

RAG para Atendimento Acadêmico: Uma Avaliação de Modelos de Embedding Compactos no Contexto Universitário

Mateus Henrique Ferreira De Sousa¹, André Castelo Branco Soares²

¹Departamento de Computação – Universidade Federal do Piauí (UFPI)
Teresina – PI – Brasil

{mathenrique, andre.soares}@ufpi.edu.br

Abstract. *This work presents an institutional chatbot based on Retrieval-Augmented Generation (RAG) for academic support in Portuguese, developed with a focus on low operational cost. The main contribution lies in the comparative evaluation of compact open-source embedding models, up to 0.6 billion parameters, and proprietary models in the task of information retrieval over institutional documents. The evaluation, conducted using the metrics Recall, MRR, and NDCG based on questions drawn from institutional documents from the academic context, indicates that compact open-source models can achieve competitive performance in this scenario. These results highlight their potential as viable and lower-cost alternatives for academic applications in Portuguese.*

Resumo. *Este trabalho apresenta um chatbot institucional baseado em RAG para suporte acadêmico em português, desenvolvido com foco em baixo custo operacional. A principal contribuição consiste na avaliação comparativa de modelos de embedding open-source compactos, com até 0.6B de parâmetros, e modelos proprietários na recuperação de informação sobre documentos institucionais. A avaliação, realizada com as métricas Recall, MRR e NDCG a partir de perguntas formuladas sobre documentos do contexto acadêmico, indica que modelos open-source compactos podem alcançar desempenho competitivo nesse cenário. Esses resultados apontam seu potencial como alternativas viáveis e de menor custo para aplicações acadêmicas em português.*

1. Introdução

[Canale et al. 2025] demonstram que, em avaliações controladas de sistemas RAG, a variação do modelo de embedding pode produzir diferenças de desempenho superiores às causadas pela variação do LLM. O componente que codifica a relação semântica entre consulta e documento condiciona a qualidade da resposta de forma ao menos tão determinante quanto o modelo que a gera. Em domínios regulados por documentação normativa, como o acadêmico institucional, essa dependência se acentua. Recuperação imprecisa produz respostas incorretas, independentemente da sofisticação do modelo generativo. Em arquiteturas RAG, o embedding não constitui apenas um componente auxiliar da pipeline, mas uma decisão de arquitetura cujo impacto se propaga diretamente para a qualidade do resultado final.

Para modelos de linguagem, o Chatbot Arena consolidou-se como parâmetro comparativo amplamente adotado [Chiang et al. 2024]. Para modelos de embedding, essa

função é exercida pelo MTEB e por sua extensão multilíngue, o MMTEB¹, que avaliam modelos em múltiplas tarefas, idiomas e domínios a partir de protocolos padronizados [Muennighoff et al. 2023, Enevoldsen et al. 2025]. Ambos os recursos servem como referência inicial para reduzir o espaço de candidatos, embora o próprio MMTEB demonstre que o desempenho de um modelo varia substancialmente conforme tarefa e idioma, o que limita sua validade como critério definitivo e impõe avaliação no domínio específico da aplicação [Enevoldsen et al. 2025]. Apesar da existência desses instrumentos, a produção brasileira sobre chatbots universitários não os incorpora de forma efetiva. Uma revisão de 30 trabalhos conduzida nesta pesquisa revela que a maioria dos estudos não identifica o modelo de embedding empregado, e que nenhum deles realiza comparação experimental entre alternativas. Nos raros casos em que o modelo é nomeado, a escolha é feita sem justificativa — registrando-se até a adoção de modelos monolíngues em inglês para recuperação em português [Layane et al. 2025, Spricigo 2026].

Este trabalho se estrutura em torno de três objetivos. O primeiro, de natureza conceitual, é discutir e aprofundar a importância da escolha do modelo de embedding em arquiteturas RAG, trazendo essa questão ao primeiro plano da discussão sobre chatbots acadêmicos. O segundo, de natureza experimental, é avaliar seis modelos de embedding open-source com até 0,6 bilhões de parâmetros, posicionados entre as melhores alternativas compactas do MMTEB multilíngue para retrieval, em comparação ao text-embedding-3-large, sobre o corpus documental do sistema proposto, mensurando a recuperação por Recall@6, MRR e NDCG@6. O terceiro, de natureza aplicada, é entregar um protótipo funcional de chatbot baseado em RAG, implementado em Streamlit, para o suporte acadêmico dos estudantes de Ciência da Computação da UFPI.

Esses objetivos são interdependentes. O protótipo não é apenas o produto final da iniciativa, mas também o ambiente experimental que viabiliza a avaliação: nele, a etapa de recuperação é isolada enquanto os demais componentes permanecem fixos, garantindo que os resultados reflitam exclusivamente o efeito do modelo de embedding. A avaliação, ao produzir evidências de que essa única variação é suficiente para gerar diferenças mensuráveis de desempenho, fundamenta empiricamente a discussão sobre a importância do embedding. Tomados em conjunto, os três objetivos visam contribuir com um artefato implantável para a instituição, com resultados experimentais que orientem a seleção de modelos em sistemas RAG de baixo custo e com uma argumentação fundamentada em dados para que o embedding receba, na produção sobre chatbots acadêmicos em português, tratamento experimental compatível com sua influência sobre o sistema.

2. Trabalhos Relacionados

2.1. Chatbots Institucionais em Universidades Brasileiras

O desenvolvimento de chatbots em universidades brasileiras abrange hoje domínios que vão de secretarias acadêmicas e bibliotecas a processos seletivos e programas de integridade [Elias 2024, do Nascimento 2024, Coelho 2024, dos Passos et al. 2025, Pantoja et al. 2025, Almeida 2024, de Oliveira 2024, Felipe et al. 2025, Chaves et al. 2024], com abordagens tecnológicas que incluem desde plataformas no-code e NLP clássico até arquiteturas RAG [do Amaral et al. 2025,

¹MTEB: *Massive Text Embedding Benchmark*; MMTEB: *Massive Multilingual Text Embedding Benchmark*.

Bordin et al. 2025, Rayzer and Santos 2025, Cabral 2025, Soares et al. 2025, de Souza et al. 2025, de Souza 2024, Layane et al. 2025, Rodrigues et al. 2025]. Parte desses trabalhos apresenta avaliações relevantes, com uso de métricas como Exact Match, F1, MRR e NDCG [de Souza 2024, Rayzer and Santos 2026], avaliação com usuários reais [Bordin et al. 2025] e verificação factual por ontologia [do Amaral et al. 2025]. No entanto, tanto nos trabalhos com avaliação mais rigorosa quanto nos demais, a atenção se concentra predominantemente na escolha do LLM, na usabilidade da interface ou na viabilidade da implantação. O tratamento dispensado ao modelo de embedding permanece, na maioria dos casos, ausente ou superficial, lacuna que a análise sistemática a seguir busca documentar de forma estruturada.

2.2. Análise Sistematizada

Para sistematizar o tratamento dispensado ao embedding na literatura analisada, foi considerado um corpus de 30 trabalhos sobre chatbots institucionais no contexto universitário, selecionados entre os resultados mais bem indexados no Google Scholar por aderência temática ao problema investigado. Com base nesse corpus, propõe-se uma escala de maturidade com cinco níveis: E0, quando tema (embedding) não é mencionado; E1, quando embeddings são referenciados genericamente sem identificação do modelo; E2, quando o modelo é nomeado, mas não justificado; E3, quando a escolha é fundamentada em critério explícito, como custo, idioma ou benchmark; e E4, quando há comparação experimental entre modelos. Aplicada ao corpus, a escala revela que 18 dos 30 trabalhos analisados se enquadram em E0², isto é, o nada sobre o embedding é mencionado.

Em sete casos, há menção genérica sem identificação do modelo (E1) [do Amaral et al. 2025, Petrin and Néto 2026, Rayzer and Santos 2025, Rayzer and Santos 2026, Rodrigues et al. 2025, Soares et al. 2025, de Souza 2024]. Apenas cinco trabalhos chegam ao nível E2: [Cabral 2025] adota AWS Bedrock Embeddings; [Bordin et al. 2025] emprega o modelo `text-embedding-004`; [de Carvalho 2025] utiliza `flan-t5-small`; por fim, dois trabalhos adotam o modelo BAAI/bge-base-en-v1.5 [Spricigo 2026, Layane et al. 2025]. Essa escolha é particularmente sensível pois esse modelo específico é treinado para ser usado em textos em inglês. Benchmarks como o MTEB (Massive Text Embedding Benchmark) e sua ampliação multilíngue, o MMTEB (Massive Multilingual Text Embedding Benchmark), demonstram que o desempenho de modelos de embedding varia de acordo com a tarefa, o idioma e o domínio de aplicação [Muennighoff et al. 2023, Enevoldsen et al. 2025]. Nesse contexto, usar uma variante monolíngue em inglês para recuperação sobre documentos acadêmicos em português, quando há alternativas multilíngues disponíveis, torna-se uma decisão metodologicamente frágil, que compromete a qualidade da recuperação dos seus documentos.

Nenhum trabalho do corpus atinge E3 ou E4. Além dos trabalhos discutidos nesta seção, o corpus sistematizado também inclui estudos como [de Melo 2023, da Silva 2021, de Andrade et al. 2020, Oliveira 2025, Silva et al. 2023], que foram considerados na classificação proposta, embora não detalhados no corpo do texto.

Fora do subdomínio universitário, existem trabalhos interessantes que avaliam o

²[Almeida 2024, de Andrade et al. 2020, Anibal 2025, Chaves et al. 2024, Coelho 2024, Elias 2024, Felipe et al. 2025, Gonçalves et al. 2025, de Melo 2023, Monteiro 2023, do Nascimento 2024, de Oliveira 2024, Oliveira 2025, Pantoja et al. 2025, dos Passos et al. 2025, da Silva 2021, Silva et al. 2023, de Souza et al. 2025].

desempenho dos modelos de embeddings: em um estudo, foram comparados cinco modelos da família Multilingual E5 e modelos proprietários da OpenAI em três bases em português com Recall@n, identificando Multilingual E5 Large como melhor custo-benefício [Medeiros and de Oliveira 2025]; em outro, três modelos foram avaliados em documentos hospitalares com MRR e NDCG@10, com intfloat/multilingual-e5-small alcançando MRR de 0,666 e taxa de falha de 1,4 [da Cunha et al. 2026].

2.3. Avaliação de Sistemas RAG

A geração aumentada por recuperação foi introduzida por Lewis et al. como forma de combinar a capacidade generativa de modelos de linguagem com recuperação explícita de informação em bases externas [Lewis et al. 2020]. Desde então, a avaliação de sistemas RAG tornou-se um campo próprio, com abordagens que podem incidir sobre a recuperação, a geração ou o pipeline completo, combinando métricas de relevância, fidelidade e correção com diferentes trade-offs entre custo, cobertura e interpretabilidade [Yu et al. 2025]. Frameworks como o RAGAs [Es et al. 2024] e o ARES [Saad-Falcon et al. 2024] oferecem avaliação automatizada do pipeline integrado, enquanto o uso de LLMs como juízes tem se difundido como alternativa prática, embora sua confiabilidade ainda esteja em discussão [Brehme et al. 2025]. Neste trabalho, a avaliação incide exclusivamente sobre a etapa de recuperação, por duas razões. Primeiro, porque isolar essa etapa permite atribuir as diferenças de desempenho diretamente ao modelo de embedding, sem a interferência de variações na geração. Segundo, porque, em domínios documentais normativos, a qualidade da recuperação constitui pré-condição da qualidade da resposta: trechos mal recuperados comprometem o resultado final independentemente do LLM utilizado.

3. Metodologia

3.1. Sistema Proposto

O sistema proposto consiste em um protótipo de assistente conversacional baseado em RAG, desenvolvido para apoiar estudantes de Ciência da Computação em consultas sobre informações institucionais e acadêmicas³. Sua finalidade é concentrar, em uma interface única implementada em Streamlit, conteúdos como regulamentos, calendários, documentos curriculares e informações administrativas do curso, que normalmente se encontram dispersos em diferentes páginas e arquivos institucionais. O protótipo integra recuperação de contexto documental e geração de respostas por um LLM acessado via OpenRouter, caracterizando uma aplicação RAG utilizável no domínio considerado. No entanto, para fins de avaliação experimental, o estudo isola a etapa de recuperação e mantém os demais componentes do sistema constantes, variando apenas o modelo de embedding, permitindo comparar seu impacto na recuperação em um domínio acadêmico real.

3.2. Corpus Institucional

O corpus institucional foi construído a partir de páginas públicas da universidade e documentos institucionais em PDF. Para a primeira, foram desenvolvidos scripts de web

³Código-fonte disponível em: <https://github.com/analyticalcompetitor/compChatUfpi>. Protótipo disponível em: <https://compchat.streamlit.app/>. Acesso em: 25 jul. 2026.

scraping que extraem automaticamente o conteúdo de páginas do site da UFPI relevantes ao contexto acadêmico do curso, como informações sobre docentes, benefícios e procedimentos de estágio. Para a segunda, documentos normativos e administrativos (regulamentos, resoluções, projetos pedagógicos, editais de pesquisa, calendário, requerimentos) foram coletados e processados por scripts de extração de texto de PDF. Todo o conteúdo resultante foi convertido em duas representações textuais, .txt e .md, o que permite investigar o impacto do formato sobre a recuperação.

3.3. Configuração Experimental

3.3.1. Recuperação Densa

O experimento avalia exclusivamente a qualidade da recuperação dos chunks relevantes, isolando o componente que determina quais trechos documentais serão disponibilizados ao modelo generativo. Foram selecionados seis modelos open-source com até 0,6B de parâmetros, escolhidos entre alternativas bem posicionadas no MMTEB Multilingual para tarefas de retrieval, conforme a Tabela 1. Todos foram lançados nos últimos doze meses, à exceção do *multilingual-e5-large-instruct*, incluído por ser frequentemente adotado em trabalhos recentes de recuperação em português. Esse recorte reflete o foco em custo-benefício e escalabilidade, investigando até que ponto modelos compactos mantêm recuperação competitiva frente a alternativas de maior porte. Como referência comparativa, soma-se o modelo proprietário *text-embedding-3-large*.

Tabela 1. Modelos open-source selecionados com base no MMTEB Multilingual, considerando desempenho em retrieval e limite máximo de 0.6B parâmetros.

Modelo	Parâmetros	Retrieval
microsoft/harrier-oss-v1-0.6b	0.6B	70.75
microsoft/harrier-oss-v1-270m	270M	66.38
Qwen/Qwen3-Embedding-0.6B	0.6B	64.65
jinaai/jina-embeddings-v5-text-nano	239M	63.26
google/embeddinggemma-300m	300M	62.49
intfloat/multilingual-e5-large-instruct	0.6B	57.12

A outra variável experimental é o formato textual do corpus. Todos os conteúdos coletados foram preparados em versões .txt e .md, permitindo investigar se diferentes formas de representação textual do mesmo conteúdo afetam a recuperação. Embora ambas as versões preservem informações estruturais do documento, o formato .md torna essa organização mais explícita por meio de marcações de títulos, listas e tabelas, enquanto o .txt oferece uma representação mais simples. Como trabalhos recentes em RAG sugerem que a explicitação da estrutura documental pode influenciar a modelagem do conteúdo recuperado [Tan et al. 2025], cada modelo é avaliado nas duas representações, a fim de verificar empiricamente qual delas produz melhor desempenho no corpus institucional analisado.

Para assegurar comparabilidade entre os testes, a infraestrutura de recuperação permanece fixa em todas as execuções. Em cada condição experimental, o corpus é reindexado em uma base ChromaDB específica para a combinação entre modelo e formato, e

a busca é realizada por um `ThresholdRetriever` com base em similaridade do cosseno, retornando até 6 chunks por consulta. A segmentação dos documentos foi realizada com base em sua estrutura textual, buscando preservar unidades semanticamente coerentes, como seções, subseções, listas e blocos temáticos. Nos documentos mais extensos, marcações `CHUNK` foram geradas automaticamente por scripts de pré-processamento para explicitar os limites entre segmentos consecutivos, permitindo maior controle sobre a indexação dos trechos relevantes no processo de avaliação. A avaliação é conduzida com a biblioteca `ranx`⁴, utilizando as métricas `mrr`, `ndcg@6` e `recall@6`. [Bassani 2022]

3.3.2. Recuperação híbrida por fusão de rankings

A literatura recente mostra que diferentes modelos de embedding apresentam comportamentos distintos dependendo do domínio, da tarefa e do idioma. Mesmo quando dois modelos apresentam desempenho semelhante nas métricas agregadas, eles podem recuperar conjuntos parcialmente diferentes de documentos relevantes. Assim, a combinação de múltiplos modelos pode explorar essa complementaridade e aumentar a cobertura da recuperação.

Para isso, foram utilizados os dois modelos open-source de melhor desempenho na avaliação preliminar: **jinaai/jina-embeddings-v5-text-nano** e **google/embeddinggemma-300m**. Cada modelo gera, de forma independente, um ranking de documentos candidatos. Os rankings são posteriormente combinados por meio do algoritmo **Reciprocal Rank Fusion (RRF)**, que atribui maior peso aos documentos posicionados nas primeiras posições de cada ranking e soma essas contribuições para produzir uma ordenação final. Ao contrário da recuperação baseada em um único modelo de embedding, essa abordagem procura aproveitar a complementaridade entre diferentes representações vetoriais do mesmo corpus, reduzindo falhas específicas de cada modelo.

3.4. Conjunto de Avaliação

O conjunto de avaliação foi particionado em oito datasets de teste temáticos, de modo a representar diferentes segmentos do corpus institucional: *Calendars*, *CGE*, *CGE_Docs*, *Others*, *PPC*, *Professors*, *Regulation* e *Research*. Esses subconjuntos abrangem, respectivamente, calendários acadêmicos, informações da página da Controladoria Geral de Estágios, documentos da CGE, conteúdos sobre TCC, requerimentos e benefícios, o Projeto Pedagógico do Curso, informações sobre docentes, o Regulamento Geral da Graduação e editais PIBIC e PIBIT. Ao todo, foram definidas 250 questões em linguagem natural, distribuídas entre os diferentes recortes temáticos do corpus.

Cada instância do conjunto associa uma consulta aos chunks anotados como relevantes, permitindo avaliar a recuperação no nível da unidade textual efetivamente indexada pelo sistema. A distribuição por dificuldade foi mantida aproximadamente balanceada no conjunto agregado, com 30,8% de questões fáceis, 40% médias e 29,2% difíceis. O subconjunto *CGE_Docs* reúne 60 questões, número superior aos demais, por corresponder à parte mais extensa do corpus em termos de documentos indexados; os outros datasets concentram entre 20 e 30 questões cada.

⁴<https://amenra.github.io/ranx/>.

O conjunto de avaliação foi construído pelos próprios autores, o que pode introduzir viés de formulação não representativo do uso real por estudantes. O corpus é restrito a um único departamento, o que limita a generalização. O total de 250 questões é suficiente para as métricas agregadas, mas restringe a análise nos subconjuntos menores. A avaliação incide exclusivamente sobre a recuperação, sem mensurar a qualidade das respostas geradas, dimensão a ser incorporada com a implantação do protótipo em uso real.

3.5. Protocolo de Avaliação

A avaliação foi conduzida exclusivamente sobre a etapa de recuperação. Para cada uma das 250 consultas do conjunto de avaliação, os chunks relevantes foram manualmente identificados e anotados previamente, de modo que os resultados esperados são conhecidos antes da execução. A avaliação é, portanto, determinística: dado um modelo de embedding e um formato textual, a mesma consulta sempre recupera os mesmos chunks na mesma ordem, o que permite comparações controladas entre configurações.

As anotações de relevância foram estruturadas como *Qrels* (associação entre cada consulta e seus chunks relevantes) e os resultados retornados pelo sistema como *Runs* (ranking dos chunks recuperados com seus escores de similaridade). A comparação entre ambos foi realizada com a biblioteca *ranx* [Bassani 2022], que calcula de forma padronizada as métricas de avaliação. Esse protocolo foi aplicado a todos os modelos de embedding nas duas representações textuais do corpus (.txt e .md). As métricas adotadas foram $Recall@k$, MRR e $NDCG@k$.

3.6. Métricas de Avaliação

O desempenho da recuperação foi mensurado com as métricas $Recall@6$, MRR e $NDCG@6$. No caso de $Recall@6$ e $NDCG@6$, considerou-se $k = 6$, isto é, apenas os seis primeiros chunks retornados pelo sistema para cada consulta foram levados em conta no cálculo. Já a MRR foi computada em sua forma padrão, considerando a posição do primeiro chunk relevante recuperado no ranking.

A $Recall@6$ mede a proporção de chunks relevantes recuperados entre os seis primeiros resultados, permitindo avaliar o grau de cobertura da resposta esperada nesse limite. A MRR privilegia a posição do primeiro item relevante retornado, atribuindo maior valor aos casos em que ele aparece mais cedo no ranking. Por sua vez, a $NDCG@6$ avalia a qualidade da ordenação dos resultados ao considerar a posição dos chunks relevantes entre os seis primeiros itens recuperados.

4. Resultados

4.1. Desempenho Geral dos Modelos

A Tabela 2 apresenta os resultados agregados de todos os modelos, ordenados por $NDCG@6$ decrescente. O modelo proprietário *text-embedding-3-large* obteve o melhor desempenho geral, seguido de perto por dois modelos open-source: *jina-embeddings-v5-text-nano* e *embeddinggemma-300m*, ambos com $NDCG@6$ acima de 0,74. O *multilingual-e5-large-instruct* registrou os menores valores em todas as métricas. Em relação ao formato textual, o TXT supera o MD na maioria dos modelos, embora a diferença seja expressiva apenas nos dois modelos de menor desempenho geral; nos demais, a variação é inferior a 0,005 pontos de $NDCG@6$.

Tabela 2. Resultados gerais dos modelos avaliados, ordenados por NDCG@6 decrescente. O melhor resultado por métrica está em negrito.

Modelo	Fmt.	NDCG@6	MRR	Recall@6
openai/text-embedding-3-large	TXT	0.7829	0.7937	0.8531
openai/text-embedding-3-large	MD	0.7783	0.7822	0.8632
jinaai/jina-embeddings-v5-text-nano	TXT	0.7487	0.7535	0.8361
jinaai/jina-embeddings-v5-text-nano	MD	0.7467	0.7483	0.8499
google/embeddinggemma-300m	TXT	0.7465	0.7555	0.8250
google/embeddinggemma-300m	MD	0.7437	0.7512	0.8217
microsoft/harrier-oss-v1-0.6b	TXT	0.7245	0.7242	0.8174
microsoft/harrier-oss-v1-0.6b	MD	0.7208	0.7276	0.8035
Qwen/Qwen3-Embedding-0.6B	MD	0.7209	0.7196	0.8189
Qwen/Qwen3-Embedding-0.6B	TXT	0.7198	0.7206	0.8165
microsoft/harrier-oss-v1-270m	TXT	0.6850	0.6745	0.7971
microsoft/harrier-oss-v1-270m	MD	0.6654	0.6581	0.7688
intfloat/multilingual-e5-large-instruct	TXT	0.5993	0.5885	0.7181
intfloat/multilingual-e5-large-instruct	MD	0.5789	0.5733	0.6807

4.2. Análise por Subconjunto do Corpus

A Figura 1 e a Figura 2 apresentam os resultados de NDCG@6 e Recall@6 por modelo e subconjunto do corpus. A distribuição cromática evidencia que nenhum modelo é uniformemente superior: o *text-embedding-3-large* apresenta o melhor desempenho agregado, mas no *CGE* e *Regulation* o *jina-embeddings-v5-text-nano* o supera em ambas as métricas. Os subconjuntos *Others* e *CGE_Docs* concentram os valores mais elevados para todos os modelos, enquanto o *PPC* registra os menores em ambas as figuras. Os subconjuntos restantes apresentam desempenho intermediário com maior variação entre modelos.

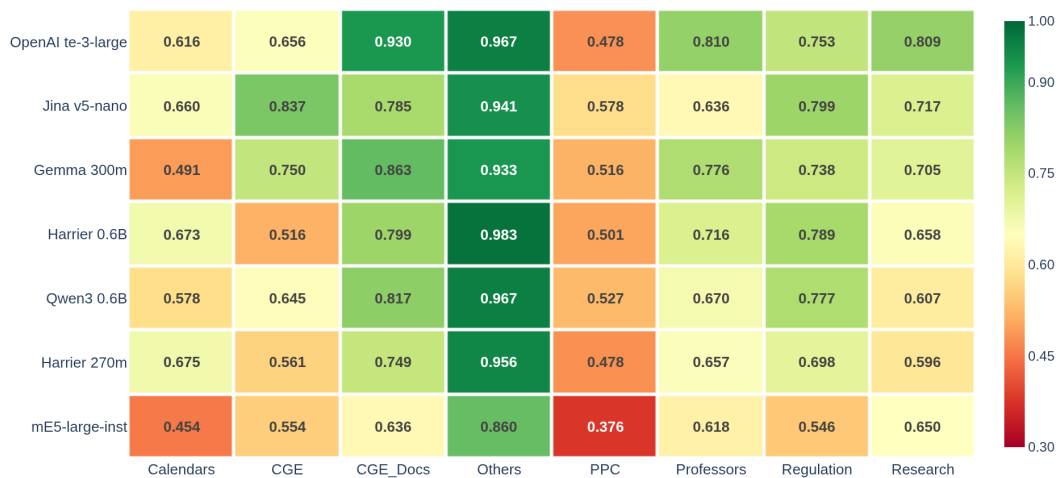


Figura 1. NDCG@6 por modelo de embedding e subconjunto do corpus (formato TXT). Escala cromática: verde indica maior desempenho; vermelho indica menor desempenho.

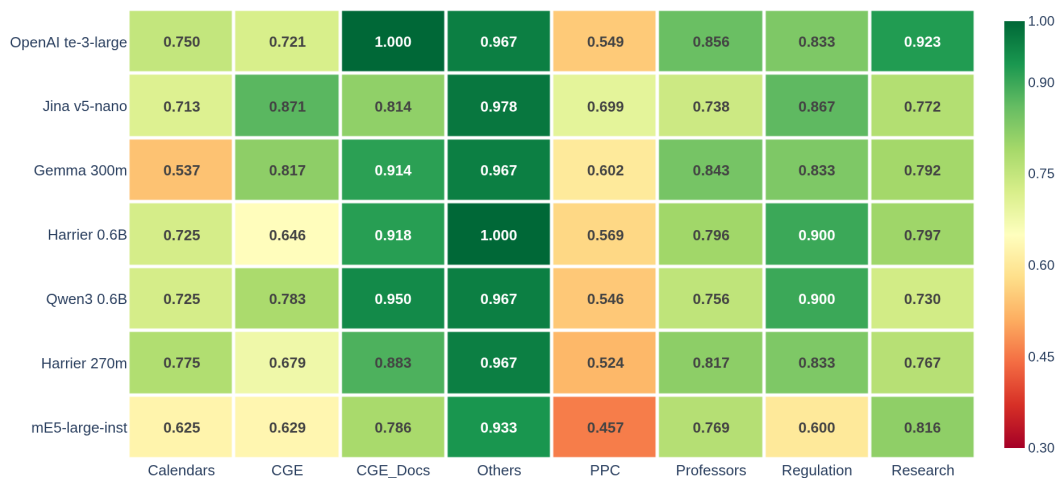


Figura 2. Recall@6 por modelo de embedding e subconjunto do corpus (formato TXT). Escala cromática: verde indica maior desempenho; vermelho indica menor desempenho.

4.3. Recuperação híbrida por fusão de rankings

Os resultados da recuperação híbrida demonstram que a combinação de múltiplos modelos de embedding pode ser mais eficaz do que a utilização de um único modelo. A estratégia baseada em Reciprocal Rank Fusion, combinando os modelos jina-embeddings-v5-text-nano e embeddinggemma-300m, alcançou **MRR de 0,8001**, **NDCG@6 de 0,7976** e **Recall@6 de 0,8829**, representando o melhor desempenho geral entre todas as configurações experimentais.

4.4. Análise de Significância Estatística

Diferenças numéricas pequenas nem sempre representam diferenças estatisticamente confiáveis. Para verificar essa questão, foi aplicado o teste de randomização de Fisher ($p \leq 0,05$). A Tabela 3 apresenta os resultados, em que os sobrescritos indicam os modelos para os quais a diferença foi estatisticamente significativa.

Tabela 3. Resultados para o corpus TXT com indicação de diferenças estatisticamente significativas segundo o teste de randomização de Fisher ($p \leq 0,05$).

#	Model	MRR	NDCG@6	Recall@6
a	Qwen_Qwen3-Embedding-0.6B	0.721 ^{dg}	0.720 ^d	0.817 ^d
b	google_embeddinggemma-300m	0.756 ^{dg}	0.747 ^{dg}	0.825 ^d
c	hybrid_rrf	0.800 ^{abdefg}	0.798 ^{abdefg}	0.883 ^{abdefg}
d	intfloat_multilingual-e5-large-instruct	0.589	0.599	0.718
e	jinaai_jina-embeddings-v5-text-nano	0.754 ^{dg}	0.749 ^{dg}	0.836 ^d
f	microsoft_harrier-oss-v1-0.6b	0.724 ^{dg}	0.724 ^{dg}	0.817 ^d
g	microsoft_harrier-oss-v1-270m	0.675 ^d	0.685 ^d	0.797 ^d
h	openai_text-embedding-3-large	0.794 ^{adfg}	0.783 ^{adfg}	0.853 ^{dg}

A estratégia híbrida baseada em Reciprocal Rank Fusion apresentou o maior desempenho em todas as métricas. Seu desempenho foi significativamente superior a todos os modelos open-source individuais. Entretanto, não foi observada diferença significativa em relação ao text-embedding-3-large, indicando desempenho estatisticamente equivalente entre ambas as abordagens.

5. Discussão

5.1. Desempenho dos Modelos e Não Transferibilidade de Benchmarks

A distância entre o melhor e o pior modelo não é marginal: apenas a troca do embedding, com todos os demais componentes fixos, produz quase 15 pontos de diferença no NDCG@6 e faz com que o modelo inferior deixe de recuperar cerca de 12% a mais dos chunks relevantes. A escolha do embedding é, portanto, uma decisão de arquitetura com impacto substancial, confirmando no domínio acadêmico institucional o que [Canale et al. 2025] demonstrou em configurações RAG de propósito geral.

Entre os open-source, *jina-embeddings-v5-text-nano* e *embeddinggemma-300m* se destacam com desempenho próximo ao proprietário, ambos com até 300M de parâmetros e executáveis localmente. Dois resultados, porém, contrariam expectativas de rankings gerais: o *multilingual-e5-large-instruct*, competitivo em outros domínios em português [Medeiros and de Oliveira 2025, da Cunha et al. 2026], registrou o pior desempenho neste corpus; os modelos Harrier, mais bem posicionados no MMTEB Multilingual entre os testados, ficaram abaixo do Jina e do Gemma, que possuem scores consideravelmente menores no benchmark. Ambos os casos ilustram que desempenho em benchmarks gerais não se transfere para domínios específicos [Enevoldsen et al. 2025, Muennighoff et al. 2023]. O MMTEB é útil como triagem, mas sua ordenação não pode ser lida como predição de desempenho em uma aplicação concreta.

A influência do formato textual mostrou-se limitada. Em cinco dos sete modelos, a diferença de NDCG@6 entre TXT e MD é inferior a 0,005 pontos. As variações mais expressivas ocorrem nos modelos de menor desempenho geral (*harrier-oss-v1-270m* e

multilingual-e5-large-instruct), o que sugere que modelos com menor capacidade de representação semântica são mais sensíveis à forma de apresentação do texto, enquanto modelos mais robustos absorvem essas diferenças.

Esses resultados podem servir como referência inicial para outros chatbots acadêmicos que operem sobre documentação semelhante. A consistência do Jina e do Gemma em subconjuntos temáticos variados sugere que modelos com esse perfil tendem a ser competitivos em corpora institucionais em português. Essa indicação, contudo, não substitui avaliação no corpus próprio de cada aplicação: como os dados deste trabalho demonstram, a relação entre posição em benchmarks e desempenho em domínio específico é frágil o suficiente para que qualquer adoção sem verificação empírica represente um risco evitável.

5.2. Análise por Subconjunto do Corpus

Nenhum modelo domina uniformemente todos os subconjuntos. O *text-embedding-3-large* mantém vantagem mais consistente no *CGE_Docs*, o subconjunto mais extenso, mas no *CGE* o *jina-embeddings-v5-text-nano* o supera com margem expressiva, demonstrando que a superioridade agregada pode mascarar fragilidades localizadas. O subconjunto *Others*, que reúne questões sobre documentos de TCC, requerimentos e benefícios da PRAEC, concentra os melhores resultados para todos os modelos. Esse desempenho era esperado. As questões formuladas para esse eixo são mais diretas do que as dos demais subconjuntos, e o corpus contém pouco material sobre esses temas, o que reduz a competição entre chunks e facilita a recuperação do trecho correto.

O *PPC* representa o caso oposto. Consultas sobre carga horária de estágio recuperam com prioridade cláusulas de termos de convênio da CGE, que compartilham o mesmo campo semântico, antes do trecho do PPC que contém a tabela curricular. Perguntas sobre avaliação retornam artigos do Regulamento Geral da Graduação antes do chunk correspondente do PPC. Diferentemente do *Others*, aqui os chunks-alvo empregam a mesma terminologia normativa que os documentos concorrentes, e o embedding, que captura proximidade semântica mas não distingue fonte documental, não consegue resolver essa competição. A escolha do modelo continua relevante nesse subconjunto, já que a distância entre o melhor e o pior permanece expressiva, mas mesmo o modelo de melhor desempenho não atinge resultados satisfatórios. Isso sugere que, para subconjuntos com esse perfil, a escolha de um bom modelo é condição necessária mas não suficiente, e que ganhos adicionais provavelmente exigem estratégias complementares para uma recuperação mais eficiente.

5.3. Recuperação Híbrida e Viabilidade Operacional

A estratégia híbrida por Reciprocal Rank Fusion, combinando *jina-embeddings-v5-text-nano* e *embeddinggemma-300m*, alcançou o melhor desempenho global. O resultado é significativo não pelo valor absoluto das métricas, mas pelo que implica: dois modelos que individualmente ficam abaixo do proprietário, combinados, atingem desempenho estatisticamente equivalente a ele ($p \leq 0,05$, teste de randomização de Fisher). Isso indica que os modelos recuperam conjuntos parcialmente complementares de chunks e que a fusão de rankings explora essa complementaridade de forma efetiva. Simultaneamente, o método híbrido foi estatisticamente superior a todos os modelos individuais open-source,

confirmando que a combinação produz ganho real. Essa equivalência com o proprietário deve ser ponderada pelo custo computacional de manter dois modelos em inferência simultânea.

Em contextos com restrições severas de hardware ou latência, a decisão entre recuperação híbrida e modelo único dependerá dos recursos disponíveis. A análise estatística também delimita com precisão quais modelos são efetivamente competitivos. O *text-embedding-3-large* não apresentou vantagem significativa sobre Jina e Gemma individualmente, indicando que a margem nas métricas agregadas não é estatisticamente robusta. Por outro lado, foi superior a todos os demais open-source. Apenas Jina e Gemma constituem, neste corpus, alternativas reais à solução proprietária. Além da equivalência de desempenho e da eliminação do custo por token, os modelos open-source preservam a possibilidade de ajuste fino sobre o corpus institucional, o que não é viável para o *text-embedding-3-large*. Essa vantagem, contudo, não foi testada pois não faz parte do escopo deste trabalho.

6. Conclusão

Este trabalho apresentou um protótipo funcional de chatbot institucional baseado em RAG para suporte acadêmico do curso de ciência da computação da UFPI, e conduziu uma avaliação comparativa de modelos de embedding compactos aplicada ao corpus documental da universidade. O protótipo, desenvolvido com foco em baixo custo operacional e viabilidade de implantação local, representa o primeiro passo de uma iniciativa voltada a oferecer soluções reais de atendimento acadêmico automatizado na instituição. Os resultados da avaliação indicam que o *jina-embeddings-v5-text-nano* e o *embeddinggemma-300m* constituem as escolhas mais adequadas para o sistema, com desempenho próximo ao do *text-embedding-3-large* e possibilidade de execução local sem custo de inferência. O experimento também mostrou que o *multilingual-e5-large-instruct*, adotado como baseline em trabalhos de recuperação em português, apresentou o pior desempenho neste corpus, reforçando que a escolha do embedding deve ser avaliada nos dados reais da aplicação. O formato textual exerceu influência limitada e não representou fator de decisão crítico para os modelos de melhor desempenho.

Como próximos passos, pretende-se implantar o protótipo em ambiente de uso real na universidade, coletando perguntas formuladas por estudantes reais para construir um conjunto de avaliação mais representativo do domínio e evoluir a análise para métricas que englobem o pipeline completo, incluindo a qualidade das respostas geradas. Este trabalho estabelece a base técnica e experimental para que essa iniciativa possa ser continuada e expandida por trabalhos futuros no contexto da instituição.

Declaração sobre uso de Inteligência Artificial

Durante a elaboração deste trabalho, ferramentas de inteligência artificial generativa foram utilizadas de forma assistida e responsável, como suporte a tarefas de escrita e implementação de código, sem delegação de julgamentos científicos ou metodológicos. O desenho da pesquisa, a definição dos experimentos, a interpretação dos resultados e as conclusões são de autoria exclusiva dos pesquisadores. Para auxílio na escrita e revisão do texto, foram utilizados os modelos Claude Sonnet 4.6 e GPT 5.4 via Perplexity. Para auxílio na implementação do código, foram utilizados os modelos Claude Opus 4.6

e Gemini 3.1 via Antigravity (editor de código do Google). Os autores assumem total responsabilidade pelo conteúdo, experimentos, análises e resultados apresentados.

Referências

- Almeida, E. J. C. (2024). Otimização da comunicação institucional na ufcg através da implementação de chatbots: uma estratégia inovadora para melhoria do atendimento ao discente. Relatório técnico (mestrado profissional em administração pública – profiap), Universidade Federal de Campina Grande.
- Anibal, E. L. d. A. (2025). Avaliação da usabilidade do chatbot minhaufc. Trabalho de conclusão de curso, Universidade Federal do Ceará.
- Bassani, E. (2022). ranx: A blazing-fast python library for ranking evaluation and comparison. In *European Conference on Information Retrieval*, pages 259–264. Springer.
- Bordin, A. C. B., Lunardi, G. M., Piveta, E. K., and Emmendorfer, L. (2025). Unichat: Arquitetura e avaliação de um agente rag de baixo custo para o contexto universitário. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 886–892. SBC.
- Brehme, L., Ströhle, T., and Breu, R. (2025). Can llms be trusted for evaluating rag systems? a survey of methods and datasets. In *2025 IEEE Swiss Conference on Data Science (SDS)*, pages 16–23. IEEE.
- Cabral, J. P. F. (2025). Chatbot inteligente para acesso a regulamentos acadêmicos: um sistema de recuperação de informações baseado em rag. Trabalho de conclusão de curso, Universidade Federal do Rio Grande do Norte.
- Canale, L., Scotta, S., Messina, A., and Farinetti, L. (2025). Bes4rag: A framework for embedding model selection in retrieval-augmented generation. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 134–142.
- Chaves, L. N. A., de Oliveira, M. A., Pereira, M. L. A., Kanda, J. Y., dos Santos, J. V. S., Soares, A. C., and de Sousa, S. R. G. (2024). Explorando a viabilidade de chatbot na comunicação do programa de integridade: o caso da ufam. *Peer Review Journals*, 16(16):242–261.
- Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., et al. (2024). Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.
- Coelho, G. S. d. S. (2024). Chatbot para o serviço de referência da biblioteca setorial berilo borba do centro de ciências sociais aplicadas da universidade federal da paraíba. Trabalho de conclusão de curso, Universidade Federal da Paraíba.
- da Cunha, M. V., Silveira, M. R., Sperb, C. B., Santana, B. S., Freitas, L. A., and Corrêa, U. B. (2026). Avaliação end-to-end de um sistema rag para documentos hospitalares em português. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)-Vol. 1*, pages 749–758.
- da Silva, T. P. (2021). Desenvolvimento de um sistema do tipo chatbot para o curso de sistemas de informação. Trabalho de conclusão de curso, Universidade Federal de Santa Catarina.

- de Andrade, G. G., Silva, G. R. S., Júnior, F., Santos, G. A., de Mendonça, F. L. L., and de Sousa Júnior, R. T. (2020). Evatalk: A chatbot system for the brazilian government virtual school. In *Proceedings of the 22nd International Conference on Enterprise Information Systems (ICEIS 2020)*.
- de Carvalho, B. G. (2025). Desenvolvimento de um chatbot educacional para apoio aos estudantes do curso de engenharia de software da ufc-campus quixadá. Trabalho de conclusão de curso, Universidade Federal do Ceará.
- de Melo, L. B. (2023). Regbot: chatbot para o regulamento de ensino de graduação da universidade federal de campina grande. Trabalho de conclusão de curso, Universidade Federal de Campina Grande.
- de Oliveira, A. J. D. (2024). Implementação de um chatbot em uma universidade pública: aprimorando a experiência do atendimento. Dissertação (mestrado profissional em administração), Universidade Federal Fluminense.
- de Souza, A. E. B. (2024). Proposta de chatbot inteligente baseado na organização acadêmica do instituto federal de pernambuco. Trabalho de conclusão de curso, Instituto Federal de Pernambuco.
- de Souza, W. F. V., Nogueira, I. L., and de Souza, A. V. (2025). Assistente virtual inteligente para instituições educacionais: Aplicação da técnica rag no ifbot. In *Workshop sobre as Implicações da Computação na Sociedade (WICS)*, pages 217–224. SBC.
- do Amaral, V. M. R., Silveira, L. E., Rigo, P. D., and Lunardi, G. M. (2025). A rag-powered academic chatbot with ontology-driven factual verification. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 404–414. SBC.
- do Nascimento, D. F. G. (2024). Projeto de aplicação chatbot web para o gerenciamento de chamados de atendimento: um estudo de caso na ufpb. Trabalho de conclusão de curso, Universidade Federal da Paraíba.
- dos Passos, K. G. F., Silva, L. M., Picalho, A. C., and Fadel, L. M. (2025). Agentes de inteligência artificial aplicados aos serviços de bibliotecas universitárias: Genspark e manusai. *XXIII Seminário Nacional de Bibliotecas Universitárias*, pages 1–13.
- Elias, G. C. (2024). Uma alternativa de comunicação: Como o chatbot pode ser uma opção para a coordenação de registros acadêmicos de um campus do ifsp. In *Congresso de Inovação, Ciência e Tecnologia do IFSP*, volume 15.
- Enevoldsen, K., Chung, I., Kerboua, I., Kardos, M., Mathur, A., Stap, D., Gala, J., Siblini, W., Krzemiński, D., Winata, G. I., et al. (2025). Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*.
- 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.
- Felipe, R. T. T., Arnoni, T. H., and Ramos, A. M. (2025). Sofiá: Um chatbot baseado em ia para suporte à gestão de estágios em universidades públicas. In *Anais do XXIV Congresso Internacional de Gestão Universitária*. INPEAU/UFSC.

- Gonçalves, W. A., dos Anjos e Santos, M. G. T., Leal, L., Rodrigues, N. S., and de Souza, E. P. (2025). Integração de chatbots com inteligência artificial ao ensino superior: estudo sobre seu desempenho na mediação da aprendizagem individualizada. *Educere et Educare*, 20(50):151–166.
- Layane, L., Machado, G., Morandi, G., Silva, N., Junior, W., and Alves, E. (2025). Chatbot customizável baseado em ia generativa para informação de conteúdos acadêmicos de instituições de ensino superior públicas. In *Latin American Symposium on Digital Government (LASDiGov)*, pages 169–180. SBC.
- 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.
- Medeiros, L. S. F. and de Oliveira, H. T. A. (2025). Comparação de modelos de embeddings e llms para geração aumentada por recuperação em português. In *Seminário Integrado de Software e Hardware (SEMISH)*, pages 429–440. SBC.
- Monteiro, A. S. (2023). A transformação digital da comunicação na administração pública: uma análise da viabilidade do uso de um chatbot na secretaria de recursos humanos da universidade federal de campina grande. Dissertação (mestrado profissional em administração pública), Universidade Federal de Campina Grande.
- Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. (2023). Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.
- Oliveira, S. C. (2025). Chatbot com inteligência artificial como ferramenta de apoio à comunicação acadêmica: Uma proposta para o departamento de ciências da administração da ufsc. Trabalho de conclusão de curso, Universidade Federal de Santa Catarina.
- Pantoja, A. S., Nascimento, K. W. S., de Lima Duarte, G., Silva, A. B. A., and de Jesus Martins, Y. (2025). Desenvolvimento de um chatbot inteligente para auxílio de atividades organizacionais durante o período de processos seletivos da ufpa. In *Workshop sobre Aspectos Sociais, Humanos e Econômicos de Software (WASHES)*, pages 143–154. SBC.
- Petrin, N. S. M. and Néto, J. C. (2026). Empathiced: Um chatbot educacional com rag e computação afetiva em um contexto de iniciação científica coordenada. *ARACÊ*, 8(3):e12683–e12683.
- Rayzer, L. P. B. and Santos, F. (2025). Um LLM com RAG para as normativas institucionais da UDESC. In *Escola Regional de Aprendizado de Máquina e Inteligência Artificial da Região Sul (ERAMIA-RS)*, pages 224–227. SBC.
- Rayzer, L. P. B. and Santos, F. (2026). Explorando o uso de LLMs com RAG para simplificar o acesso às normativas da universidade do estado de santa catarina. *Regae: Revista de Gestão e Avaliação Educacional*.
- Rodrigues, G. B., Beal, F., Marcon, M., Souza, A. C. C., Ortoncelli, A. R., Souza, F. C. M., and Silva, R. A. (2025). Sabiá: Um chatbot de inteligência artificial generativa para suporte no dia a dia do ensino superior. *arXiv preprint arXiv:2511.10787*.

- Saad-Falcon, J., Khattab, O., Potts, C., and Zaharia, M. (2024). Ares: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 338–354.
- Silva, A. L. S., da Silva, F. R. P., Lopes, C. H. S., Freire, A. R., da Penha Francisco, L., and de Campos da Silva, G. (2023). A implantação do whatsapp institucional da unicamp: a universidade ao alcance das mãos. In *Resumo dos trabalhos do SIMTEC Simpósio dos Profissionais da UNICAMP*.
- Soares, R. R. P., Cardoso, F. E. A., Costa, N. H. F., and Melo, R. F. d. (2025). Desenvolvimento de um assistente virtual como suporte acadêmico do curso de licenciatura em computação usando ia generativa personalizada e rag. *Revista Sítio Novo*, 9:e1757.
- Spricigo, V. E. (2026). Chatbot RAG personalizado de dados em modelos de linguagem natural generativos. Trabalho de conclusão de curso, Instituto Federal Goiano.
- Tan, J., Dou, Z., Wang, W., Wang, M., Chen, W., and Wen, J.-R. (2025). Htmlrag: Html is better than plain text for modeling retrieved knowledge in rag systems. In *Proceedings of the ACM on Web Conference 2025, WWW '25*, page 1733–1746. Association for Computing Machinery.
- Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., and Liu, Z. (2025). Evaluation of retrieval-augmented generation: A survey. In *Big Data*, pages 102–120, Singapore. Springer Nature Singapore.