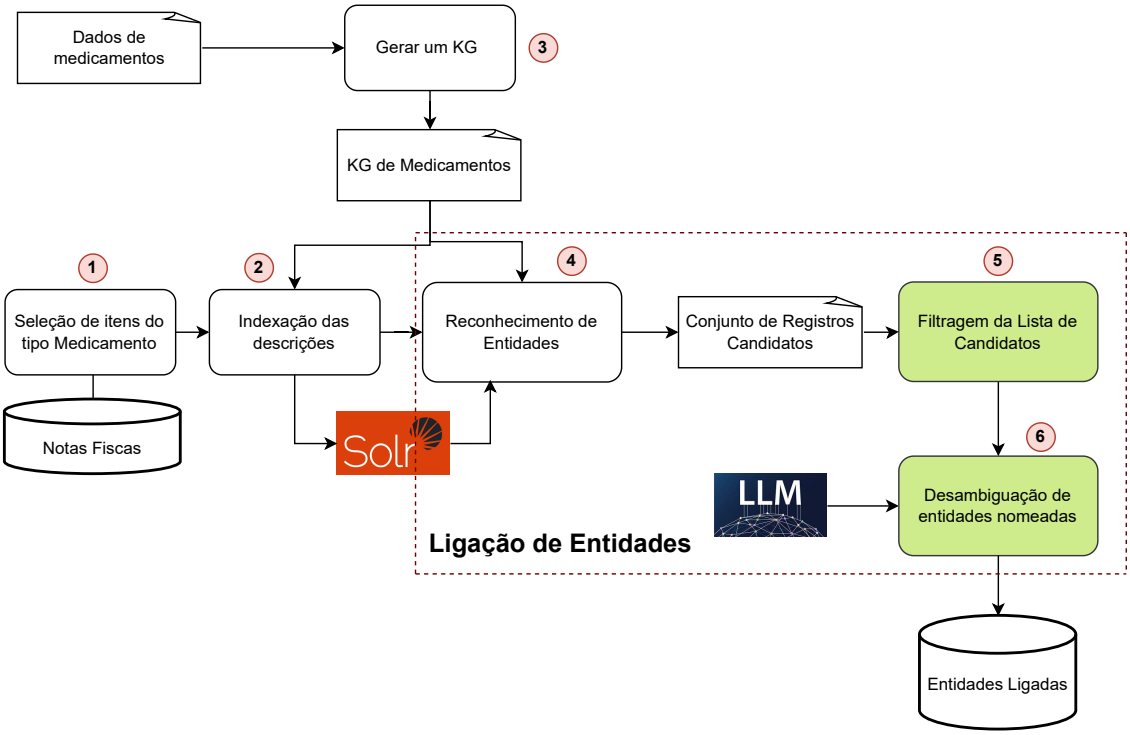
A **etapa 3** representa a construção da base de conhecimento a ser utilizada no processo. Nesta fase, dados sobre medicamentos são carregados em um grafo de conhecimento, responsável por armazenar as informações sobre os registros candidatos à ligação de entidades. O uso de um grafo permite a modelagem de relações entre diferentes atributos e entidades, contribuindo melhor análise e recuperação de informação.

A **etapa 4** é a etapa de reconhecimento de entidades. Nesta etapa, são cruzados os dados provenientes das etapas 2 e 3. Para cada descrição, são recuperados os registros de medicamentos que apresentam maior compatibilidade léxica, gerando uma lista inicial de correspondências.

A **etapa 5** corresponde à filtragem da lista de candidatos gerada pela etapa anterior. É feita a minimização da lista de registros candidatos para possibilitar e diminuir o custo computacional da etapa subsequente.

O **etapa 6** representa a tarefa final do processo. Para cada par de descrição e lista de candidatos, estes são apresentados a um LLM com o objetivo de selecionar o candidato mais compatível. As entidades ligadas são, enfim, armazenadas no banco de dados, e é efetuada a avaliação dos resultados dos modelos.

Figura 4 – Processo proposto para reconhecimento e ligação de entidades



Fonte: Elaborada pelo autor.

5 EXPERIMENTOS

Este capítulo define as variáveis a serem utilizadas neste trabalho, com o intuito de implementar o processo proposto no capítulo anterior. Nas seções seguintes, serão definidos o conjunto de dados, o método de seleção de entidades candidatas, a amostragem, o prompt utilizado e os modelos selecionados.

5.1 CONJUNTO DE DADOS

Foram utilizados três diferentes datasets para a execução dos experimentos. Os registros a serem ligados são descrições de notas fiscais de medicamentos relacionadas a compras públicas do Governo do Estado de Santa Catarina, constando com 1.565 registros, salvos em um banco de dados relacional PostgreSQL.

A Figura 5 mostra alguns exemplos de descrições de notas fiscais. Nota-se a despadronização da informação, com alguns medicamentos contendo informações como laboratório, concentração e volume, enquanto os mesmos dados estão faltantes em outras descrições. Além disso, todas as informações estão representadas em uma única string, sem ordem ou estrutura definida.

Para os medicamentos que farão parte da população da base de conhecimento, são utilizadas duas fontes: (i) lista de medicamentos aprovados pela Anvisa, obtidos via API e (ii) lista de preços de medicamentos oferecida pela CMED, em formato CSV. Estes dados foram utilizados para popular um grafo de conhecimento, que abriga todo o conhecimento sobre medicamentos a ser utilizado no processo de ligação de entidades. O uso do grafo de conhecimento possibilita que, em trabalhos posteriores, seja feita uma extensão com aplicação de regras lógicas, contribuindo para um processo de recuperação mais preciso e capaz de identificar relações implícitas entre os dados.

A Figura 6 apresenta exemplos de informações de medicamentos recuperadas do grafo de conhecimento. São recuperados os valores de nome do medicamento, princípio ativo, laboratório e apresentação do remédio. O campo de apresentação também abriga variadas informações de maneira desestruturada.

5.2 SELEÇÃO DE ENTIDADES

O algoritmo de Seleção de Entidades Candidatas escolhido para definir os dados de entrada dos experimentos deste trabalho foi o de similaridade léxica. As descrições de medicamentos de NFes, indexadas na base de dados Solr, foram utilizadas para recuperar informação do Grafo de Conhecimento através de buscas por similaridade léxica. Foi avaliado o casamento de tokens pertencentes aos campos **nome do medicamento** ou **princípio ativo**. Os registros com alta correspondência foram selecionados como candidatos para a etapa de ligação de entidades.

Figura 5 – Exemplos de descrições de notas fiscais

Produto
sertralina 50mg c/30 gen
ancoron 100mg c/30 comp
plenance 10 mg c/90 cpr
anti septico spray 50ml #
losartana potassica 50mg cx/960cpr generico prati donaduzzi
profenid im - ap 2ml 100mg
vast 40mg 30comp rev
combiron folico c45 cpr
depakote er 500mg c/30cp +
to lrest 100mg 20cpr c1 tg
esomex 40mg c/28 comp

Fonte: Elaborada pelo autor.

Figura 6 – Exemplos de medicamentos recuperados do grafo de conhecimento

nome do medicamento	princípio ativo	laboratório	apresentação
CETOPROFENO	CETOPROFENO	CRISTALIA PRODUTOS QUIMICOS FARMACEUTICOS LTDA.	100 MG PO LIOF P/ SOL INJ CX 50 FA VD TRANS
ARTRINID	CETOPROFENO	UNIAO QUIMICA FARMACEUTICA NACIONAL S/A	50 MG/ML SOL INJ IM CT 50 AMP VD AMB X 2 ML
ARTRINID	CETOPROFENO	UNIAO QUIMICA FARMACEUTICA NACIONAL S/A,	100 MG PO LIOF SOL INJ IV CT 50 FA VD TRANS

Fonte: Elaborada pelo autor.

5.3 DEFINIÇÃO DA AMOSTRA

Para a conferência dos resultados oferecidos pelos modelos, é necessária a ligação manual das entidades. Para a definição de uma amostra representativa, foi considerada a metodologia determinada por (Rea; Parker, 2012), que define um guia para amostragens em populações pequenas.

$$amostra = \frac{Z^2 \cdot (0.25) \cdot N}{Z^2 \cdot (0.25) + (N-1) \cdot M_E^2}.$$

Nesta equação, M_E representa a margem de erro, Z o Z-score relacionado à margem de erro, e N o tamanho do dataset original.

A amostra foi definida, portanto, como 150 descrições aleatórias, extraídas do dataset original de 1.565 descrições, contando com uma margem de erro de 10% e confiabilidade de 99%.

5.4 DEFINIÇÃO DO PROMPT

Ao serem eleitas as entidades candidatas, atribuímos aos LLMs a tarefa de fazer a ligação dessas entidades. Foi definida uma abordagem zero-shot com o seguinte prompt:

> Para o seguinte medicamento: `{medicamento}`, qual das seguintes descrições é mais provável de se referir ao medicamento descrito?
>
> `{lista_candidatos}`
>
> Por favor, responda somente com o código (número) correspondente à resposta correta, sem texto adicional. Decida de acordo com as seguintes prioridades:
>
> 1. Laboratório
> 2. Concentração
> 3. Apresentação
>
> Escolha a string que apresenta a maior quantidade de informações compatíveis com a descrição do medicamento. A resposta PRECISA estar entre as opções fornecidas.

O prompt descreve as regras definidas para ligação: O laboratório responsável pelo medicamento deve ser o parâmetro mais importante para associar registros, seguido pela concentração do medicamento (ex: 50MG/ML), e apresentação (ex: 50 CAPSULAS).

5.5 MODELOS SELECIONADOS

Para a execução dos experimentos, foram selecionados cinco diferentes Grandes Modelos de Linguagem. Foram eleitos modelos pequenos, que pudessem ser executados localmente, dado a natureza sigilosa dos dados de licitações. A seleção considerou a relevância dos modelos no estado da arte.

A Tabela 2 apresenta o nome e o tamanho da janela de contexto de cada um dos modelos selecionados.

Tabela 2 – Modelos utilizados

Modelo	**Nome da API do Modelo**	**Context Window**
Mistral	`mistral:7b`	`8k tokens`
Gemma	`gemma3:27b`	`128k tokens`
Qwen	`qwen2.5:72b`	`128k tokens`
LLaMa	`llama3.3:70b`	`128k tokens`

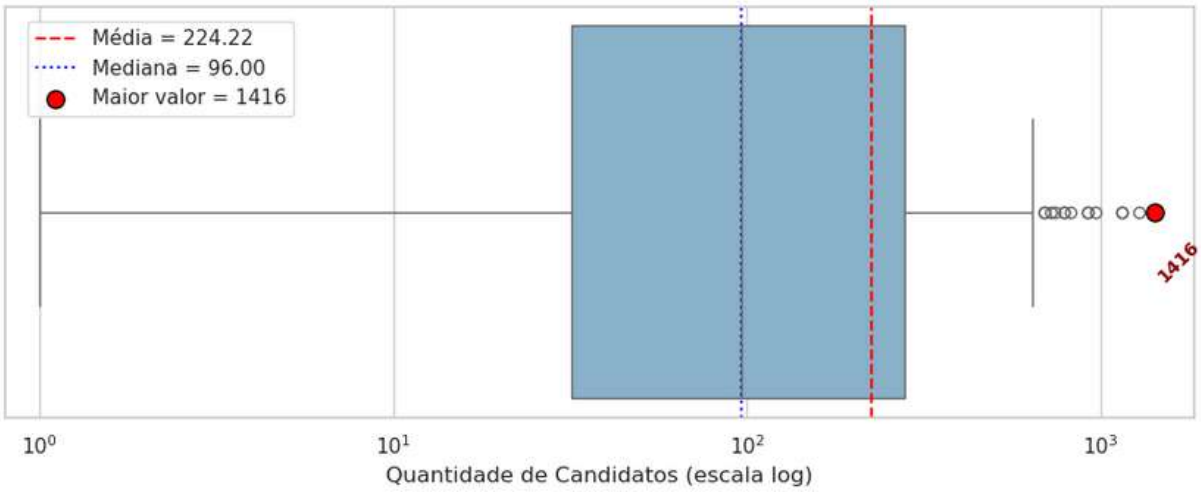
6 RESULTADOS

6.1 FILTRAGEM DE REGISTROS CANDIDATOS

Nesta seção, será feita uma análise das listas de registros candidatos e das heurísticas de filtragem aplicadas sobre elas, em uma tentativa de reduzir o número de registros. O objetivo desta tarefa é reduzir o número de tokens na entrada dos LLMs, possibilitando a adequação dentro da janela de contexto dos modelos, bem como reduzindo o custo computacional e facilitando o processo de decisão dos modelos.

A Figura 7 apresenta um gráfico boxplot com a distribuição da quantidade de registros candidatos por descrição, em escala logarítmica para melhor visualização. Nota-se que a maioria das descrições possui até 100 correspondências na base de conhecimento, mas com muitos *outliers* que tiveram até mais de 3500 medicamentos considerados similares e candidatos à desambiguação.

Figura 7 – Distribuição da quantidade de candidatos por registro

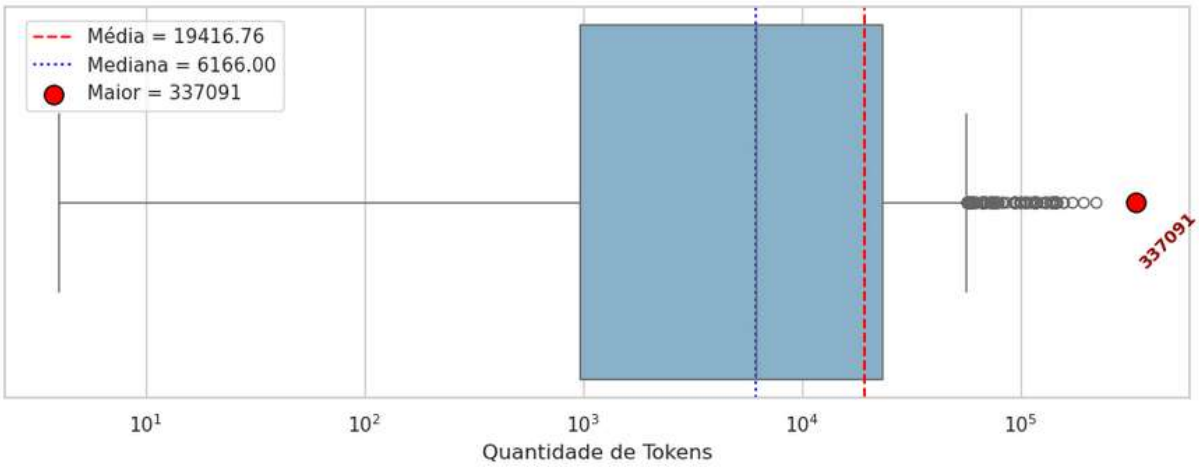


Fonte: Elaborada pelo autor.

Um melhor entendimento do problema enfrentado pode ser obtido a partir da Figura 8. Nela, observa-se a distribuição da quantidade de tokens que compõe cada lista de registros candidatos, em que muitos possuem mais de 50.000 tokens, com valores que podem chegar a até mais de 300.000. Esta quantidade altamente elevada de tokens implica em um alto grau de complexidade, ou até impossibilita a resolução da tarefa por LLMs com janelas de contexto de tamanhos insuficientes.

Dado o exposto, percebe-se a necessidade de encontrar maneiras de reduzir a quantidade de opções a serem entregues aos LLMs para o processo de desambiguação. Com esse objetivo, foi elaborado o passo de filtragem para a lista de registros candidatos, baseado nas informações presentes nas descrições dos medicamentos. Ao analisar os padrões presentes na representação textual dos dados nas descrições, puderam ser construídas expressões regulares que identificam essas informações no

Figura 8 – Distribuição da quantidade de tokens por registro



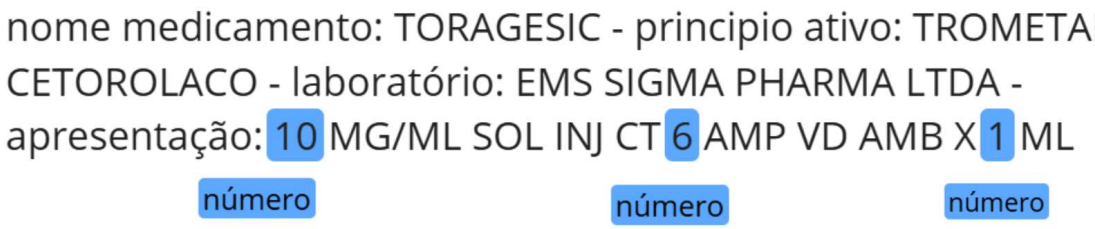
Fonte: Elaborada pelo autor.

texto para, depois, avaliar a compatibilidade entre uma descrição e cada um de seus candidatos, com o objetivo de descartar candidatos incompatíveis.

Valores como concentração ou volume foram considerados importantes valores numéricos a serem observados no processo de eleição dos registros corretos. Portanto, foram testadas duas diferentes heurísticas: (I) a extração de quaisquer valores numéricos que estejam presentes tanto na descrição da nota fiscal quanto na do candidato, e (II) a extração de valores numéricos junto de suas unidades de medida (ml, mg, etc.). A abordagem (I) propõe uma alternativa mais generalizada, enquanto (II) apresenta maior sensibilidade à maneira como os dados são apresentados.

A expressão que implementa a primeira heurística busca todos as substrings compostas exclusivamente por dígitos ([0-9]) que fazem parte de uma descrição. A Figura 9 apresenta a identificação dos valores baseado na heurística (I), capturando todos os valores numéricos presentes no texto.

Figura 9 – Exemplo da heurística número I



Fonte: Elaborada pelo autor.

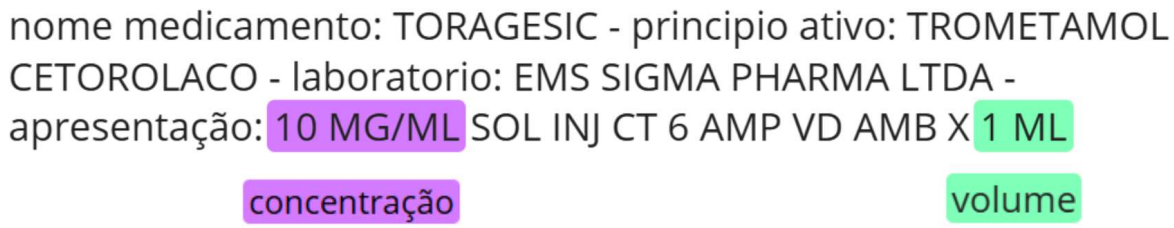
Para a implementação da segunda heurística, foi necessária uma análise de como os dados eram normalmente estruturados e suas unidades de medida. Foram identificadas as principais unidades de medida para concentração (MGG, G, MG, etc.) e volume (ML) para a construção de expressões regulares que pudessem reconhecer estes valores no texto. As expressões construídas são apresentadas a seguir:

```
Concentração
([0-9]+([\.,][0-9]+)*\s*(MGG|G|MG|UI|%|MCG)\s*(/\s*[0-9]*?\s*(G|ML|MG|UI|%)+)?)

Volume
\s+(C/)?[0-9]+([\.,][0-9]+)*\s?(ML)(?!/)
```

A identificação dos valores efetuada pelas expressões regulares anteriores são representados pela Figura 10, considerando as unidades de medida mais comuns juntamente dos valores numéricos, em uma tentativa de incorporar uma maior restrição dos valores capturados.

Figura 10 – Exemplo da heurística número II

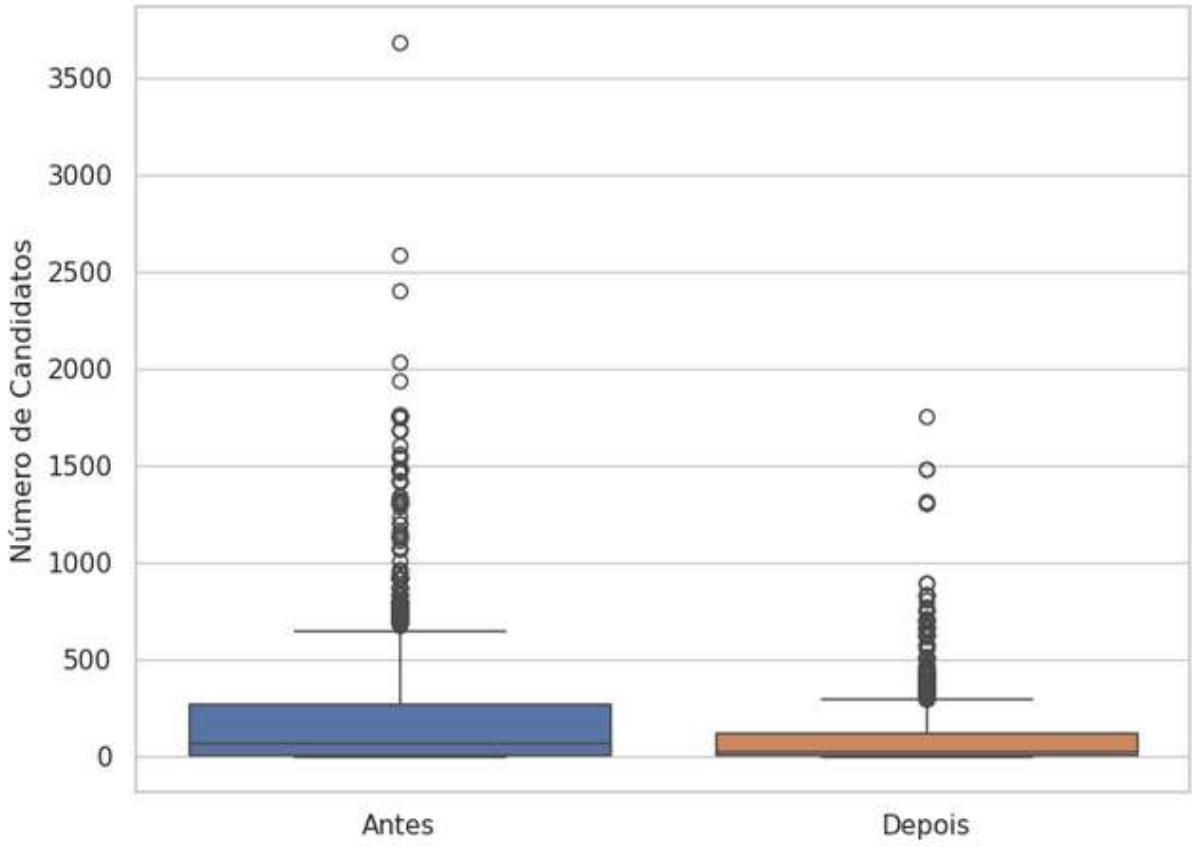


Fonte: Elaborada pelo autor.

Para a comparação do desempenho entre as duas alternativas, as métricas selecionadas foram a redução do número de candidatos por registro, e a quantidade de registros considerados corretos que foram equivocadamente descartados (falsos negativos).

A Figura 11 apresenta a redução do tamanho da lista de candidatos efetuada pela heurística (I). Observa-se um evidente sucesso na filtragem de candidatos para os valores outliers, drasticamente diminuindo a quantidade de tokens necessária para desambiguação destas notas fiscais. O outlier mais extremo, que antes constava com mais de 3.500 candidatos, teve este número reduzido para menos de 2.000 através desta heurística.

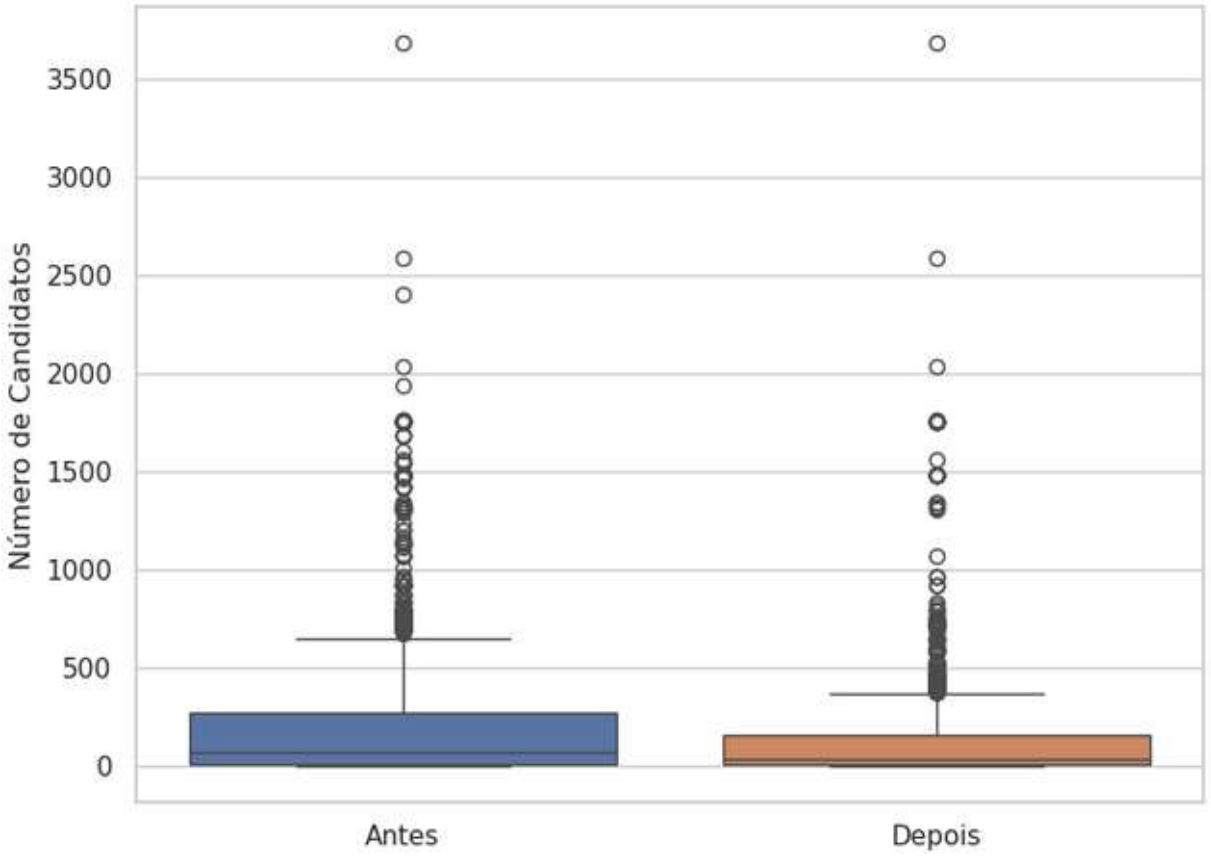
Figura 11 – Antes e depois do tamanho das listas de candidatos com heurística I



Fonte: Elaborada pelo autor.

A Figura 12 apresenta a redução efetuada pela heurística (II). Nota-se que, apesar de haver uma diminuição dos quartis e limites superiores, esta heurística teve menor desempenho para os outliers, não conseguindo bom desempenho para os casos mais extremos. Os dois registros mais graves não apresentaram qualquer diminuição no seu número de candidatos através desta heurística.

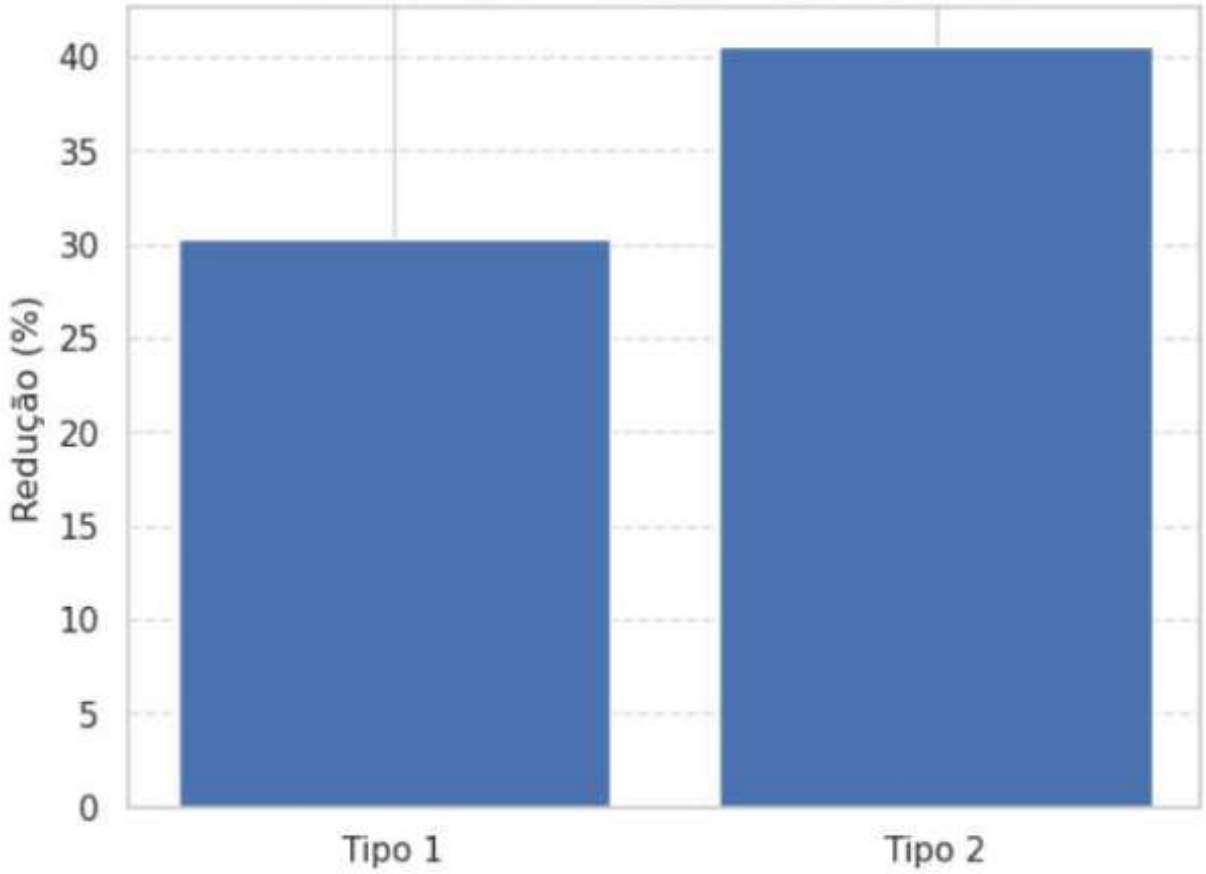
Figura 12 – Antes e depois do tamanho das listas de candidatos com heurística II



Fonte: Elaborada pelo autor.

O desempenho final da redução efetuada pelas duas heurísticas é comparado pela Figura 13. A abordagem (II) reduziu em apenas 26% o número médio de candidatos para cada registro, comparado com 40% para a heurística (I), e apresentou desempenho pior para os casos mais críticos (como observado na Figura 12). Isso se dá pois os registros com quantidades mais elevadas de registros costumam ser misturas de dois princípios ativos diferentes, contendo padrões de apresentação mais complexos de serem contemplados por expressões regulares tão específicas.

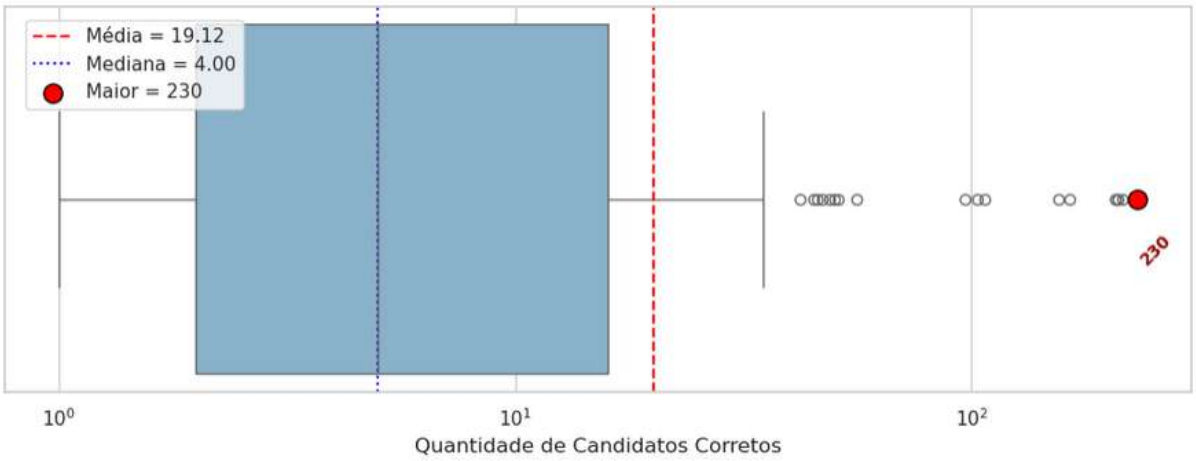
Figura 13 – Comparação da redução de candidatos entre as duas heurísticas



Fonte: Elaborada pelo autor.

Para análise dos falsos positivos, foi utilizada a amostra de 150 registros mencionada no capítulo anterior para ser desambiguada manualmente, listando os registros considerados como respostas corretas para o processo de desambiguação. A Figura 14 mostra a distribuição logarítmica da quantidade de registros considerados corretos pela desambiguação manual, para cada lista de candidatos. Em muitos dos casos, existem múltiplos registros considerados corretos entre as opções, com um valor médio de 19 candidatos, com máximo de 230.

Figura 14 – Quantidade de candidatos considerados corretos por registro



Fonte: Elaborada pelo autor.

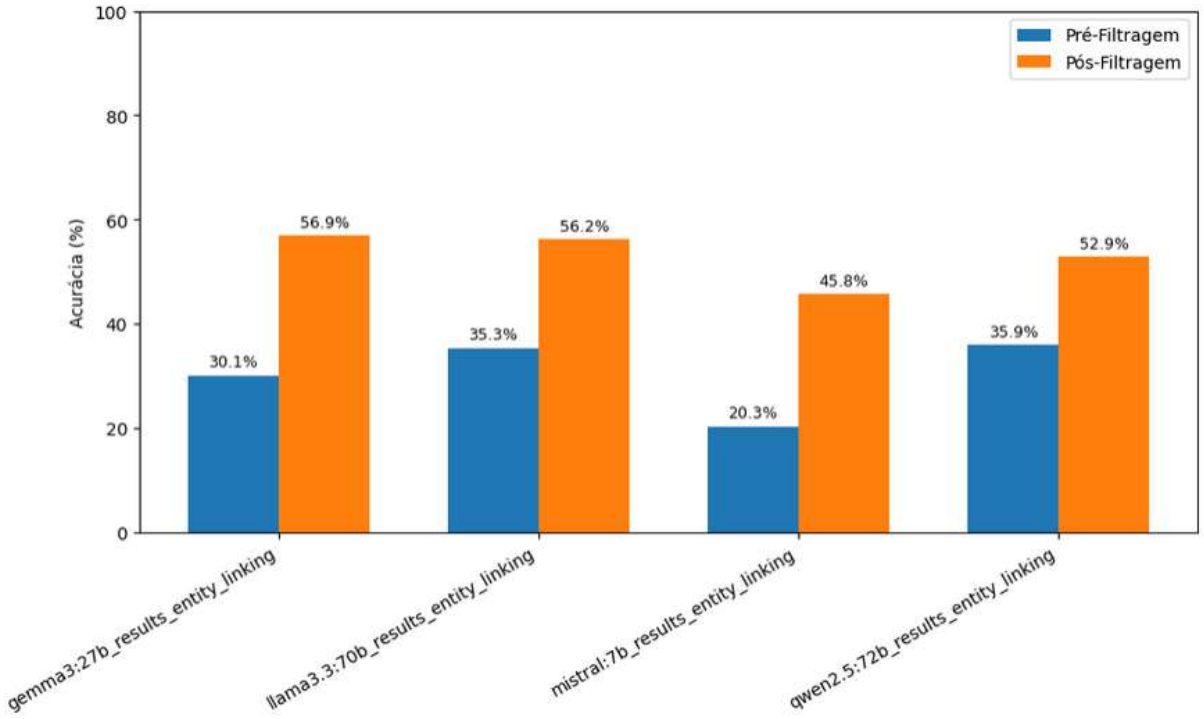
Finalmente, a execução dos dois algoritmos de filtragem na amostra resultou que a abordagem (I) não descartou nenhum registro considerado correto, enquanto a abordagem (II) descartou um ou mais registros corretos em 13% das descrições presentes na amostra. Essa observação conclui que a abordagem (II), apesar de mais generalizada, desempenha melhor a função de filtro para a lista de candidatos, conseguindo maior porcentagem de filtração e com menos incidência de falsos negativos.

6.2 DESAMBIGUAÇÃO DE ENTIDADES

Esta seção apresenta os resultados dos resultados da tarefa de desambiguação de entidades usando o processo proposto. É analisada a performance dos modelos selecionados de acordo com sua acurácia e tempo médio de resposta. Para demonstração da eficácia da etapa anterior, são executados os experimentos sobre os dados em sua forma original e sobre os dados após filtragem, utilizando a filtragem considerada mais bem sucedida na etapa anterior.

A melhoria no desempenho dos modelos é evidente na Figura 15, que apresenta a acurácia de cada modelo para a tarefa de desambiguação, com os dados pré e pós filtragem. Todos os modelos testados tiveram ganhos altamente significativos de acurácia ao lidar com uma lista de candidatos reduzida, comprovando a eficiência da etapa de filtragem para impulsionar o processo de ligação de entidades. Também é possível perceber o destaque dos modelos gemma3:27b e llama3.3:70b, que alcançaram a melhor acurácia para a tarefa. Ademais, é importante ressaltar a capacidade do modelo Gemma de conseguir os melhores resultados tendo uma quantidade de parâmetros menor que a metade dos modelos Llama e Qwen, indicando uma alta performance para a tarefa. O modelo mistral:7b, com o menor número de parâmetros, apresentou resultados inferiores aos outros modelos experimentados.

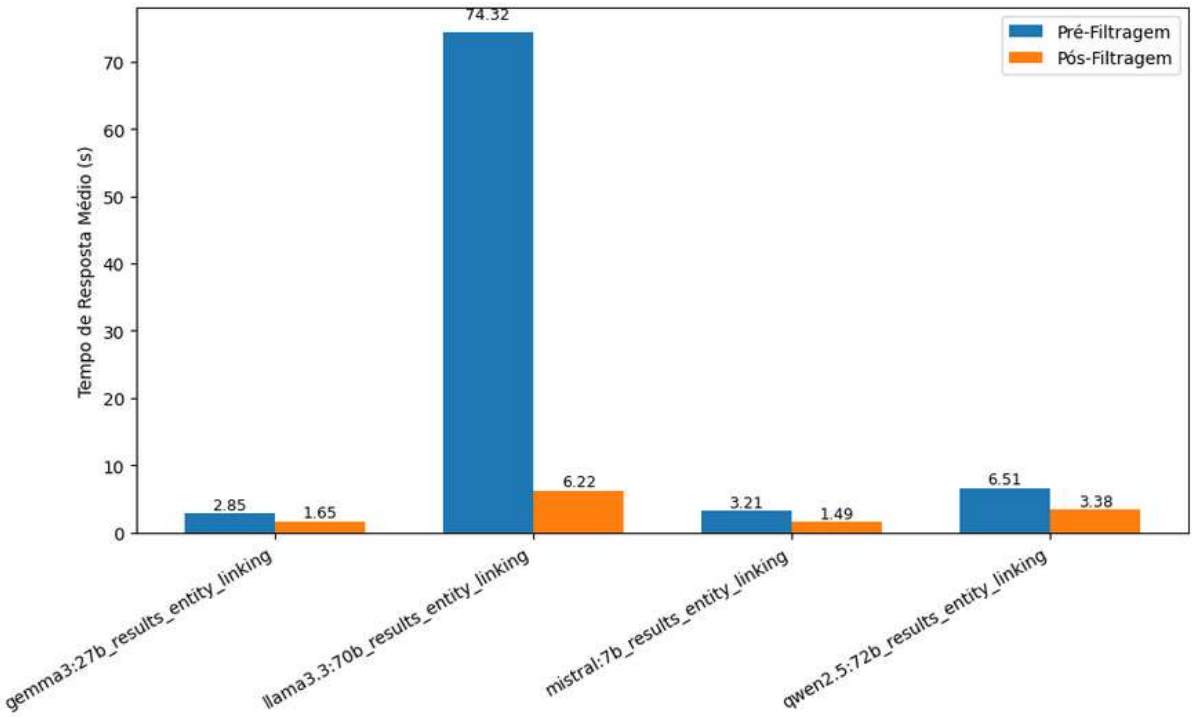
Figura 15 – Resultado da Desambiguação de Entidades



Fonte: Elaborada pelo autor.

O tempo de resposta médio de cada modelo pode ser observado na Figura 16. O gráfico demonstra uma queda significativa do tempo de resposta para todos os modelos após a etapa de redução da quantidade de registros candidatos, destacando a importância da tarefa dentro do processo proposto. Os modelos mistral:7b e gemma3:27b se destacaram neste âmbito, tendo os menores tempos antes e depois da redução do número de candidatos, chegando a valores menores que 2 segundos. Evidencia-se um enorme tempo de resposta do modelo llama3.3:70b para os dados pré-filtragem, destacando a dificuldade deste modelo para lidar com a enorme quantidade de opções em relação aos outros selecionados.

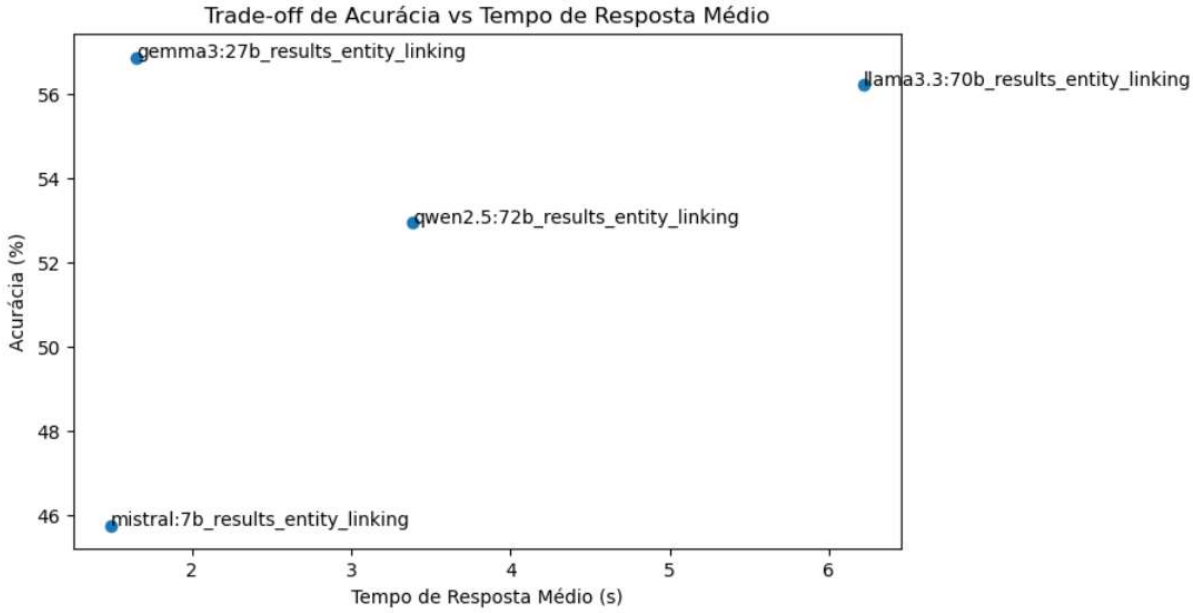
Figura 16 – Comparação do Tempo de Resposta entre Modelos



Fonte: Elaborada pelo autor.

Unindo os resultados dos dois parâmetros avaliados previamente, temos o gráfico de trade-off entre acurácia e tempo de resposta médio na Figura 17. Para esta comparação, foi utilizada a versão reduzida da lista de candidatos, para produzir os resultados finais de todo o processo de ligação de entidades. O desempenho mais satisfatório pertence ao modelo gemma3:27b, que aparece no canto superior esquerdo do gráfico, demonstrando um dos menores tempos de resposta enquanto simultaneamente tem os melhores resultados em termos de acurácia. O outro modelo com maior acurácia, llama3.3:70b, teve o tempo de resposta mais demorado, atingindo os seis segundos, além da queda drástica de desempenho para listas maiores que foi constatada na Figura 16. A quantidade inferior de parâmetros do modelo mistral:7b pode justificar sua posição no canto inferior esquerdo do gráfico.

Figura 17 – Trade-off de Acurácia vs Tempo de Resposta Médio



Fonte: Elaborada pelo autor.

7 CONCLUSÃO E TRABALHOS FUTUROS

Descrições de notas fiscais são comumente organizadas de maneira despadronizada, o que dificulta o acesso à informação e a integração desta informação com dados de outras origens. Este trabalho avalia uma abordagem mista, usando LLMs e conhecimento de domínio sobre medicamentos registrados pela Anvisa, integrado a informações da CMED em um grafo de conhecimento, para extrair informação e desambiguar entidades de notas fiscais de medicamentos. A utilização exclusiva de LLMs para desambiguar entidades candidatas teve resultados de acurácia de no máximo 35,9%. Por outro lado, os experimentos revelaram que se pode alcançar até 56,9% de acurácia ao combinar o uso de LLMs com uma etapa prévia que usa o grafo de conhecimento e expressões regulares para reduzir a quantidade de candidatos. Esta abordagem também contribui de maneira relevante para reduzir custos computacionais e latência.

Entre os modelos avaliados, o gemma3:27b apresentou o melhor equilíbrio entre acurácia e tempo de resposta, tendo resultados competitivos com modelos bastante maiores, como o llama3.3:70b, enquanto apresenta custo computacional significativamente menor. Esse resultado demonstra o potencial de LLMs de pequeno porte para a resolução de tarefas, especialmente quando unidos à outras técnicas mais tradicionais.

7.1 TRABALHOS FUTUROS

Para aprimorar a solução proposta, trabalhos futuros terão foco na ampliação e o aprimoramento das etapas desenvolvidas neste estudo. Uma análise dos resultados presentes neste trabalho, seguida de um aprofundado estudo da estrutura dos dados, pode promover a aprimoração das heurísticas de redução de candidatos apresentadas, contribuindo para uma contínua otimização da etapa de filtragem. Esta melhoria pode reduzir ainda mais o custo computacional do processo, influenciando positivamente na qualidade dos resultados do processo de desambiguação de entidades. Paralelo a este trabalho, também está sendo realizada a extensão do uso do grafo de conhecimento, incorporando a identificação de nomes de superfície na descrição de medicamentos, bem como agregando o uso de regras para a realização de inferência lógica.

Adicionalmente, considera-se a realização de experimentos adicionais com diferentes modelos de linguagem, abrangendo variadas arquiteturas e quantidades de parâmetros. Essa ampliação tem como objetivo avaliar o desempenho de distintos LLMs dentro do processo proposto, contribuindo para uma análise mais abrangente e para a construção de uma solução final mais robusta e otimizada.

Finalmente, também prevê-se a implementação do processo utilizando indexa-

ção reversa, isto é, do conhecimento de domínio, ao invés das descrições do itens das notas fiscais. Essa alteração tem o potencial de proporcionar aumentos de desempenho e precisão na recuperação de entidades candidatas, especialmente em bases de dados de grande volume. Atualmente, outros membros do nosso grupo de pesquisa estão trabalhando nestes temas, com mais avanços, incluindo resultados parciais um pouco melhores e perspectivas promissoras de viabilizar o processamento eficiente de dezenas de milhares de notas fiscais mensalmente.

REFERÊNCIAS

BECKHAUSER, William; FILETO, Renato. Boosting not so Large Language Models by using Knowledge Graphs and Reinforcement Learning. In*:* ANAIS do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. Belém/PA: SBC, 2024. p. 165–175.

CHOUDHARY, Shivani; LUTHRA, Tarun; MITTAL, Ashima; SINGH, Rajat. **A Survey of Knowledge Graph Embedding and Their Applications**. [S.l.*:* s.n.], 2021. arXiv: 2107.07842 [cs.IR]. Disponível em: https://arxiv.org/abs/2107.07842.

DING, Yifan; POUDEL, Amrit; ZENG, Qingkai; WENINGER, Tim; VEERAMANI, Balaji; BHATTACHARYA, Sanmitra. EntGPT: Linking Generative Large Language Models with Knowledge Bases. **arXiv preprint arXiv:2402.06738**, 2024.

IBM. **Natural language processing**. IBM Documentation. Disponível em: https://www.ibm.com/docs/en/rpa/21.0?topic=automation-natural-language-processing. Acesso em: 5 jul. 2025.

JURAFSKY, Daniel; MARTIN, James H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models**. 3rd. [S.l.*:* s.n.], 2025. Online manuscript released January 12, 2025.

KANDPAL, Nikhil; DENG, Haikang; ROBERTS, Adam; WALLACE, Eric; RAFFEL, Colin. **Large Language Models Struggle to Learn Long-Tail Knowledge**. [S.l.*:* s.n.], 2023. arXiv: 2211.08411 [cs.CL]. Disponível em: https://arxiv.org/abs/2211.08411.

LIU, Xukai; LIU, Ye; ZHANG, Kai; WANG, Kehang; LIU, Qi; CHEN, Enhong. OneNet: A Fine-Tuning Free Framework for Few-Shot Entity Linking via Large Language Model Prompting. **arXiv preprint arXiv:2410.07549**, 2024.

MA, Youmi; HIRAOKA, Tatsuya; OKAZAKI, Naoaki. **Named Entity Recognition and Relation Extraction using Enhanced Table Filling by Contextualized Representations**. [S.l.*:* s.n.], 2022. arXiv: 2010.07522 [cs.CL]. Disponível em: https://arxiv.org/abs/2010.07522.

OPENAI. **Introducing ChatGPT: A large-scale conversational AI model**. [S.l.*:* s.n.], nov. 2022. Blog post on the OpenAI website. Research preview released November 30, 2022. Disponível em: https://openai.com/blog/chatgpt.

PAN, Shirui; LUO, Linhao; WANG, Yufei; CHEN, Chen; WANG, Jiapu; WU, Xindong. Unifying Large Language Models and Knowledge Graphs: A Roadmap. **IEEE Transactions on Knowledge and Data Engineering**, Institute of Electrical e Electronics Engineers (IEEE), v. 36, n. 7, p. 3580–3599, jul. 2024. ISSN 2326-3865.

REA, L.M.; PARKER, R.A. **Designing and Conducting Survey Research: A Comprehensive Guide**. [S.l.]: Wiley, 2012. ISBN 9780787981259.

SARKER, Iqbal. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. **SN Computer Science**, v. 2, ago. 2021.

SONG, Zirui; YAN, Bin; LIU, Yuhan; FANG, Miao; LI, Mingzhe; YAN, Rui; CHEN, Xiuying. Injecting Domain-Specific Knowledge into Large Language Models: A Comprehensive Survey, 2025. arXiv: 2502.10708 [cs.CL].

XIAO, Zilin; GONG, Ming; WU, Jie; ZHANG, Xingyao; SHOU, Linjun; JIANG, Daxin. Instructed Language Models with Retrievers Are Powerful Entity Linkers. In*:* PROCEEDINGS of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP). Singapore: Association for Computational Linguistics, 2023. p. 2267–2282.

XU, Zhenran; CHEN, Yulin; HU, Baotian. Improving Biomedical Entity Linking with Cross-Entity Interaction. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 37, p. 13869–13877, jun. 2023.

YUAN, Hongyi; YUAN, Zheng; YU, Sheng. **Generative Biomedical Entity Linking via Knowledge Base-Guided Pre-training and Synonyms-Aware Fine-tuning**. [S.l.*:* s.n.], 2022. arXiv: 2204.05164 [cs.CL]. Disponível em: https://arxiv.org/abs/2204.05164.

ZHAO, Wayne Xin et al. **A Survey of Large Language Models**. [S.l.*:* s.n.], 2025. arXiv: 2303.18223 [cs.CL]. Disponível em: https://arxiv.org/abs/2303.18223.

Apêndices

APÊNDICE A – ARTIGO DO TCC

Extração de Informação de Notas Fiscais apoiada em Conhecimento

Pedro Nack Martins[1]

[1]Departamento de Informática e Estatística
Universidade Federal de Santa Catarina (UFSC)
Florianópolis – SC – Brazil

pedro.nack@grad.ufsc.br

Abstract. *The product description in an invoice is an unstructured and non-standardized text field, which makes it difficult to identify the product and its characteristics, and consequently to analyze and compare information such as the prices of the same products across different invoices. Extracting the information contained in these descriptions and identifying the correspondence between the described product and products registered in official databases enables the exploration of this information. This monograph explores the use of Natural Language Processing (NLP) techniques for extracting information and performing entity linking from electronic invoices (NF-es) to data on medicines approved by Anvisa stored in a Knowledge Graph (KG).*

Resumo. *A descrição de produto em uma nota fiscal é um campo de texto não estruturado e não padronizado, dificultando a identifição do produto e suas características, e consequentemente a análise e a comparação de informações como preço dos mesmos produtos em diferentes notas fiscais. A extração da informação contida nessas descrições e a identificação da correspondência do produto descrito com produtos cadastrados em bases de dados oficiais possibilita a exploração dessa informação. A presente monografia explora o uso de técnicas de Processamento de Linguagem Natural (PLN) para a extração da informação e ligação de entidades de notas fiscais eletrônicas (NF-es), com dados de medicamentos aprovados pela Anvisa armazenados em um Grafo de Conhecimento (KG).*

1. Introdução

Na era contemporânea, o vasto volume de dados disponíveis em formato digital representa um potencial imensurável para análises estatísticas e investigações. Contudo, o crescimento acelerado da quantidade e da variedade desses dados, além de sua complexidade, traz grandes desafios. Nesse cenário, a extração de informação de textos e sua integração com bases de conhecimento pré-existentes se apresenta como uma parte crucial do processamento e análise de dados. No contexto de notas fiscais, vendedores costumam preencher as descrições textuais de produtos em itens de NF-es sem padronização, gerando problemas no processo de estudo desses dados em grandes quantidades, e dificultando a união desses dados com informações de outras fontes.

A utilização de grandes modelos de linguagem (do inglês *Large Language Models* - LLMs) é uma abordagem recente, porém muito poderosa, para realizar tarefas de Processamento de Linguagem Natural. Este projeto tem como objetivo utilizar LLMs para

desambiguar e ligar entidades pertencentes a diferentes fontes de informação, identificando pares de itens correspondentes e permitindo a união destes dados desestruturados, gerando assim um aumento no potencial analítico desta informação. Ao final, espera-se que esse trabalho demonstre o potencial desta ferramenta em processos de integração e análise de dados textuais em grande escala, contribuindo para o avanço de soluções baseadas em inteligência artificial no campo de medicamentos.

2. Fundamentação Teórica

Este capítulo apresenta os conceitos fundamentais ao entendimento do trabalho, bem como descreve as técnicas e ferramentas utilizadas para o desenvolvimento da ideia e estrutura proposta.

2.1. Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) é uma área da Inteligência Artificial que estuda a capacidade de um computador de interpretar texto e fala de maneira análoga à humana [IBM]. A área de PLN abrange uma grande variedade de processos, como extração de informação, tradução, reconhecimento de fala, etc.

Para que um texto possa ser mais facilmente analisado por máquina, é conveniente que seja feita uma padronização da informação. O processo de tokenização, portanto, descreve a etapa de divisão de um texto em partes menores denominadas *tokens*, que podem ser pequenas frases, palavras, subpalavras ou até mesmo letras [Jurafsky and Martin 2025]. Cada token agrega um valor semântico à sentença, que é utilizado pelo computador para interpretar o sentido presente na informação.

2.2. Extração de Informação

A tokenização é base para diversas tarefas de PLN, incluindo tarefas de extração de informação. Extração de Informação (do inglês, Information Extraction - IE) se refere ao processo de extrair e estruturar dados de forma automatizada a partir de fontes não estruturadas [Jurafsky and Martin 2025]. Este trabalho lidará com dados textuais, em linguagem natural, presentes em descrições de notas fiscais.

2.3. Reconhecimento de Entidades Nomeadas

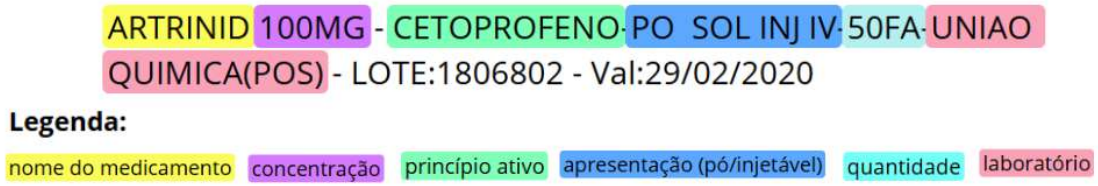
Entidades nomeadas são qualquer objeto a qual se pode atribuir um nome próprio: pessoas, cidades, organizações, etc. O reconhecimento e categorização dessas entidades é importante para identificação de relacionamentos entre entidades [Ma et al. 2022], solucionando problemas de desambiguação ou análises de sentimento.

A Figura 1 apresenta um exemplo de extração de informações na descrição do medicamento apresentado na descrição apresentada, categorizando os diferentes dados presentes. Na legenda, temos os diferentes tipos de entidades mais comumente presentes no conjunto de dados, como nome do medicamento, concentração e quantidade.

2.4. Desambiguação de Entidades

Em muitos casos, ao reconhecer as entidades de um texto, se torna interessante a incorporação de informação externa, por exemplo, presente em uma base de dados, para acrescentar semântica ao contexto de uma sentença. A recuperação desta informação

Figura 1. Exemplo de NER em Descrição de Medicamento



Elaborada pelo autor.

pode se tornar difícil em cenários onde existe ambiguidade, como é o caso de palavras homônimas, ou contextos em que instâncias possuem definições muito similares, como na área biomédica [Yuan et al. 2022]. Para isto, muitas dessas entidades precisam passar por um processo de desambiguação, orientado a definir a qual exato objeto de uma base de conhecimento aquela entidade se refere.

3. Processo Proposto

A abordagem proposta é centrada em diferentes técnicas de PLN. Para a etapa de Filtragem, foram desenvolvidas expressões regulares baseadas nos padrões de representação encontrados nos dados, com o objetivo de capturar informação que pudesse diminuir o número de candidatos com um baixo custo computacional. Na etapa de Desambiguação, para lidar com a complexidade e despadronização da informação, são utilizados LLMs pequenos, que possam ser executados localmente nos servidores da UFSC, devido à privacidade dos dados.

A Figura 2 apresenta o processo desenvolvido, desde a definição da origem e geração dos dados utilizados nos experimentos deste trabalho até a etapa final da resolução do problema apresentado. As etapas do processo em cor branca representam tarefas previamente desenvolvidas por outros estudantes do projeto, enquanto em verde temos as etapas 5 e 6, efetivamente realizadas e analisadas neste trabalho.

A **etapa 5** corresponde à filtragem da lista de candidatos gerada pela etapa anterior. É feita a minimização da lista de registros candidatos para possibilitar e diminuir o custo computacional da etapa subsequente.

O **etapa 6** representa a tarefa final do processo. Para cada par de descrição e lista de candidatos, estes são apresentados a um LLM com o objetivo de selecionar o candidato mais compatível. As entidades ligadas são, enfim, armazenadas no banco de dados, e é efetuada a avaliação dos resultados dos modelos.

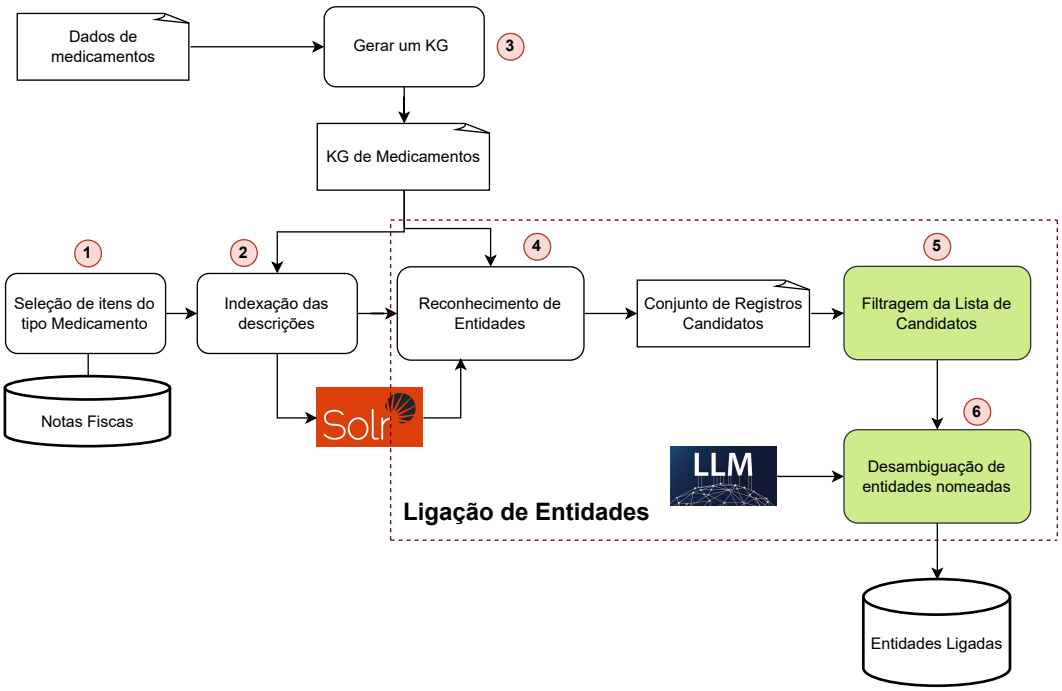
4. Experimentos

4.1. Conjunto de Dados

Foram utilizados três diferentes datasets para a execução dos experimentos. Os registros a serem ligados são descrições de notas fiscais de medicamentos relacionadas a compras públicas do Governo do Estado de Santa Catarina, constando com 1.565 registros, salvos em um banco de dados relacional PostgreSQL.

Para os medicamentos que farão parte da população da base de conhecimento, são utilizadas duas fontes: (i) lista de medicamentos aprovados pela Anvisa, obtidos via API e

Figura 2. Processo proposto para reconhecimento e ligação de entidades



Elaborada pelo autor.

(ii) lista de preços de medicamentos oferecida pela CMED, em formato CSV. Estes dados foram utilizados para popular um grafo de conhecimento, que abriga todo o conhecimento sobre medicamentos a ser utilizado no processo de ligação de entidades.

4.2. Seleção de Entidades

O algoritmo de Seleção de Entidades Candidatas escolhido para definir os dados de entrada dos experimentos deste trabalho foi o de similaridade léxica. As descrições de medicamentos de NFes, indexadas na base de dados Solr, foram utilizadas para recuperar informação do Grafo de Conhecimento através de buscas por similaridade léxica. Foi avaliado o casamento de tokens pertencentes aos campos **nome do medicamento** ou **princípio ativo**. Os registros com alta correspondência foram selecionados como candidatos para a etapa de ligação de entidades.

4.3. Definição da Amostra

Para a conferência dos resultados oferecidos pelos modelos, é necessária a ligação manual das entidades. Para a definição de uma amostra representativa, foi considerada a metodologia determinada por [Rea and Parker 2012], que define um guia para amostragens em populações pequenas. A amostra foi definida, portanto, como 150 descrições aleatórias, extraídas do dataset original de 1.565 descrições, contando com uma margem de erro de 10% e confiabilidade de 99%.

5. Definição do Prompt

Ao serem eleitas as entidades candidatas, atribuímos aos LLMs a tarefa de fazer a ligação dessas entidades. Foi definida uma abordagem zero-shot com o seguinte prompt:

Para o seguinte medicamento: {medicamento}, qual das seguintes descrições é mais provável de se referir ao medicamento descrito?
{lista_candidatos}
Por favor, responda somente com o código (número) correspondente à resposta correta, sem texto adicional. Decida de acordo com as seguintes prioridades:

1. Laboratório
2. Concentração
3. Apresentação

Escolha a string que apresenta a maior quantidade de informações compatíveis com a descrição do medicamento. A resposta PRECISA estar entre as opções fornecidas.

6. Modelos Selecionados

Para a execução dos experimentos, foram selecionados cinco diferentes Grandes Modelos de Linguagem. Foram eleitos modelos pequenos, que pudessem ser executados localmente, dado a natureza sigilosa dos dados de licitações. A seleção considerou a relevância dos modelos no estado da arte.

A Tabela 1 apresenta o nome e o tamanho da janela de contexto de cada um dos modelos selecionados.

Tabela 1. Modelos utilizados

Modelo	Nome da API do Modelo	Context Window
Mistral	`mistral:7b`	`8k tokens`
Gemma	`gemma3:27b`	`128k tokens`
Qwen	`qwen2.5:72b`	`128k tokens`
LLaMa	`llama3.3:70b`	`128k tokens`

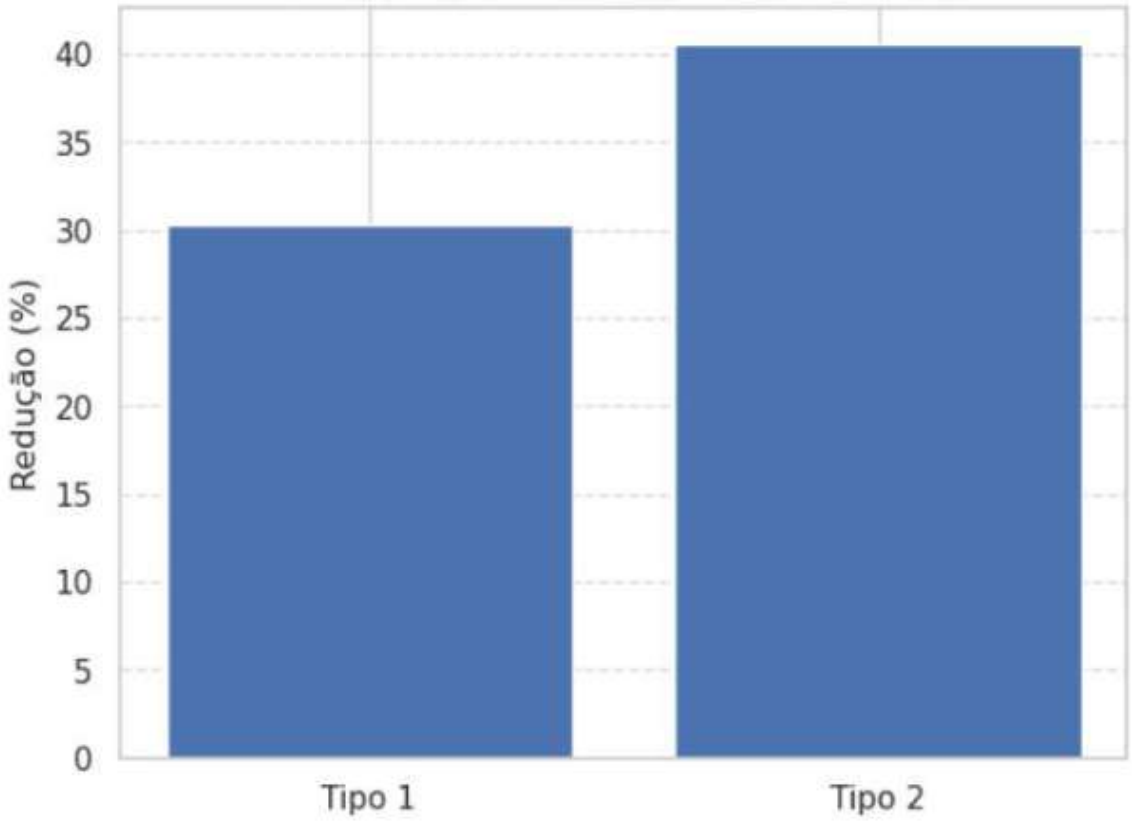
7. Resultados

7.1. Filtragem de Registros Candidatos

Valores como concentração ou volume foram considerados importantes valores numéricos a serem observados no processo de eleição dos registros corretos. Portanto, foram testadas duas diferentes heurísticas: (I) a extração de quaisquer valores numéricos que estejam presentes tanto na descrição da nota fiscal quanto na do candidato, e (II) a extração de valores numéricos junto de suas unidades de medida (ml, mg, etc.). A abordagem (I) propõe uma alternativa mais generalizada, enquanto (II) apresenta maior sensibilidade à maneira como os dados são apresentados.

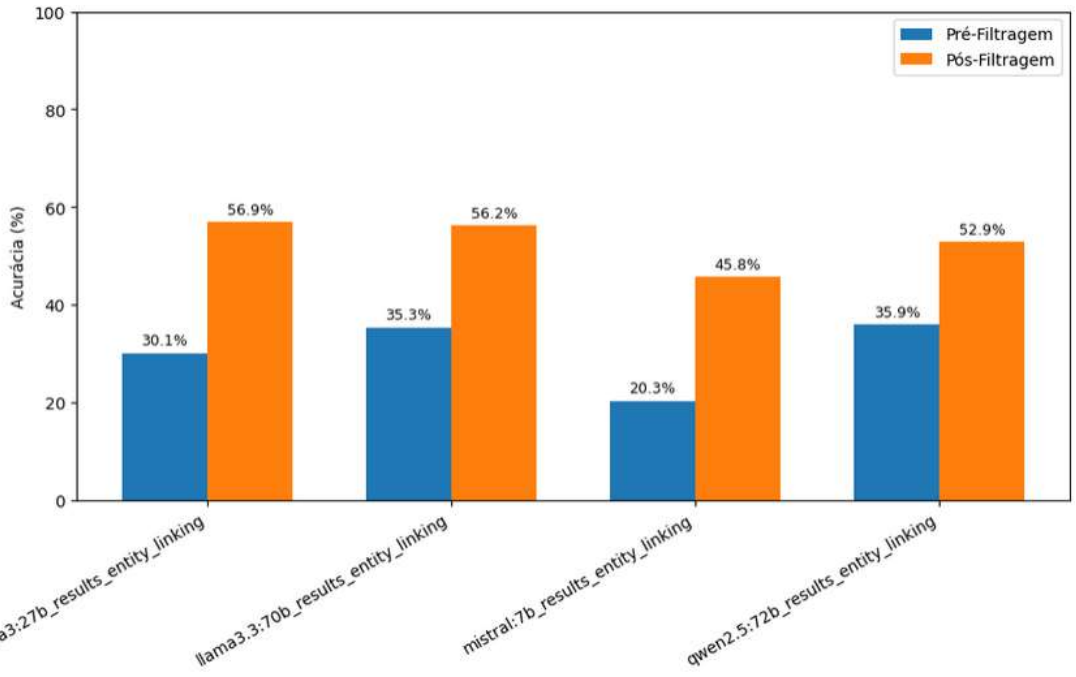
O desempenho final da redução efetuada pelas duas heurísticas é comparado pela Figura 3. A abordagem (II) reduziu em apenas 26% o número médio de candidatos para cada registro, comparado com 40% para a heurística (I), e apresentou desempenho pior para os casos mais críticos. Isso se dá pois os registros com quantidades mais elevadas de registros costumam ser misturas de dois princípios ativos diferentes, contendo padrões de apresentação mais complexos de serem contemplados por expressões regulares tão específicas.

Figura 3. Comparação da redução de candidatos entre as duas heurísticas



Elaborada pelo autor.

Figura 4. Resultado da Desambiguação de Entidades



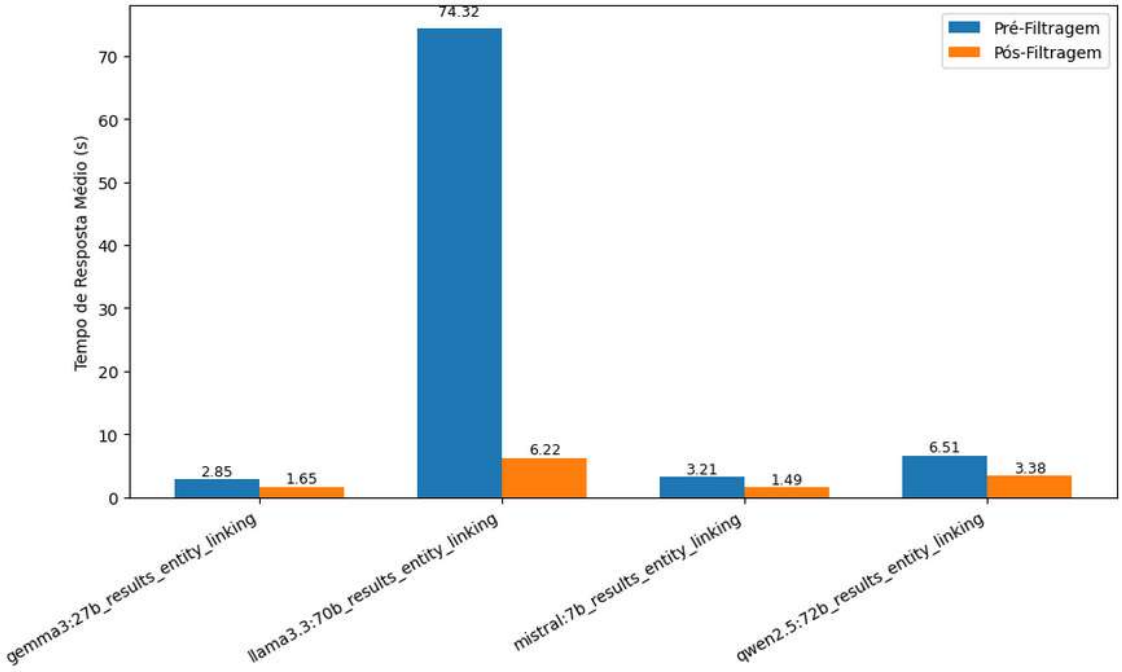
Elaborada pelo autor.

7.2. Desambiguação de Entidades

A melhoria no desempenho dos modelos é evidente na Figura 4, que apresenta a acurácia de cada modelo para a tarefa de desambiguação, com os dados pré e pós filtragem. Todos os modelos testados tiveram ganhos altamente significativos de acurácia ao lidar com uma lista de candidatos reduzida, comprovando a eficiência da etapa de filtragem para impulsionar o processo de ligação de entidades. Também é possível perceber o destaque dos modelos gemma3:27b e llama3.3:70b, que alcançaram a melhor acurácia para a tarefa. Ademais, é importante ressaltar a capacidade do modelo Gemma de conseguir os melhores resultados tendo uma quantidade de parâmetros menor que a metade dos modelos Llama e Qwen, indicando uma alta performance para a tarefa. O modelo mistral:7b, com o menor número de parâmetros, apresentou resultados inferiores aos outros modelos experimentados.

O tempo de resposta médio de cada modelo pode ser observado na Figura 5. O gráfico demonstra uma queda significativa do tempo de resposta para todos os modelos após a etapa de redução da quantidade de registros candidatos, destacando a importância da tarefa dentro do processo proposto. Os modelos mistral:7b e gemma3:27b se destacaram neste âmbito, tendo os menores tempos antes e depois da redução do número de candidatos, chegando a valores menores que 2 segundos. Evidencia-se um enorme tempo de resposta do modelo llama3.3:70b para os dados pré-filtragem, destacando a dificuldade deste modelo para lidar com a enorme quantidade de opções em relação aos outros selecionados.

Figura 5. Comparação do Tempo de Resposta entre Modelos



Elaborada pelo autor.

8. Conclusão

Descrições de notas fiscais são comumente organizadas de maneira despadronizada, o que dificulta o acesso à informação e a integração desta informação com dados de outras origens. Este trabalho avalia uma abordagem mista, usando LLMs e conhecimento de domínio sobre medicamentos registrados pela Anvisa, integrado a informações da CMED em um grafo de conhecimento, para extrair informação e desambiguar entidades de notas fiscais de medicamentos. A utilização exclusiva de LLMs para desambiguar entidades candidatas teve resultados de acurácia de no máximo 35,9%. Por outro lado, os experimentos revelaram que se pode alcançar até 56,9% de acurácia ao combinar o uso de LLMs com uma etapa prévia que usa o grafo de conhecimento e expressões regulares para reduzir a quantidade de candidatos. Esta abordagem também contribui de maneira relevante para reduzir custos computacionais e latência.

Entre os modelos avaliados, o gemma3:27b apresentou o melhor equilíbrio entre acurácia e tempo de resposta, tendo resultados competitivos com modelos bastante maiores, como o llama3.3:70b, enquanto apresenta custo computacional significativamente menor. Esse resultado demonstra o potencial de LLMs de pequeno porte para a resolução de tarefas, especialmente quando unidos à outras técnicas mais tradicionais.

Referências

IBM. Natural language processing.

Jurafsky, D. and Martin, J. H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition. Online manuscript released January 12, 2025.

Ma, Y., Hiraoka, T., and Okazaki, N. (2022). Named entity recognition and relation extraction using enhanced table filling by contextualized representations.

Rea, L. and Parker, R. (2012). *Designing and Conducting Survey Research: A Comprehensive Guide*. Wiley.

Yuan, H., Yuan, Z., and Yu, S. (2022). Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning.