

RAG: Avaliação de um sistema de recuperação da informação de medicamentos Rename baseado em LLMs

Nathalia Adriele de Lima^{1*}; Felipe Pinto da Silva²

¹ Escola Superior de Agricultura "Luiz de Queiroz" – ESALQ/USP. Estudante do curso MBA em Data Science e Analytics. Marília, São Paulo, Brasil

² Escola Superior de Agricultura "Luiz de Queiroz" – ESALQ/USP. Professor do curso de MBA em Data Science e Analytics. Rua Alexandre Herculano, 120 – Vila Monteiro; 13418-445. Piracicaba, São Paulo, Brasil

*autor correspondente: natyadriele@gmail.com

RAG: Avaliação de um sistema de recuperação da informação de medicamentos Rename baseado em LLMs

Resumo

Este trabalho de conclusão propôs-se a estudar métodos de recuperação da informação baseados em LLMs e RAG, com o propósito de analisar as potencialidades desse tipo emergente de sistema, bem como contribuir para um melhor acesso a informações de medicamentos da lista RENAME. Desta forma, um sistema protótipo foi projetado para responder dúvidas de usuários sobre medicamentos da referida lista. Este sistema RAG foi avaliado utilizando quatro métricas principais do framework Ragas: são elas as métricas de precisão de contexto, “recall” de contexto, fidelidade e relevância de resposta. Os resultados da avaliação demonstraram um desempenho promissor do sistema, com pontuações médias elevadas em precisão de contexto (0,93%) e “recall” de contexto (0,87%), indicando uma capacidade robusta de identificar e recuperar informações contextuais relevantes. A métrica de fidelidade alcançou 0,83%, o que sugere que as respostas geradas pelo sistema foram majoritariamente fiéis às informações fornecidas.

Palavras-chave: LLM; RAG; RENAME; Medicamentos; Ragas.

Introdução

A evolução da Inteligência Artificial [IA] tem proporcionado avanços significativos em diversas áreas, com destaque para o campo da medicina, saúde e informação. Nesse contexto, os sistemas baseados em “Generative Artificial Intelligence” [GenAI] oferecem oportunidades sem precedentes para aprimorar a qualidade dos cuidados médicos e otimizar processos de recuperação da informação médica.

Um dos principais desafios nessa área tem sido recuperar informações de dados médicos. Isso se deve, de acordo com Lima (2023), ao grande volume de dados de textos e à sua natureza não estruturada, que tem criado barreiras para aqueles que necessitam acessar e entender rapidamente os textos, especialmente sobre medicamentos.

No Brasil, a Relação Nacional de Medicamentos Essenciais [RENAME] é uma lista de medicamentos essenciais no Sistema Único de Saúde [SUS]. Embora esta lista seja um instrumento fundamental, o acesso a informações detalhadas sobre os medicamentos listados pode ser desafiador. O documento RENAME, em formato PDF, é extenso e semiestruturado, dificultando pesquisas aprofundadas sobre os medicamentos.

Tradicionalmente, a recuperação da informação médica se baseou em sistemas de indexação por palavras-chave, que muitas vezes não fornecem respostas contextualizadas. Recentes avanços em IA, particularmente em GenAI, “Retrieval-Augmented Generation” [RAG] e “Large Language Models” [LLMs], demonstraram potencial para revolucionar esse processo. De acordo com Nazi e Peng (2024), essas tecnologias emergem como soluções promissoras para transpor barreiras históricas, proporcionando métodos mais eficientes para

indexar, recuperar e gerar informações contextualmente relevantes a partir de textos não estruturados.

O sistema RAG representa uma evolução na forma como os modelos de linguagem interagem com bases de conhecimento externas. Conforme Gao et al. (2023), a abordagem RAG combina o conhecimento parametrizado dos LLMs com conhecimento externo não parametrizado, o que ajuda a mitigar problemas de alucinação e permite a atualização de informações em tempo real via sistemas de recuperação da informação baseados em busca vetorial, o que possibilita um aumento na precisão das respostas. Assim, ao integrar um mecanismo de recuperação de informações com a capacidade generativa dos LLMs, o RAG permite que modelos de LLMs acessem e utilizem informações relevantes de forma mais objetiva e contextualizada. Esta abordagem é particularmente promissora para aplicações que exigem acesso a informações atualizadas e específicas, como é o caso da RENAME.

Logo, a pesquisa e desenvolvimento de um sistema RAG para a recuperação de informações da RENAME apresentou-se como uma oportunidade única para melhor compreender a análise desses tipos de sistemas e para contribuir para um melhor acesso a informações da lista RENAME, de maneira a melhorar o acesso e a compreensão das informações sobre medicamentos essenciais no SUS. Ao combinar a capacidade de processamento de linguagem natural dos LLMs com um mecanismo eficiente de recuperação de dados e informação, foi possível desenvolver um sistema que não apenas fornece informações de modo eficiente e rápido, mas também de forma contextualizada e relevante para usuários, sejam estudantes, pacientes ou outros.

Ao desenvolver e avaliar um sistema de indexação e recuperação da informação baseado em LLMs e RAG para otimizar a extração e sumarização de informação de dados textuais não estruturados em contextos de medicamentos, o presente estudo teve como objetivo principal investigar e avaliar a eficácia dos sistemas RAG. E, nesse processo, como estudo de caso, contribuir para um melhor acesso a informações de medicamentos da lista, em comparação com métodos tradicionais. Especificamente, a vertente aplicada da pesquisa visou o desenvolvimento de um protótipo funcional de sistema de recuperação da informação baseado em RAG denominado “ACMR - AI de Consulta de Medicamentos RENAME”, focado em responder questões e dúvidas comuns sobre medicamentos. Por fim, a pesquisa objetivou contribuir com o desenvolvimento de métodos e boas práticas para implementar e avaliar sistemas baseados em LLMs e RAG em fluxos de trabalho de indexação, recuperação e avaliação de informações, identificando limitações e contribuições do campo da Ciência de Dados para a recuperação da informação e do conhecimento médico não estruturado, além de sugerir direções para pesquisas futuras.

Metodologia, Material e Métodos

Esta pesquisa, necessariamente, enquadrou-se em uma abordagem que combinou elementos de natureza teórica e aplicada, podendo também ser caracterizada como descritiva-exploratória. Realizou-se um aprofundamento da literatura de estado da arte para ampliar o entendimento acerca da aplicação dos métodos e ferramentas atuais para o desenvolvimento e análise de sistemas de Busca Vetorial, RAG e LLMs, voltados para a indexação e recuperação da informação no domínio da saúde. Por fim, direcionou-se a um estudo de caso aplicado com o desenvolvimento de um aplicativo protótipo de RAG.

A escolha pelo método de pesquisa bibliográfica foi respaldada pela necessidade de explorar um amplo espectro de conhecimentos existentes sobre os temas abordados. Nesse contexto, conforme Lima e Miotto (2007), a pesquisa e o aprofundamento bibliográfico implicaram em um conjunto ordenado de procedimentos para buscar soluções relacionadas ao objeto de estudo, evitando ser aleatório. Para Morandi e Camargo (2015), o aprofundamento bibliográfico, enquanto estudos secundários, consolida e integra os resultados de estudos primários relevantes sobre um tópico ou questão de pesquisa específica. Este foi um método assertivo para aprofundar a compreensão de uma nova abordagem de indexação, recuperação de contexto, geração e análise de textos relacionados aos medicamentos listados na RENAME, uma vez que clarificou métodos, soluções e ferramentas emergentes e fundamentais para o projeto.

Além disso, a pesquisa adotou uma postura descritiva-exploratória, alinhada à definição de Gil (2002), que argumentou que pesquisas desse tipo visam proporcionar maior familiaridade com o problema abordado, explicitando-o ou gerando hipóteses. Nesse sentido, este trabalho de pesquisa visou também fornecer uma visão objetiva de métodos e tecnologias emergentes, como busca vetorial, RAG e LLMs, aplicadas à recuperação da informação no domínio da saúde e dos medicamentos.

Ademais, a pesquisa aplicada é intimamente ligada à busca por soluções práticas, justificando, assim, a direção desta pesquisa para um estudo de caso prático que, de acordo com Yin (2001), demonstra sua utilidade ao investigar a aplicação prática e a eficácia de novos conceitos teóricos. Além disso, o autor enfatizou que não é imprescindível ter observações diretas e minuciosas como única fonte de evidência em um estudo de caso, pois este pôde integrar tanto evidências quantitativas quanto qualitativas. Como já foi pontuado, o estudo de caso desta pesquisa voltou-se ao desenvolvimento e à análise de um protótipo de sistema de recuperação de informação de medicamentos da lista RENAME.

Com relação à etapa aplicada da pesquisa, como apresentado na Figura 1, optou-se pela adoção de um método sequencial que balizou o desenvolvimento do aplicativo RAG protótipo. Esse método ou fluxo dividiu-se em cinco principais passos ou etapas.



Figura 1. Fluxo de trabalho com as principais fases de aplicação prática
Fonte: Dados originais da pesquisa

Os cinco passos foram:

1. **Formação do Dataset RENAME:** Etapa de aquisição, organização e pré-preparação dos dados, baseando-se no documento de Relação Nacional de Medicamentos Essenciais (RENAME), que, como já foi pontuado, é um importante instrumento orientador do uso de medicamentos e insumos no SUS, disponibilizado no site do Ministério da Saúde.
2. **Limpeza e Padronização dos Dados:** Etapa na qual os dados (textos estruturados e não estruturados de bulas de medicamentos e da lista RENAME) foram revisados e padronizados para garantir que fossem consistentes e de alta qualidade para o sistema RAG.
3. **Conversão e Armazenamento dos Dados Vetoriais:** Refere-se ao processamento e construção de uma base de dados de medicamentos (vetorização e armazenamento em banco de dados vetoriais).
4. **Recuperação da Informação e Geração de Contexto:** Indica a etapa da implementação do pipeline de códigos do sistema de chat de perguntas e respostas associadas à lista RENAME, por meio do uso das tecnologias de GenAI, Busca Vetorial, RAG e LLMs.
5. **Avaliação do Sistema:** Se concentra apenas na avaliação do sistema, por meio de métricas comumente usadas para avaliar soluções de arquitetura RAGs.

RENAME - Relação Nacional de Medicamentos Essenciais

A Relação Nacional de Medicamentos Essenciais (RENAME), de acordo com Brasil MS (2022), é uma lista que conta com um abrangente conjunto de medicamentos regulamentados no Brasil, determinada pelo Ministério da Saúde. A lista abrange medicamentos considerados fundamentais e de alto custo. O propósito da RENAME é disciplinar e gerir o acesso e a utilização desses medicamentos, assegurando sua

disponibilidade e promovendo o uso criterioso, visando a proteção dos pacientes, especialmente no âmbito do Sistema Único de Saúde (SUS). A elaboração da lista RENAME decorre de um trabalho colaborativo entre diversas áreas da saúde, com o objetivo de obter o melhor resultado terapêutico e otimizar os recursos públicos (MS, 2022, p. 5-12). Como mostra a Tabela 1, os pontos-chave da lista RENAME, conforme MS (2022), são:

Tabela 1. Pontos-chave da lista Rename

Pontos-chave da lista Rename	Descrição
1. Definição e objetivos da Relação Nacional de Medicamentos Essenciais (Rename)	A Rename é uma lista de medicamentos controlados e essenciais estabelecida pelo Ministério da Saúde do Brasil. Seu objetivo é regular o acesso e uso desses medicamentos, garantindo a sua disponibilidade, promovendo o uso racional e a segurança dos pacientes, especialmente no Sistema Único de Saúde (SUS) (Rename, 2022).
2. Composição e categorias da Rename	A Rename é composta por diferentes categorias de medicamentos, como imunobiológicos, antineoplásicos, medicamentos hospedeiros e outros. A lista é atualizada periodicamente com base em evidências científicas e recomendações de especialistas (Rename, 2022).
3. Importância da Rename para o SUS	A Rename é um instrumento orientador fundamental para o uso de medicamentos e insumos no SUS. Ela apresenta os medicamentos oferecidos em todos os níveis de atenção e linhas de cuidado do SUS, proporcionando transparência sobre o acesso aos medicamentos (Rename, 2022).
4. Adoção da classificação "AWaRe" - ("Access, Watch, Reserve") - Traduzidos na Rename 2022 como Acesso, Alerta e Reservado	Uma inovação importante na edição da Rename ano 2022 foi a adoção da classificação "AWaRe" da Organização Mundial da Saúde (OMS) para medicamentos antimicrobianos, visando contribuir para a redução do desenvolvimento de bactérias resistentes a esses medicamentos (Rename, 2022).
5. Histórico e regulamentação da Rename	A lista Rename, desde sua criação em 1964, destaca os atos normativos que reafirmam sua importância para o SUS, como a Política Nacional de Medicamentos e a Política Nacional de Assistência Farmacêutica (Rename, 2022).
6. Desafios e atualização contínua da Rename	Diante da complexidade das necessidades de saúde da população, da rápida incorporação de novas tecnologias e dos diferentes modelos de organização e financiamento do sistema de saúde. Existem desafios contínuos relacionados à atualização constante da lista (Rename, 2022).

Fonte: Rename (2022)

A lista RENAME serviu como base para o desenvolvimento do estudo de caso e possibilitou mapear todos os dados de medicamentos utilizados nesse trabalho.

Em continuidade, o tópico a seguir apresenta com detalhes as ferramentas e tecnologias que viabilizaram a execução desses processos ou etapas. Logo, são descritas as principais soluções técnicas que foram aplicadas para a coleta e processamento dos dados, destacando-se as APIs, frameworks, modelos e outras, além de justificar-se a escolha dessas tecnologias em função das especificidades do projeto.

Ferramentas e Tecnologias Utilizadas

Ao longo da execução do estudo, fez-se imperativo o emprego de um amplo grupo de ferramentas e tecnologias usadas desde a obtenção dos dados por meio da elaboração de scripts de raspagem/extração, bem como a utilização de modelos de LLMs e de “embedding” ou incorporação de textos. Ademais, também foram empregadas APIs e ferramentas para o desenvolvimento da aplicação e para a aferição da qualidade do sistema RAG. A Tabela 2 exibe os principais recursos que foram empregados no desenrolar da construção do aplicativo de indexação e recuperação de informações da lista RENAME.

Tabela 2. Conjunto de ferramentas e tecnologias utilizadas e testadas

Ferramenta	Descrição
Modelos LLMs (Llama3.1, gemma2:9b e Mixtral:7B)	<ul style="list-style-type: none">Modelos de linguagem foram usados para o processamento do texto de RENAME, sendo capazes de gerar textos, responder perguntas e auxiliar em tarefas complexas de NLP.
Modelos de “Embedding” (all-MiniLM-L6-v2 e nomic-embed-text)	<ul style="list-style-type: none">Modelos de “Embedding” (all-MiniLM-L6-v2 e nomic-embed-text) foram utilizados para converter texto em vetores numéricos, visando as análises semânticas e a busca de similaridade dos textos da RENAME.
Ambientes de Desenvolvimento (Visual Studio Code, Jupyter Notebooks e Colab)	<ul style="list-style-type: none">Ambientes foram utilizados para escrever, executar e depurar código, tanto no desenvolvimento local quanto em nuvem (Google Colab).
APIs de plataformas de Desenvolvimento de LLM (Ollama, Hugging Face e Groq API)	<ul style="list-style-type: none">APIs foram empregadas para acesso a modelos de linguagem, o que permitiu a integração dos LLMs, pipelines e scripts usados no desenvolvimento do aplicativo.
Avaliação da Aplicação RAG (RAGAS)	<ul style="list-style-type: none">Ferramentas de avaliação da aplicação de sistemas “Retrieval-Augmented Generation” (RAG) foram utilizadas para medir a qualidade dos modelos de resposta com base em dados recuperados.
Outras Ferramentas (Langchain, Pinecone e Streamlit)	<ul style="list-style-type: none">Frameworks e bibliotecas para a construção de pipelines de LLMs (Langchain). Armazenamento dos vetores e busca de similaridade (Pinecone). Para o desenvolvimento da aplicação interativa de AI e dados RENAME (Streamlit).

Fonte: Dados originais da pesquisa

Nesse ponto, entende-se que é necessário um aprofundamento informativo maior sobre os modelos de LLM utilizados no sistema desenvolvido: Llama 3, Gemma 2-9B, Mixtral 8-7B, assim como o modelo de "embedding" all-MiniLM-L6-v2, e a ferramenta de avaliação de sistemas RAG, o framework Ragas, empregado para avaliar a fidelidade da resposta, a relevância da resposta, a precisão do contexto e o "recall" do contexto.

Posto isso, segue mais detalhes sobre os modelos de LLM utilizados.

O gemma 2-9B, como descrito em seu repositório oficial no Hugging Face (2024), faz parte da família Gemma, que engloba modelos abertos, leves e de última geração, desenvolvidos e disponibilizados pelo Google. Trata-se de um modelo de linguagem do tipo "decoder-only", projetado para tarefas de texto para texto. Os modelos Gemma, ainda segundo o repositório, são indicados para tarefas de geração de texto, como resposta a perguntas, resumo e raciocínio. No contexto do protótipo da pesquisa, ele é o agente/LLM que tem a responsabilidade de reescrever a pergunta feita pelo usuário, para que esta seja mais clara e objetiva, visando fornecer ao usuário uma resposta mais precisa e relevante.

Já conforme Mistral (2023), o Mixtral 8-7B é um modelo licenciado sob Apache 2.0, que superou o Llama 2 70B em benchmarks, com inferência 6 vezes mais rápida. Utilizando apenas 12,9B dos seus 46,7B de parâmetros por token, ainda de acordo com Mistral (2023), ele oferece um desempenho eficiente em tarefas multilíngues, além de ter forte capacidade de seguir instruções, com desempenho comparável ao GPT-3.5. No contexto do aplicativo desenvolvido nessa pesquisa, ele foi o LLM responsável pela geração das respostas apresentadas ao usuário, tendo como base os vetores recuperados por similaridade de cosseno do banco de dados vetorial que armazena dados da lista RENAME.

Por sua vez, de acordo com a Meta (2024), o Llama 3 é um modelo de linguagem natural avançado, pertencente à nova geração de modelos abertos da empresa. Ele foi projetado para oferecer capacidades de última geração em várias tarefas, como raciocínio, uso de ferramentas, tradução multilíngue e síntese de texto. Ainda de acordo com a Meta (2024), o modelo possui suporte a oito idiomas, além da capacidade de processar até 128K tokens de contexto. O Llama 3 é adequado para aplicações complexas, incluindo agentes conversacionais multilíngues e assistentes de programação, sendo competitivo com outros grandes modelos de linguagem, tanto abertos quanto fechados.

No protótipo, o Llama 3 age como modelo crítico/avaliador, usado via Ragas, para avaliar o pipeline RAG e dar feedback sobre o sistema e dados. Esse referido conjunto de dados, gerado sinteticamente para esse procedimento de avaliação, continha as seguintes colunas: perguntas, que foram as perguntas com base nas quais o pipeline RAG foi avaliado;

o contexto, que é o contexto recuperado e passado para o LLM gerador responder à pergunta; e a resposta correta para as referidas perguntas.

Ainda no que se refere aos modelos, especificamente, o modelo de "embedding" usado tanto no pipeline de tokenização, "chunks" e vetorização dos dados quanto no aplicativo/sistema para a busca e recuperação vetorial do banco de dados vetorial, foi o modelo denominado all-MiniLM-L6-v2. Que, de acordo com a página oficial do modelo no repositório Hugging Face, o sentence-transformers/all-MiniLM-L6-v2 (2024) é um modelo "transformer" para vetorização de textos, que mapeia sentenças e parágrafos para um espaço vetorial denso de 384 dimensões, sendo adequado para tarefas como agrupamento semântico e pesquisa vetorial.

Por fim, o Ragas, que, conforme descrito na página oficial do Ragas (2024), é uma estrutura projetada para ajudar na avaliação de pipelines de sistema RAG. Ainda de acordo com Ragas (2024), ele fornece técnicas e ferramentas que facilitaram o aprendizado contínuo de aplicativos RAG, uma vez que permite a geração sintética de conjuntos diversificados de dados de teste e a aplicação de métricas de avaliação orientadas por LLMs para mensurar objetivamente o desempenho do sistema, gerando insights como o número de alucinações, que ajudaram a iterar e aprimorar o aplicativo.

Conforme ilustrado no diagrama da Figura 2, intitulado "ragas score", o Ragas fornece métricas personalizadas para avaliar cada componente do pipeline RAG separadamente. Esta avaliação se divide em duas abordagens principais: "geração" e "recuperação". Na abordagem de geração, o Ragas avalia dois aspectos: a "fidelidade", que se refere à precisão factual da resposta gerada, e a "relevância da resposta", que diz respeito à pertinência da resposta em relação à pergunta.

Já na abordagem de recuperação, dois critérios são considerados: a "precisão do contexto", definida como a proporção entre sinal e ruído no contexto recuperado, e a "recuperação do contexto", que mede a capacidade do sistema em recuperar todas as informações relevantes para responder à pergunta. Assim, o Ragas se concentra tanto na qualidade da geração quanto na eficácia da recuperação da informação.

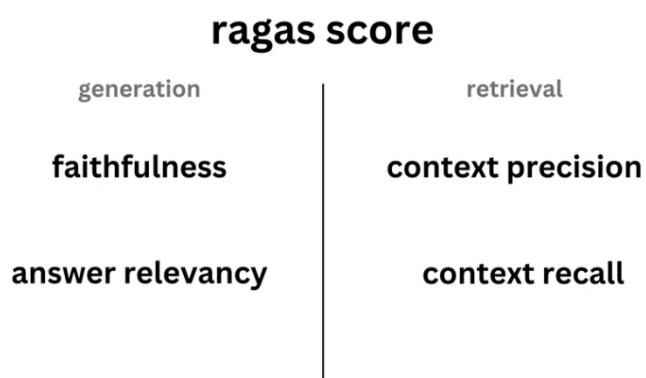


Figura 2. Diagrama ragas score
Fonte: Ragas (2024)

Posto isso, no contexto do app RENAME, o Ragas foi usado para a aplicação de métricas de avaliação orientadas por LLM. Atualmente, o Ragas disponibiliza 8 tipos de métricas de avaliação, sendo que as métricas de avaliação usadas no projeto foram: "faithfulness", "answer_relevancy", "context_recall" e "context_precision".

Conforme a página oficial do Ragas (2024) e com o propósito de detalhar um pouco mais cada uma das métricas usadas, segue:

"Faithfulness", ou Fidelidade, mede, como já foi pontuado, a consistência factual da resposta gerada em relação ao contexto fornecido, sendo calculado a partir da resposta e do contexto recuperado, com uma pontuação que varia entre 0 e 1, onde valores mais altos indicam melhor fidelidade. A resposta é considerada fiel se todas as alegações contidas nela poderem ser inferidas do contexto apresentado. Para determinar isso, um conjunto de alegações da resposta é inicialmente identificado e, em seguida, cada uma delas é verificada quanto à possibilidade de inferência a partir do contexto (Ragas, 2024). A pontuação de fidelidade é calculada por meio da form. (1):

$$\text{Pontuação de fidelidade} = \frac{|\text{Nº de afirmações na resposta gerada que podem ser inferidas do contexto fornecido}|}{|\text{Nº total de afirmações na resposta gerada}|} \quad (1)$$

De acordo com Ragas (2024), a métrica de avaliação denominada "Answer Relevancy" (Relevância da Resposta) concentra-se em medir a pertinência da resposta em relação ao "prompt" fornecido. Respostas incompletas ou com informações redundantes recebem uma pontuação mais baixa, enquanto respostas altamente relevantes obtêm uma pontuação mais elevada. Essa métrica é calculada utilizando a pergunta, o contexto e a resposta dada pelo sistema. A relevância da resposta é expressa pela similaridade média do consenso entre a pergunta original e uma série de perguntas sintéticas, geradas a partir da resposta (Ragas, 2024). Essa pontuação é calculada pela form. (2):

$$\text{Relevância da resposta} = \frac{1}{N} \sum_{i=1}^N \cos(E_{gi}, E_o) \quad (2)$$

Onde: E_{gi} é a incorporação da questão gerada i ; E_o é a incorporação da questão original; e N é o número de perguntas geradas, que é 3 por padrão.

Conforme descrito por Ragas (2024), o "Context Recall", ou Recall do Contexto, avalia o grau em que o contexto recuperado correspondeu à resposta anotada, considerada como a verdade fundamental. Ele é calculado com base na questão e no contexto recuperado, com valores variando entre 0 e 1 — quanto mais próximo de 1, melhor o desempenho. O recall é estimado verificando se cada afirmação na resposta anotada pôde ser associada ao contexto recuperado. Idealmente, todas as afirmações da resposta devem ser atribuídas ao contexto recuperado (Ragas, 2024). O recall do contexto é calculado por meio da form. (3):

$$\text{Recall do contexto} = \frac{|\text{Afirmações GT (ground truth context) que podem ser atribuídas ao contexto}|}{|\text{Nº de afirmações em GT}|} \quad (3)$$

Finalmente, conforme descrito na página oficial do Ragas (2024), o "Context Precision", ou Precisão de Contexto, é uma métrica que mede se todos os itens relevantes da verdade fundamental presentes no contexto foram classificados nas posições mais altas. O ideal é que todos os blocos relevantes apareçam nas primeiras posições do ranking. Essa métrica é calculada usando a pergunta, a verdade fundamental e o contexto, com valores variando entre 0 e 1 — pontuações mais altas indicam maior precisão (Ragas, 2024). As fórmulas para calcular a precisão, utilizada na fórmula anteriormente descrita, e a precisão de contexto, form. (4) e form. (5), são:

$$\text{Precisão de contexto@K} = \frac{\sum_{k=1}^K (\text{Precisão@k} \times v_k)}{\text{Nº total de itens relevantes no topo K resultados}} \quad (4)$$

$$\text{Precisão@k} = \frac{\text{verdadeiros positivos@k}}{(\text{verdadeiros positivos@k} + \text{falsos positivos@k})} \quad (5)$$

Em que: k é o número total de blocos em "retrieved_contexts"; e $v_k \in \{0, 1\}$ é o indicador de relevância no rank k .

Considerando as métricas descritas e sua relevância para a avaliação de desempenho, e com o intuito de aprofundar e discutir um pouco mais sobre o panorama do atual estágio dos sistemas RAGs, segue uma breve explanação.

LLMs, Busca Vetorial e Retrieval-augmented Generation

A ascensão recente dos grandes modelos de linguagem, conhecidos como LLMs ("Large Language Models"), como o GPT-2, GPT-3 e GPT-4, desenvolvidos pela OpenAI, bem como os modelos de código aberto Llama, Llama 2 e Llama 3, criados pela Meta, marcou o início de uma nova era no campo da inteligência artificial. Conforme descreveram Xu e Yiqiang (2024), esses modelos de LLMs foram treinados em grandes volumes de dados e demonstraram uma capacidade sem precedentes de gerar textos coerentes e contextualmente relevantes, ampliando os horizontes das possibilidades da AI em tarefas de linguagem e geração de imagens. Essa característica específica dos modelos de LLMs tem possibilitado a criação de inúmeras aplicações, dentre elas os sistemas de indexação e recuperação de dados não estruturados, para os mais variados propósitos e domínios. Até então, a tarefa de recuperação desses dados puramente textuais era muito complexa, problemática e limitada.

Essa evolução, como ilustrado na Figura 3, começou em 2017 com a introdução da arquitetura Transformer, apresentada no artigo seminal Attention Is All You Need, de Vaswani et al. (2017). Nessa obra, os autores propuseram uma arquitetura de rede neural baseada em um mecanismo chamado "self-attention" (ou "auto-atenção" em português). De acordo com Vaswani et al. (2023), essa inovação permitiu que o modelo atribuísse pesos diferentes a cada parte dos dados de entrada, eliminando a necessidade de mecanismos como recorrência e convoluções neurais, que até então eram as bases dos modelos de processamento de linguagem natural. Ainda conforme Vaswani et al. (2023), esses padrões anteriores, quando comparados à arquitetura Transformer, mostraram-se menos eficientes em termos de qualidade e capacidade de processamento.

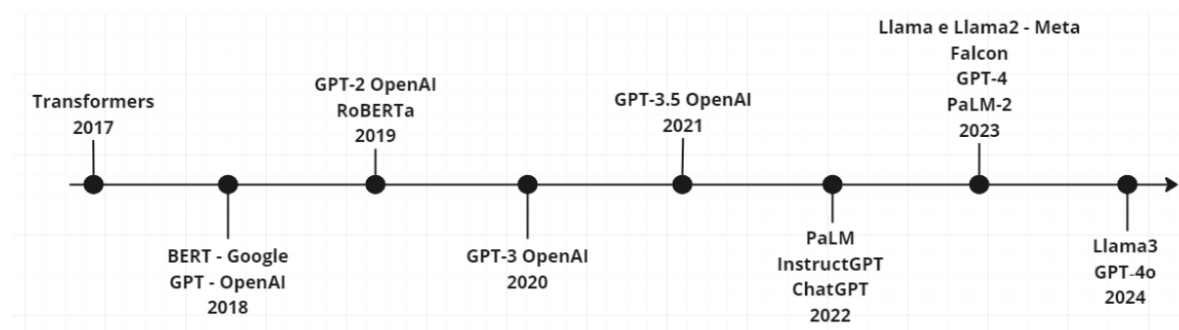


Figura 3. Linha de evolução dos modelos de LLMs

Fonte: Resultados originais da pesquisa

Optou-se por usar nesta pesquisa somente modelos de código aberto, utilizados via Ollama e API Groq. O modelo Mixtral-8x7B foi usado para geração, o Gemma-7B para

reformular as perguntas, o Llama 3 como modelo crítico para avaliação do sistema RAG e o all-MiniLM-L6-v2, para a vetorização e recuperação dos vetores. Essa escolha foi feita devido aos altos custos de uso dos modelos proprietários da OpenAI. Essa escolha possibilitou a realização de um número maior de testes ao longo do desenvolvimento do protótipo.

Na esteira da evolução dos LLMs, emergiram os bancos de dados vetoriais, que empregam vetores como unidades primárias de armazenamento e recuperação da informação e do conhecimento, além dos sistemas RAG, que auxiliam e otimizam os LLMs e as pesquisas semânticas, recuperando informações para gerar respostas com menos eventos de alucinações. Essas duas ferramentas formaram a base da indexação e recuperação de dados do sistema desenvolvido neste projeto.

A busca vetorial, ou busca semântica, conforme Schwaber (2023), envolve encontrar vetores semelhantes a um vetor de consulta, representando o que se busca saber ou encontrar. Uma pergunta é convertida em um vetor de consulta, e a similaridade entre vetores é quantificada por funções como a distância euclidiana ou a similaridade de cosseno, determinando o quão próximos os vetores estão, e, assim, permitindo a recuperação eficiente de informações.

Além disso, as estruturas de coleções de vetores permitem a recuperação eficiente de objetos semelhantes em um espaço métrico, utilizando métodos como a recuperação aproximada e a busca por top-k semelhantes. Essas abordagens são amplamente exploradas na construção de sistemas baseados em modelos de LLMs, devido à rapidez, escalabilidade, versatilidade e utilidade das representações vetoriais na recuperação de informações (Bruch, p. 1-6, 2024).

Neste contexto, os bancos de dados vetoriais desempenharam um papel crítico, pois provêm uma infraestrutura pronta, escalável e robusta para armazenar e recuperar essas representações vetoriais. Assim, eles são projetados para lidar com grandes volumes de dados vetoriais, otimizando operações de similaridade e busca semântica. Exemplos notáveis de bancos incluem o Pinecone, Weaviate, Milvus e o Chroma, que oferecem soluções eficientes para consultas em alta dimensionalidade, utilizando técnicas de indexação e aproximação, especialmente a similaridade de cosseno.

Todas essas novas aplicações, de acordo com Schwaber (2023), dependem de incorporações vetoriais, denominadas "embeddings", que carregam informações semânticas essenciais para compreensão e memória de longo prazo, necessárias para tarefas complexas. Esses "embeddings", conforme Schwaber (2023), são gerados por modelos de IA desenvolvidos com o propósito específico de criar vetores de matrizes numéricas que capturam o significado semântico.

Devido às múltiplas características desses vetores, sua gestão se tornou complexa. Desta forma, para lidar com esses dados, foram necessários bancos de dados especializados, como os anteriormente citados, que ofereceram recursos específicos para armazenamento e consulta de "embeddings", algo que os bancos de dados tradicionais não conseguiram fazer.

Logo, a representação vetorial com bancos de dados especializados aprimorou significativamente a capacidade de armazenamento, busca e recuperação de informações de textos não estruturados. Nesse contexto de tecnologias e técnicas emergentes para geração, representação, armazenamento e recuperação da informação e do conhecimento, surgiu também a técnica, ou método, de "Retrieval-Augmented Generation" (RAG).

O termo RAG foi criado por Lewis et al. (2020) e introduzido em seu artigo intitulado "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". Conceitualmente, RAG foi definida como uma técnica que utiliza uma base de conhecimento externa para complementar o contexto dos LLMs e gerar respostas. Conforme Gao et al. (2023), a abordagem RAG combina o conhecimento parametrizado dos LLMs com conhecimento externo não parametrizado, o que ajuda a mitigar problemas de alucinação (comuns nos atuais modelos LLM) e potencializa a recuperação, aumentando a precisão das respostas.

Como mostra o fluxograma da Figura 4, a arquitetura genérica RAG funciona ou flui da seguinte maneira: as consultas ("queries") dos usuários, que podem ter diferentes modalidades, servem como entrada ("input") tanto para o mecanismo de recuperação ("retriever") quanto para o gerador LLM ("generator"). O mecanismo de recuperação busca fontes de dados relevantes no armazenamento (bancos de dados vetoriais), enquanto o gerador interage com os resultados da recuperação e, finalmente, produz resultados para o usuário ("output"), de acordo com a consulta.

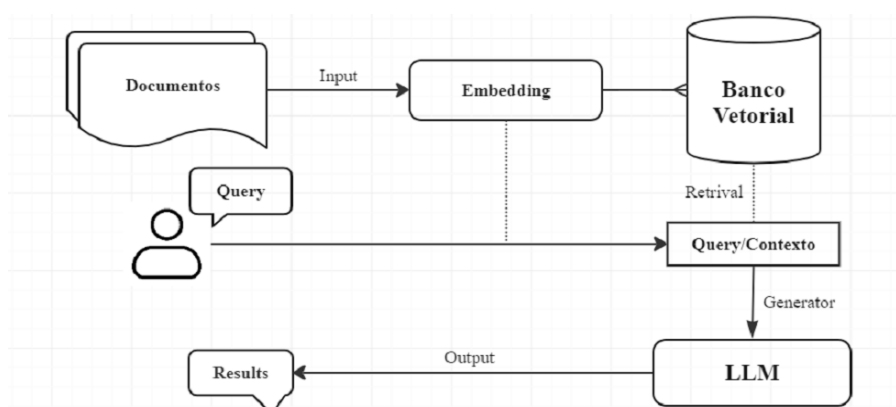


Figura 4. Arquitetura genérica de RAG
Fonte: Resultados originais da pesquisa

Desta forma, a técnica de RAG tem sido importante no contexto de aplicação dos modelos de LLMs, uma vez que, como mostrou a Figura 4, combinou o poder dos LLMs com uma base de conhecimento externa ao modelo, melhorando a abrangência, o contexto e a precisão das respostas, o que ajudou a mitigar problemas de alucinação (informações incorretas ou irrelevantes). Além disso, ao integrar fontes externas de conhecimento, os sistemas RAGs permitem uma atualização contínua das informações. Essa atualização mantém as respostas relevantes e aumenta a confiabilidade do modelo.

Com o objetivo de esclarecer como os dados foram obtidos para alimentar essas soluções, no próximo tópico, é descrito os processos e procedimentos para a aquisição do conjunto de dados, abordando as fontes de coleta, os critérios de seleção e os métodos empregados para garantir a qualidade e relevância das informações. Assim, pretende-se fornecer uma visão clara de como os dados foram integrados às ferramentas/métodos mencionadas anteriormente, para alcançar os resultados esperados.

Aquisição do Conjunto de Dados

O “dataset” coletado e usado neste projeto apresenta uma vasta e detalhada coleção de informações sobre 534 medicamentos cuja denominação genérica está incluída na lista RENAME, que, como já foi observado, é o principal guia de referência para medicamentos considerados indispensáveis ao atendimento das necessidades básicas de saúde no Brasil. Para cada um desses medicamentos, o “dataset” trouxe dados abrangentes, como o princípio ativo, que é a substância responsável pela ação terapêutica. Além disso, são descritas as indicações, especificando para quais condições clínicas o medicamento é recomendado, assim como as contraindicações, que apontam os casos em que seu uso deve ser evitado.

O “dataset” também aprofunda na posologia, detalhando as doses recomendadas e a frequência de administração, e incluiu informações cruciais sobre como utilizar corretamente o medicamento, desde o método de aplicação até o ajuste de doses em casos especiais, como para idosos ou pacientes com doenças crônicas. Outra seção importante refere-se às reações adversas, onde são listados os possíveis efeitos colaterais que podem ocorrer durante o tratamento, variando desde os mais comuns e leves até os mais graves e raros, proporcionando uma compreensão clara dos riscos associados.

Além disso, são apresentadas as interações medicamentosas, descrevendo como o medicamento em questão pode vir a interagir com outros fármacos, alimentos ou substâncias, o que é essencial para evitar efeitos indesejados ou diminuição da eficácia terapêutica. A farmacocinética e a farmacodinâmica também foram detalhadas, fornecendo insights sobre

como o medicamento é absorvido, distribuído, metabolizado e excretado pelo corpo, bem como sua ação no organismo em termos de mecanismos biológicos.

Essas informações foram cuidadosamente extraídas via raspagem de bulas originais dos medicamentos, garantindo precisão e confiabilidade nos dados apresentados. As bulas são de fontes formais e confiáveis, que contêm informações exaustivas sobre o uso seguro dos medicamentos, seus efeitos terapêuticos e as precauções que devem ser tomadas, assegurando que o usuário ou profissional de saúde tenha uma compreensão clara e completa dos fármacos.

Além dos dados específicos de cada medicamento, o dataset também incluiu textos originais que contextualizam a RENAME, explicando seu propósito como uma lista oficial de medicamentos essenciais que orienta a política de saúde pública no Brasil via SUS.

Logo, o conjunto de dados do sistema RAG oferece uma visão abrangente do papel da lista na promoção do acesso a medicamentos essenciais, destacando sua importância na garantia de tratamentos eficazes e acessíveis para toda a população. Desta forma, o conjunto de dados não só pode oferecer uma análise técnica e detalhada dos medicamentos, como também um panorama do papel estratégico da RENAME na organização e no fornecimento de cuidados de saúde no Brasil.

Após a aquisição do dataset, foi realizada uma pré-análise e o processamento desses dados puramente textuais. Posto isso, o próximo tópico descreve com detalhes o fluxo de processamento dos dados utilizados visando garantir a integridade, precisão e usabilidade desses dados, e, como essas abordagens foram essenciais para a construção de uma base de dados vetorizada, robusta e confiável para usos subsequentes.

Processamento dos Dados

O objetivo deste tópico é descrever detalhadamente o fluxo de processamento de dados utilizado para o projeto de RAG, com foco na descrição da preparação dos dados para a construção da base de conhecimento armazenada em vetores. Foi implementado um processo robusto de indexação e vetorização dos dados extraídos dos documentos de bulas de medicamentos, de modo que permitisse consultas eficientes e contextualizadas.

O processamento dos dados utilizados no projeto foi conduzido para assegurar que as informações extraídas fossem devidamente estruturadas e preparadas para indexação e busca vetorial ou consulta semântica. Esse processo começou com a instalação e configuração de várias bibliotecas Python essenciais, como LangChain (um framework de código aberto com estrutura flexível que facilita o desenvolvimento de aplicativos baseados

em LLMs e voltados para raciocínio e recuperação e reconhecimento de contexto), "Pinecone" (banco de dados vetorial especializado em indexação e consulta de vetores de alta dimensão), "Sentence-Transformers" ou SBERT (modelos pré-treinados de "embeddings" de texto, especificamente, o modelo all-MiniLM-L6-v2 para este projeto) e o NLTK (conjunto de bibliotecas e programas para processamento de linguagem natural), que forneceram as ferramentas necessárias para manipulação, segmentação e vetorização dos textos presentes nas bulas dos medicamentos. Como apresentado na Figura 5, o processamento e/ou vetorização dos dados seguiu um fluxo bem definido.

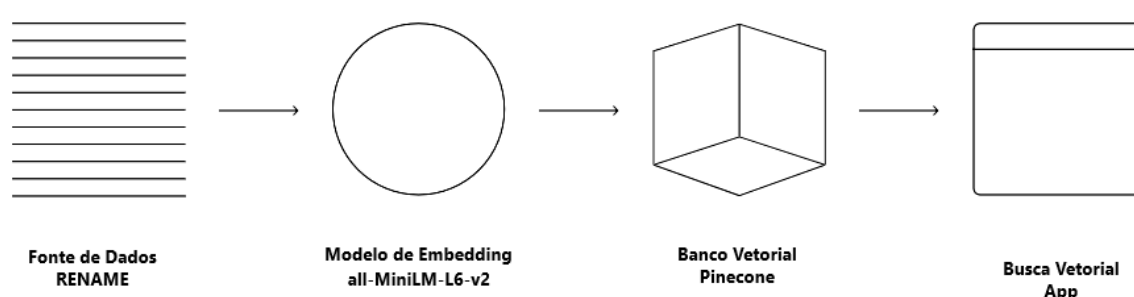


Figura 5. Fluxo para vetorização dos dados
Fonte: Adaptado de Pinecone (2024)

A primeira etapa envolveu o carregamento dos documentos em formato de texto não estruturado. Esses documentos, armazenados em diretórios, foram lidos utilizando a biblioteca LangChain, que automatizou o processo de ingestão dos arquivos. Como já foi pontuado, o conjunto de dados inclui informações detalhadas sobre 534 medicamentos da lista RENAME. Após o carregamento, verificou-se o número total de documentos, garantindo que todos os dados esperados estavam disponíveis para o processamento subsequente.

Devido ao tamanho e à complexidade dos textos, foi necessário dividir os documentos em blocos menores, denominados "chunks". Essa divisão foi feita de forma a manter a continuidade e a integridade das informações, utilizando sobreposições controladas para garantir que nenhum conteúdo relevante fosse perdido ou cortado. Esse procedimento padrão no desenvolvimento de RAG facilitou o processo de vetorização, que converteu os textos em representações numéricas em um espaço vetorial.

A vetorização foi realizada por meio de um modelo de "embeddings", que converteu cada pedaço de texto em um vetor de dimensão fixa. Esses vetores numéricos foram utilizados para calcular similaridades semânticas (similaridade de cosseno) entre os documentos, possibilitando buscas eficientes e precisas. O modelo selecionado para a vetorização foi o "all-MiniLM-L6-v2", uma escolha "open source" que ofereceu um bom

equilíbrio entre desempenho e eficiência, dado o tamanho e a natureza dos dados textuais do projeto.

Com os textos vetorizados, os vetores foram então armazenados no banco de dados Pinecone. Foi criado um índice específico para o projeto, que utilizou a métrica do cosseno para medir a similaridade entre os vetores. Esse índice foi hospedado no ambiente de nuvem AWS em modo "Serverless", o que garantiu uma escalabilidade eficiente para o volume de dados e para consultas (testes) com um bom tempo de latência.

Após a criação do índice de vetores, o sistema ficou pronto para realizar buscas semânticas, permitindo a realização de consultas contextualizadas baseadas no conteúdo dos documentos, como, por exemplo, a busca por uma descrição específica da lista RENAME dentro dos documentos vetorizados, onde o modelo foi capaz de retornar rapidamente as seções mais relevantes dos textos via busca por similaridade.

Desta forma, o processamento de dados descrito permitiu que as informações extraídas das bulas fossem estruturadas e disponibilizadas para consulta de forma robusta. O uso de técnicas avançadas de vetorização semântica e a integração com o banco de dados vetoriais Pinecone possibilitou que o sistema fosse capaz de realizar consultas complexas e com uma precisão aceitável neste estágio do projeto, garantindo que os dados dos medicamentos da lista RENAME pudessem ser acessados de forma contextualizada, rápida e relevante. Esse processo contribuiu significativamente para a eficiência do projeto, assegurando que os dados coletados fossem utilizados de maneira otimizada e organizada.

Após o processamento do conjunto de dados, no qual os dados textuais foram transformados em dados vetoriais, a etapa subsequente foi aferir a precisão e consistência dos resultados. Logo, o próximo tópico descreve a validação, onde foram apontados o método e as métricas usadas para verificar a qualidade do sistema RAG.

Validação

A validação do aplicativo RAG baseou-se em uma avaliação sistemática que utilizou um conjunto de dados para testar a precisão e a relevância das respostas fornecidas pelos modelos usados para vetorização e geração textual do sistema. Para isso, foi utilizado um conjunto de dados de teste que foi gerado sinteticamente, a partir dos dados originais, por meio do uso do framework Ragas especializado em avaliação RAG, estes dados consistem em perguntas e respostas relacionadas aos medicamentos da lista RENAME.

Essa abordagem permitiu avaliar o desempenho de modelos de LLMs quanto à precisão da recuperação, geração e aderência ao conteúdo original da lista, validando as

métricas de fidelidade, relevância da resposta, precisão do contexto e recuperação do contexto.

Os resultados da avaliação foram coletados e analisados. Essa etapa foi crucial para identificar as áreas de maior desempenho e aquelas que necessitavam de melhorias. O uso de um subconjunto menor do “dataset” permitiu uma validação mais rápida e eficaz.

O uso de métricas objetivas e padronizadas como as do framework Ragas visou assegurar que a validação fosse conduzida de forma científica e consistente, oferecendo uma visão clara das capacidades atuais do sistema e das melhorias necessárias para garantir que a implementação da RAG do sistema proposto atendesse aos altos padrões de qualidade e confiança exigidos pelo setor de saúde.

Após a conclusão da etapa de validação, que evidenciou a eficácia e a confiabilidade do sistema e dos dados processados e vetorizados, avançou-se para a fase de implementação. O próximo tópico aborda a aplicação das soluções desenvolvidas e utilizadas, integrando a base de dados ao sistema e aos modelos de LLMs. A implementação assegurou a aplicação eficiente em um ambiente prático, possibilitando a obtenção de resultados e “insights” alinhados aos objetivos do projeto.

Implementação

O processo de desenvolvimento, implementação e “deploy” do aplicativo RAG denominado ACMR (AI de Consulta de Medicamentos RENAME) seguiu um fluxo bem definido, como demonstrado na Figura 6. E, envolveu, como já foi apontado, a utilização de diversas tecnologias e processos, com o objetivo de criar um sistema eficiente e uma interface amigável para consultas sobre medicamentos incluídos na lista.

A aplicação foi desenvolvida em Python, utilizando o framework de prototipagem Streamlit para a criação da interface de usuário de aplicações de Ciência de Dados e AI. O protótipo de aplicativo permite que “usuários” façam perguntas relacionadas aos medicamentos listados na RENAME e recebam respostas contextuais e refinadas com base nos dados extraídos e vetorizados das bulas com dados reais.

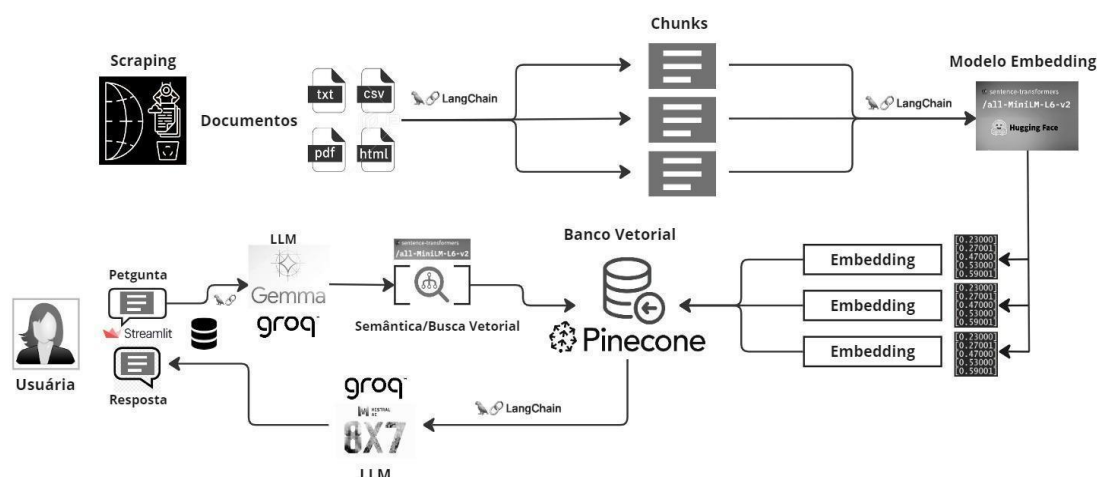


Figura 6. Fluxograma de representação do sistema RAG de medicamentos RENAME
Fonte: Dados originais da pesquisa

Depois do processamento/vetorização dos dados, uma das primeiras etapas na implementação do protótipo foi a integração com os modelos de LLMs. Para isso, foi utilizada a API Groq, que, entre outras, fornece serviços de inferência rápida de IA/LLMs "open source" na nuvem via API. Essa solução permitiu o uso dos LLMs neste projeto, provendo a geração de respostas semânticas com base no contexto fornecido.

Para garantir que a aplicação pudesse lidar com múltiplas interações e manter o histórico/encadeamento das conversas, foi utilizada uma memória de conversação, a "ConversationBufferWindowMemory" do Langchain, que armazenou as interações mais recentes, permitindo que o sistema respondesse de forma mais natural e coerente às perguntas subsequentes.

Para refinar as perguntas dos usuários e melhorar a relevância das respostas, o aplicativo tem uma função de refino de consultas, baseada em um modelo de linguagem adicional (Gemma-7B-IT), fornecido pela plataforma/API Groq. Essa abordagem ajuda o sistema, a partir da pergunta do usuário, a formular perguntas mais adequadas sem perder o sentido original das perguntas, aumentando a precisão das respostas retornadas.

Além disso, a aplicação incorpora uma ferramenta de busca semântica através da integração com a API do banco vetorial Pinecone. Esta ferramenta permitiu a criação e armazenamento de um índice vetorial dos textos das bulas dos medicamentos, o que possibilita que o sistema localize informações de forma mais eficiente e baseada em similaridade semântica. Para isso, como já foi mencionado, é utilizado o modelo de "embedding" all-MiniLM-L6-v2, da SentenceTransformer, que, através da comparação por similaridade das perguntas dos usuários e dos textos das bulas de medicamentos em vetores

numéricos, recupera k números mais similares, para então o LLM de geração retornar uma resposta semântica baseada nos dados recuperados.

O aplicativo também inclui uma funcionalidade de armazenamento em banco de dados relacional, que armazena o histórico de conversas em um banco de dados PostgreSQL. Isso foi implementado de forma eficiente com a utilização de consultas SQL que inseriram cada pergunta e resposta no banco de dados, garantindo que o histórico de interações possa ser acessado posteriormente. Essa integração é facilitada pelo uso da biblioteca `psycopg2` para a conexão e manipulação dos dados no PostgreSQL, permitindo a persistência das interações entre o usuário e o assistente.

O “deploy” da aplicação exigiu uma configuração dos serviços em nuvem disponibilizados via APIs, onde a infraestrutura de banco de dados vetorial e a API de busca semântica via modelos LLMs foram disponibilizadas. Para garantir o melhor desempenho do sistema, a aplicação foi implantada utilizando provedores de serviços em nuvem disponibilizados pela Groq e uma versão gratuita do Pinecone, usando uma instância do AWS para hospedar o índice de vetores. Também foi desenvolvida uma interface de usuário com o Streamlit (framework para ciência de dados e AI/ML), como apresentado nas Figuras 7 e 8.

Deploy ⓘ


ACMR - AI de Consulta de Medicamentos RENAME.

ACMR é um assistente inteligente para consultas rápidas sobre medicamentos da lista RENAME.

Histórico de Consultas:

Consulta 1

Pergunta: Ex: Quais as contraindicações do Cloridrato de Clorpromazina?

 Por favor, elabore sua pergunta relacionada a um medicamento da lista RENAME!

Faça uma pergunta!

Digite sua pergunta abaixo: Ex: O que é lista RENAME?

Sua pergunta

Ex: Quais as contraindicações do Acetato de Fludrocortis

Enviar

ACMR 🍌 - AI de Consulta de Medicamentos RENAME © 2024 | Desenvolvido com ❤️ e 🧠 usando Streamlit

Figura 7. Interface de usuário do sistema de consulta de medicamentos RENAME

Fonte: Dados originais da pesquisa

Deploy

Histórico de Consultas:

Consulta 1

Pergunta: Ex: Quais as contraindicações do Cloridrato de Clorpromazina?

Por favor, elabore sua pergunta relacionada a um medicamento da lista Renamel

Consulta 2

Pergunta: Quais são as reações adversas mais comuns ao usar Eculizumabe?

As reações adversas mais comuns ao usar Eculizumabe incluem reações no local da infusão, como dor, vermelhidão, inchaço e calor. Outros efeitos adversos comuns incluem cefaleia, náusea, diarreia e prisão de ventre. Algumas reações alérgicas também podem ocorrer, mas são menos comuns. É importante lembrar que cada pessoa pode reagir de maneira diferente a este medicamento, e é essencial que qualquer sintoma ou reação adversa seja relatado imediatamente ao seu médico ou equipe de saúde.

Faça uma pergunta!

Digite sua pergunta abaixo: Ex: O que é lista RENAME?

Sua pergunta

Quais são as reações adversas mais comuns ao usar Ecul

Enviar

Consulta refinada pelo Agente

Quais são os principais efeitos colaterais relacionados ao uso do Eculizumabe?

Figura 8. Interface do usuário com pergunta, reformulação da pergunta e resposta
Fonte: Dados originais da pesquisa

Esse ambiente integrado e otimizado permitiu que o sistema funcionasse de maneira consistente e eficiente, proporcionando respostas rápidas, relevantes e detalhadas sobre medicamentos da lista RENAME. Desta forma, a implementação do RAG destaca-se por sua abordagem modular, onde cada componente, desde o refinamento de consultas até a busca semântica, trabalha de forma harmônica para oferecer uma experiência de usuário intuitiva e informativa.

Com a implementação das soluções descritas, foi viável examinar os impactos e a eficácia dos métodos aplicados. Assim, no próximo tópico, são apresentados os resultados obtidos com a aplicação dessas soluções, discutindo-se detalhadamente os achados. Finalmente, é feita uma análise crítica acerca dos LLMs e RAGs, e dos resultados obtidos, avaliando-se o sucesso das abordagens utilizadas no contexto do projeto.

Resultados e Discussão

Finalmente, os resultados obtidos nesta pesquisa forneceram uma análise do desempenho do sistema de recuperação de informações de medicamentos da lista RENAME, baseado em LLMs e RAG, que foi desenvolvido como estudo de caso. Como pontuado, a avaliação foi realizada utilizando métricas disponíveis no framework Ragas. Essas métricas ofereceram “insights” sobre a capacidade do sistema em identificar contextos relevantes, recuperar informações e formular respostas precisas.

Conforme introduzido anteriormente e, de acordo com Ragas (2024). A precisão de contexto ("context_precision") é uma medida que avalia a proporção de resultados relevantes

entre todos os retornados. A recuperação de contexto ("context_recall") avaliou quantos registros relevantes em um conjunto de dados de teste foram recuperados. A fidelidade ("faithfulness") mediu a consistência factual das respostas em comparação ao contexto recuperado. Por fim, a relevância da resposta ("answer_relevancy") avaliou a pertinência das informações sobre medicamentos retornadas em relação à pergunta original.

O conjunto de dados utilizado na avaliação do sistema foi gerado sinteticamente no Ragas, a partir da extração de trechos de textos originais das bulas de medicamentos. Os dados das colunas "question", "contexts" e "ground_truth" foram extraídos diretamente dos dados de teste, enquanto as respostas ("answer") foram geradas pelo sistema via busca vetorial e geração aumentada do LLM. A Figura 9 mostra o resultado do teste do conjunto de dados sintéticos gerados, a partir de textos não estruturados extraídos das bulas.



	question	contexts	ground_truth	evolution_type	metadata	episode_done
0	O que é a lista RENAME?	[A lista RENAME, ou Relação Nacional de Medic...	A lista RENAME regula o acesso a medicamentos ...	simple	[{"source": "/content/data/dataset-texto-renam...	True
1	Qual a dosagem recomendada de Cabergolina para...	[Doses de Cabergolina de até 4,5 mg por seman...	A dose de Cabergolina pode ser de até 4,5 mg p...	simple	[{"source": "/content/data/dataset-texto-renam...	True
2	O Eltrombopague Olamina pode ser utilizado par...	[Eltrombopague Olamina está indicado para o t...	Eltrombopague Olamina é indicado para tratar a...	reasoning	[{"source": "/content/data/dataset-texto-renam...	True
3	O que deve ser feito se um paciente esquecer d...	[Se um paciente esquecer de tomar uma dose de...	Se o paciente esquecer uma dose de Hemifumarat...	multi_context	[{"source": "/content/data/dataset-texto-renam...	True
4	Quais cuidados devem ser tomados ao usar Hemit...	[Hemitartarato de Norepinefrina pode causar g...	Hemitartarato de Norepinefrina não deve ser us...	reasoning	[{"source": "/content/data/dataset-texto-renam...	True
5	Qual a função da Norepinefrina no corpo humano?	[Norepinefrina é liberada nas terminações das...	Norepinefrina é um neurotransmissor que aument...	simple	[{"source": "/content/data/dataset-texto-renam...	True
6	Como deve ser administrada a Heparina Sódica d...	[Durante procedimentos dialíticos, a Heparina...	A Heparina Sódica deve ser administrada na lin...	simple	[{"source": "/content/data/dataset-texto-renam...	True
7	Hidroclorotiazida pode ser usada durante a gra...	[A Hidroclorotiazida é classificada na catego...	A Hidroclorotiazida não deve ser usada durante...	reasoning	[{"source": "/content/data/dataset-texto-renam...	True
8	Quais efeitos adversos estão associados ao uso...	[O uso prolongado de Hidroxiuréia está associ...	O uso prolongado de Hidroxiuréia pode causar l...	multi_context	[{"source": "/content/data/dataset-texto-renam...	True
9	Quando um paciente deve buscar orientação médi...	[Após o tratamento com soro antirrábico, o pa...	O paciente deve buscar orientação médica após ...	simple	[{"source": "/content/data/dataset-texto-renam...	True
10	Quais cuidados são recomendados para pacientes...	[Em pacientes com insuficiência hepática grav...	Doses mais baixas de Cabergolina são recomenda...	simple	[{"source": "/content/data/dataset-texto-renam...	True
11	Qual a ação da Eltrombopague Olamina em pacien...	[Eltrombopague Olamina estimula a produção de...	Eltrombopague Olamina é indicado para estimular	reasoning	[{"source": "/content/data/dataset-texto-renam...	True

Figura 9. Recorte do teste Ragas do dataset

Fonte: Resultados originais da pesquisa

A Figura 10, recorte da tabela com os resultados, exibe os resultados da avaliação, destacando os dados e métricas obtidos.



	question	answer	contexts	ground_truth	context_precision	faithfulness	answer_relevancy	context_recall
0	O que é a lista RENAME?	A lista RENAME é a Relação Nacional de Medicam...	[A lista RENAME, ou Relação Nacional de Medic...	A lista RENAME regula o acesso a medicamentos ...	1.0	0.666667	0.672497	1.0
1	Qual a dosagem recomendada de Cabergolina para...	A dose de Cabergolina para pacientes hiperprol...	[Doses de Cabergolina de até 4,5 mg por seman...	A dose de Cabergolina pode ser de até 4,5 mg p...	1.0	1.000000	0.650395	1.0
2	O Eltrombopague Olamina pode ser utilizado par...	O Eltrombopague Olamina não deve ser utilizado...	[Eltrombopague Olamina está indicado para o t...	Eltrombopague Olamina é indicado para tratar a...	1.0	1.000000	0.768823	1.0
3	O que deve ser feito se um paciente esquecer d...	Se um paciente esquecer de tomar uma dose de H...	[Se um paciente esquecer de tomar uma dose de...	Se o paciente esquecer uma dose de Hemifumarat...	0.0	0.500000	0.600668	0.0
4	Quais cuidados devem ser tomados ao usar Hemit...	Hemitartarato de Norepinefrina deve ser utiliz...	[Hemitartarato de Norepinefrina pode causar g...	Hemitartarato de Norepinefrina não deve ser us...	1.0	1.000000	0.611224	1.0
5	Qual a função da Norepinefrina no corpo humano?	A Norepinefrina é um neurotransmissor que elev...	[Norepinefrina é liberada nas terminações das...	Norepinefrina é um neurotransmissor que aument...	1.0	0.750000	0.618575	0.0
6	Como deve ser administrada a Heparina Sódica d...	A Heparina Sódica deve ser administrada na lin...	[Durante procedimentos dialíticos, a Heparina...	A Heparina Sódica deve ser administrada na lin...	1.0	1.000000	0.729551	1.0
7	Hidroclorotiazida pode ser usada durante a gra...	A Hidroclorotiazida não deve ser usada durante...	[A Hidroclorotiazida é classificada na catego...	A Hidroclorotiazida não deve ser usada durante...	1.0	1.000000	0.732925	1.0
8	Quais efeitos adversos estão associados ao uso...	O uso prolongado de Hidroxiureia pode estar as...	[O uso prolongado de Hidroxiureia está associ...	O uso prolongado de Hidroxiureia pode causar l...	1.0	0.666667	0.461255	1.0
9	Quando um paciente deve buscar orientação médi...	O paciente deve buscar orientação médica imedi...	[Após o tratamento com soro antirrábico, o pa...	O paciente deve buscar orientação médica após ...	1.0	0.500000	0.453540	1.0

Figura 10. Resultados da avaliação do sistema RAG do RENAME

Fonte: Resultados originais da pesquisa

A Figura 11 mostra as pontuações médias alcançadas pelo sistema para as diferentes métricas de avaliação. As métricas "context_precision" e "context_recall" obtiveram as pontuações mais altas (0,93 e 0,86, respectivamente). Embora o recall tenha sido bom, ele ficou abaixo da precisão, o que sugere que o sistema não recuperou todas as informações contextuais relevantes, mesmo que 86% dos contextos relevantes tenham sido retornados. A precisão de contexto, que mediu a proporção de resultados relevantes ou precisão, foi de 93%. No entanto, o volume de testes foi pequeno, e alcançar uma precisão perfeita com o grande volume de dados deste projeto seria inviável, tanto pelo alto custo de processamento quanto por não ser o objetivo do projeto, nem da abordagem do framework Ragas.

A pontuação da métrica "faithfulness" foi de 0,82, o que indica que algumas respostas apresentaram menor consistência factual em comparação ao conteúdo recuperado, sugerindo espaço para melhoria na fidelidade das respostas. A métrica "answer_relevancy", com uma pontuação intermediária de 0,63, mostrou que a relevância das respostas foi apenas satisfatória, mostrando que algumas respostas foram avaliadas como incompletas ou redundantes. Por fim, o sistema alcançou uma boa média total de 81% em um conjunto de dados com 15 perguntas e respostas.

```

Média das avaliações por métrica:
context_precision: 0.9333
context_recall: 0.8667
faithfulness: 0.8274
answer_relevancy: 0.6333

Média total das avaliações: 0.8152

```

Figura 11. Resultado geral da avaliação Ragas da RAG do RENAME

Fonte: Resultados originais da pesquisa

O gráfico da Figura 12 apresenta as métricas "context_precision", "context_recall", "faithfulness" e "answer_relevancy". Observou-se que o "context_precision" atingiu 1,0 em quase todas as perguntas, demonstrando alta precisão na identificação do contexto relevante. O "context_recall", por sua vez, variou de 0,0 a 1,0, com algumas perguntas atingindo recall perfeito, enquanto duas delas apresentaram desempenho de 0.

A métrica de "faithfulness", de modo geral, foi boa, com muitos valores atingindo 1,0, embora algumas perguntas tenham registrado pontuações entre 0,50 e 1,0. A "answer_relevancy" apresentou variações de 0,50 a 0,77, com a maioria das pontuações situando-se entre 0,60 e 0,73.



Figura 12. Avaliação das métricas para várias perguntas geradas no sistema
Fonte: Resultados originais da pesquisa

O desempenho individual foi satisfatório, com muitas células verdes e verde-claras. No entanto, houve ainda algumas áreas de preocupação, particularmente em "context_recall" e "faithfulness". Em especial, os valores 0 de "context_recall" sugeriram que o sistema perdeu informações contextuais importantes para algumas perguntas, como as perguntas 3 e 5. Já os valores de "faithfulness" indicaram que, para certas perguntas, as respostas não estavam totalmente alinhadas com as informações fornecidas. Houve ainda um espaço em branco, devido a uma falha corriqueira do framework Ragas, como mostra a Figura 13.

WARNING:ragas.llms.output_parser:Failed to parse output. Returning None.

Figura 13. Bug comum do Ragas da avaliação de métricas
Fonte: Resultados originais da pesquisa

Por meio da análise do gráfico de "box plot" das métricas de avaliação da Figura 14, foi possível observar que o "context_precision" e o "context_recall" mostraram a menor variabilidade entre todas as métricas, com suas medianas em 1.0. Ambos apresentaram outliers próximos a 0, indicando alguns casos raros de desempenho extremamente baixos.

A métrica "faithfulness" apresentou a maior variabilidade, com a mediana próxima de 1.0, mas o quartil inferior desceu até aproximadamente 0.67. O valor mínimo chegou a cerca de 0.5, indicando uma variação considerável no desempenho, com alguns casos de fidelidade significativamente mais baixa.

Já a métrica "answer_relevancy" teve uma distribuição mais compacta em comparação com "faithfulness", com a mediana em torno de 0.62. O quartil superior alcançou aproximadamente 0.7, e houve dois outliers visíveis, um próximo a 0.45 e outro próximo a 0.77, mostrando que, de modo geral, a relevância das respostas foi relativamente consistente, mas com alguns casos de desempenho acima e abaixo da média.

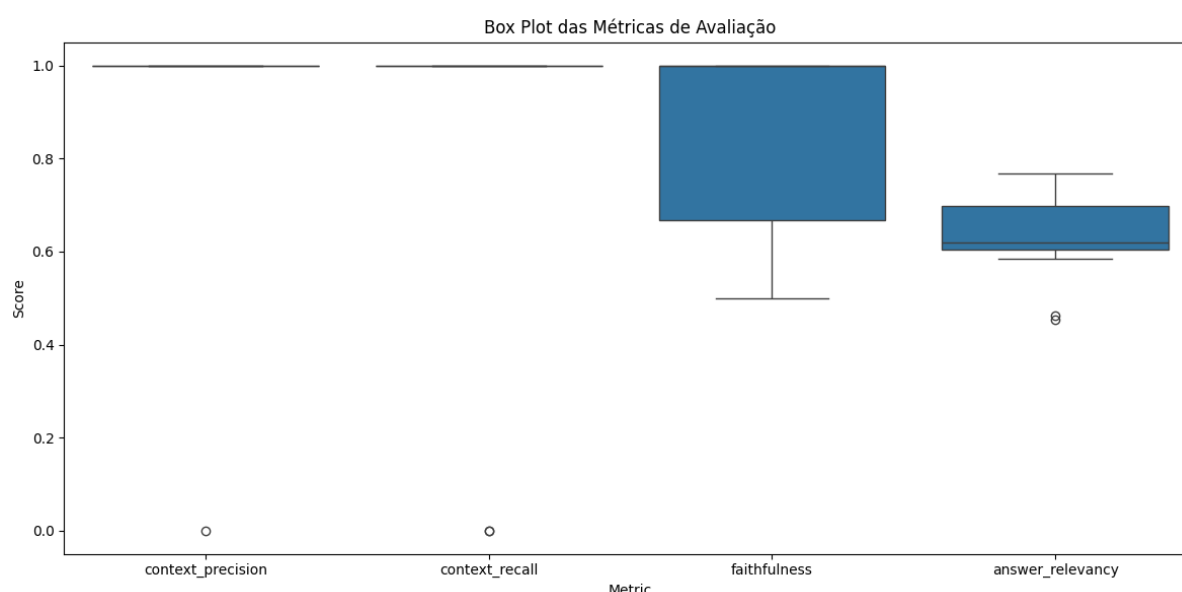


Figura 14. Diagrama de caixa comparativo das quatro métricas de avaliação

Fonte: Resultados originais da pesquisa

Analisando o gráfico de barras com as pontuações médias por métrica da Figura 15, foi possível confirmar que o "context_precision" teve a pontuação média mais alta, de 0.93, indicando que o sistema foi muito preciso na identificação do contexto relevante, embora não tenha sido perfeito, como pontuado anteriormente.

O "context_recall" teve a segunda maior pontuação, com 0.87, sugerindo um bom desempenho na recuperação de informações contextuais importantes.

Enquanto a métrica "faithfulness" mostrou uma pontuação média de 0.83, a terceira mais alta, demonstrando que as respostas foram geralmente fiéis às informações fornecidas, mas ainda houve espaço para melhorias.

Por fim, a "answer_relevancy", com a pontuação média mais baixa, de 0.63, indicou que a relevância das respostas foi a menos consistente entre todas as métricas avaliadas. O que é um ponto que requer total atenção. Uma das possibilidades para isso ter ocorrido é o comprimento reduzido dos tokens das respostas, por questões de custo de processamento.

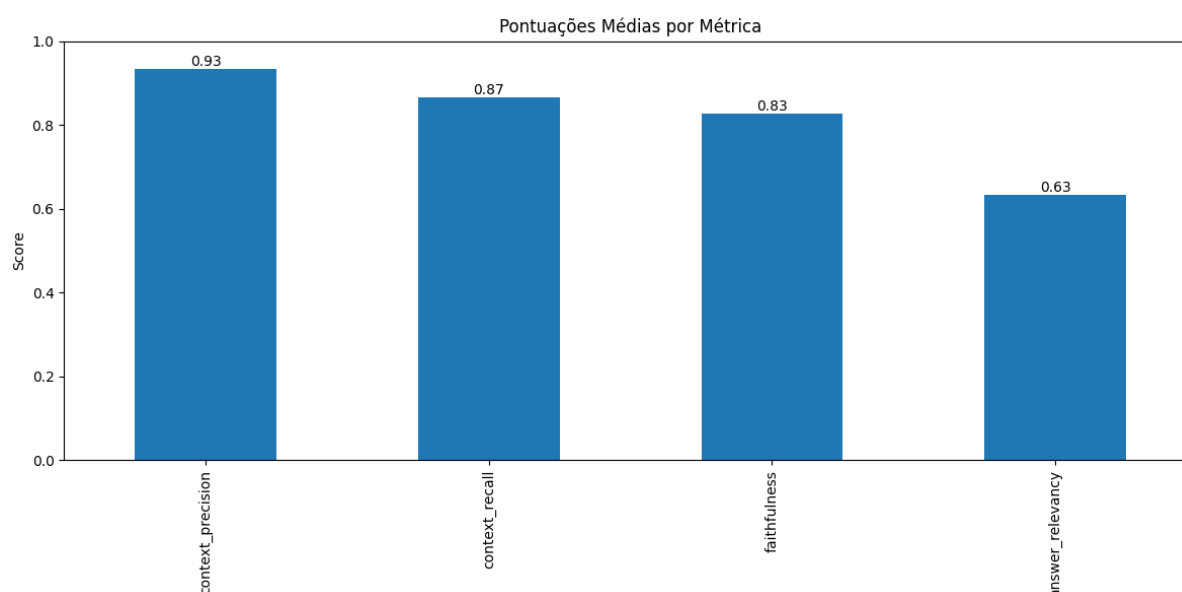


Figura 15. Gráfico de barras para comparação das pontuações médias das quatro métricas
Fonte: Resultados originais da pesquisa

Em conclusão, o sistema demonstrou um bom desempenho geral, com excelência em precisão de contexto. Para melhorar, o foco principal deve ser pequenos ajustes no "context_recall", seguido por ajustes na fidelidade e, sobretudo, na métrica de relevância das respostas, o que poderia resultar em um sistema mais equilibrado e eficaz em todas as métricas. Embora existam algumas variações, as métricas indicaram que o sistema performou consistentemente bem para a maioria das perguntas.

Nesse ponto, é importante frisar que, embora as métricas geradas pelo framework Ragas tenham sido um bom indicador de qualidade (nesse caso como prova de conceito), faz-se necessário uma curadoria humana multidisciplinar. Por se tratar de informações sensíveis, como medicamentos e seus efeitos, o sistema precisa funcionar no mais alto nível de qualidade, sem as alucinações comuns aos LLMs.

Conclusões

Os sistemas de indexação e recuperação da informação baseados em LLMs e RAG emergem como solução para antigos e persistentes problemas, como a recuperação de dados textuais não estruturados. Baseando-se nas análises realizadas e no contexto das métricas de avaliação do protótipo RAG deste estudo de caso, foi possível concluir que os sistemas de recuperação de informação baseados nessas tecnologias são promissores. As análises destacaram que essas abordagens podem aumentar a precisão na identificação de contextos relevantes, possibilitando uma busca semântica e contextualizada. Isso sugere que, quando informações são recuperadas, tendem a ser altamente pertinentes à consulta. Ou seja, essa constatação indica que tais sistemas são capazes de gerar respostas que, geralmente, se alinham bem com as informações fornecidas e são relevantes para as consultas dos usuários. O protótipo desenvolvido oferece uma base sólida para o desenvolvimento futuro de uma aplicação viável de recuperação e geração de informações sobre medicamentos de forma precisa, com oportunidades claras para aprimoramentos que poderiam elevar ainda mais sua eficácia e confiabilidade. Para o avanço desse sistema/pesquisa, é crucial focar em uma curadoria multidisciplinar de dados e informações, além de implementar técnicas de expansão de consulta e/ou aprimoramento das respostas via o uso de agentes de refinamento contínuo na geração de respostas.

Referências

BRASIL. Ministério da Saúde. Relação Nacional de Medicamentos Essenciais - Rename 2022. Conass, v.1, p. 1-10, 2024. Disponível em: <<https://shorturl.at/f77o6>>. Acesso em: 1 mar. 2024.

GAO, Ya.; MOEN, Hans.; KOIVUSALO, Salla.; KOSKINEN, Miika.; MARTTINEN, Pekka. Query-Guided Self-Supervised Summarization of Nursing Notes. ArXiv, v.1, p. 1-28, 2024. Disponível em: <<https://arxiv.org/pdf/2407.04125>>. Acesso em: 22 jul. 2024.

GIL, Antonio Carlos. Como elaborar projetos de pesquisa. 4. ed. São Paulo: Atlas, p. 41-55, 2002. Disponível em: <<https://shorturl.at/qp1eK>>. Acesso em: 4 mar. 2024.

HUGGING FACE. google/gemma-2-9b. Huggingface.co. Disponível em: <<https://huggingface.co/google/gemma-2-9b>>. Acesso em: 20 jul. 2024.

HUGGING FACE. sentence-transformers/all-MiniLM-L6-v2. Huggingface.co. Disponível em: <<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>>. Acesso em: 20 maio 2024.

LEWIS, Patrick.; PEREZ, Ethan.; PIKTUS, Aleksandra. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. v.4, p. 1-19, 2021.

LIMA, Telma Cristiane Sasso de.; MIOTO, Regina Célia Tamaso. Procedimentos metodológicos na construção do conhecimento científico: a pesquisa bibliográfica. Revista Katálisis, Scielo, p. 37-45, 2007. Disponível em: <<https://shorturl.at/LhIF1>>. Acesso em: 4 mar. 2024.

META. Introducing Llama 3.1: Our most capable models to date. Meta, 2024. Disponível em: <<https://ai.meta.com/blog/meta-llama-3-1/>>. Acesso em: 24 jul. 2024.

MORANDI, Maria Isabel W. Motta.; CAMARGO, Luis F. Riehs. Revisão sistemática da literatura - Design Science Research: método e pesquisa para avanço da ciência e da tecnologia. Porto Alegre: Bookman, 2015.

MINISTÉRIO DA SAÚDE. Relação Nacional de Medicamentos Essenciais - Rename. Gov.br. Disponível em: <<https://www.gov.br/saude/pt-br/composicao/sectics/rename>>. Acesso em: 1 mar. 2024.

MISTRAL AI TEAM. Mixtral of experts: A high quality Sparse Mixture-of-Experts. Mistral, 2023. Disponível em: <<https://mistral.ai/news/mixtral-of-experts/>>. Acesso em: 20 maio 2024.

RAGAS. Doc - Introduction. Ragas, 2024. Disponível em: <<https://docs.ragas.io/en/stable/>>. Acesso em: 10 jul. 2024.

NAZI, Zabir Al.; PENG, Wei. Large language models in healthcare and medical domain: a review. ArXiv, v.1, p. 1-22, 2024. Disponível em: <<https://shorturl.at/lysvE>>. Acesso em: 22 jul. 2024.

SCHWABER, Roie. What is a Vector Database & How Does it Work?. Pinecone, 2023. Disponível em: <<https://www.pinecone.io/learn/vector-database>>. Acesso em: 11 jul. 2024.

VASWANI, Ashish.; SHAZEER, Noam.; PARMAR, Niki. Attention Is All You Need. v.7, p. 1-15, 2023.

XU, Guo.; YIQIANG, Chen. Generative AI for Synthetic Data Generation: Methods, Challenges and the Future. v.1, p. 1-8, 2024.

YIN, Robert K. Estudo de caso, planejamento e métodos. 2. ed. São Paulo: Bookman, 2001. Disponível em: <<https://shorturl.at/O6v4B>>. Acesso em: 4 mar. 2024.