Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística e Departamento de Informática
Programa de Especialização em *Data Science* e *Big Data*

Henrique Pereira Tesser Ortiz

Melhorando o Acesso a Informações Farmacêuticas Contextuais

Curitiba
2025

Henrique Pereira Tesser Ortiz

Melhorando o Acesso a Informações Farmacêuticas Contextuais

Artigo apresentado ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Wagner Hugo Bonat

Curitiba

2025

Melhorando o Acesso a Informações Farmacêuticas Contextuais

Improving Access to Contextual Pharmaceutical Information

Henrique Pereira Tesser Ortiz[1], Wagner Hugo Bonat[2]

[1] Aluno do programa de Especialização em Data Science & Big Data, henriqueortiz.1@gmail.com

[2] Professor do Departamento de Estatística - DEST/UFPR, wbonat@gmail.com

Resumo

A complexidade e a extensão das bulas de medicamentos representam uma barreira significativa para a acessibilidade da informação farmacêutica. Para mitigar este problema, este trabalho apresenta o desenvolvimento de um sistema de *Retrieval-augmented Generation* (RAG) capaz de responder a consultas em linguagem natural. O sistema opera sobre um corpus de 8.479 bulas de medicamentos coletadas da ANVISA, que são processadas e armazenadas em um banco de dados vetorial. A arquitetura RAG utiliza um *Large Language Model* (LLM) para sintetizar respostas a partir de trechos (chunks) relevantes recuperados de forma semântica. A avaliação da acurácia semântica foi conduzida contra um *Golden Standard* de 180 perguntas e respostas, e os resultados foram mensurados através da *Similaridade de Cossenos*. A performance geral do sistema revelou uma similaridade média de 0.63, confirmando sua capacidade de gerar respostas pertinentes. Contudo, a variação da similaridade entre diferentes categorias de perguntas sugere oportunidades de aprimoramento no componente de recuperação e na estratégia de chunking para garantir a consistência e a relevância das respostas, especialmente em tópicos de maior complexidade.
Palavras-chave: Retrieval-augmented generation (RAG), Large Language Model (LLM), Bulas de medicamentos

Abstract

The complexity and length of drug package inserts pose a significant barrier to the accessibility of pharmaceutical information for the general public. To mitigate this issue, this work presents the development of a *Retrieval-Augmented Generation* (RAG) system capable of responding to natural language queries. The system operates on a corpus of 8,479 drug package inserts collected from ANVISA, which are processed and stored in a vector database. A *Large Language Model* (LLM) is utilized to synthesize responses by leveraging relevant text chunks retrieved semantically from this database. The system's semantic accuracy was evaluated against a *Golden Standard* of 180 questions and answers, with results measured using *Cosine Similarity*. The overall performance of the system revealed an average similarity of 0.63, confirming its ability to generate pertinent responses. However, the variation in similarity across different question categories suggests opportunities for improvement in the retrieval component and the chunking strategy to enhance the consistency and relevance of responses, particularly for more complex topics.
Keywords: Retrieval-augmented generation (RAG), Large Language Model (LLM), Medicine Leaflets

1 Introdução

A bula de medicamento é a principal fonte de informação para pacientes e profissionais de saúde, mas sua extensão e linguagem técnica constituem uma barreira significativa para a compreensão por parte do público geral. Isso pode levar à dificuldade de encontrar informações críticas, como contraindicações e posologia, de forma ágil e precisa. Perguntas cotidianas, como "Sou lactante, posso tomar paracetamol?" ou "A nimesulida é indicada para dor muscular?", evidenciam a necessidade de uma ferramenta que traduza e contextualize essas informações de maneira eficiente.

O objetivo deste trabalho é desenvolver e avaliar um sistema robusto que torne as informações das bulas de medicamentos mais acessíveis através de consultas em linguagem natural, utilizando um pipeline de RAG. A adoção da arquitetura RAG justifica-se por sua capacidade de combinar a robustez de um LLM com a precisão de informações factuais provenientes de uma base de conhecimento externa, minimizando alucinações e fornecendo respostas contextualmente embasadas.

2 Fundamentação Teórica

2.1 Retrieval-Augmented Generation (RAG)

O *Retrieval-Augmented Generation* (RAG) é uma arquitetura de inteligência artificial que combina a capacidade generativa de *Large Language Models* (LLMs) com um mecanismo de recuperação de informações [1]. O seu principal objetivo é gerar

respostas mais precisas, factuais e contextualizadas, minimizando o risco de "alucinações" que podem ocorrer em LLMs que dependem apenas de seus dados de treinamento. O pipeline de RAG funciona em três etapas: primeiro, a indexação de uma base de conhecimento externa; segundo, a recuperação de trechos relevantes (chunks) dessa base com base na consulta do usuário; e terceiro, a geração de uma resposta pelo LLM, que utiliza os chunks recuperados como contexto para a síntese da informação. Essa abordagem garante que as respostas sejam baseadas em dados verificáveis e específicos para a tarefa, como as bulas de medicamentos.

2.2 Large Language Models (LLMs)

LLMs são modelos de linguagem pré-treinados em vastos volumes de dados de texto e código [2]. Eles são capazes de entender, gerar e processar linguagem natural, realizando tarefas como tradução, resumo e resposta a perguntas. No contexto do RAG, o LLM atua como o componente generativo. Ele recebe a pergunta do usuário juntamente com os trechos de texto recuperados da base de dados e, a partir desse contexto, formula uma resposta coerente e informativa. A utilização da API do ChatGPT neste projeto exemplifica como um LLM pode ser empregado para sintetizar e apresentar informações de forma acessível e em linguagem natural.

2.3 Bancos de Dados para Armazenamento e Recuperação

A eficiência do sistema RAG depende fundamentalmente de uma infraestrutura de dados robusta. Neste projeto, dois tipos de bancos de dados foram empregados:

MongoDB: Um banco de dados *NoSQL* orientado a documentos, o *MongoDB* foi escolhido para armazenar os dados brutos do projeto. Sua flexibilidade permite a organização dos dados cadastrais dos medicamentos e o armazenamento das bulas em formato binário (BSON), facilitando a gestão e o acesso aos documentos originais. A sua natureza não-relacional foi ideal para a estruturação dos dados heterogêneos coletados da ANVISA.

ChromaDB: Este é um banco de dados vetorial de código aberto, otimizado para a busca por similaridade [3]. No pipeline do RAG, o *ChromaDB* [4] armazena as representações vetoriais (embeddings) dos chunks de texto extraídos das bulas. Quando uma consulta é feita, o sistema busca no banco os vetores mais próximos semanticamente ao vetor da pergunta, permitindo a recuperação rápida e eficiente dos trechos de texto mais relevantes. A busca por *similaridade de cossenos* é a técnica subjacente que garante a pertinência da recuperação de informação.

3 Materiais e Métodos

3.1 Aquisição e Gerenciamento do Conjunto de Dados

A base de conhecimento para este projeto foi constituída por bulas de medicamentos em formato PDF, obtidas do site da ANVISA (Agência Nacional de Vigilância Sanitária). Para a coleta, foi desenvolvido um script em *Python* que automatizou o processo de download em massa. As bibliotecas *requests* [5] e *BeautifulSoup* [6] foram utilizadas para interagir com a API do site e coletar as informações necessárias. Este processo resultou na coleta de 8.479 bulas, juntamente com dados cadastrais dos respectivos produtos. Os dados foram organizados e persistidos em um banco de dados *NoSQL*, o *MongoDB*. A escolha deste banco de dados foi estratégica devido à sua capacidade de armazenar dados como documentos e, de forma binária, os próprios arquivos PDF das bulas, o que facilitou o gerenciamento e a recuperação dos dados.

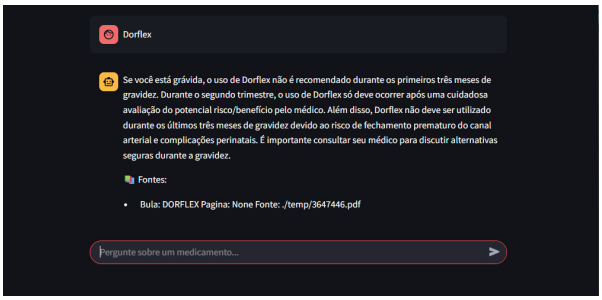
3.2 Estrutura e Processamento do Pipeline RAG

Para a etapa de prova de conceito, uma amostra de 80 bulas foi selecionada para a realização de testes de *embedding* e armazenamento no banco vetorial. As bulas foram carregadas do MongoDB, quebradas em "chunks" de texto e processadas para criar suas representações vetoriais (embeddings). Esses vetores foram então armazenados em um banco de dados vetorial, o *ChromaDB*. Essa etapa possibilitou buscas por similaridade semântica, que servem como a primeira camada do sistema RAG.

A aplicação final foi desenvolvida utilizando a API do *ChatGPT* , a biblioteca *Langchain* [7] para a orquestração do pipeline e o framework *Streamlit* [8] para a interface de usuário. O fluxo do sistema opera da seguinte forma: primeiro, o agente garante que a pergunta do usuário inclua o nome do medicamento. Em seguida, a busca é realizada no banco de dados vetorial para recuperar os "Top K Chunks" mais relevantes. Por fim, esses chunks são passados para o LLM, que os utiliza como contexto para gerar a resposta final ao usuário.

Figura 1: Pergunta realizada na Interface de Usuário

Figura 2: Resposta Obtida na Interface de Usuário



4 Resultados

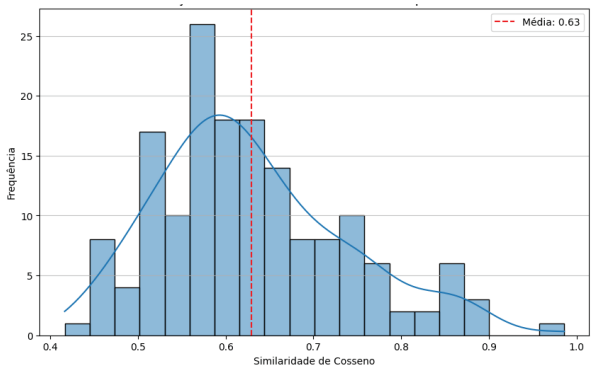
4.1 Metodologia de Avaliação

Para avaliar o desempenho do sistema, foi desenvolvido um *Golden Standard* (Conjunto de Dados de Ouro) [9]. Este conjunto consistiu na criação de 18 tipos de perguntas a serem realizadas para uma amostragem de medicamentos. Um total de 10 produtos foi selecionado para criar este conjunto de respostas de referência. No total, o Golden Standard foi composto por 180 perguntas e suas respostas ideais. A métrica utilizada para a avaliação da performance foi a *similaridade de cossenos* [10], que mede a precisão semântica das respostas do RAG em relação às respostas do conjunto de ouro.

4.2 Análise da Performance

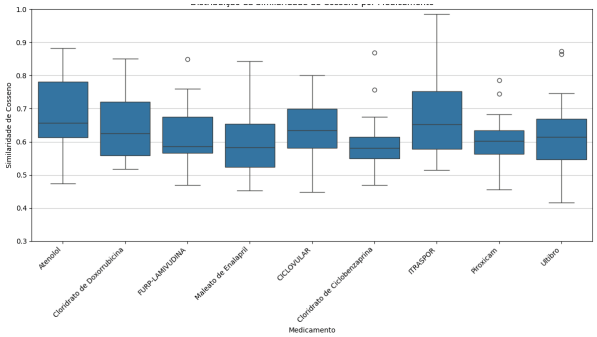
A análise geral dos resultados mostrou que a maioria das respostas geradas pelo RAG apresentou alta similaridade com as respostas ideais. A média de similaridade de cossenos foi de 0.63, indicando que o sistema geralmente entrega informações relevantes e semanticamente próximas ao esperado. A distribuição da similaridade de cossenos demonstrou que uma parte significativa das respostas se concentra em um patamar que sugere adequação e utilidade.

Figura 3: Distribuição da Similaridade de Cosseno das Respostas RAG



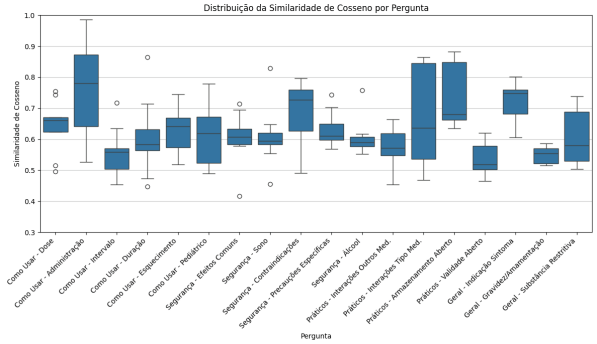
A performance do RAG variou ligeiramente entre diferentes medicamentos, o que sugere que a complexidade de cada bula pode influenciar a qualidade da resposta. Conforme ilustrado no gráfico de distribuição por medicamento, a similaridade de cossenos apresentou variações, com alguns medicamentos, como o Atenolol e o Cloridrato de Doxorrubicina, apresentando distribuições com medianas mais altas.

Figura 4: Distribuição da Similaridade de Cosseno por Medicamento



A avaliação por categoria de pergunta (e.g., posologia, contra indicações) revelou que o sistema é muito eficaz em questões sobre administração e armazenamento, com medianas de similaridade mais altas e menor variabilidade, conforme mostra o gráfico de distribuição por tipo de pergunta. Por outro lado, há espaço para melhoria em tipos de perguntas mais complexas, como duração ou interações, onde a similaridade foi mais variável e as medianas foram menores. As similaridades mais baixas foram frequentemente relacionadas à falha na recuperação de chunks relevantes para certas perguntas.

Figura 5: Distribuição da Similaridade de Cosseno por Tipo de Pergunta



5 Discussão

A similaridade de cossenos média de 0.63 demonstra que a arquitetura RAG é uma abordagem viável para a democratização da informação farmacêutica. No entanto, os resultados também revelaram oportunidades claras de otimização. A principal limitação identificada reside no componente de retrieval. As pontuações de similaridade mais baixas foram frequentemente correlacionadas à falha do sistema em recuperar os trechos de texto mais relevantes para a consulta. Isso sugere a necessidade de aprimorar o modelo de recuperação, a estratégia de segmentação de texto (chunking) ou a indexação para

garantir que as informações pertinentes sejam sempre acessadas.

Além disso, a análise comparativa entre as respostas do RAG e o Golden Standard revelou que o estilo de resposta difere. As respostas no Golden Standard tendem a ser mais curtas e diretas, enquanto as geradas pelo RAG podem ser mais prolixas. Essa diferença na forma de escrita pode influenciar os scores de similaridade, mesmo quando a informação essencial está presente. A otimização do LLM para gerar respostas mais concisas e objetivas poderia não apenas melhorar as métricas de similaridade, mas também alinhar a resposta às expectativas do usuário. Essa adaptação é crucial para aprimorar a experiência do usuário e a confiabilidade do sistema.

6 Conclusão

Os resultados da avaliação do sistema RAG para bulas de medicamentos, utilizando a similaridade de cossenos como métrica de precisão semântica em relação a um Golden Standard, demonstram um desempenho promissor na geração de respostas relevantes e acuradas. A distribuição geral da similaridade de cossenos, com uma média de 0.63, indica que, na maioria dos casos, o sistema conseguiu gerar respostas semanticamente próximas às respostas de referência. A curva de distribuição, embora apresente variância, concentra uma parte significativa das respostas em um patamar de similaridade que sugere adequação e utilidade. No entanto, a variância observada nos resultados, especialmente em perguntas mais complexas, aponta para a necessidade de otimização do componente de Retrieval e do refinamento do estilo de geração do LLM. As similaridades mais baixas estão frequentemente relacionadas à não recuperação de chunks relevantes pelo sistema para certas perguntas. Isso aponta para a necessidade de otimizar o modelo de recuperação, a estratégia de segmentação de texto (chunking) ou a indexação. Além disso, as respostas do Golden Standard tendem a ser mais curtas e diretas do que as geradas pelo RAG, o que pode influenciar os scores de similaridade. A adaptação do estilo de geração do RAG para maior concisão e objetividade pode melhorar as métricas e alinhar-se melhor com as expectativas de respostas. Futuros trabalhos devem focar nessas áreas para aprimorar a consistência e a qualidade das respostas, consolidando o potencial da abordagem RAG para a acessibilidade da informação em saúde.

7 Referências

[1] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive NLP tasks." Advances in Neural Information Processing Systems, 2020

[2] Brown, Tom B., et al. "Language models are few-shot learners." Advances in Neural Information Processing Systems, 2020

[3] Johnson, Jeff, Matthijs Douze, and Hervé Jégou. "Billion-scale similarity search with gpus." IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019

[4] CHROMA. ChromaDB. Versão 0.4.1.

[5] REITZ, K. Requests: HTTP for Humans. Versão 2.32.3.

[6] RICHARDSON, L. Beautiful Soup 4. Versão 4.12.3.

[7] LANGCHAIN. LangChain. Versão 0.1.0.

[8] STREAMLIT. Streamlit. Versão 1.25.0. 2023.

[9] Gao, T., et al. (2024). "RAG vs. Fine-tuning: Which One to Use?"

[10] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.