Comparação de Modelos de *Embeddings* e LLMs para Geração Aumentada por Recuperação em Português

Luiz Sabiano Ferreira Medeiros[1], **Hilário Tomaz Alves de Oliveira**[1]

[1] Programa de Pós-Graduação em Computação Aplicada – PPComp
Instituto Federal do Espírito Santo (IFES)
Av. dos Sabiás 330 – Morada de Laranjeiras – Serra – ES – Brazil – 29166-630

luizsabiano@gmail.com, hilario.oliveira@ifes.edu.br

Abstract. *Large language models (LLMs) represent a breakthrough in natural language processing, boosting performance in tasks such as text generation and question answering. However, they face challenges such as hallucinations and lack of access to updated information. The Retrieval-Augmented Generation (RAG) technique seeks to mitigate these problems by integrating external information retrieval into text generation, improving the accuracy and timeliness of the answers. This work investigated several open-source and proprietary embedding models and LLMs applied to the RAG technique, considering three databases containing documents written in Brazilian Portuguese. The experimental results demonstrated that the Multilingual E5 large and Gemma 2 9B models obtained the best performance among the evaluated models based on different evaluation measures.*

Resumo. *Os modelos de linguagem de larga escala (LLMs) representam um avanço para a área de processamento de linguagem natural, impulsionando o desempenho em tarefas como geração de texto e resposta a perguntas. No entanto, eles enfrentam desafios como alucinações e falta de acesso a informações atualizadas. A técnica de geração aumentada por recuperação (RAG) busca mitigar esses problemas ao integrar recuperação de informações externas à geração de texto, melhorando a precisão e a atualidade das respostas. Este trabalho realizou uma investigação de diversos modelos embeddings e LLMs de código aberto e proprietários aplicados à técnica RAG considerando três bases de dados contendo documentos escritos em português do Brasil. Os resultados experimentais demonstraram que os modelos Multilingual E5 large e Gemma 2 9B obtiveram o melhor desempenho dentre os modelos avaliados com base em diferentes medidas de avaliação.*

1. Introdução

O surgimento dos modelos de linguagem de larga escala (LLMs, do inglês *Large Language Models*) representou um importante avanço para a área de Processamento de Linguagem Natural (PLN), permitindo o desenvolvimento de sistemas capazes de executar tarefas como geração de texto, criação automática de resumos e a capacidade de responder a perguntas com grande fluência [Zhao et al. 2023]. Esses modelos são treinados com grandes quantidades de dados para aprender padrões linguísticos e gerar respostas semelhantes às humanas. Mesmo com alta capacidade em responder perguntas, os LLMs enfrentam

dois desafios importantes. Primeiro, eles são propensos a gerar informações incorretas ou enganosas, um fenômeno conhecido como alucinação [Ji et al. 2023]. Segundo, seu conhecimento é limitado aos dados usados durante o treinamento, o que pode resultar em respostas desatualizadas ou incompletas quando questionados sobre informações recentes ou conteúdo específico de um domínio [Lewis et al. 2020].

Para mitigar essas limitações, a técnica de geração aumentada por recuperação (RAG, do inglês *Retrieval-Augmented Generation*) surgiu como uma abordagem que combina a tradicional área de Recuperação de Informação (RI) com o uso de LLM para geração de respostas [Lewis et al. 2020]. Ao incorporar um mecanismo de recuperação, a técnica de RAG permite que os LLMs acessem fontes de conhecimento externas dinamicamente, reduzindo a dependência de dados de treinamento estáticos e aprimorando a precisão factual das respostas geradas [Fan et al. 2024]. Essa abordagem híbrida é particularmente relevante em domínios que exigem informações atualizadas, como aplicações no domínio jurídico [Wiratunga et al. 2024], médico [Xiong et al. 2024] e financeiro [Iaroshev et al. 2024]. Além disso, usando RAG é possível mitigar o problema de alucinação ao fundamentar as respostas geradas nos documentos recuperados, melhorando sua confiabilidade e alinhamento com o conhecimento factual [Chen et al. 2024].

A arquitetura básica de um sistema de RAG consiste em dois componentes principais: o módulo de recuperação e o módulo de geração, conforme ilustrado na Figura 1. O módulo de recuperação é responsável por buscar documentos relevantes em uma ou mais bases de dados externas, geralmente utilizando modelos de *embeddings* para representar tanto a consulta do usuário quanto os documentos, além de algoritmos de similaridade para identificar as correspondências mais relevantes. O módulo de geração, por sua vez, gera a resposta com base na consulta do usuário e nos documentos recuperados, produzindo respostas que integram o conhecimento extraído. A escolha dos modelos para ambos os módulos impacta diretamente a eficácia do sistema, tornando sua seleção uma decisão crítica no projeto e implementação de um sistema de RAG.

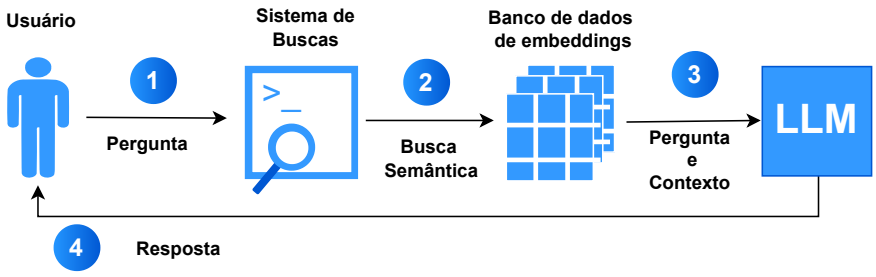


Figura 1. Fluxo básico de um sistema tradicional de RAG.

Diversos estudos investigaram o desempenho de abordagens baseadas em RAG, principalmente na língua inglesa [Fan et al. 2024]. Apesar da disponibilidade de diversos modelos de *embedding* e LLMs multilíngues, ainda há poucos estudos comparativos que avaliem sua eficácia no contexto da técnica de RAG para o português do Brasil [Passinato et al. 2024]. Compreender o desempenho de diferentes modelos em tarefas de recuperação e geração é fundamental para melhorar o desempenho dos modelos e adaptá-los a diferentes cenários de aplicação.

Neste contexto, este trabalho tem como objetivo realizar uma análise comparativa de modelos de *embeddings* e LLMs no âmbito da técnica de RAG, com foco no desempenho dos módulos de recuperação e geração, em documentos escritos em português do Brasil. Na etapa de recuperação de informação, foram avaliados cinco modelos de *embeddings* de texto, compreendendo o modelo de código aberto Multilingual E5 [Wang et al. 2024] e o modelo proprietário da OpenAI[1], ambos em diferentes variações em tamanho de arquitetura. Para a etapa de geração, foram investigados os modelos de código aberto do Llama [Grattafiori et al. 2024] e Gemma [Team et al. 2024], bem como os modelos proprietários Sabiá-3 e Sabiazinho-3 [Abonizio et al. 2024].

Os experimentos foram conduzidos utilizando três corporas (Pirá [Paschoal et al. 2021], FairytaleQA PT-BR [Leite et al. 2024] e Squad 2 PT-BR [Rajpurkar et al. 2018]) de perguntas e respostas disponíveis na literatura. A avaliação dos modelos foi realizada com as tradicionais métricas do ROUGE e BERTScore, além das medidas de fidelidade (*faithfulness*) e relevância da resposta, computadas por meio da ferramenta RAGAS [Es et al. 2024]. Os experimentos realizados consideraram modelos com diferentes escalas de parâmetros para comparar suas capacidades em tarefas específicas e identificar soluções viáveis em cenários com recursos computacionais abundantes ou limitados.

Duas questões de pesquisa guiaram os experimentos realizados. Primeiramente, buscamos determinar qual modelo de *embedding* apresenta o melhor desempenho entre os avaliados, analisando seu impacto na qualidade da recuperação. Em seguida, investigamos qual LLM gera respostas mais precisas e contextualmente relevantes. Ao abordar essas questões, este trabalho busca identificar as configurações mais eficazes, contribuindo para futuras pesquisas voltadas ao desenvolvimento de aplicações de RAG em português. O código-fonte desenvolvido neste trabalho está disponível em `https://github.com/luizsabiano/rag_eval_ptbr`.

2. Trabalhos Relacionados

Desde a introdução da técnica de RAG por Lewis et al. [Lewis et al. 2020], diversos trabalhos têm investigado e aprimorado essa abordagem, com o objetivo de desenvolver soluções capazes de responder às perguntas dos usuários de maneira mais precisa e contextualmente relevante. Embora vários estudos tenham explorado diferentes estratégias baseadas em RAG [Fan et al. 2024], ainda há poucas pesquisas focadas em documentos escritos em português do Brasil. Nesta seção, apresentamos uma análise dos trabalhos que abordaram o uso de RAG para o português.

O trabalho de Costa e Souza Filho [da Costa and e Souza Filho 2024] investigou abordagens de ajuste fino e RAG para a adaptação de LLMs no contexto de perguntas e respostas. Os autores avaliaram os modelos PTT5 e Llama 3 com 8 bilhões de parâmetros (8B). Para o módulo de Recuperação de Informação (RI), foram analisados os métodos *Best Match 25* (BM25), *Dense Passage Retriever* (DPR) e ColBERT. Os modelos foram avaliados utilizando as métricas do ROUGE-1, ROUGE-L, BERTScore e GPTScore. Os experimentos foram conduzidos com as bases de dados Pirá e Databricks-Dolly, e os resultados demonstraram que a incorporação de dados externos relevantes por meio de RAG melhorou a qualidade das respostas geradas pelos LLMs, em comparação com o uso exclusivo

[1]https://platform.openai.com/docs/guides/embeddings/embedding-models

do ajuste fino. Além disso, observou-se que o modelo PTT5, quando ajustado, apresentou desempenho semelhante ao Llama 3 8B sem ajuste.

Passinato et al. [Passinato et al. 2024] propuseram o desenvolvimento de um chatbot para o domínio de informações oftalmológicas, utilizando três técnicas de RAG: *Naive RAG*, HYDE e *Rewrite-Retrieve-Read*. Os modelos analisados foram o Mistral, com 7 bilhões de parâmetros (7B), para geração de texto, e o *Multilingual E5*, para representações de *embeddings*. O corpus utilizado foi coletado manualmente na internet a partir de artigos publicados em blogs mantidos por hospitais e clínicas especializadas. A avaliação dos resultados foi realizada com o *framework* RAGAS, utilizando as métricas de fidelidade de contexto, relevância da resposta e relevância do contexto. Os experimentos demonstraram que a técnica *Rewrite-Retrieve-Read* apresentou melhor desempenho, especialmente em perguntas formuladas de maneira mais informal e em casos em que o contexto recuperado continha entre 500 e 750 caracteres.

O estudo apresentado em [Paranhos et al. 2024] avaliou o impacto de diferentes arquiteturas RAG no contexto jurídico, analisando cinco abordagens: *Naive RAG*, *HyDE*, *Corrective RAG*, *Self-RAG* e *RAG-Fusion*. O modelo de *embedding* do Google foi utilizado para indexação e recuperação, enquanto o modelo Gemini 1.5 Pro foi usado para geração das respostas. A avaliação, realizada com o *framework* RAGAS, considerou métricas como fidelidade, revocação e precisão do contexto. Os resultados indicam que o *Self-RAG* é a abordagem mais robusta para perguntas que exigem fidelidade e relevância, enquanto o *RAG-Fusion* se destacou em cenários que demandam alta precisão na recuperação.

Kuratomi et al. [Kuratomi et al. 2024] projetaram e avaliaram um assistente virtual baseado em RAG, desenvolvido especificamente para a Universidade de São Paulo (USP). Os autores avaliaram diferentes modelos de recuperação e geração, ajustando hiperparâmetros como tamanho do bloco (*chunks*) e número de documentos recuperados. O conjunto de dados utilizado foi coletado de 866 documentos disponíveis no site da USP, abrangendo o período entre janeiro de 2023 até maio de 2024. Foram utilizados cinco modelos de *embeddings*: *Paraphrase MiniLM*, *Paraphrase Mpnet*, *Distiluse base* e *BM25*. Como modelos de LLMs, foram avaliados o *GPT 3.5*, *Llama 3*, *Mixtral* e *Sabiá-2*.

A Tabela 1 apresenta as principais características de cada trabalho relacionado: modelo de *embedding* utilizado para recuperação de informação, modelo generativo (LLM), base de dados e métricas de avaliação utilizados. Em geral, observou-se que os trabalhos existentes avaliam o desempenho dos modelos em um único domínio e, com exceção do trabalho de [Kuratomi et al. 2024], pouca diversidade de modelos foi considerada nos experimentos. Por isso, este estudo expandiu trabalhos anteriores ao investigar a eficácia de diversos modelos de *embeddings* e LLMs em três bases de dados públicas de domínios distintos, possibilitando uma avaliação em diferentes contextos e estilos textuais. Além disso, a avaliação foi realizada por meio das métricas tradicionais do ROUGE-L e BERTScore, bem como de métricas derivadas da biblioteca RAGAS, que usa uma abordagem de LLM como juiz.

3. Materiais e Métodos

3.1. Conjuntos de Dados

A avaliação de um sistema de RAG requer um conjunto de dados anotados com, no mínimo, três elementos: **(i)** uma pergunta, **(ii)** o contexto no qual a resposta pode ser

Tabela 1. Comparativo dos trabalhos relacionados.

Trabalho	Modelo de *Embedding*	LLM	Base de Dados	Avaliação
[da Costa and e Souza Filho 2024]	BM25, DPR e ColBERT	PPT5 e Llama 3 8B	Pirá e Databricks Dolly	ROUGE-1, ROUGE-L, BERTScore e Métrica própria
[Passinato et al. 2024]	E5	Mistral 7B	Artigos de blogs	RAGAS
[Paranhos et al. 2024]	*Embeddings* da OpenAI	Gemini 1.5 pro	Documentos do TCE-GO	RAGAS
[Kuratomi et al. 2024]	Paraphrase MiniLM, Paraphrase Mpnet base, Distiluse base	GPT 3.5, Llama 3, Mixtral e Sabia 2	Coletado no site da USP	F1-score, Similaridade de cosseno e Métrica própria utlizando LLM
Trabalho Proposto	E5 (small, base e large), *embeddings* da OpenAI (small e large)	Llama, Gemma, Sabiá 3 e Sabiazinho 3	Pirá, FairytaleQA PT-BR e Squad 2 PT-BR	*Recall*, ROUGE-L, BERTScore e RAGAS

encontrada e **(iii)** a resposta correspondente. Neste trabalho, foram utilizadas as bases de dados do Pirá [Paschoal et al. 2021], SQuAD 2 [Rajpurkar et al. 2018] e FairytaleQA PT-BR [Leite et al. 2024].

O Pirá é um conjunto de dados bilíngue (português e inglês) de perguntas e respostas sobre o oceano, biodiversidade, mudanças climáticas e a costa brasileira. É o primeiro corpus com suporte ao português, contendo 2.261 registros extraídos de (i) resumos de artigos científicos sobre o litoral brasileiro e (ii) trechos de relatórios da ONU sobre o oceano. As perguntas e respostas foram geradas e avaliadas manualmente por voluntários, incluindo estudantes e pesquisadores da Universidade de São Paulo [Paschoal et al. 2021].

O FairytaleQA é um conjunto de dados contendo 10.580 pares de perguntas e respostas em inglês, criado por especialistas educacionais a partir de 278 histórias infantis, visando avaliar a compreensão narrativa de alunos do jardim de infância à oitava série [Xu et al. 2022]. Com intuito de suprir a escassez de bases de dados em outros idiomas, Leite et al. [Leite et al. 2024] traduziram o FairytaleQA para espanhol, francês e português, incluindo suas variantes europeia e brasileira. Neste trabalho, utilizamos a versão da base de dados traduzida para o português do Brasil.

O *Stanford Question Answering Dataset 2.0* (SQuAD2) é um conjunto de dados desenvolvido para a compreensão de texto por máquinas, combinando 100.000 perguntas e respostas de sua primeira versão (SQuAD1.1), extraídas de 536 artigos da Wikipédia, com mais de 50.000 perguntas sem resposta, formuladas por *crowdworkers* para se assemelharem às respondíveis [Rajpurkar et al. 2018]. Neste estudo, foi utilizada uma versão traduzida automaticamente do SQuAD2 pelo Tradutor do Google e comumente usada em trabalhos para o português do Brasil.

A Tabela 2 apresenta a média e o desvio padrão do número de palavras por componente (contexto, pergunta e resposta) nas bases de dados PIRÁ, FairytaleQA PT-BR e SQuAD2 PT-BR. Destaca-se que apenas os exemplos do conjunto de teste de cada base foram utilizados, uma vez que não foi realizado treinamento de modelos neste trabalho. Embora contenha um número significativamente menor de perguntas, a base de dados Pirá foi utilizada devido à alta qualidade de suas perguntas e respostas, bem como por abranger um domínio de conhecimento mais desafiador em comparação com as bases do FairyTaleQA PT-BR e SQuAD2 PT-BR.

Tabela 2. Estatística das bases de dados.

Conjunto Teste			Palavras	
Base de Dados	**Registros**	**Componente**	**Média**	**Desvio Padrão**
Pirá	216	Contexto	284,51	149,11
		Pergunta	14,73	5,94
		Resposta	16,07	15,16
FairyTaleQA PT-BR	1.007	Contexto	197,96	86,92
		Pergunta	7,95	3,35
		Resposta	6,98	5,57
SQuAD2 PT-BR	5.930	Contexto	153,24	68,25
		Pergunta	11,91	4,06
		Resposta	3,62	3,83

3.2. Modelos de *Embeddings* e LLMs Avaliados

Nesta seção, são apresentados os modelos de *embeddings* e LLMs utilizados nos experimentos realizados. Os modelos selecionados incluem tanto alternativas de código aberto quanto modelos proprietários, variando em termos do tamanho da representação vetorial, número de parâmetros e capacidade de geração de respostas.

Os modelos *Multilingual E5* pertencem a uma família de modelos de *embeddings* de código aberto desenvolvida para tarefas de recuperação de informações e classificação semântica [Wang et al. 2024]. Eles estão disponíveis em três configurações: *Small*, Base e *Large*. O *Multilingual E5 Small* possui um espaço vetorial de 384 dimensões e 118 milhões de parâmetros, enquanto o *Multilingual E5 Base* tem 768 dimensões e 278 milhões de parâmetros. O modelo *Multilingual E5 Large* oferece uma representação vetorial de 1.024 dimensões e 560 milhões de parâmetros. Além disso, foram considerados dois modelos proprietários da OpenAI (*text-embedding-3-small* e *text-embedding-3-large*), com 1.536 e 3.072 dimensões, respectivamente.

Os experimentos avaliaram diversos LLMs de código aberto e proprietários, como o Gemma 2, da Google, em variantes de 2 e 9 bilhões de parâmetros [Team et al. 2024]; o Llama 3, da Meta, sendo avaliados os modelos de 1, 3, 8 e 70 bilhões de parâmetros [Grattafiori et al. 2024]; e os modelos Sabiá 3 e Sabiazinho 3, da Maritaca AI, otimizados para o português do Brasil [Abonizio et al. 2024].

3.3. Metodologia Experimental

A avaliação dos modelos de *embeddings* e LLM foi realizada por meio de dois experimentos, ambos utilizando as bases de dados do Pirá, FairytaleQA PT-BR e SQuAD2 PT-BR.

O primeiro experimento avaliou a eficácia dos modelos de *embeddings* de texto na tarefa de recuperação de informação, focando na assertividade do contexto recuperado. Para cada base de dados, todos os contextos foram indexados no banco de dados vetorial do ChromaDB[2], com o objetivo de verificar se, dada uma pergunta de entrada, os modelos eram capazes de recuperar o contexto correto, considerando diferentes valores no *top-n* contextos recuperados. A métrica *recall@n* foi utilizada para medir a capacidade dos modelos de recuperar o contexto real, quantificando a porcentagem de casos em que o

[2]`https://www.trychroma.com/`

contexto correto estava presente entre os n contextos recuperados. Além disso, o custo computacional foi analisado, correlacionando a melhoria na métrica de *recall@n* com o aumento da quantidade média de *tokens* nos top-n contextos recuperados. O modelo de *embedding* com o melhor custo-benefício, considerando os três conjuntos de dados, foi selecionado para o módulo de recuperação utilizado no Experimento 2.

O segundo experimento focou na avaliação das respostas geradas pelos modelos de LLM. As métricas utilizadas foram a fidelidade (*faithfulness*) e a relevância das respostas (*answer relevance*), avaliadas com o framework RAGAS [Es et al. 2024]. O modelo *GPT-4o-mini* da OpenAI foi usado como LLM juiz para computar as métricas de avaliação. A fidelidade quantifica a consistência das respostas em relação ao contexto recuperado, enquanto a relevância mede o grau em que as respostas atendem à pergunta original. Além disso, foram incorporadas as métricas do ROUGE-L [Lin 2004] e do BERTScore [Zhang et al. 2019] para quantificar a similaridade léxica e semântica entre as respostas geradas e as respostas de referência. O *prompt* usado para geração das perguntas foi:

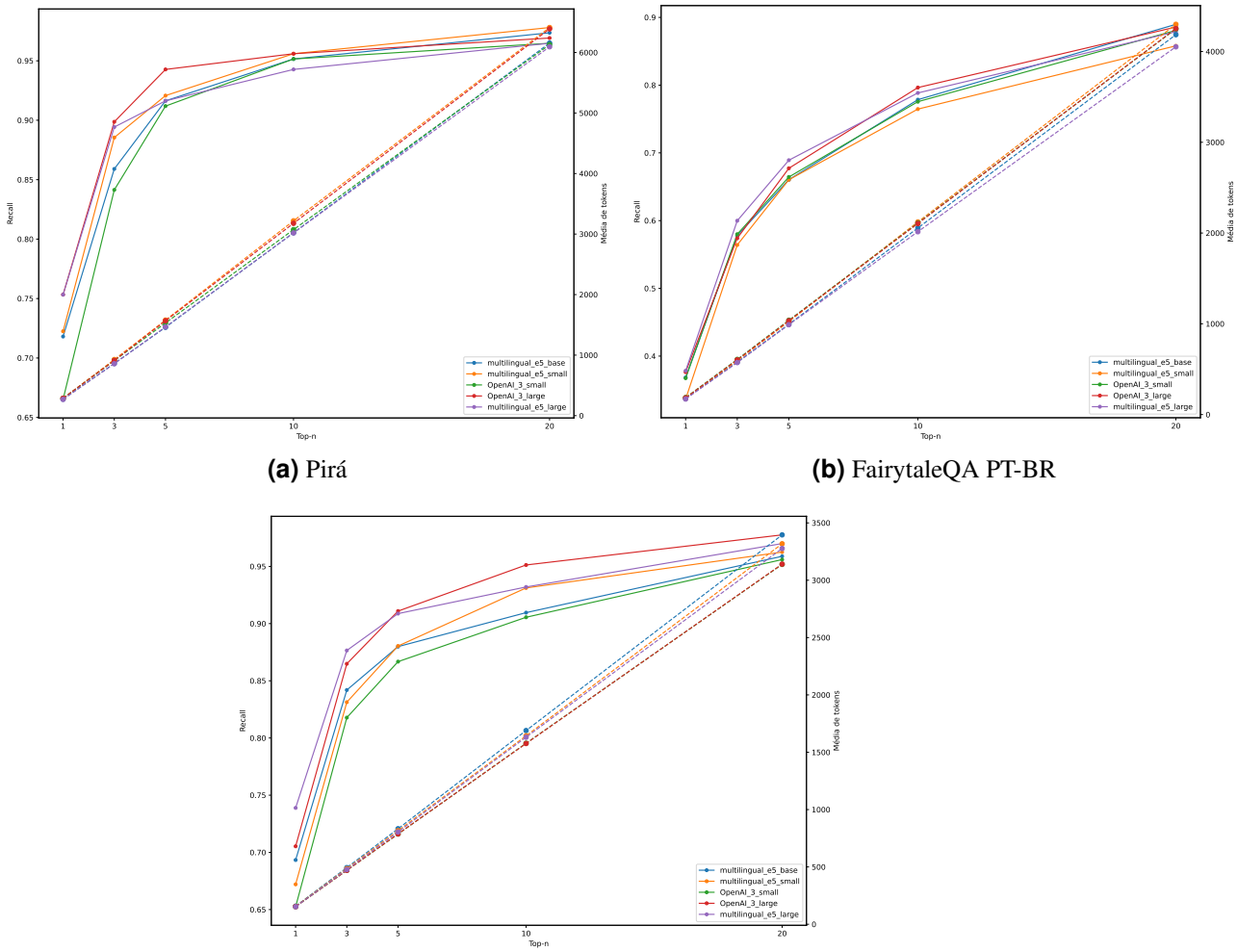
> *Você é um assistente virtual responsável por responder às perguntas dos usuários. Responda à PERGUNTA a seguir com base apenas no CONTEXTO fornecido e nenhum conhecimento a mais. Forneça uma resposta detalhada e precisa escrita usando a norma culta do português do Brasil. Não forneça informações não mencionadas no CONTEXTO. Caso não seja possível responder à pergunta usando somente o contexto fornecido, retorne à mensagem "NONE".*
> *CONTEXTO:* {*contexto*}
> *PERGUNTA:* {*pergunta*}

4. Resultados

4.1. Experimento 1 - Avaliação dos Modelos de *Embedding*

A Figura 2 apresenta os resultados do primeiro experimento, visando avaliar diferentes modelos de *embeddings* para compor o módulo de recuperação da informação. O eixo *x* representa a quantidade de contextos recuperados (*top-n*) do banco de dados vetorial, enquanto o eixo *y* à esquerda indica o valor da métrica de *recall@n* e o eixo *y* à direita mostra a quantidade média de *tokens* do *prompt* construído usando os contextos recuperados. A relação entre essas variáveis evidencia o compromisso entre o aumento da quantidade de *tokens* do *prompt* (representado por linhas pontilhadas) e a melhoria da métrica de *recall@n* (representado por linhas sólidas).

A análise do aumento na quantidade de *tokens* (linhas pontilhadas) revela uma relação diretamente proporcional entre a quantidade de contextos recuperados (*top-n*) e a quantidade média de *tokens*. A homogeneidade entre os modelos indica que, dentro do intervalo analisado, essa variável não influencia significativamente a escolha do modelo de recuperação. Por outro lado, a evolução da métrica de *recall@n* (linhas sólidas) não segue a mesma linearidade. Há um aumento acentuado do *recall@n* entre *top-1* e *top-3*, seguido por um crescimento moderado até *top-5*, com estabilização subsequente e ganhos mínimos além desse ponto. A comparação entre a quantidade média de *tokens* e o *recall@n* mostra que, após *top-5*, o aumento de *tokens* gera benefícios marginais, mas acarreta um maior custo computacional por conta da quantidade de *tokens* no *prompt*. Assim, o *top-5* foi definido como o melhor ponto de equilíbrio identificado neste experimento.



(a) Pirá

(b) FairytaleQA PT-BR

(c) SQuAD2 PT-BR

Figura 2. Resultados do experimento de avaliação dos modelos de *embeddings*.

Considerando o *top-5*, os modelos *Multilingual E5 large* e o *text-embedding-3-large* da OpenAI apresentam desempenho semelhante, alcançando $0,689$ e $0,677$ no FairytaleQA PT-BR e $0,909$ e $0,911$ no SQuAD2 PT-BR, respectivamente. No PIRÁ, o modelo *large* da OpenAI obteve o melhor desempenho, com *recall* de $0,940$, enquanto os modelos *Multilingual E5 small*, *base* e *large* apresentaram valores ligeiramente inferiores, de $0,920$, $0,910$ e $0,910$, respectivamente.

Os modelos de *embeddings* da OpenAI foram utilizados como referência devido à sua ampla adoção e desempenho reconhecido. No entanto, por se tratarem de modelos proprietários acessíveis via a *Application Programming Interface* (API) da OpenAI, levantam preocupações quanto à segurança de dados sensíveis e envolvem custos de uso. Por outro lado, o *Multilingual E5 large* apresentou resultados competitivos, tendo o melhor desempenho no FairytaleQA PT-BR, o segundo no SQuAD2 PT-BR (superado apenas pelo modelo *large* da OpenAI) e o terceiro no PIRÁ (atrás do modelo *large* da OpenAI e do *Multilingual E5 small*). Devido a esse desempenho e ao fato de ser de código aberto, o *Multilingual E5 large* foi considerado o melhor modelo neste experimento.

4.2. Experimento 2 - Avaliação dos LLMs

O objetivo deste segundo experimento é avaliar diferentes LLMs para a geração da resposta para a pergunta do usuário, considerando os contextos recuperados. Com base nos resultados do Experimento 1, utilizamos o modelo de *embedding Multilingual E5 large* e os cinco contextos mais relevantes (*top-5*) do banco de dados vetorial para compor o *prompt*. A Tabela 3 apresenta os resultados obtidos utilizando a métrica de *recall* computada das medidas do ROUGE-L e BERTScore nos conjuntos de dados: Pirá, FairytaleQA PT-BR e SQuAD2 PT-BR. O modelo com melhor desempenho em cada base de dados e medida de avaliação está destacado em negrito.

Tabela 3. Resultados considerando as medidas do ROUGE-L e BERTScore.

Modelos / Medidas Avaliação	**Pirá**		**FairytaleQA PT-BR**		**SQuAD2 PT-BR**	
	ROUGE-L	BERTScore	ROUGE-L	BERTScore	ROUGE-L	BERTScore
LLama 3.2 1B	0,147	0,725	0,105	0,715	0,118	0,708
Llama 3.2 3B	0,218	0,768	0,127	0,729	0,138	0,728
Llama 3.1 8B	0,172	0,755	0,079	0,701	0,073	0,712
Llama 3.1 70B	0,267	0,793	0,158	0,753	0,207	0,761
Gemma 2 2B	0,287	0,766	0,188	0,755	0,269	0,763
Gemma 2 9B	**0,394**	**0,799**	**0,268**	**0,779**	**0,453**	**0,799**
Sabiá 3	0,195	0,779	0,143	0,751	0,177	0,751
Sabiazinho 3	0,172	0,781	0,119	0,743	0,129	0,736

Os resultados obtidos pelos modelos apresentam uma certa variabilidade. No entanto, o modelo Gemma 2 9B demonstrou desempenho superior em todas as bases de dados e métricas de avaliação. Apesar das diferenças de domínio entre as bases, seu desempenho manteve-se estável. O tamanho da arquitetura influenciou o desempenho dos modelos Llama, com o modelo de 70B superando, em geral, os de 1B, 3B e 8B. Esse padrão também foi observado entre os modelos do Gemma de 2B e 9B. Uma análise manual revelou que o modelo Llama 3.2 1B frequentemente gerava respostas indicando impossibilidade de responder à pergunta, mesmo quando o contexto correto estava presente no *prompt*. Esse comportamento também foi observado nos modelos Llama de 3B e 8B.

De modo geral, o baixo desempenho dos modelos pode ser atribuído a dois fatores principais: **(i)** As respostas presentes nas bases de dados são, em sua maioria, curtas, enquanto os LLMs, especialmente os modelos maiores (Sabiá 3, Sabiazinho 3 e LLama 70B), apresentaram uma tendência de gerar respostas mais longas e detalhadas; e **(ii)** As métricas BERTScore e, em especial, ROUGE-L apresentam limitações por não capturarem nuances como variações lexicais ou estruturais que mantêm o mesmo significado. Adicionalmente, essas métricas tendem a favorecer respostas mais curtas e com maior similaridade em relação às de referência.

Para complementar a avaliação dos LLMs analisados, foram utilizadas as métricas de fidelidade e relevância da resposta, computadas por meio da biblioteca RAGAS. Devido a restrições orçamentárias relacionadas ao uso do modelo GPT-4o-mini para o cálculo dessas métricas, a avaliação foi realizada apenas nas bases de dados do Pirá e FairyTaleQA PT-BR. Os resultados obtidos estão apresentados na Tabela 4, com os melhores valores em cada métrica e base de dados destacados em negrito.

Tabela 4. Resultados considerando as medidas de Fidelidade e Relevância da Resposta.

Modelos / Medidas Avaliação	Pirá		FairytaleQA PT-BR	
	Fidelidade	Relevância da Resposta	Fidelidade	Relevância da Resposta
LLama 3.2 1B	0,571	0,599	0,512	0,480
Llama 3.2 3B	0,781	0,704	0,665	0,556
Llama 3.1 8B	0,667	0,576	0,544	0,411
Llama 3.1 70B	0,781	**0,716**	0,669	**0,581**
Gemma 2 2B	**0,789**	0,690	**0,672**	0,553
Gemma 2 9B	**0,789**	0,692	0,670	0,555
Sabiá 3	0,775	0,710	0,668	0,574
Sabiazinho 3	0,727	0,648	0,631	0,506

Os resultados das métricas de fidelidade e relevância da resposta apresentaram menor variabilidade em comparação com ROUGE-L e BERTScore. Com exceção dos modelos Llama 3.2 1B e Llama 3.1 8B, que tiveram desempenho significativamente inferior, os demais modelos obtiveram resultados mais equilibrados. O Gemma 2 9B destacou-se como a melhor opção, superando até mesmo o Llama 3.1 70B, que possui uma arquitetura significativamente maior. Além disso, por ser um modelo de código aberto, o Gemma 2 9B demonstrou alto desempenho em relação ao Sabiá-3, que é proprietário. Para cenários com recursos computacionais mais limitados, o Gemma 2 2B surge como uma alternativa interessante, pois apresentou resultados similares aos obtidos pelo Gemma 9B, mas com uma arquitetura consideravelmente mais leve.

5. Conclusões

Este trabalho avaliou o desempenho de diferentes modelos de *embeddings* e LLMs na construção de um sistema de recuperação aumentada por geração (RAG) para documentos escritos em português do Brasil. Foram realizados experimentos com três bases de dados da literatura, diversas métricas de avaliação e diferentes modelos de código aberto e proprietários. O primeiro experimento identificou que o modelo *Multilingual E5 large* apresentou o melhor custo-benefício entre a métrica de *recall* e o custo computacional e financeiro. Já o segundo experimento revelou que o Gemma 2 9B obteve um desempenho mais consistente em todas as bases de dados, destacando-se nas métricas ROUGE-L, BERTScore, fidelidade e relevância da resposta. Os resultados experimentais demostraram que modelos de código aberto podem ser alternativas viáveis em cenários onde a privacidade dos dados é um aspecto importante.

Uma limitação deste trabalho está relacionada ao fato de as bases de dados utilizadas conterem, em geral, respostas relativamente curtas, o que pode ter restringido a análise de respostas mais longas e detalhadas geradas por modelos como o Sabiá 3 e o Sabiazinho 3. Dessa forma, como proposta para trabalhos futuros, sugere-se a realização de uma análise qualitativa das respostas geradas, com o objetivo de identificar cenários em que os modelos produziram respostas mais completas e factualmente corretas. Outra linha de pesquisa contempla a exploração de novas arquiteturas de RAG, como a abordagem Hyde, bem como a investigação do impacto da inclusão de etapas de reescrita de consultas e de revisão das respostas geradas.

Agradecimentos

Os autores agradecem à FAPES/UnAC (Nº FAPES 1228/2022 P 2022-CD0RQ, Nº SIAFEM 2022-CD0RQ) pelo apoio financeiro concedido por meio do sistema UniversidadES.

Referências

Abonizio, H., Almeida, T. S., Laitz, T., Junior, R. M., Bonás, G. K., Nogueira, R., and Pires, R. (2024). Sabiá-3 technical report. *arXiv preprint arXiv:2410.12049*.

Chen, J., Lin, H., Han, X., and Sun, L. (2024). Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

da Costa, L. and e Souza Filho, J. O. (2024). Adapting llms to new domains: A comparative study of fine-tuning and rag strategies for portuguese qa tasks. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 267–277, Porto Alegre, RS, Brasil. SBC.

Es, S., James, J., Anke, L. E., and Schockaert, S. (2024). Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158.

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., and Li, Q. (2024). A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Iaroshev, I., Pillai, R., Vaglietti, L., and Hanne, T. (2024). Evaluating retrieval-augmented generation models for financial report question and answering. *Applied Sciences*, 14(20):9318.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Kuratomi, G., Pirozelli, P., Cozman, F., and Peres, S. (2024). A rag-based institutional assistant. In *Anais do XXI Encontro Nacional de Inteligência Artificial e Computacional*, pages 755–766, Porto Alegre, RS, Brasil. SBC.

Leite, B., Osório, T. F., and Cardoso, H. L. (2024). Fairytaleqa translated: Enabling educational question and answer generation in less-resourced languages. In *European Conference on Technology Enhanced Learning*, pages 222–236. Springer.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Paranhos, S., Tomazini, J., Junior, C. C., and de Oliveira, S. T. (2024). Avaliação do impacto de diferentes padrões arquiteturais rag em domínios jurídicos. In *Anais da XII Escola Regional de Informática de Goiás*, pages 99–108, Porto Alegre, RS, Brasil. SBC.

Paschoal, A. F., Pirozelli, P., Freire, V., Delgado, K. V., Peres, S. M., José, M. M., Nakasato, F., Oliveira, A. S., Brandão, A. A., Costa, A. H., et al. (2021). Pirá: A bilingual portuguese-english dataset for question-answering about the ocean. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4544–4553.

Passinato, E. B., Rios, W. S., and Galvão Filho, A. R. (2024). Integração de modelos de linguagem e rag na criação de chatbots oftalmológicos. In *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, pages 354–365. SBC.

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. (2024). Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F. (2024). Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Wiratunga, N., Abeyratne, R., Jayawardena, L., Martin, K., Massie, S., Nkisi-Orji, I., Weerasinghe, R., Liret, A., and Fleisch, B. (2024). Cbr-rag: case-based reasoning for retrieval augmented generation in llms for legal question answering. In *International Conference on Case-Based Reasoning*, pages 445–460. Springer.

Xiong, G., Jin, Q., Lu, Z., and Zhang, A. (2024). Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251.

Xu, Y., Wang, D., Yu, M., Ritchie, D., Yao, B., Wu, T., Zhang, Z., Li, T. J.-J., Bradford, N., Sun, B., Hoang, T. B., Sang, Y., Hou, Y., Ma, X., Yang, D., Peng, N., Yu, Z., and Warschauer, M. (2022). Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. Association for Computational Linguistics.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).