RAG em Domínio Normativo: Avaliação de Recuperadores e Diagnóstico de Viés em Dados Sintéticos do PPC-CC

Maurício Catanio [1]
Guilherme Dal Bianco [2]

Resumo:

Sistemas de Recuperação Aprimorada Generativa (RAG) aumentam a confiabilidade de Modelos de Linguagem de Grande Escala (LLMs) ao fundamentar suas respostas em fontes externas. Contudo, a eficácia do RAG depende criticamente do componente de recuperação de informação (RI), cujo comportamento em domínios técnicos permanece pouco explorado. Este trabalho realiza uma investigação experimental e qualitativa dos fatores que influenciam o desempenho de recuperadores léxicos (BM25, TF-IDF), densos (*embeddings* especializados) e híbridos (SPLADE) no contexto de um caso real: um sistema de perguntas e respostas sobre o Projeto Pedagógico do Curso (PPC) de Ciência da Computação da UFFS. Os resultados demonstram que, neste domínio normativo, métodos léxicos tradicionais apresentaram desempenho superior ou equivalente ao de abordagens neurais mais complexas. A análise revelou ainda: (i) vieses estruturais em *datasets* sintéticos de avaliação, que inflam artificialmente as métricas de métodos baseados em termos; e (ii) limitações práticas significativas na transferência do modelo híbrido SPLADE para o português brasileiro, devido a expansões lexicais ruidosas. O estudo evidencia que a escolha do método de RI deve considerar as características intrínsecas do domínio-alvo, desafiando a pressuposição de superioridade automática das abordagens neurais.

[1]Ciência da Computação, Universidade Federal da Fronteira Sul (UFFS)
mauricatanio@yahoo.com.br
[2]Ciência da Computação, Universidade Federal da Fronteira Sul (UFFS)
guilherme.dalbianco@uffs.edu.br

Abstract: Retrieval-Augmented Generation (RAG) systems enhance the reliability of Large Language Models (LLMs) by grounding responses in external sources. However, RAG effectiveness critically depends on the information retrieval (IR) component, whose behavior in specific technical domains remains underexplored. This work conducts an experimental and qualitative investigation of the factors influencing the performance of lexical (BM25, TF-IDF), dense (specialized embeddings), and hybrid (SPLADE) retrievers in a real-world case: a question-answering system for the Computer Science Course Pedagogical Project (PPC) at UFFS. Results demonstrate that, in this normative domain, traditional lexical methods performed equally or superiorly to more complex neural approaches. The analysis further revealed: (i) structural biases in synthetic evaluation datasets, which artificially inflate metrics for term-based methods; and (ii) significant practical limitations in transferring the hybrid SPLADE model to Brazilian Portuguese, due to noisy lexical expansions. The study highlights that the choice of IR method must consider the intrinsic characteristics of the target domain, challenging the assumed automatic superiority of neural approaches.

***P*alavras-chave:** Recuperação Aprimorada Generativa, Recuperação de Informação, Modelos de Linguagem, Documentos Institucionais.

***K*eywords:** Retrieval-Augmented Generation, Information Retrieval, Language Models, Institutional Documents.

1 Introdução

O avanço dos Modelos de Linguagem de Grande Escala (LLMs), como GPT, BERT e LLaMA, consolidou novas formas de interação humano-máquina, com aplicações em tradução, geração de conteúdo e sistemas de busca. Com a popularização de interfaces como o ChatGPT, esses modelos passaram a integrar rotinas acadêmicas e corporativas, inclusive entre estudantes de ciência da computação, que relatam uso frequente e receptividade positiva a essas tecnologias (1).

Apesar desse impacto, LLMs generalistas apresentam limitações em domínios técnicos com poucos dados; por exemplo, na extração de informações de documentos institucionais. Entre os desafios recorrentes estão a geração de informações incorretas (alucinações), a dificuldade de adaptação sem retreinamento e a ausência de rastreabilidade das fontes utilizadas na geração das respostas (2). Nesse cenário, a abordagem de Recuperação Aprimorada Generativa (RAG) surge como uma alternativa promissora, combinando a geração textual baseada em LLMs com mecanismos de recuperação de trechos relevantes, de modo a melhorar a precisão, a auditabilidade e o alinhamento factual das respostas (3).

A técnica RAG funciona de forma modular em três fases principais: recuperação e geração. Primeiro, os documentos são preprocessados e segmentados em trechos curtos (*chunks*). Depois, são indexados em estruturas adequadas, de acordo com a arquitetura do sistema (e.g. banco de dados vetorial). Durante a recuperação, é passada ao sistema de RAG uma consulta; o sistema então recupera os trechos mais relevantes segundo algum modelo de recuperação da informação (RI). Em seguida, na fase de geração, esses trechos são fornecidos como contexto adicional ao LLM, que gera uma resposta fundamentada nas evidências recuperadas. Essa integração permite que o modelo utilize conhecimento externo sem necessidade de ajuste fino.

Para a etapa de RI, destacam-se três famílias metodológicas amplamente utilizadas: (i) a recuperação esparsa, baseada na contagem e correspondência de termos; (ii) a recuperação densa, baseada em representações vetoriais contínuas de tamanho fixo; e (iii) métodos híbridos, que combinam aspectos das abordagens esparsas e densas. Nesta terceira família inserem-se, por exemplo, modelos baseados em expansão lexical neural, como o SPLADE, que podem ser entendidos como híbridos do ponto de vista representacional, por empregar modelos densos para induzir representações esparsas compatíveis com sistemas léxicos tradicionais.

Em um trabalho recente (4), foi proposta uma aplicação de RAG na Universidade Federal da Fronteira Sul (UFFS), com o objetivo de auxiliar estudantes a esclarecer dúvidas sobre o Projeto Pedagógico de Curso de Ciência da Computação (PPC-CC). Essa versão inicial, construída com arquitetura simplificada e métodos densos de recuperação, demonstrou viabilidade funcional e estabeleceu resultados de base sólidos para investigações posteriores.

O objetivo inicial deste trabalho foi investigar experimentalmente o desempenho de recuperadores léxicos (TF–IDF, BM25), densos (incorporações baseadas em *transformers*) e híbridos, com foco no *Sparse Lexical and Expansion Model for First Stage Ranking* (SPLADE), no domínio específico. Contudo, ao longo do desenvolvimento, três fatores alteraram o rumo do estudo: (i) o conjunto de consultas sintéticas apresentou viés lexical, favorecendo métodos esparsos tradicionais; (ii) o treinamento limitado do SPLADE, realizado com poucas épocas e *dataset* ruidoso, produziu expansões pouco informativas; e (iii) representações densas generalistas mostraram tendência a recuperar trechos conceituais amplos, perdendo precisão lexical em documentos longos e fragmentados.

Diante desses desdobramentos empíricos, o foco do trabalho deslocou-se de uma comparação puramente quantitativa para uma análise qualitativa das falhas, dos vieses e dos comportamentos característicos de cada abordagem de recuperação. Assim, foi realizado um estudo para compreender como as características do domínio, do conjunto de dados sintético e do treinamento influenciam o desempenho dos recuperadores. Essa análise, por consequência, oferece contribuições práticas para o desenvolvimento de sistemas RAG aplicados a documentos institucionais e normativos.

2 Fundamentação Teórica

2.1 Representações Léxicas e Semânticas em Recuperação da Informação

A Recuperação da Informação (RI) é uma área da ciência da computação voltada à identificação de documentos relevantes a partir de consultas textuais. Tradicionalmente, os métodos de recuperação dividem-se em abordagens baseadas em correspondência exata de termos (esparsas) e abordagens baseadas em correlações semânticas (densas).

A abordagem mais simples de representação esparsa é o *Bag of Words* (BoW), onde a frequência do termo t no documento d é denotada por $tf(t, d)$. Contudo, a contagem bruta tende a favorecer termos comuns e pouco informativos (como artigos e preposições). Para mitigar isso, o modelo *Term Frequency–Inverse Document Frequency* (TF-IDF) introduz um fator de penalidade para termos muito frequentes na coleção. O peso $w_{t,d}$ é dado por:

$$w_{t,d} = tf(t, d) \times \log\left(\frac{N}{df(t)}\right) \tag{1}$$

onde N é o número total de documentos e $df(t)$ é a quantidade de documentos que contêm o termo t.

O BM25 (*Best Matching 25*) é uma evolução probabilística desse conceito e é considerado o estado da arte para recuperadores esparsos (5). Ele refina o TF-IDF ao introduzir a normalização pelo tamanho do documento e a saturação da frequência do termo. A pontuação de um documento d para uma consulta Q contendo termos q_i é dada por:

$$\text{score}(d, Q) = \sum_{q_i \in Q} IDF(q_i) \cdot \frac{tf(q_i, d) \cdot (k_1 + 1)}{tf(q_i, d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{avgdl}\right)} \tag{2}$$

Nesta Equação, $|d|$ é o tamanho do documento, $avgdl$ é o tamanho médio dos documentos na coleção, e k_1 e b são hiperparâmetros ajustáveis. O parâmetro b controla o impacto da normalização de comprimento, penalizando documentos muito longos que poderiam ter mais correspondências apenas por sua extensão, uma característica importante ao lidar com a fragmentação de documentos normativos como o PPC.

Apesar de sua eficácia e simplicidade, representações léxicas possuem limitações importantes. Elas não capturam relações semânticas entre palavras. Termos como “cachorro” e “canino” são tratados como completamente distintos, e os vetores tendem a ser altamente esparsos, uma vez que a maioria dos termos do vocabulário não aparece em cada documento (6).

Para superar essas limitações, surgiram os métodos de representação densa, também conhecidos como incorporações de palavras (*word embeddings*). Esses métodos mapeiam palavras para vetores em espaços contínuos, de forma que distâncias ou direções no espaço reflitam características semânticas. Modelos como o *word2vec* (7) consolidaram essa linha de pesquisa ao representar palavras de maneira a preservar relações sintáticas e semânticas por meio de operações vetoriais.

Com o avanço dos modelos de linguagem, surgiram representações contextuais, como no BERT (*Bidirectional Encoder Representations from Transformers*) (8). Diferentemente dos embeddings fixos, o BERT gera vetores dependentes do contexto em que cada palavra aparece. Assim, palavras polissêmicas, como "banco" (instituição financeira) e "banco" (assento), recebem vetores distintos conforme o uso no texto. Esse avanço possibilitou o desenvolvimento de recuperadores densos em que consultas e documentos são comparados por similaridade vetorial, frequentemente por distância de cosseno.

Em suma, cada abordagem possui vantagens distintas. Representações esparsas são mais interpretáveis, eficientes e mantêm alta precisão para consultas literais. Por outro lado, representações densas oferecem maior flexibilidade semântica, permitindo recuperar documentos relevantes mesmo sem coincidência exata de termos, embora a um custo computacional mais elevado e com um potencial aumento de ruído nos resultados.

2.2 Comparando Vetores: Similaridade e Distância

Modelos de recuperação densa fundamentam-se na hipótese de que as representações densas ocupam um espaço vetorial semântico, onde a posição de cada vetor codifica propriedades linguísticas aprendidas durante o treinamento(9). Nesse espaço, assume-se que a proximidade geométrica reflete a similaridade de significado. Para quantificar essa proximidade, utilizam-se métricas geométricas. A escolha da métrica impacta diretamente como o sistema interpreta a relevância.

A Similaridade do Cosseno indica o alinhamento angular entre dois vetores, ignorando suas magnitudes (comprimentos). Conforme ilustrado na Figura 1, o foco recai exclusivamente na abertura do ângulo θ entre os vetores A e C. É definida matematicamente por:

$$\text{sim}_{\cos}(A, B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{3}$$

Essa métrica assume valores no intervalo $[-1, 1]$, onde 1 indica vetores com a mesma orientação (alta similaridade semântica), enquanto valores próximos de 0 indicam vetores aproximadamente ortogonais, sugerindo ausência de relação semântica relevante.

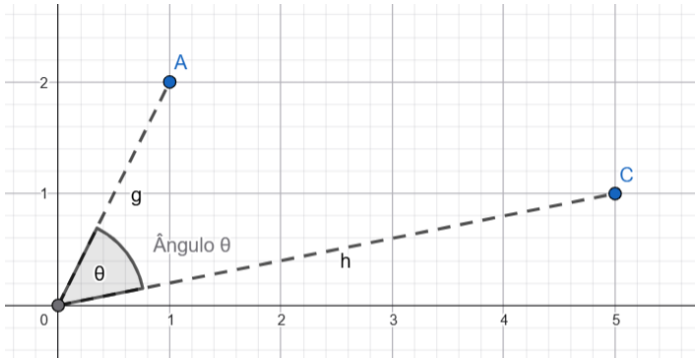

Figura 1. Representação geométrica da Similaridade do Cosseno. A métrica avalia a proximidade semântica com base no ângulo θ entre os vetores A e C, sendo indiferente ao comprimento (magnitude) das retas g e h Fonte: Elaborado pelo autor.

Por sua vez, a **Distância Euclidiana (L^2)** mede a separação linear absoluta entre dois vetores. Considerando os mesmos pontos apresentados na Figura 1, a Figura 2 ilustra que essa métrica corresponde ao comprimento do segmento de reta que conecta os vetores A e B:

$$d_{\text{euc}}(A, B) = \|A - B\| = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} \tag{4}$$

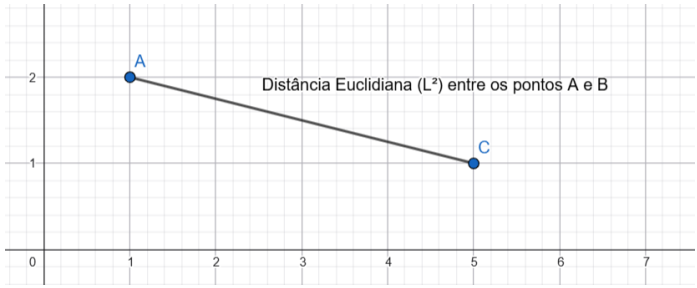

Figura 2. Visualização da Distância Euclidiana aplicada aos mesmos vetores. Diferente do cosseno, a relevância é determinada pela distância linear (linha continua) Fonte: Elaborado pelo autor.

A interpretação geométrica do significado, na qual relações semânticas emergem da disposição relativa dos vetores no espaço, foi popularizada por modelos como o *word2vec* (7)

e permanece como fundamento conceitual dos métodos modernos de recuperação densa.

2.3 MLMs e treinamento de modelos de linguagem

No treinamento baseado em Masked Language Modeling (MLM), parte dos tokens da sequência de entrada é substituída por um marcador especial (geralmente [MASK]), e o modelo é treinado para prever os tokens originais a partir do contexto restante. Essa formulação deriva da tarefa de Cloze(10), na qual o leitor deve preencher lacunas em um texto com base em conhecimento contextual, e permite que o modelo agregue simultaneamente informações à esquerda e à direita do token mascarado, resultando em representações totalmente bidirecionais (8).

Ao final do pré-treinamento, o modelo aprende um mecanismo de predição chamado *prediction head*, capaz de projetar cada posição da sequência em uma distribuição sobre todo o vocabulário. Essa capacidade de prever pesos lexicais é central para métodos de recuperação baseados em expansão lexical e é fundamental para o entendimento do *SPLADE* (11).

2.4 Recuperação Aprimorada Generativa (RAG)

Modelos de Linguagem de Grande Escala (LLMs) representam o estado da arte em tarefas gerais de processamento de linguagem natural e demonstram desempenho humano ou superior em diversos *benchmarks*. Apesar disso, quando aplicados a domínios que exigem conhecimento factual específico, como documentos técnicos, normativos ou institucionais. Esses modelos frequentemente produzem informações imprecisas, descontextualizadas ou incorretas, fenômeno conhecido como alucinação (3).

Tradicionalmente, a adaptação de LLMs a novos domínios é realizada por meio do ajuste fino (*fine-tuning*) de seus pesos com dados específicos. Embora eficaz, essa estratégia é computacionalmente custosa e limitada pela necessidade de reconstruir o modelo sempre que o conhecimento externo é atualizado (2). Para contornar essas limitações, surgiu a técnica de Recuperação Aprimorada Generativa (*Retrieval-Augmented Generation* - RAG), que integra mecanismos de recuperação de informação à geração textual(3).

Em um sistema de RAG, a resposta não depende exclusivamente do conhecimento armazenado nos pesos do LLM; o sistema busca trechos relevantes em uma coleção externa de documentos e os utiliza como contexto adicional durante a geração. Esse mecanismo tem demonstrado redução de alucinações, aumento da transparência das respostas e atualização contínua do conhecimento, sem necessidade de retreinamento do modelo.

Arquiteturas de RAG costumam operar em três etapas principais (3), conforme ilustrado na Figura 3:

1. **Indexação**: os documentos são segmentados e representados de forma esparsa ou densa, sendo armazenados em índices invertidos ou em bancos vetoriais.

2. **Recuperação**: dada uma consulta, o sistema seleciona os K trechos mais relevantes com base em um modelo de similaridade, como o BM25 (esparso) ou embeddings semânticos (denso).

3. **Geração**: a consulta e os trechos recuperados são concatenados e fornecidos como entrada para a LLM, que gera uma resposta fundamentada nas evidências.

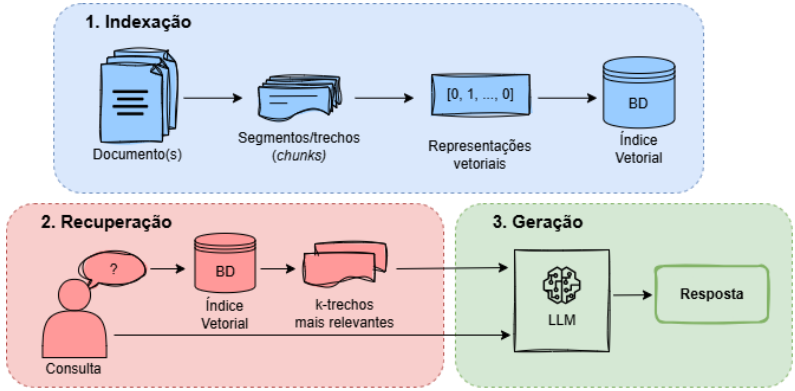

Figura 3. Fluxo de funcionamento do sistema RAG dividido em três macro-etapas. Na **Indexação (azul)**, documentos são fragmentados e convertidos em vetores; na **Recuperação (vermelho)**, a consulta do usuário busca os trechos mais relevantes no índice vetorial; na **Geração (verde)**, o LLM utiliza a consulta e o contexto recuperado para sintetizar a resposta final. Fonte: Elaborado pelo autor.

A modularidade do RAG permite variações em cada etapa. Recuperadores podem ser esparsos, densos ou híbridos, e a etapa de geração pode incorporar estratégias de agregação de trechos, múltiplas iterações de consulta ou técnicas avançadas de engenharia de prompts. Pesquisas recentes exploram essas variações para melhorar eficiência, escalabilidade e qualidade das respostas (2, 12).

2.5 SPLADE

O SPLADE representa uma classe de modelos de recuperação de informação baseados em expansão lexical esparsa, nos quais modelos de representação densos são utilizados para gerar vetores esparsos interpretáveis. Do ponto de vista representacional, o SPLADE

pode ser entendido como um modelo híbrido, pois emprega representações densas para produzir vetores esparsos compatíveis com sistemas léxicos tradicionais. Sua principal função é transformar consultas e documentos em representações vetoriais esparsas, preservando a eficiência de sistemas de busca tradicionais e, ao mesmo tempo, incorporando a riqueza semântica capturada por modelos de linguagem modernos.

Esse método de expansão lexical tem fundamento na arquitetura de modelos baseados em MLM, conforme discutido na Seção 2.3. Em modelos MLM, quando um *token* é mascarado, uma camada final (denominada cabeça de predição) produz uma distribuição de probabilidades sobre todo o vocabulário, indicando quais termos são mais prováveis para preencher a lacuna, conforme ilustrado na Figura 4. Esse mecanismo de predição é explorado para gerar representações esparsas interpretáveis, entendidas como expansões de modelos do tipo BoW.

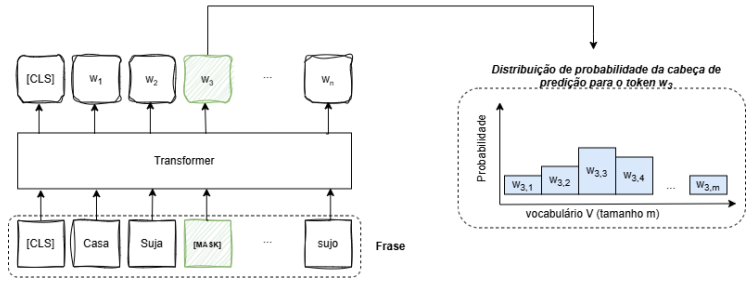

Figura 4. Mecanismo de predição em modelos MLM, base para a expansão do SPLADE. O modelo projeta o estado oculto de um token (neste caso, w_3) sobre todo o vocabulário, atribuindo pesos a termos semanticamente relacionados. No SPLADE, esse processo é aplicado a todos os tokens da entrada (sem mascaramento) e os resultados são agregados.
Fonte: Elaborado pelo autor.

Os primeiros métodos baseados nessa ideia, como o SparTerm (13), já utilizavam a projeção probabilística sobre o vocabulário como mecanismo de expansão lexical. O SPLADE aprofunda essa abordagem ao integrar diretamente a cabeça de predição do modelo MLM a um módulo específico de agregação, denominado *SpladePooling*.

A operação do modelo segue o seguinte fluxo:

1. **Processamento da entrada:** um Transformer pré-treinado para MLM (como BERT, RoBERTa, DistilBERT ou ModernBERT) processa o texto de entrada.
2. **Geração de logits:** a cabeça de predição produz, para cada posição do texto, logits associados a todos os termos do vocabulário, representando seus respectivos pesos de relevância.

3. **Agregação via *SpladePooling*:** o módulo agrega essas distribuições ao longo dos tokens, resultando em um único vetor esparso com dimensionalidade igual ao tamanho do vocabulário.

O vetor resultante concentra, de forma interpretável, os termos que o modelo considera mais informativos para representar a consulta ou o documento. A esparsidade dessas representações permite compatibilidade com indexadores tradicionais baseados em modelos léxicos, mantendo eficiência computacional mesmo em cenários de grande escala.

A escolha do SPLADE como objeto central desta análise comparativa fundamenta-se em três eixos principais:

1. o SPLADE representa de forma clara uma classe específica de métodos em recuperação de informação: os modelos esparsos baseados em expansão lexical. Essa abordagem combina a capacidade expressiva dos *Transformers* com a interpretabilidade e a compatibilidade de estruturas tradicionais de indexação, funcionando como um elo entre dois métodos consolidados da área.

2. o desempenho do SPLADE em métricas amplamente adotadas evidencia sua relevância e maturidade metodológica. No estudo de revisão comparativa (14), que consolida resultados de diversos métodos de recuperação de informação baseados em BERT no conjunto *MS MARCO Passage Ranking*(15), o SPLADE figura consistentemente entre os modelos com melhor desempenho.

3. o SPLADE dispõe de implementações públicas e documentação acessível (16). Apesar de seu desempenho promissor e da ampla adoção em língua inglesa, não há, até o momento, uma versão pública treinada especificamente para o português brasileiro, o que abre espaço para investigações adicionais.

3 Trabalhos relacionados

A seleção dos trabalhos relacionados apresentada nesta seção tem como objetivo fundamentar as escolhas metodológicas e contextualizar os resultados obtidos. O estudo apresentado em 3.1 atua como estudo base, fornecendo os conjuntos de dados, parâmetros e resultados de referência iniciais para o experimento. Em 3.2, a discussão expande-se para o cenário da Recuperação de Informação (RI) em língua portuguesa, com resultados experimentais esparsos ou densos. Por fim, em 3.3, aborda-se a especificidade de domínios técnicos e normativos, onde a literatura sugere uma convergência de desempenho entre abordagens densas e léxicas.

3.1 Aplicação de RAG ao PPC-CC da UFFS

O estudo conduzido por (4) constitui o principal referencial experimental deste trabalho, servindo tanto como base metodológica quanto como conjunto inicial de resultados para comparação. Nesse estudo, foi proposta e avaliada um sistema de RAG aplicado aos PPC-CC da UFFS. A arquitetura segue o formato canônico de sistemas RAG, contemplando as etapas de pré-processamento do corpus, indexação, recuperação de informação e geração de respostas, modularização essa que permite a avaliação isolada de seus componentes.

O corpus utilizado consiste nos PPC-CC da UFFS, versões de 2018 e 2024. Os documentos em formato PDF foram convertidos para texto, preservando seções textuais e tabelas. Em seguida, o conteúdo foi segmentado em trechos (*chunks*) de tamanho fixo, com sobreposição entre segmentos consecutivos. Especificamente, adotou-se uma estratégia de segmentação baseada em janelas de 512 caracteres, com uma sobreposição de 48 caracteres entre os segmentos, de modo a preservar a continuidade semântica entre segmentos adjacentes.

Para treinar e avaliar os recuperadores densos, o autor construiu um **conjunto de dados sintético para recuperação de informação inicial (DSRI-INIC)**. Nessa abordagem, para cada trecho segmentado do documento, um LLM então gerou quatro perguntas relacionadas aos conteúdos centrais. Assim, para um total de N trechos, obtiveram-se $4N$ perguntas, cada uma automaticamente anotada com o identificador do trecho de origem correspondente. Esse conjunto serviu como base inicial para a avaliação dos modelos densos, trazendo métricas de desempenho dos diferentes modelos de recuperadores antes da integração ao sistema de RAG.

Adicionalmente, o autor elaborou manualmente um **Conjunto de Dados de Avaliação do Sistema de RAG (DSRAG-E2E)**. Seu desenvolvimento envolveu: (i) a formulação manual de perguntas que cobrem aspectos centrais dos documentos PPC-CC, simulando dúvidas factuais de um estudante; e (ii) a elaboração manual de respostas ideais, redigidas com base no conteúdo exato dos documentos, que serviram como padrão de referência para avaliar

a completude e a correção factual das respostas geradas pelo sistema.

O estudo avaliou dezoito modelos abertos de modelos densos, priorizando aqueles otimizados para tarefas de similaridade semântica. As métricas empregadas incluem Hit Rate, MRR, MAP e nDCG, amplamente utilizadas na avaliação de sistemas de recuperação de informação e adequadas a tarefas de *ranking* baseado em relevância.

Com base nos resultados obtidos para valores de $k = 1, 3, 5, 10$, o modelo *multi-qa-mpnet-base-dot-v1* foi selecionado em função do equilíbrio entre desempenho e custo computacional. Posteriormente, esse modelo foi ajustado por 30 épocas utilizando duas funções de perda: (i) *Multiple Negatives Ranking Loss*, para reforçar a distinção entre pares positivos e negativos; e (ii) *Matryoshka Loss*, visando induzir hierarquias de granularidade nas representações vetoriais. Após o ajuste fino, observou-se melhora consistente em todas as métricas, com destaque para o aumento do MRR@5 de 51,51% para 73,35%, estabelecendo um recuperador denso especializado para o domínio dos PPCs.

Com base nesse recuperador especializado, o autor implementou uma variante da arquitetura RAG denominada no estudo de *Pipeline de RAG com Embeddings Especializados* (PREE). Essa implementação mantém a estrutura geral da abordagem básica, diferenciando-se pela utilização de embeddings ajustados ao domínio na etapa de recuperação. Sua operação compreende duas etapas principais: (i) **Recuperação**, na qual os trechos mais relevantes à consulta do usuário são selecionados por meio do cálculo da similaridade do cosseno entre embeddings; e (ii) **Geração**, em que os trechos recuperados são concatenados à pergunta em um *prompt* fornecido a um LLM, responsável pela geração da resposta final.

3.2 Desempenho Comparativo em um Dataset Nativo em Português

O dataset Quati (17), composto por consultas nativas e um corpus em português brasileiro, oferece alguns resultados sobre a performance relativa de recuperadores esparsos, densos e híbridos. Em seus experimentos, que medem a capacidade de sistemas distintos em recuperar passagens únicas, o BM25 destacou-se, sendo responsável por 50,6% das recuperações exclusivas no corpus de 1 milhão de passagens. Em contraste, modelos densos como o E5-large (24,4%) e o híbrido neural SPLADE v2 em português (30,2%) apresentaram contribuições menores. Esse resultado, em um domínio aberto e diverso, antecipa a efetividade persistente de métodos léxicos tradicionais, fenômeno observado no presente trabalho, e que parece se acentuar em domínios técnicos com vocabulário de baixa ambiguidade, como o PPC. Além disso, a metodologia do Quati utiliza um LLM para anotação de relevância em um conjunto de passagens diversificado, recuperado por múltiplos sistemas; Isso contrasta com o método de geração sintética de consultas a partir de um trecho único, adotada na construção do DSRI-INIC e do DSRI-LEXVAR.

3.3 Eficácia de Abordagens Híbridas em Domínios Normativos

Estudos em domínios com caracteristicas semelhantes ao do PPC, como o jurídico-regulatório, frequentemente defendem abordagens híbridas. (18), ao aplicar RAG a textos regulatórios financeiros, mostrou que um recuperador híbrido (BM25 + representações densas ajustadas) superou métodos isolados em métricas como *Recall@10* (0.8333) e MAP@10(0.7016). Por outro lado, a melhoria foi baixa, e o BM25 puro manteve-se como base sólida, com *Recall@10* (0.7611) e MAP@10(0.6237). Este resultado parece demonstrar uma proximidade dos resultados observados entre métodos léxicos e densos, indicando que, em domínios técnico-normativos, o ganho semântico de modelos densos sobre a precisão léxica pode ser menor. Vale notar que o conjunto de avaliação foi anotado manualmente, possivelmente atenuando o viés lexical de conjuntos de dados sintéticos.

4 Metodologia

Esta seção detalha o desenho experimental adotado para analisar o comportamento de diferentes métodos de RI no sistema de RAG aplicado ao PPC-CC da UFFS. O estudo evoluiu de uma comparação quantitativa padrão para uma análise qualitativa das falhas e dos vieses dos recuperadores, motivada por resultados empíricos iniciais. Todo o código, scripts de experimento e instruções para reprodução estão disponíveis publicamente no repositório GitHub do projeto[3].

4.1 Conjuntos de Dados

A análise utilizou dois conjuntos de dados de consultas sobre o mesmo corpus de documentos (PPC-CC UFFS 2018 e PPC-CC UFFS 2024).

Tabela 1. Resumo dos conjuntos de dados e escopo de avaliação na arquitetura RAG.

Dataset	Etapa Avaliada	Objetivo Principal	Métricas Chave
DSRI-INIC	**2. Recuperação**	Avaliar precisão da busca em cenário base (viés léxico).	HR@k, nDCG
DSRI-LEXVAR	**2. Recuperação**	Testar robustez da busca contra variação lexical.	HR@k, nDCG
DSRAG-E2E	**2. Recuperação + 3. Geração**	Avaliar a qualidade final da resposta ao usuário.	Corretude, Fidelidade

O DSRI-INIC gerado no trabalho apresentado na Seção 3.1 é utilizado para avaliar isoladamente o módulo de Recuperação (etapa central, em vermelho na Figura 3).Composto por $4N$ perguntas geradas automaticamente por um LLM a partir de N trechos do documento, com anotação automática do trecho-alvo. **Para a geração da base de dados foi utilizado o seguinte Prompt:**

Você é um assistente especializado no Plano Pedagógico do Curso (PPC) de Ciência da Computação da UFFS. Sua função é fornecer metadados sobre trechos extraídos do PPC. Os tópicos comumente são sobre: grades curriculares, ementas, regulamentos, regras, objetivos, infraestrutura, corpo docente e assuntos relacionados. Analise o seguinte texto e forneça os seguintes metadados: Tópico, Palavras-chave, Possíveis perguntas

Texto: {trecho}"

[3]Repositório: <https://github.com/Catanio/tcc-comparacao-metodos-ri-uffs>.

Subsequentemente, foi gerado um novo dataset denominado **DSRI-LEXVAR**, com o objetivo de mitigar o viés lexical identificado no DSRI-INIC. O objetivo foi criar um conjunto em que a similaridade semântica fosse necessária para a recuperação correta, e não apenas a coincidência de palavras. Cada conjunto de perguntas foi novamente associado ao trecho que o originou no estudo original. Na geração, foi utilizada API do Google Gemini 2.5 Flash Lite, sendo instruída a criar perguntas que *(i)* Evitassem repetir o vocabulário específico do trecho-origem. *(ii)* Utilizassem paráfrases, reformulações sintáticas e variações estilísticas. *(iii)* Mantivessem o alinhamento semântico com o conteúdo original.

Para a avaliação do sistema de RAG,englobando Recuperação e Geração (fluxo vermelho e verde na Figura 1), foi utilizado o conjunto **DSRAG-E2E**(visto em 3.1), onde a qualidade da resposta depende da precisão da busca e da capacidade de síntese do LLM.

4.2 Recuperação da Informação

Para análise comparativa, implementamos métodos representativos das três principais famílias metodológicas de recuperadores:

- **BM25**[4]: Recuperador esparso de referência, utilizado como linha de base lexical. Sua performance estabelece o patamar mínimo esperado para qualquer método mais sofisticado.
- **TF-IDF**[5]: Recuperador léxico tradicional, empregado como contraponto ao BM25 pela facilidade na verificação do conjunto de palavras utilizado na recuperação.
- **Embeddings Especializados**[6]: Recuperador denso otimizado para o domínio-alvo, denominado FT-E. Tem como objetivo avaliar o ganho potencial de representações semânticas especializadas em comparação com métodos léxicos e modelos genéricos.
- **SPLADE**[7]: Recuperador híbrido que gera representações esparsas interpretáveis. O checkpoint utilizado foi treinado por 2 épocas em um subconjunto de 300 mil passagens do dataset mMARCO(19) em português, utilizando a implementação simple-splade[8]. Objetiva-se investigar se a expansão lexical guiada por um modelo de linguagem pode

[4]Implementado utilizando a biblioteca `rank_bm25`.
[5]Implementado utilizando `TfidfVectorizer` do `scikit-learn`.
[6]Checkpoint: `winderfeld/cc-uffs-ppc-ft-test-multiqa`, derivado do modelo `sentence-transformers/multi-qa-mpnet-base-dot-v1` com ajuste fino para o domínio do PPC.
[7]Checkpoint: `mauricatanio/splade-bertimbau`, baseado no `neuralmind/bert-base-portuguese-cased` e treinado na arquitetura SPLADE.
[8]Disponível em: https://github.com/marevol/simple-splade

combinar a precisão léxica do BM25 e TF-IDF com a generalização semântica dos métodos densos. Seu comportamento constitui um dos focos centrais da análise qualitativa realizada.

4.3 Métricas de Avaliação

A avaliação do sistema considerou duas famílias complementares de métricas: (i) métricas de Recuperação de Informação (RI), que quantificam a precisão e ordenação dos trechos recuperados; e (ii) métricas de avaliação de respostas, que medem a qualidade final das respostas geradas pelo sistema de RAG, comparando-as com as respostas de referência do conjunto DSRAG-E2E.

Para (*i*), as métricas avaliam a capacidade do sistema de recuperar e ordenar corretamente os trechos relevantes para cada consulta, assumindo como "resposta correta"(20) o trecho-origem da pergunta no conjunto sintético. Foram calculadas para os pontos de corte $k = 1, 3, 5, 10$.

- **Taxa de acerto** (**Hit Rate HR@k)**: Mede a porcentagem de consultas em que o trecho correto aparece entre os k primeiros resultados.

- **Classificação Recíproca Média** (**Mean Reciprocal Rank - MRR@k)** (21): Avalia a qualidade do ranking considerando a posição do primeiro trecho correto. Calcula-se o inverso da posição desse trecho (1 para primeiro lugar, 1/2 para segundo, etc.) e faz-se a média sobre todas as consultas, privilegiando sistemas que colocam a resposta correta no topo da lista.

- **Ganho Cumulativo Descontado Normalizado** (**Normalized Discounted Cumulative Gain -**
  nDCG@k) (22): Avalia a ordenação dos resultados, penalizando os sistemas que colocam o trecho correto em posições muito baixas na lista.

Em (*ii*), para avaliar a qualidade das respostas geradas pelo sistema de RAG, adotou-se o paradigma LLM como avaliador automatizado (*LLM-as-a-Judge*) (23). As métricas seguem o mesmo padrão do trabalho anterior 3.1, baseando-se na implementação do DeepEval[9] e avaliadas pelo modelo Mistral Large[10].

[9]Documentação do DeepEval: <https://docs.confident-ai.com/docs/getting-started>

[10]Vale notar que, na implementação original, a métrica de Recuperação Contextual utilizava o modelo Llama-3.2-90B. Devido a uma alteração na disponibilidade da API da GROQ durante a execução dos experimentos, optou-se por padronizar o avaliador, substituindo-o pelo Mistral Large, já utilizado nas demais métricas.

- **Corretude (*Correctness*)**: Mede o grau de alinhamento factual entre a resposta gerada e a resposta de referência (*ground truth*). Um escore alto indica que a resposta contém as mesmas informações-chave sem erros factuais ou omissões graves.

- **Relevância da Resposta (*Answer Relevancy*)**: Avalia a proporção de afirmações na resposta que são diretamente pertinentes para responder à consulta. É calculada como:

$$\text{Relevância} = \frac{\text{Nº de afirmações relevantes}}{\text{Total de afirmações na resposta}}$$

- **Recuperação Contextual (*Contextual Recall*)**: Mede a completude da recuperação. Avalia qual fração das informações presentes na resposta ideal está contida no contexto recuperado e fornecido ao LLM. É calculada como:

$$\text{Recuperação Contextual} = \frac{\text{Nº de afirmações da resposta ideal presentes no contexto}}{\text{Total de afirmações na resposta ideal}}$$

- **Fidelidade (*Faithfulness*)**: Verifica se todas as afirmações da resposta gerada são suportadas pelo contexto fornecido, sem introduzir alucinações ou contradições. É calculada como:

$$\text{Fidelidade} = \frac{\text{Nº de afirmações verificáveis no contexto}}{\text{Total de afirmações na resposta gerada}}$$

4.4 Protocolo Experimental e Análise

A avaliação dos novos experimentos foi conduzida em duas fases interligadas: análise quantitativa e análise qualitativa.

A **análise quantitativa** teve como objetivo estabelecer uma linha de base comparativa, reproduzindo os resultados do estudo anterior e mensurando o desempenho inicial dos novos métodos testados. Foram calculadas as métricas de Recuperação de Informação: HR, MRR e nDCG); nos pontos de corte $k = 1, 3, 5, 10$. O foco principal foi comparar o desempenho agregado dos diferentes recuperadores (densos, esparsos e lexical) tanto no conjunto DSRI-INIC quanto no DSRI-LEXVAR, identificando padrões gerais e variações de desempenho entre as abordagens.

Já a **análise qualitativa**, busca explicar os comportamentos observados na fase quantitativa por meio de duas análises manuais focadas:

1. **Diagnóstico do SPLADE**: Para casos em que o TF-IDF recuperava com sucesso o trecho correto, mas o SPLADE falhava, realizou-se uma inspeção manual das expansões

lexicais geradas pelo SPLADE. O objetivo era identificar a introdução de termos irrelevantes ou a ausência de termos-chave na representação esparsa, buscando explicar as falhas de recuperação.

2. **Análise de Viés Lexical**: Para pares (consulta, trecho-alvo) do conjunto DSRI-LEXVAR nos quais o TF-IDF obtinha sucesso e o modelo de embeddings densos (FT-E) falhava, foi avaliada manualmente a sobreposição de termos entre a consulta e o trecho. Essa análise procurou quantificar o grau de viés lexical presente no conjunto de dados, isto é, o quanto o sucesso do método lexical dependia de uma correspondência superficial de palavras.

5 Resultados

Esta seção apresenta os resultados obtidos nas três etapas experimentais delineadas na metodologia: (1) avaliação dos métodos de Recuperação de Informação (RI) com o conjunto de dados sintético inicial (DSRI-INIC); (2) avaliação dos métodos de RI com o novo conjunto de dados sintético lexicamente variado (DSRI-LEXVAR); e (3) avaliação do sistema de RAG com perguntas manuais validadas (DSRAG-E2E). A interpretação integrada e a análise das causas subjacentes a esses resultados são discutidas na Seção 6.

5.1 Avaliação dos Métodos de RI com DSRI-INIC e DSRI-LEXVAR

O conjunto DSRI-INIC 3.1, permitiu um teste inicial dos métodos de RI. Os resultados consolidados para os pontos de corte $k = 1, 3, 5, 10$ são apresentados na Tabela 2.

Tabela 2. Resultados de Recuperação de Informação (RI) no conjunto DSRI-INIC. Métricas reportadas para $k = 1, 3, 5, 10$.

Métrica	BM25	TF-IDF	FT-Embeddings	BM25+SPLADE	TF-IDF+SPLADE
HR@1	**0.656**	0.640	0.650	0.500	0.577
MRR@1	**0.656**	0.640	0.650	0.500	0.577
nDCG@1	**0.656**	0.640	0.650	0.500	0.577
HR@3	**0.830**	0.788	0.815	0.690	0.749
MRR@3	**0.733**	0.705	0.725	0.586	0.654
nDCG@3	**0.758**	0.726	0.748	0.613	0.678
HR@5	**0.868**	0.830	0.859	0.755	0.793
MRR@5	**0.742**	0.715	0.735	0.601	0.664
nDCG@5	**0.774**	0.744	0.766	0.640	0.696
HR@10	0.899	0.878	**0.922**	0.833	0.848
MRR@10	**0.746**	0.721	0.744	0.611	0.671
nDCG@10	0.784	0.759	**0.787**	0.665	0.713

Os resultados na Tabela 2 mostram um desempenho elevado para todos os métodos, com valores de Hit Rate@5 superiores a 75%; i.e., os métodos foram capazes de recuperar o trecho correto entre os cinco primeiros resultados para a maioria das consultas. Os métodos léxicos puros (BM25, TF-IDF) e o método denso especializado (FT-Embeddings) apresentaram performance muito similar e competitiva, com o BM25 liderando em HR@5 (86.8%). Em contraste, as combinações híbridas que utilizam o SPLADE (BM25+SPLADE e TF-IDF+SPLADE) apresentaram uma queda de desempenho consistente, ficando entre 7.7%

e 12.1% abaixo do FT-Embeddings em HR@5. Este padrão inicial motivou a investigação sobre um possível viés lexical no DSRI-INIC, levando à criação do conjunto DSRI-LEXVAR.

Para testar a robustez dos métodos diante da variação lexical, os experimentos foram repetidos com o conjunto DSRI-LEXVAR. Os resultados completos são apresentados na Tabela 3.

Tabela 3. Resultados de Recuperação de Informação (RI) no conjunto DSRI-LEXVAR. Métricas reportadas para $k = 1, 3, 5, 10$.

Métrica	BM25	TF-IDF	FT-Embeddings	BM25+SPLADE	TF-IDF+SPLADE
HR@1	**0.436**	0.373	0.409	0.239	0.309
MRR@1	**0.436**	0.373	0.409	0.239	0.309
nDCG@1	**0.436**	0.373	0.409	0.239	0.309
HR@3	**0.575**	0.557	0.561	0.430	0.475
MRR@3	**0.499**	0.456	0.477	0.321	0.381
nDCG@3	**0.518**	0.482	0.498	0.348	0.405
HR@5	**0.636**	0.625	0.620	0.518	0.566
MRR@5	**0.513**	0.471	0.490	0.340	0.402
nDCG@5	**0.544**	0.509	0.523	0.385	0.443
HR@10	**0.734**	0.711	**0.734**	0.627	0.657
MRR@10	**0.525**	0.482	0.506	0.355	0.414
nDCG@10	**0.575**	0.537	0.560	0.420	0.472

Como esperado, a dificuldade do conjunto DSRI-LEXVAR resultou em uma redução geral das métricas para todos os métodos, com Hit Rates@5 caindo para a faixa de 51.8% a 63.6%. A ordem relativa dos métodos, no entanto, manteve-se: BM25 e TF-IDF permaneceram no topo, seguidos de perto pelo FT-Embeddings, enquanto os métodos com SPLADE continuaram em desvantagem. A queda percentual do SPLADE em relação ao FT-Embeddings aumentou para 16.5% (BM25+SPLADE) e 8.8% (TF-IDF+SPLADE). Este resultado confirma que o viés lexical do DSRI-INIC inflava artificialmente os escores absolutos, mas não alterava a conclusão fundamental sobre a classificação dos métodos para este domínio.

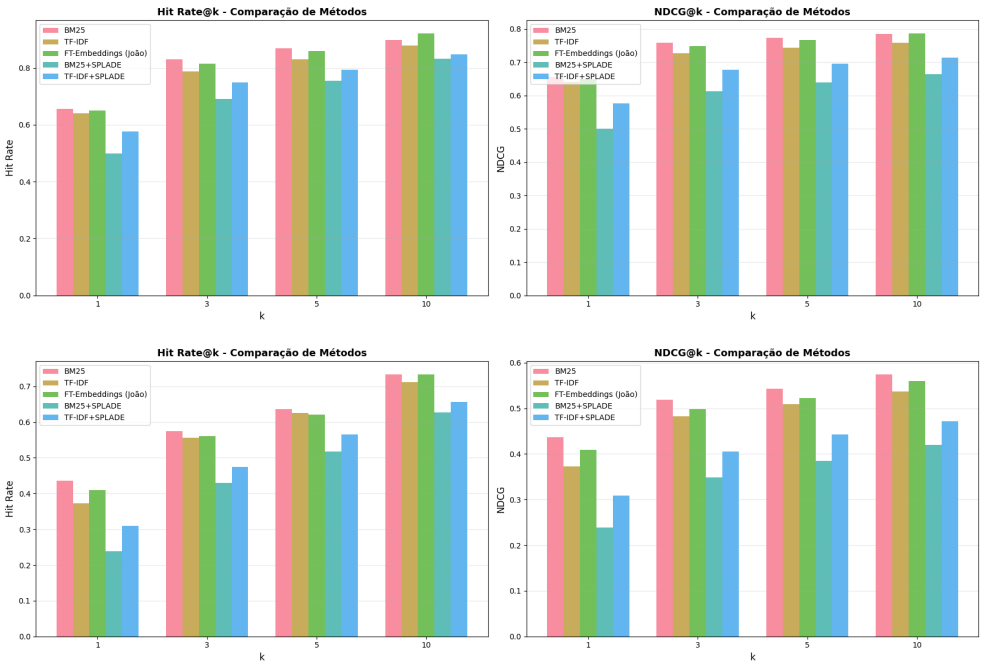

Figura 5. Comparativo visual do impacto do viés lexical nas métricas de recuperação. O gráfico superior apresenta o desempenho no DSRI-INIC, inflado pela sobreposição de termos. O gráfico inferior evidencia a retração no desempenho (DSRI-LEXVAR) para todos os métodos quando a correspondência exata é removida. Fonte: Elaborado pelo autor.

5.2 Avaliação do sistema de RAG

A avaliação final do sistema RAG foi realizada com o conjunto DSRAG-E2E, utilizando o paradigma *LLM-as-a-judge* conforme detalhado na Seção 4.3. Os resultados das quatro métricas de qualidade de resposta, bem como a taxa de aprovação final, são apresentados nas Tabelas 4 e 5.

Tabela 4. Resultados da avaliação do sistema de RAG no conjunto DSRAG-E2E (N=120). Métricas contínuas (média ± desvio padrão).

Recuperador	Corretude	Relevância	ContextualRecall	Faithfulness
BM25	0.916 ± 0.269	0.938 ± 0.212	**0.858 ± 0.312**	0.908 ± 0.234
TF-IDF	0.855 ± 0.344	0.913 ± 0.267	0.785 ± 0.392	0.888 ± 0.259
FT-Embeddings	0.857 ± 0.335	0.912 ± 0.253	0.820 ± 0.339	0.889 ± 0.267
BM25+SPLADE	0.533 ± 0.494	0.842 ± 0.350	0.492 ± 0.455	**0.973 ± 0.117**
TF-IDF+SPLADE	0.699 ± 0.450	0.914 ± 0.268	0.665 ± 0.434	0.919 ± 0.236

Tabela 5. Taxa de aprovação das respostas no conjunto DSRAG-E2E (N=120).

Recuperador	Aprovação
BM25	**69.2%**
TF-IDF	63.3%
FT-Embeddings	60.0%
BM25+SPLADE	37.5%
TF-IDF+SPLADE	52.5%

Os resultados das Tabelas 4 e 5 revelam um padrão consistente com as fases de RI isoladas:

- O **BM25** obteve o melhor desempenho geral, liderando em *Correctness* (0.916), *Contextual Recall* (0.858) e taxa de *Aprovação* (69.2%).
- **TF-IDF** e **FT-Embeddings** apresentaram resultados muito próximos e competitivos em todas as métricas, com ligeira vantagem do TF-IDF na taxa de aprovação (63.3% vs. 60.0%).
- Os métodos que incorporam o **SPLADE** apresentaram uma queda significativa na qualidade, particularmente nas métricas de *Correctness* e *Contextual Recall*. Notavelmente, o BM25+SPLADE atingiu o pior valor de *Correctness* (0.533) e a menor taxa de aprovação (37.5%). A métrica de *Faithfulness*, no entanto, foi a mais alta para estes

métodos, sugerindo que as respostas, ainda que incompletas ou incorretas, raramente alucinavam informações fora do contexto fornecido.

Estes resultados consolidam as observações das etapas anteriores e fornecem uma base sólida para a análise qualitativa e a discussão que se seguem.

6 Discussão

Esta seção analisa os resultados apresentados na Seção 5 com base nos objetivos do estudo e da literatura existente. A discussão é organizada em quatro eixos interpretativos principais, que emergiram da análise empírica: (1) o comportamento discrepante do SPLADE em português brasileiro; (2) a persistência da eficácia dos recuperadores léxicos; (3) a influência estrutural do *dataset* sintético na avaliação; e (4) a avaliação integrada do desempenho de todos os recuperadores no contexto do sistema de RAG. Por fim, são apresentadas as limitações metodológicas do estudo.

6.1 Desempenho do SPLADE em português brasileiro

Os resultados experimentais revelaram um desempenho consistentemente inferior dos métodos que incorporaram o SPLADE (Tabelas 2, 3, 4 e 5). Enquanto em benchmarks internacionais o SPLADE se consolida como estado da arte em recuperação híbrida (16)(14), sua aplicação ao português brasileiro mostrou-se problemática.

Este trabalho indicou a ausência de um *checkpoint* público, funcional e de alta qualidade para o SPLADE em português. A implementação citada no trabalho do dataset QUATI (17) não está publicamente disponível. A tentativa de adaptação realizada neste estudo, utilizando o BERTimbau como base e treinando com um subconjunto limitado do mMARCO, produziu um modelo cujas expansões lexicais eram frequentemente ruidosas, genéricas ou semanticamente irrelevantes. Esse comportamento reflete uma dependência crítica da arquitetura SPLADE em relação a um pré-treinamento robusto e em grande escala, alinhado ao idioma do modelo base (11). Sem essa base, o mecanismo de expansão do vocabulário tende a gerar ruído em vez de proporcionar a generalização semântica controlada que caracteriza o método. Consequentemente, a queda de 12.1% a 16.5% no Hit Rate@5 em relação ao FT-Embeddings e a taxa de aprovação de apenas 37.5% para o BM25+SPLADE no teste *end-to-end* (Tabelas 4 e 5) são um sintoma da barreira prática da aplicação de métodos do estado da arte em contextos multilíngues (24).

6.2 Eficácia Persistente dos Métodos Léxicos

Um dos achados mais robustos deste estudo foi o desempenho competitivo e frequentemente superior dos recuperadores léxicos BM25 e TF-IDF. Esse resultado se manteve tanto na avaliação isolada de RI (onde lideraram as tabelas) quanto na avaliação do sistema de RAG (onde o BM25 obteve a maior taxa de aprovação, 69.2%). Sugerimos que essa eficácia pode ser atribuída a características intrínsecas do domínio de aplicação:

1. **Dominância Lexical**: O PPC é um documento normativo com um vocabulário técnico estável e repetitivo (e.g., "ementa", "componentes curriculares", "TCC I", "créditos"). Recuperadores baseados em correspondência de termos são naturalmente favorecidos nesse contexto, em que a presença de palavras-chave específicas é um forte indicador de relevância.

2. **Baixa Ambiguidade Semântica**: As consultas típicas sobre o PPC buscam informações factuais e definições precisas, raramente exigindo a resolução de ambiguidades ou inferência pragmática profunda. Isso minimiza a vantagem teórica dos embeddings densos em capturar relacionamentos semânticos complexos.

3. **Fragmentação do *Corpus***: A segmentação do documento em trechos orientados à estrutura resultou em trechos curtos, muitos dos quais são identificáveis por um conjunto reduzido de termos distintivos. O BM25, que privilegia termos raros e discriminativos (5), é particularmente eficaz nesse cenário, enquanto modelos densos podem diluir a relevância dos termos de consulta em representações contínuas de significado mais amplo.

Portanto, o bom desempenho dos métodos esparsos parece decorrer da natureza do domínio-alvo. Esse achado ressalta a importância de considerar as características do *corpus* e da tarefa ao escolher um método de recuperação, o que traz um contraponto à noção de que as abordagens neurais são intrinsecamente superiores em todos os contextos.

6.3 Impacto do *Dataset* Sintético e seu Viés Estrutural

A análise comparativa entre os resultados do DSRI-INIC e do DSRI-LEXVAR (Seções 5.1 e 5.1) evidencia o impacto do conjunto de dados de avaliação nas conclusões sobre o desempenho dos modelos. Para ilustrar concretamente esses fenômenos, a análise qualitativa detalhada (apresentada no Apêndice **Apêndice A)** categorizou as falhas e os vieses observados em tipologias principais, como o *overfitting* léxico e o viés de rótulo único. A seguir, detalham-se as duas fontes de viés estrutural identificadas:

6.3.1 *Overfitting* Léxico do Conjunto de Consultas O DSRI-INIC foi gerado por um LLM instruído a criar perguntas a partir de cada trecho, um processo que frequentemente levou à reprodução literal de termos e construções sintáticas presentes no trecho-origem. Esse viés parece ter inflado artificialmente as métricas dos métodos léxicos, como demonstrado pela queda geral de aproximadamente 23 pontos percentuais no Hit Rate@5 ao se migrar para o DSRI-LEXVAR (Tabelas 2 e 3). Embora a ordem relativa dos métodos tenha se mantido, a diferença de pontuação foi drasticamente alterada, o que alerta para o risco de se tirar conclusões gerais a partir de conjuntos de dados sintéticos.

6.3.2 Problema do Rótulo Único (*Single-Label Bias*) A anotação automática que associa cada pergunta sintética apenas ao trecho que a gerou cria um falso negativo para trechos alternativos que também poderiam responder à pergunta de maneira completa ou até mais informativa. Por exemplo, para a pergunta "O que será detalhado nos planos de ensino das disciplinas de TCC?", o trecho apontado pelo DSRI-LEXVAR era um parágrafo burocrático, enquanto um trecho recuperado por um método denso continha a descrição completa dos objetivos e estrutura do TCC. Esse viés penaliza desproporcionalmente os recuperadores densos, que, por buscarem similaridade semântica, podem recuperar conteúdo relevante que diverge lexicalmente do "rótulo oficial".

Assim, parte da aparente inferioridade dos métodos densos nos experimentos de RI isolada pode ser atribuída a limitações na construção do *dataset* de avaliação, e não a uma falha fundamental da abordagem.

6.4 Avaliação do sistema de RAG e o Impacto dos Recuperadores

A avaliação do sistema de RAG (4 e 5) corrobora e amplia as observações das etapas anteriores. O BM25 manteve a liderança, produzindo respostas com a maior Correctness (0.916) e Contextual Recall (0.858), demonstrando que, para este domínio, a precisão léxica se traduz diretamente em qualidade final da resposta. Os recuperadores TF-IDF e FT-Embeddings apresentaram desempenho muito próximo e competitivo, com o FT-Embeddings mostrando um Contextual Recall ligeiramente superior (0.820 vs. 0.785). Essa proximidade sugere que, uma vez especializado para o domínio, um recuperador denso pode alcançar a efetividade dos métodos léxicos. Por fim, as combinações com o SPLADE apresentaram novamente o desempenho mais fraco, com uma Correctness drasticamente reduzida (0.533 para BM25+SPLADE). É notável, contudo, que sua métrica de Faithfulness foi a mais alta (0.973), indicando que o modelo gerador raramente alucinava informações fora do contexto (insuficiente) que lhe era fornecido.

Em síntese, este experimento indica que, para domínios técnicos normativos e com vocabulário estável, como o PPC, recuperadores léxicos maduros, como o BM25, não apenas são competitivos, mas também podem ser a escolha mais eficaz e de menor custo computaci-

onal.

6.5 Limitações do Estudo

Esta pesquisa está sujeita a limitações que devem ser consideradas na interpretação dos resultados e que apontam direções para trabalhos futuros:

Treinamento do SPLADE: O modelo híbrido foi treinado com recursos limitados (poucas épocas e um subconjunto do mMARCO), não atingindo seu potencial pleno. Um treinamento mais extensivo e com um conjunto de triplas anotadas específicas para o português poderia alterar seus resultados.

Natureza do *Corpus*: O PPC é um único documento, altamente estruturado e com vocabulário técnico repetitivo. As conclusões sobre a superioridade dos métodos léxicos podem não se generalizar para domínios com maior ambiguidade semântica ou linguagem menos específica.

Anotação Automática: Os *datasets* sintéticos de RI (DSRI-INIC e DSRI-LEXVAR) foram construídos com anotação automática, que associa cada pergunta a um único trecho-origem. Essa abordagem não captura cenários de relevância parcial ou a existência de múltiplos trechos corretos, introduzindo um viés na avaliação que pode penalizar métodos com capacidade de recuperação semântica mais ampla.

Avaliação Automática (*LLM-as-a-Judge*): Embora as métricas baseadas em LLM tenham sido validadas e sejam amplamente utilizadas, elas não substituem integralmente a avaliação humana. Os escores podem refletir vieses inerentes ao modelo avaliador (*Mistral Large*) ou do framework de avaliação (*DeepEval*).

7 Conclusão

Este trabalho realizou uma avaliação comparativa de métodos de RI no contexto de um sistema RAG aplicado ao Projeto Pedagógico do Curso de Ciência da Computação da UFFS. A análise evoluiu de uma abordagem puramente quantitativa para uma investigação qualitativa dos resultados que contrariaram expectativas iniciais.

Os experimentos demonstraram que neste domínio normativo de vocabulário técnico estável os métodos léxicos tradicionais (especialmente o BM25) mantiveram desempenho superior ou equivalente a abordagens densas e híbridas. Essa eficácia é atribuída à natureza do PPC: baixa ambiguidade semântica, terminologia repetitiva e trechos identificáveis por palavras-chave. Na avaliação do sistema RAG, o BM25 produziu respostas de maior qualidade, com taxa de aprovação de 69,2

A investigação do SPLADE como recuperador híbrido revelou desafios práticos para sua aplicação em português brasileiro, gerando expansões lexicais genéricas, demonstrando desempenho consistentemente inferior.

A análise dos conjuntos de dados sintéticos identificou dois vieses estruturais: (i) *overfitting* lexical, que inflou métricas de métodos baseados em termos; e (ii) viés do rótulo único, que penalizou recuperadores que traziam trechos semanticamente relevantes, mas lexicalmente distantes do "rótulo correto" automático.

Com base nas limitações identificadas, futuros trabalhos poderiam: *(i)* regenerar conjuntos de dados com anotações supervisionadas para mitigar o viés do rótulo único; e *(ii)* investigar treinamentos mais extensivos do SPLADE para português, utilizando conjuntos como mMARCO e QUATI.

Em síntese, este estudo reforça que a escolha do método de recuperação em sistemas RAG deve considerar as características do domínio-alvo. Para documentos normativos como o PPC, métodos léxicos mostraram-se vantajosos, desafiando a noção de superioridade intrínseca das abordagens densas, enquanto apontam para a necessidade de avanços em recuperadores híbridos para o português.

Apêndice A Diagnóstico Qualitativo de Falhas e Vieses

Este apêndice apresenta registros de execução (*logs*) extraídos durante a validação dos modelos. Os casos foram selecionados para ilustrar, com dados reais, as três principais patologias identificadas na discussão: a injeção de ruído pelo SPLADE, o favorecimento léxico em detrimento do semântico e as falhas de anotação no conjunto de dados.

Apêndice A.1 Fenômeno 1: Ruído Semântico na Expansão (SPLADE)

Os dados demonstram que o SPLADE, treinado com conjuntos curtos de triplas (query, par positivo, par negativo) e por poucas épocas, tende a uma expansão genérica, transformando termos acadêmicos em conceitos corporativos ou educacionais amplos, desviando o foco da recuperação.

Consulta (ID: 75f75ea8...): *"Como é formada a equipe que avaliará o trabalho final?"*

Análise sobre a Expansão de Termos: A tabela abaixo mostra os termos que o SPLADE adicionou à consulta e seus respectivos pesos. Note a deriva semântica de "Banca Acadêmica" para "Gestão Empresarial".

Termo Adicionado	Peso	Impacto na Recuperação
`competências`	1.94	O recuperador priorizou documentos sobre "Perfil do Egresso" e "Matriz Curricular", onde estes termos são frequentes, ignorando o Regulamento do TCC (Art. 22).
`gestão`	1.67	
`estratégia`	1.54	
`empresarial`	1.34	

Resultado Prático:

- **TF-IDF (Sem expansão):** Recuperou o Art. 21/22 corretamente (Posição 2).
- **SPLADE:** Falhou. O Top-1 foi um documento administrativo contendo: *"...Sistemas de informação | Pesquisa operacional | Empreendedorismo..."* devido à atração pelos termos "empresarial" e "gestão".

Outro Caso Crítico (ID: d508c5...): *"...Tópicos Especiais em Computação V"*

- **Expansão SPLADE:** Adicionou `criança (0.96), idade (1.33), escolar (1.44).`
- **Diagnóstico:** O modelo associou "ensino/computação" ao contexto de educação básica infantil, falhando em recuperar a ementa técnica da disciplina universitária.

Apêndice A.2 Fenômeno 2: Vantagem Mecânica dos Métodos Léxicos no DSRI-INIC

A análise dos casos onde o TF-IDF superou o FT-Embeddings revela uma ancoragem rígida das perguntas sintéticas sobre os termos exatos do trecho-origem. O processo de geração criou questões que frequentemente incorporam a própria resposta em sua formulação, replicando marcadores estruturais e vocabulário específico de forma que não condiz com o comportamento natural de dúvida de um usuário real. Isso reduz a necessidade de inferência semântica e favorece artificialmente métodos de correspondência exata.

Caso 1: Ancoragem lexical exata (Vício de Origem)
Consulta: *"De acordo com o § 1º, qual é a preferência na formação da banca avaliadora para o TCC I e TCC II?"*
Documento Alvo: *"...Art. 22º ... § 1º Será formada uma banca para TCC I e TCC II, preferencialmente idênticas."*

Diagnóstico: Este caso ilustra a inversão de causalidade no dataset sintético. O modelo gerador utilizou o endereço da informação (*§ 1º*) — que o usuário desconhece e busca encontrar — como parte da própria pergunta.

- **Impacto na Avaliação:** A presença literal do token "*§ 1º*" infla o score do TF-IDF, garantindo a recuperação correta por motivos mecânicos, e não semânticos.
- **Irrealidade:** Penaliza métodos densos que buscam o conceito abstrato de "regras de formação de banca", mas que não supervalorizam a string do parágrafo específico.

Caso 2: Alta Sobreposição Lexical (60%)
Consulta: *"O que será detalhado nos planos de ensino das disciplinas de TCC?"*
Documento Alvo: *"...tarão nos respectivos planos de ensino dos componentes curriculares TCC I e II."*

Diagnóstico:

- **TF-IDF:** Sucesso (Posição 1). A coincidência de múltiplos tokens raros (*planos, ensino, tcc*) garantiu o topo do ranking, dispensando o entendimento do contexto.
- **FT-Embeddings:** Falha. O modelo recuperou trechos sobre "objetivos do TCC", semanticamente próximos, mas perdeu a especificidade burocrática exigida pela correspondência exata dos termos da pergunta.

Apêndice A.3 Fenômeno 3: Viés de Rótulo Único (Single-Label Bias)

Este fenômeno ilustra uma limitação da métrica de avaliação e não do modelo. O log abaixo prova que o modelo recuperou uma resposta correta, mas foi penalizado porque o *dataset* considerava apenas um ID como "gabarito".

Consulta: *"Como é formada a equipe que avaliará o trabalho final?"*
ID do Gabarito Oficial: `62a88...` (Art. 22)

Resultado do TF-IDF (Classificado como ERRO pela métrica):

1. **[Recuperado na Posição 1] ID: 66d32... (Art. 16)**
   Conteúdo: "**Art. 16** A banca avaliadora será composta por, pelo menos, três membros, sendo presidida pelo orientador do trabalho."
2. **[Recuperado na Posição 2] ID: 62a88... (Art. 22 - Gabarito)**
   Conteúdo: "**Art. 22** A banca avaliadora será composta por, pelo menos, três membros, sendo presidida pelo orientador do trabalho."

Conclusão do Diagnóstico: O Art. 16 e o Art. 22 possuem texto idêntico. O modelo encontrou a informação correta na primeira posição, mas a métrica *Hit Rate@1* contabilizou como falha (0), pois esperava exclusivamente o ID do Art. 22. Isso sugere que a performance real dos modelos é superior à reportada nas tabelas quantitativas.

Referências

1 BUDHIRAJA, R. et al. "it's not like jarvis, but it's pretty close!" - examining chatgpt's usage among undergraduate students in computer science. In: *Proceedings of the 26th Australasian Computing Education Conference*. ACM, 2024. (ACE 2024). Disponível em: <http://dx.doi.org/10.1145/3636243.3636257>.

2 GAO, Y. et al. *Retrieval-Augmented Generation for Large Language Models: A Survey*. 2024.

3 LEWIS, P. et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. 2021.

4 BUSSOLOTTO, J. V. W. Llms como ferramenta para consultas acadêmicas no ensino superior: Uma análise de RAG aplicado ao PPC-CC da UFFS. Chapecó, SC, Brasil, jun. 2024. Trabalho de Conclusão de Curso (Ciência da Computação).

5 ROBERTSON, S.; ZARAGOZA, H. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, v. 3, p. 333–389, 01 2009.

6 NASEEM, U. et al. *A Comprehensive Survey on Word Representation Models: From Classical to State-Of-The-Art Word Representation Language Models*. 2020.

7 MIKOLOV, T. et al. *Efficient Estimation of Word Representations in Vector Space*. 2013.

8 DEVLIN, J. et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019.

9 JHA, R. et al. *Harnessing the Universal Geometry of Embeddings*. 2025. Disponível em: <https://arxiv.org/abs/2505.12540>.

10 TAYLOR, W. L. "cloze procedure": A new tool for measuring readability. *Journalism & Mass Communication Quarterly*, v. 30, p. 415 – 433, 1953. Disponível em: <https://api.semanticscholar.org/CorpusID:206666846>.

11 FORMAL, T.; PIWOWARSKI, B.; CLINCHANT, S. *SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking*. 2021.

12 RAJPUT, V. *RAG 2.0: Retrieval Augmented Language Models*. Medium - AIGuys, 2024. Accessed: 06-07-2024. Disponível em: <https://medium.com/aiguys/rag-2-0-retrieval-augmented-language-models-3762f3047256>.

13 BAI, Y. et al. *SparTerm: Learning Term-based Sparse Representation for Fast Text Retrieval*. 2020.

14 WANG, J. et al. *Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges*. 2024.

15 BAJAJ, P. et al. *MS MARCO: A Human Generated MAchine Reading COmprehension Dataset*. 2018. Disponível em: <https://arxiv.org/abs/1611.09268>.

16 LASSANCE, C. et al. *SPLADE-v3: New baselines for SPLADE*. 2024. Disponível em: <https://arxiv.org/abs/2403.06789>.

17 BUENO, M. et al. *Quati: A Brazilian Portuguese Information Retrieval Dataset from Native Speakers*. 2024. Disponível em: <https://arxiv.org/abs/2404.06976>.

18 RAYO, J.; ROSA, R. de la; GARRIDO, M. *A Hybrid Approach to Information Retrieval and Answer Generation for Regulatory Texts*. 2025. Disponível em: <https://arxiv.org/abs/2502.16767>.

19 BONIFACIO, L. et al. *mMARCO: A Multilingual Version of the MS MARCO Passage Ranking Dataset*. 2022. Disponível em: <https://arxiv.org/abs/2108.13897>.

20 MANNING, C. D.; RAGHAVAN, P.; SCHüTZE, H. *Introduction to Information Retrieval*. USA: Cambridge University Press, 2008. ISBN 0521865719.

21 CRASWELL, N. Mean reciprocal rank. In: ____. *Encyclopedia of Database Systems*. Boston, MA: Springer US, 2009. p. 1703–1703. ISBN 978-0-387-39940-9. Disponível em: <https://doi.org/10.1007/978-0-387-39940-9_488>.

22 JÄRVELIN, K.; KEKÄLÄINEN, J. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, v. 20, p. 422–446, 2002. Disponível em: <https://api.semanticscholar.org/CorpusID:1981391>.

23 GU, J. et al. *A Survey on LLM-as-a-Judge*. 2025. Disponível em: <https://arxiv.org/abs/2411.15594>.

24 WU, S. et al. *Not All Languages are Equal: Insights into Multilingual Retrieval-Augmented Generation*. 2024. Disponível em: <https://arxiv.org/abs/2410.21970>.