CORPO DE BOMBEIROS MILITAR DO DISTRITO FEDERAL
DEPARTAMENTO DE ENSINO, PESQUISA, CIÊNCIA E TECNOLOGIA
DIRETORIA DE ENSINO
CENTRO DE ESTUDOS DE POLÍTICA, ESTRATÉGIA E DOUTRINA
CURSO DE APERFEIÇOAMENTO DE OFICIAIS

Cap. QOBM/Comb. PEDRO **MATIAS** DOS SANTOS

USO DE TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL (PLN) PARA CONSULTA NORMATIVA DO CBMDF: PROPOSTA DE ARQUITETURA LEVE E ADAPTÁVEL COMO PROVA DE CONCEITO

BRASÍLIA
2026

Cap. QOBM/Comb. PEDRO **MATIAS** DOS SANTOS

USO DE TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL (PLN) PARA CONSULTA NORMATIVA DO CBMDF: PROPOSTA DE ARQUITETURA LEVE E ADAPTÁVEL COMO PROVA DE CONCEITO

Artigo científico apresentado ao Centro de Estudos de Política, Estratégia e Doutrina como requisito para conclusão do Curso de Aperfeiçoamento de Oficiais do Corpo de Bombeiros Militar do Distrito Federal.

Orientador: Ten-Cel. QOBM/Comb. PABLO FEDERICO **BAIGORRI**

BRASÍLIA
2026

Cap. QOBM/Comb. PEDRO **MATIAS** DOS SANTOS

USO DE TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL (PLN) PARA CONSULTA NORMATIVA DO CBMDF: PROPOSTA DE ARQUITETURA LEVE E ADAPTÁVEL COMO PROVA DE CONCEITO

Artigo científico apresentado ao Centro de Estudos de Política, Estratégia e Doutrina como requisito para conclusão do Curso de Aperfeiçoamento de Oficiais do Corpo de Bombeiros Militar do Distrito Federal.

Aprovado em: _____/_____/_____.

BANCA EXAMINADORA

_____
Igor **Muniz** Da Silva – Ten-Cel QOBM/Comb.
Presidente

_____
Lucas Araújo **Pereira** – Ten-Cel QOBM/Compl.
Membro

_____
Emilia Bernardes da Silva – Ten-Cel QOBM/Comb. RRm.
Membro

_____
Pablo Federico **Baigorri** – Ten-Cel QOBM/Comb.
Orientador

USO DE TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL (PLN) PARA CONSULTA NORMATIVA DO CBMDF: PROPOSTA DE ARQUITETURA LEVE E ADAPTÁVEL COMO PROVA DE CONCEITO

RESUMO

O Corpo de Bombeiros Militar do Distrito Federal (CBMDF) atua com base em um amplo e heterogêneo conjunto de normas institucionais, cuja consulta, em contextos administrativos e operacionais, demanda rapidez, precisão e segurança jurídica. Apesar da existência de sistemas corporativos de acesso documental, a recuperação da informação normativa ainda ocorre predominantemente por meio de buscas lexicais tradicionais, o que limita a identificação de trechos relevantes de forma contextualizada. Nesse cenário, este trabalho tem como objetivo propor e avaliar uma arquitetura leve e adaptável baseada em técnicas de Processamento de Linguagem Natural (PLN) para consulta normativa no CBMDF, concebida como prova de conceito e compatível com infraestrutura computacional simples. A metodologia adotada possui caráter exploratório e experimental, envolvendo a construção de um corpus normativo representativo, a aplicação de diferentes estratégias de segmentação textual (*chunking*), a indexação por métodos lexicais e semânticos e a realização de experimentos comparativos de recuperação de informação. Os resultados indicam que, embora abordagens clássicas como o BM25 apresentem desempenho competitivo em determinados cenários, técnicas baseadas em PLN exigem cuidados específicos quanto à estruturação dos textos normativos e à definição dos parâmetros de recuperação. Conclui-se que a adoção de arquiteturas leves de PLN é tecnicamente viável no contexto institucional do CBMDF e que a proposta apresentada fornece uma base flexível para investigações futuras e para a evolução de soluções de consulta normativa assistida.

Palavras-chave: Processamento de Linguagem Natural. Consulta normativa. Recuperação de informação. Arquitetura leve. CBMDF.

Use of Natural Language Processing (NLP) Techniques for Normative Consultation in the CBMDF: Proposal of a Lightweight and Adaptable Architecture as a Proof of Concept

ABSTRACT

The Federal District Military Fire Department (CBMDF) operates based on an extensive and heterogeneous set of institutional regulations, whose consultation in administrative and operational contexts requires speed, accuracy, and legal certainty. Despite the availability of internal document management systems, access to normative information is still predominantly based on traditional lexical search methods, which limits the contextual retrieval of relevant excerpts. In this context, this study aims to propose and evaluate a lightweight and adaptable architecture based on Natural Language Processing (NLP) techniques for normative consultation within the CBMDF, designed as a proof of concept and compatible with modest computational infrastructure. An exploratory and experimental methodology was adopted, involving the construction of a representative normative corpus, the application of different text segmentation (chunking) strategies, indexing through lexical and semantic retrieval methods, and comparative information retrieval experiments. The results show that, although classical approaches such as BM25 achieve competitive performance in certain scenarios, NLP-based techniques require specific attention to the structuring of normative texts and to the definition of retrieval parameters. It is concluded that the adoption of lightweight NLP architectures is technically feasible in the institutional context of the CBMDF and that the proposed solution provides a flexible foundation for future research and the evolution of assisted normative consultation systems.

Keywords: Natural Language Processing. Normative consultation. Information retrieval. Lightweight architecture. CBMDF.

1 INTRODUÇÃO

O Corpo de Bombeiros Militar do Distrito Federal (CBMDF) desenvolve suas atividades administrativas e operacionais com base em um arcabouço normativo composto por leis, regulamentos, portarias, instruções normativas, planos e demais atos administrativos internos, os quais disciplinam sua organização, funcionamento e processos decisórios, conforme previsto em seu Regimento Interno (CBMDF, 2020). Esse conjunto normativo sustenta tanto a atuação administrativa da Corporação quanto o emprego operacional de seus recursos, conferindo segurança jurídica e padronização às decisões institucionais.

Nesse contexto, o Planejamento Estratégico do CBMDF para o período de 2025 a 2030 (PLANES 2025–2030) estabelece como diretrizes institucionais o fortalecimento da gestão, a modernização de processos e a adoção de soluções inovadoras que ampliem a eficiência organizacional e apoiem a tomada de decisão em todos os níveis da Corporação (CBMDF, 2024). Entre essas diretrizes, destaca-se a valorização da informação como ativo estratégico, especialmente no apoio às atividades administrativas e operacionais.

Apesar da relevância desse acervo normativo para o funcionamento do CBMDF, o acesso às normas institucionais ainda ocorre, em grande medida, de forma dispersa e manual, com documentos distribuídos entre diferentes plataformas e sistemas internos. Essa fragmentação pode comprometer a agilidade e a precisão das decisões, sobretudo em contextos operacionais que exigem respostas rápidas e fundamentadas, bem como no exercício de atividades administrativas que demandam constante consulta a atos normativos atualizados.

Avanços recentes na área de Inteligência Artificial (IA), em especial no campo do Processamento de Linguagem Natural (PLN), têm ampliado as possibilidades de automação e recuperação inteligente de informações textuais. Essas técnicas permitem que sistemas computacionais interpretem consultas formuladas em linguagem natural e recuperem informações relevantes a partir de grandes volumes documentais. Contudo, quando aplicadas a domínios institucionais altamente especializados, como o acervo normativo de organizações públicas, tais

soluções enfrentam desafios relacionados à precisão das respostas, à confiabilidade das informações recuperadas e à adequação ao contexto organizacional.

No âmbito do CBMDF, estudos prévios já evidenciaram a relevância do uso de tecnologias da informação e do conhecimento para apoiar a gestão e a tomada de decisão. O trabalho de Baigorri (2020) analisou a gestão do conhecimento na Corporação, destacando a importância de soluções tecnológicas para organizar, sistematizar e disponibilizar informações estratégicas de forma eficiente. De modo complementar, Alvin (2025) investigou a aplicação de técnicas de Inteligência Artificial no contexto do CBMDF, demonstrando o potencial dessas tecnologias para apoiar atividades institucionais que dependem de informação normativa estruturada. Esses estudos reforçam a pertinência do tema e indicam a existência de uma base conceitual e institucional favorável ao aprofundamento da pesquisa.

Diante desse cenário, a adoção de arquiteturas leves baseadas em técnicas de Processamento de Linguagem Natural emerge como uma possibilidade a ser investigada no âmbito do CBMDF. Diferentemente de soluções dependentes de infraestrutura em nuvem ou de elevados recursos computacionais, tais arquiteturas permitem a implementação local de mecanismos de recuperação da informação. Nesse contexto, coloca-se como questão relevante verificar se uma estrutura dessa natureza é capaz de atender, de forma adequada, às necessidades institucionais de consulta normativa, sem comprometer requisitos de sigilo, segurança da informação e viabilidade operacional.

Assim, o objetivo geral deste trabalho é propor e avaliar uma arquitetura leve e modular de Processamento de Linguagem Natural (PLN) para consulta normativa no âmbito do Corpo de Bombeiros Militar do Distrito Federal (CBMDF), analisando sua viabilidade técnica, eficiência computacional e aplicabilidade institucional.

Os seguintes objetivos específicos foram então definidos para atingir o propósito geral mencionado:

I. Realizar uma revisão bibliográfica sobre PLN, busca semântica e o emprego de arquiteturas leves em contextos institucionais;

II. Selecionar, organizar e preparar o corpus normativo do CBMDF a ser utilizado nos experimentos;

III. Elaborar um questionário de referência para a validação técnica e qualitativa dos resultados;

IV. Implementar uma prova de conceito da arquitetura proposta fundamentada na técnica de Geração Aumentada de Recuperação (RAG);

V. Avaliar a qualidade das respostas geradas por meio de métricas de precisão, revocação, fidelidade e similaridade semântica;

VI. Mensurar indicadores de eficiência computacional, especificamente latência e vazão do sistema;

VII. Comparar o desempenho da arquitetura leve com modelos robustos de IA, evidenciando os *trade-offs* entre qualidade e custo computacional;

VIII. Identificar as implicações práticas da adoção da solução no CBMDF.

A relevância deste estudo se manifesta em três dimensões complementares. Cientificamente, contribui para o campo do PLN ao conduzir análises experimentais aplicadas a um domínio institucional restritivo e de alta especificidade. Institucionalmente, propõe uma solução inovadora para modernizar o acesso às normas do CBMDF. Socialmente, apoia a atuação da Corporação na prestação de serviços públicos de segurança e salvamento, oferecendo aos militares acesso rápido, contextualizado e confiável às informações normativas essenciais à execução de suas missões.

Por fim, o trabalho organiza-se em cinco seções: (i) introdução e contextualização do problema; (ii) desenvolvimento; (iii) metodologia; (iv) análise e discussão dos resultados; e (v) considerações finais e referências.

2 DESENVOLVIMENTO

2.1 Referencial teórico

2.1.1 Conceito e História do Processamento de Linguagem Natural (PLN)

O Processamento de Linguagem Natural (PLN) é um campo da Inteligência Artificial que investiga métodos computacionais para a análise, interpretação e geração de textos em linguagem humana. De acordo com Caseli e Nunes (2024), o objetivo central do PLN é permitir que sistemas computacionais lidem de forma estruturada com fenômenos linguísticos complexos, aproximando a comunicação entre humanos e máquinas.

As primeiras abordagens do PLN foram fortemente influenciadas pela linguística teórica, em especial pela gramática gerativa proposta por Chomsky (1957). Nesse período, predominavam modelos simbólicos baseados em regras explícitas, nos quais o conhecimento linguístico era codificado manualmente por especialistas. Embora essas abordagens apresentassem elevado grau de controle e interpretabilidade, sua aplicação prática mostrou-se limitada diante da complexidade e variabilidade da linguagem natural, sobretudo em grandes volumes de texto.

Posteriormente, a partir das décadas de 1970 e 1980, o campo passou a incorporar métodos estatísticos e probabilísticos, viabilizando avanços significativos na análise automática de textos. Conforme destacam Manning e Schütze (1999), técnicas baseadas em frequência e probabilidade permitiram maior escalabilidade e adaptação a diferentes domínios, ainda que permanecessem fortemente dependentes da correspondência lexical entre consultas e documentos.

O avanço mais recente do PLN está associado ao uso de redes neurais profundas e, em especial, às arquiteturas do tipo *Transformers*. O trabalho de Vaswani *et al*. (2017) introduziu o mecanismo de “autoatenção” (*self-attention)*, possibilitando o tratamento eficiente de dependências de longo alcance em sequências textuais. Segundo Jurafsky e Martin (2026), esse paradigma é atualmente a arquitetura padrão para construção de Modelos de Linguagem de Grande Escala (*Large Language Models – LLMs*), capazes de executar múltiplas tarefas linguísticas com elevado desempenho sem a necessidade de ajustes extensivos.

2.1.2 Processamento de Linguagem Natural em domínios jurídicos e normativos

Apesar do potencial expressivo, a aplicação de técnicas de Processamento de Linguagem Natural (PLN) em domínios jurídicos e normativos apresenta características e desafios específicos (Costa; Dantas, 2023; Polo et al., 2021). Segundo os mesmos autores, estes documentos são produzidos segundo padrões formais rigorosos, com vocabulário técnico específico, uso recorrente de jargões e estruturas gramaticais complexas.

Lima (2024), em seu estudo sobre a utilização de técnicas de PLN aplicadas à Constituição Federal brasileira, destaca que normas jurídicas possuem elevada densidade informacional, diferentemente da linguagem ordinária, em que a redundância semântica é comum. Ainda segundo o autor, cada termo nesse domínio é deliberadamente escolhido para produzir efeitos jurídicos específicos, de modo que pequenas variações lexicais podem resultar em alterações significativas de sentido.

Similarmente, Costa e Dantas (2023) indicam em seus estudos que o processamento inadequado de termos ou radicais pode aproximar representações semânticas de conceitos juridicamente opostos, comprometendo a interpretação automática do texto normativo.

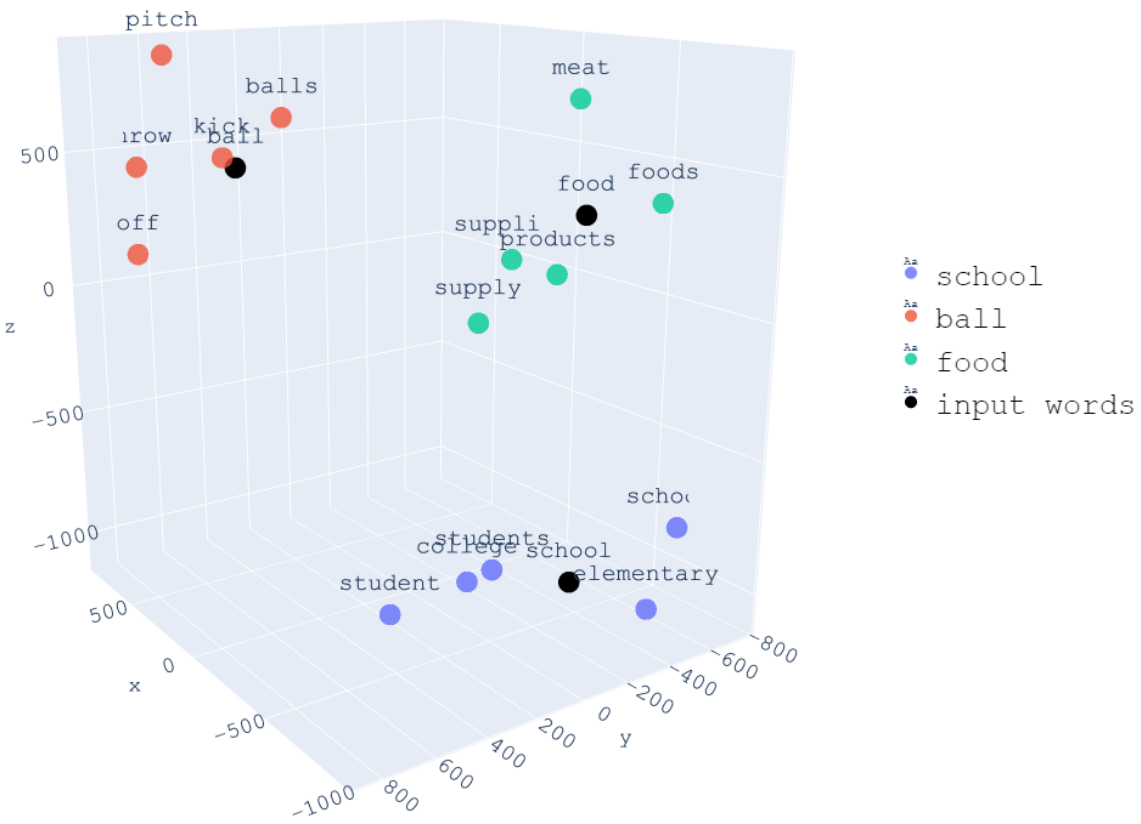
No campo da recuperação de informação jurídica, abordagens baseadas exclusivamente em correspondência lexical mostram-se eficientes e robustas do ponto de vista computacional, mas apresentam limitações na captura do significado contextual das normas. Tais limitações tornam-se evidentes em cenários de sinonímia, polissemia ou variação terminológica, nos quais a ausência de coincidência literal entre consulta e documento prejudica a recuperação da informação relevante (Lima, 2024; Pedroso; Ladeira; Faleiros, 2019).

Em contrapartida, métodos baseados em semântica distribuída ampliam a eficácia da recuperação de informações ao modelarem o significado textual de forma contextualizada (Caseli; Nunes, 2024; Jurafsky; Martin, 2023). Nessas abordagens, unidades linguísticas (palavras, sentenças ou trechos documentais, por exemplo) são convertidas em vetores numéricos de alta dimensionalidade, denominados *embeddings*, cujo posicionamento em um espaço multidimensional permite que a

proximidade geométrica indique afinidade conceitual entre os termos (Lima, 2024; Caseli; Nunes, 2024).

A Figura 1 ilustra, de forma simplificada, a organização de palavras em um espaço vetorial semântico, no qual, segundo Winastwan (2020), termos conceitualmente relacionados tendem a se agrupar em regiões próximas. Cada ponto representa a projeção tridimensional de um vetor de *embedding* originalmente calculado em um espaço de alta dimensionalidade, sendo essa visualização obtida por meio de técnicas de redução de dimensionalidade, como a *Principal Component Analysis* (PCA), que permitem preservar, de forma aproximada, as relações de proximidade semântica entre os vetores (Winastwan, 2020).

Figura 1 - Representação simplificada de palavras em um espaço vetorial semântico.



Fonte: adaptado de Winastwan (2020).

Esse mecanismo permite que documentos conceitualmente relacionados sejam identificados como semanticamente próximos, mesmo quando utilizam vocabulários distintos ou formulações diferentes. Como consequência, reduz-se a dependência estrita da correspondência lexical típica das buscas baseadas em palavras-chave, possibilitando a recuperação de informações relevantes mesmo na ausência de termos idênticos, aspecto particularmente relevante em domínios normativos e jurídicos, caracterizados por variações terminológicas e equivalências conceituais (Lima, 2024).

Contudo, a adoção de modelos generativos de larga escala em contextos normativos introduz desafios adicionais, notadamente relacionados à opacidade dos processos de inferência. Tais arquiteturas, frequentemente fundamentadas em técnicas de aprendizado profundo (*deep learning*), operam como "caixas-pretas", nas quais o conhecimento apreendido pelos algoritmos não é recuperável de forma compreensível para humanos, tornando o comportamento do sistema não determinístico e de difícil auditoria (Caseli; Nunes, 2024; Autoridade Nacional De Proteção De Dados, 2024). Essa característica compromete a auditoria das respostas geradas, uma vez que a baixa transparência decorre da complexidade envolvida na compreensão das operações matemáticas que conduzem a um determinado resultado (Autoridade Nacional de Proteção de Dados, 2024).

Destaca-se ainda, nesse cenário, o risco de ocorrência de "alucinações", fenômeno no qual sistemas produzem respostas linguisticamente plausíveis, porém factualmente incorretas ou desvinculadas das fontes oficiais (Autoridade Nacional De Proteção De Dados, 2024; Rodrigues; Marcacini, 2024). Em ambientes institucionais, especialmente na administração pública militar, tais limitações reforçam a necessidade de abordagens que conciliem compreensão semântica, previsibilidade dos resultados e aderência estrita às normas vigentes.

2.1.3 Representação e recuperação da informação normativa

A recuperação da informação normativa e jurídica consiste no processo de identificar, em um acervo de documentos legais ou administrativos, itens relevantes que satisfaçam a necessidade informacional expressa em uma consulta pelo usuário (Moreira, 2024; Pedroso; Ladeira; Faleiros, 2019). Tradicionalmente, esse processo

fundamenta-se em modelos de correspondência lexical, nos quais a relevância é calculada a partir de evidências estatísticas, como a frequência e a distribuição dos termos presentes nos documentos e na consulta (Lima, 2024; Moreira, 2024).

Entre os modelos lexicais mais consolidados, destacam-se o TF-IDF (*Term Frequency–Inverse Document Frequency*) e o BM25 (*Best Matching 25*) (Moreira, 2024; Pedroso; Ladeira; Faleiros, 2019). Essas abordagens são amplamente reconhecidas por apresentarem um bom equilíbrio entre desempenho e simplicidade, além de oferecerem elevada eficiência computacional e previsibilidade, sendo frequentemente adotadas como linhas de base (*baselines*) em ambientes institucionais (Pedroso; Ladeira; Faleiros, 2019). Entretanto, sua eficácia é limitada pelo problema da incompatibilidade de vocabulário (*vocabulary mismatch*), uma vez que a dependência da coincidência literal entre termos impede a captura de relações semânticas como sinonímia e polissemia (Jurafsky; Martin, 2026; Lima, 2024; Moreira, 2024).

Com o avanço das técnicas de representação vetorial, modelos semânticos passaram a ser empregados para ampliar a capacidade de recuperação conceitual (Autoridade Nacional De Proteção De Dados, 2024; Caseli; Nunes, 2024). Métodos como o Word2Vec representam unidades linguísticas em espaços vetoriais densos, nos quais a proximidade entre vetores reflete a similaridade semântica aprendida a partir do contexto de uso (Jurafsky; Martin, 2026). Tais vetores densos (*embeddings*) superam as limitações dos vetores esparsos (lexicais) ao capturar o significado coletivo e reduzir problemas associados à variação terminológica, comuns em contextos normativos complexos (Jurafsky; Martin, 2026).

A aplicação dessas representações em escala é viabilizada por mecanismos de busca vetorial, como a biblioteca FAISS (*Facebook AI Similarity Search*), que utiliza algoritmos de busca aproximada de vizinhos mais próximos para garantir a recuperação eficiente em grandes coleções de dado (Jurafsky; Martin, 2026).

Dentre abordagens essenciais distintas, a escolha dos modelos de representação e recuperação da informação deve estar alinhada ao contexto no qual

é inserida e buscar equilibrar eficiência computacional, capacidade semântica e previsibilidade dos resultados.

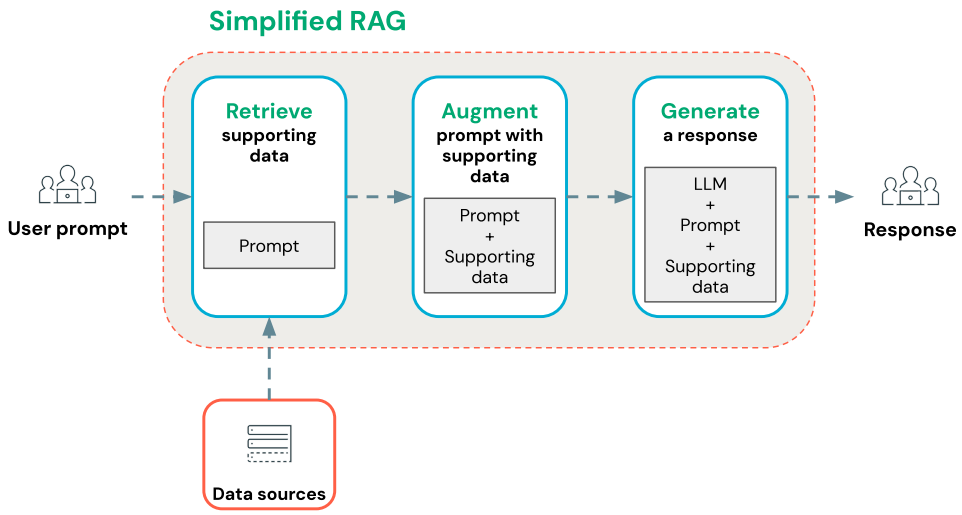
2.1.4 Arquiteturas de geração aumentada por recuperação (RAG)

Modelos de geração de linguagem apresentam elevada capacidade de produzir respostas fluentemente estruturadas (Caseli; Nunes, 2024), porém sua aplicação direta sobre bases normativas institucionais envolve riscos significativos. Em especial, esses modelos podem gerar respostas linguisticamente plausíveis, mas desvinculadas das fontes oficiais, o que compromete a confiabilidade da informação em contextos que exigem precisão**,** rastreabilidade e aderência normativa (Jurafsky; Martin, 2026; Sakiyama Et Al., 2024).

Nesse cenário, as arquiteturas de geração aumentada por recuperação (RAG) configuram-se como uma estratégia para mitigar tais riscos (Rodrigues, 2024). Essa abordagem combina mecanismos de recuperação da informação, responsáveis por selecionar trechos relevantes do acervo documental, com modelos de geração que utilizam o conteúdo recuperado como base contextual para formular respostas (Jurafsky; Martin, 2026; Lima, 2024).

De forma simplificada, o funcionamento da geração aumentada por recuperação pode ser compreendido em três etapas principais, conforme ilustrado na Figura 2. Inicialmente, a solicitação do usuário é utilizada para recuperar informações de fontes externas, como bases documentais ou repositórios vetoriais. Em seguida, esses dados de suporte são combinados à consulta original para compor o contexto da interação. Por fim, o modelo de linguagem gera a resposta com base nesse conjunto de informações recuperadas, e não apenas em conhecimento previamente aprendido.

Figura 2 – Fluxo simples de uma arquitetura de RAG



Fonte: adaptado de Microsoft (2025)

Dessa forma, a adoção de arquiteturas de geração aumentada por recuperação apresenta-se como alternativa tecnicamente adequada para aplicações institucionais. Ao equilibrar compreensão semântica, controle do conteúdo gerado e fidelidade às fontes oficiais, essa abordagem visa atender a requisitos essenciais de confiabilidade, segurança jurídica e adequação operacional.

2.2 Metodologia

A pesquisa classifica-se como aplicada, de natureza exploratória e descritiva, com abordagem qualiquantitativa. Conforme Gil (2008), a pesquisa aplicada tem como objetivo gerar conhecimento voltado à solução de problemas específicos, enquanto as pesquisas exploratórias e descritivas permitem compreender melhor um fenômeno e caracterizá-lo em seus diversos aspectos.

De acordo com Marconi e Lakatos (2007), o caráter exploratório contribui para a familiarização com o problema e a formulação de hipóteses mais precisas, enquanto o caráter descritivo busca retratar fielmente as características observáveis de um fenômeno.

Já segundo Prodanov e Freitas (2013), a abordagem qualiquantitativa combina técnicas qualitativas e quantitativas, permitindo integrar a análise interpretativa à mensuração objetiva dos resultados.

O delineamento metodológico contempla seis etapas principais:

i. Criação do corpus normativo e preparação dos dados;

ii. criação do questionário de referência;

iii. execução dos experimentos comparativos;

iv. análise crítica do desempenho dos modelos;

v. desenvolvimento do protótipo experimental e

vi. aspectos éticos da pesquisa.

2.2.1 Criação do corpus normativo e preparação dos dados

A construção do corpus normativo constituiu etapa fundamental para a condução dos experimentos, uma vez que a qualidade e a representatividade da base documental influenciam diretamente o desempenho dos mecanismos de recuperação da informação. O processo de criação do corpus foi estruturado em três fases: coleta documental, curadoria e padronização, e segmentação textual.

2.2.1.1 Fase de coleta documental

A coleta foi realizada a partir do repositório público de documentos normativos disponível no sítio institucional do CBMDF. Foram coletados documentos de naturezas diversas, incluindo portarias internas, instruções normativas, boletins técnicos profissionais, procedimentos operacionais padrão, manuais operacionais e administrativos, normas técnicas, normas de emprego operacional, bem como leis e decretos federais e distritais diretamente aplicáveis ao contexto da Corporação.

O objetivo dessa abrangência foi construir uma representação aproximada do universo normativo consultado pelos militares do CBMDF em suas atividades administrativas e operacionais, permitindo avaliar o comportamento dos algoritmos de recuperação em condições próximas ao uso real.

2.2.1.2 Fase de curadoria e padronização

Após a coleta inicial, os documentos foram submetidos a um processo de curadoria que envolveu a verificação de integridade dos arquivos, a conversão para formato textual padronizado e a aplicação de critérios de inclusão e exclusão. Foram excluídos do corpus os seguintes tipos de documentos: (i) portarias expressamente revogadas por versões posteriores, de modo a evitar a presença de informações desatualizadas que pudessem comprometer a qualidade das respostas; (ii) documentos com falhas de conversão que inviabilizassem a extração textual adequada; e (iii) documentos duplicados ou versões preliminares substituídas por versões definitivas.

Ao término dessa fase, o corpus consolidado totalizou 1.227 documentos, cuja distribuição por tipo documental é apresentada na Tabela 1.

Tabela 1 - Composição do corpus normativo por tipo documental

Tipo	Quantidade	% do total
Portaria	1118	91,1%
Boletim Técnico Profissional	32	2,6%
Instrução Normativa	28	2,3%
Decreto Distrital	10	0,8%
Manual	10	0,8%
Lei Federal	9	0,7%
Procedimento Operacional Padrão	5	0,4%
Norma de Emprego	5	0,4%
Lei Distrital	3	0,2%
Decreto Federal	2	0,2%
Norma Técnica	2	0,2%
Outros (regulamentos, planos etc.)	3	0,3%

Fonte: o autor.

A predominância de portarias (91,1%) reflete a estrutura normativa do CBMDF, na qual esse instrumento é amplamente utilizado para regulamentar procedimentos internos, definir competências e disciplinar atividades administrativas e operacionais. As demais categorias, embora numericamente menos expressivas, contribuem para a diversidade temática do corpus e representam documentos de elevada relevância institucional, como manuais operacionais e procedimentos padronizados.

2.2.1.3 Fase de segmentação textual (*chunking*)

A segmentação textual consiste na divisão dos documentos em fragmentos menores, denominados *chunks*, que constituem as unidades básicas de indexação e recuperação. A definição da estratégia de segmentação é uma decisão metodológica relevante em arquiteturas de RAG, uma vez que o tamanho e a sobreposição dos fragmentos influenciam tanto a qualidade da recuperação quanto o custo computacional do sistema (Sinha, 2025).

Neste trabalho, adotou-se uma estratégia de segmentação baseada em sentenças, com os seguintes parâmetros: tamanho máximo de 512 *tokens* por fragmento e sobreposição (*overlap*) de 64 *tokens* entre fragmentos consecutivos. A escolha desses valores fundamentou-se em práticas consolidadas na literatura sobre sistemas de RAG, que indicam essa faixa como adequada para modelos de linguagem baseados em arquiteturas *Transformer*, cujo desempenho tende a degradar em contextos excessivamente longos ou fragmentados (Sinha, 2025).

A aplicação uniforme dessa estratégia a todos os tipos documentais teve como objetivo assegurar a comparabilidade entre os cenários experimentais, permitindo isolar o efeito das variáveis de interesse, os métodos de recuperação e modelos de geração, sem a introdução de vieses decorrentes de tratamentos diferenciados por categoria documental.

Ao término do processo de segmentação, o corpus resultou em aproximadamente 11.714 fragmentos indexáveis, que constituíram a base para os experimentos de recuperação e geração descritos nas seções subsequentes.

2.2.2 Criação do questionário de referência

A avaliação objetiva de sistemas de recuperação da informação requer a existência de um conjunto de referência — tecnicamente denominado coleção de teste — composto por documentos, consultas e seus respectivos julgamentos de relevância, o que permite comparar as respostas produzidas pelo sistema com validações prévias (Moreira, 2024). Na literatura, esse conjunto de respostas corretas é frequentemente denominado *ground-truth*, constituindo a base de verdade contra a qual a eficácia do sistema é mensurada (Rodrigues, 2024).

Neste trabalho, foi elaborado um questionário de referência composto por 88 pares de consulta e resposta esperada, construídos manualmente a partir do corpus normativo do CBMDF e disponíveis em repositório *online*[1]. O processo de construção seguiu três etapas: definição dos critérios de elaboração, formulação das consultas e validação das respostas.

2.2.2.1 Critérios de elaboração

A construção do questionário de referência foi orientada por critérios voltados à representatividade e à aplicabilidade das consultas no contexto do CBMDF, a saber:

a) Cobertura temática: inclusão de consultas associadas a diferentes tipos documentais do corpus, abrangendo normas administrativas, procedimentos operacionais e normas técnicas;

b) Variação de complexidade: contemplação de perguntas objetivas e de consultas que exigem articulação de informações distribuídas em múltiplos trechos normativos;

c) Realismo das formulações: redação das consultas em linguagem natural, compatível com a forma como militares do CBMDF expressam dúvidas no contexto de trabalho;

d) Verificabilidade das respostas: fundamentação direta das respostas esperadas em trechos específicos do corpus, assegurando a rastreabilidade da informação.

2.2.2.2 Formulação das consultas

As consultas foram elaboradas pelo autor com base na análise do corpus normativo e no conhecimento das demandas informacionais típicas do contexto institucional. Buscou-se formular perguntas que refletissem situações recorrentes de consulta, tais como: identificação da norma aplicável a determinada situação,

[1] Disponível em: https://github.com/Pedro-M-Santos/cbmdf-rag-experiments/blob/main/data/goldset/goldset.json. Acesso em: 13 mar. 2026.

localização procedimentos e prazos específicos, esclarecimento de competências e atribuições, verificação de requisitos para concessão de direitos ou benefícios e até mesmo questionamentos técnicos operacionais relacionados à atuação bombeiro militar.

A Tabela 2 apresenta a distribuição das consultas por tipo documental de origem, evidenciando a cobertura alcançada em relação às diferentes categorias normativas do corpus.

Tabela 2 - Divisão de questões do conjunto de referência entre tipos de normas

Tipo documental	Sigla	Nº de questões
Portaria	POR	35
Boletim Técnico Profissional	BITP	11
Manual	MAN	7
Instrução Normativa	IN	7
Lei Federal	LF	6
Procedimento Operacional Padrão	POP	8
Lei Distrital	LD	3
Decreto Distrital	DECD	2
Decreto Federal	DECF	2
Norma de Emprego	NEO	2
Norma Técnica	NT	1
Regulamento Interno	RUBM	2
Plano Estratégico	PLANES	2
Instrução Normativa Geral	INSGER	1

Fonte: o autor.

A distribuição das consultas reflete, de forma aproximada, a composição do corpus normativo, com predominância de questões baseadas em portarias (39,8%), seguidas por boletins técnicos profissionais (12,5%) e procedimentos operacionais padrão (9,1%). Essa proporcionalidade buscou assegurar que os tipos documentais mais representativos do acervo institucional fossem adequadamente contemplados na avaliação, sem desconsiderar categorias menos numerosas, porém de elevada relevância operacional.

2.2.2.3 Validação das respostas

Para cada consulta formulada, foi redigida uma resposta esperada que sintetiza a informação normativa correta, acompanhada da indicação do(s) documento(s) e trecho(s) específico(s) que a fundamentam. Esse procedimento teve

como objetivo permitir não apenas a avaliação da similaridade semântica entre a resposta do sistema e a resposta esperada, mas também a verificação da capacidade do mecanismo de recuperação em identificar os trechos normativos pertinentes.

2.2.3 Execução dos experimentos comparativos

Os experimentos foram conduzidos com o objetivo de avaliar comparativamente diferentes configurações de recuperação da informação e geração de respostas, permitindo identificar as combinações mais adequadas ao contexto de consulta normativa do CBMDF.

Visando garantir a reprodutibilidade dos resultados apresentados, bem como a transparência nos procedimentos de pré-processamento e parametrização dos modelos, o código-fonte desenvolvido para a execução destes experimentos foi disponibilizado publicamente em repositório remoto[2].

2.2.3.1 Ambiente computacional e estrutura dos experimentos

O sistema foi implementado em linguagem Python, versão 3.10, utilizando bibliotecas de código aberto voltadas ao processamento de linguagem natural e à recuperação da informação. Para a geração de representações vetoriais (*embeddings*), foi empregada a biblioteca *Sentence-Transformers*. A indexação e busca lexical foram realizadas por meio da biblioteca BM25, enquanto a indexação e busca vetorial utilizaram a biblioteca FAISS (*Facebook AI Similarity Search*). A inferência com modelos de linguagem foi conduzida por meio da biblioteca *Transformers* e, nos cenários com modelos quantizados, pelo *framework* Llama.cpp. Para organização, manipulação e análise dos dados experimentais, foram utilizadas as bibliotecas Pandas e Matplotlib.

A infraestrutura computacional foi configurada para operar em ambiente local, sem o uso de unidades de processamento gráfico (GPU) dedicadas, refletindo as restrições típicas de ambientes institucionais. Essa escolha teve como objetivo

[2] Disponível em: https://github.com/Pedro-M-Santos/cbmdf-rag-experiments. Acesso em: 13 mar. 2026.

avaliar a viabilidade de execução da arquitetura proposta em condições compatíveis com a realidade do CBMDF.

A arquitetura foi projetada de forma modular, permitindo o chaveamento controlado entre diferentes métodos de recuperação e modelos de geração. Foram definidos sete cenários experimentais, organizados em dois blocos: o primeiro voltado à comparação de métodos de recuperação da informação (cenários S1, S2 e S3) e o segundo à comparação de modelos de geração de respostas (cenários S4, S5, S6 e S7). A Tabela 3 sintetiza a configuração de cada cenário.

Tabela 3 - Relação dos experimentos conduzidos

Cenário	Bloco	Método de recuperação	Método de geração
S1	Recuperação	BM25	Extrativo
S2	Recuperação	FAISS	Extrativo
S3	Recuperação	BM25 + FAISS	Extrativo
S4	Geração	BM25	Qwen3-0.6B
S5	Geração	BM25	GPT-4o mini
S6	Geração	BM25	Qwen3-0.6B (Q8 via Llama.cpp)
S7	Geração	BM25	Qwen3-1.7B (Q8 via Llama.cpp)

Fonte: o autor.

Para garantir a comparabilidade entre os cenários, foram adotados os seguintes procedimentos:

a) Uniformidade do corpus e da segmentação: todos os cenários utilizaram o mesmo corpus normativo, submetido à mesma estratégia de segmentação textual descrita na seção 2.2.1;

b) Questionário de referência único: as 88 consultas do questionário de referência foram submetidas a todos os cenários, permitindo a comparação direta das respostas obtidas;

c) Parâmetros de recuperação padronizados: em todos os cenários que empregaram recuperação, foi definido um número fixo de 11714 trechos a serem recuperados por consulta;

d) Controle de aleatoriedade: nos cenários com modelos generativos, os parâmetros de amostragem foram configurados de modo a minimizar a variabilidade das respostas entre execuções.

2.2.3.1.1 Cenário de recuperação da informação

O experimento S1 é o ensaio mais simples dentre os executados e utiliza um modelo de recuperação de informação a partir do *corpus* normativo e devolve esses trechos como saída. O modelo probabilístico de recuperação de informação usado é o BM25 (*Best Match* 25), uma abordagem clássica e amplamente utilizada em recuperação de informações (Caseli; Nunes, 2024).

O BM25 leva em conta a frequência em que cada termo aparece no documento, o tamanho do documento e normaliza os valores para melhorar a precisão em coleções com documentos de tamanhos variados. A escolha do BM25 é justificada por possuir um bom equilíbrio entre desempenho e simplicidade (Pedroso; Ladeira; Faleiros, 2019), sendo uma referência em tarefas de recuperação e frequentemente usado em pesquisas como modelo base para comparação com técnicas modernas (Caseli; Nunes, 2024).

Sendo assim, este primeiro experimento visou verificar como um modelo clássico porém eficiente se comportaria, sem se preocupar neste momento com modelos de geração de texto.

2.2.3.1.2 Cenário S2: recuperação semântica

O experimento S2 buscou analisar o comportamento de um algoritmo puramente extrativo, porém baseado em recuperação semântica, em contraste com a abordagem lexical adotada no cenário S1. Na recuperação semântica, os trechos de texto são representados por vetores densos (*embeddings*) gerados por algoritmos de codificação vetorial (*encoding*), permitindo capturar o significado semântico dos fragmentos textuais (Jurafsky; Martin, 2026).

Essa representação vetorial possibilita a comparação entre consulta e documentos com base em proximidade semântica, reduzindo a dependência da

correspondência literal de termos e mitigando o problema de incompatibilidade de vocabulário (Lima, 2024). Para a geração dos vetores, foi utilizado um modelo de codificação vetorial de linguagem natural adequado a cenários de recuperação semântica, enquanto a indexação e a busca vetorial foram realizadas por meio da biblioteca FAISS, escolhida por sua eficiência computacional, escalabilidade e ampla adoção em sistemas de recuperação baseados em representações vetoriais (Ong, 2024).

2.2.3.1.3 Cenário S3: recuperação híbrida

O experimento S3 avaliou uma abordagem híbrida de recuperação da informação, combinando métodos de recuperação lexical e semântica em um mesmo fluxo extrativo. A motivação central para essa escolha reside no reconhecimento de que consultas normativas apresentam naturezas distintas, variando entre perguntas objetivas, com terminologia específica, e formulações mais descritivas, expressas em linguagem natural.

Enquanto métodos baseados em correspondência lexical, como o BM25, tendem a apresentar bom desempenho em consultas que dependem de termos normativos explícitos, abordagens baseadas em recuperação semântica, como aquelas apoiadas em FAISS, mostram-se mais adequadas para capturar similaridade conceitual entre trechos semanticamente relacionados. Trabalhos recentes sobre sistemas de pergunta–resposta aplicados a documentos técnicos indicam que a combinação dessas estratégias pode reduzir limitações individuais e aumentar a robustez do processo de recuperação da informação (Hakdağli, 2024).

Dessa forma, o cenário S3 foi estruturado para investigar se a integração entre BM25 e FAISS é capaz de oferecer um equilíbrio mais adequado entre precisão e cobertura, mantendo-se compatível com os requisitos de simplicidade arquitetural, viabilidade computacional e confiabilidade exigidos em contextos institucionais, como o do CBMDF.

2.2.3.2 Cenários de geração de respostas

Nos cenários do bloco de geração (S4 a S7), o método de recuperação foi mantido constante (BM25), variando-se apenas o modelo de linguagem responsável pela síntese da resposta. Essa configuração permitiu avaliar o desempenho comparativo dos modelos de geração sob condições equivalentes de recuperação.

2.2.3.2.1 Cenário S4: Geração com modelo leve de linguagem

O experimento S4 foi concebido para avaliar a viabilidade do uso de um modelo de linguagem gerador de texto de pequeno porte, especificamente o Qwen3-0.6B, como componente de geração de texto em uma arquitetura RAG voltada ao contexto institucional. Diferentemente dos cenários anteriores, que adotam um resultado estritamente extrativo, este cenário investiga se um modelo generativo leve é capaz de produzir respostas coerentes e contextualizadas quando condicionado aos trechos normativos recuperados.

A escolha do Qwen3-0.6B (Tabela 4) fundamenta-se em sua proposta de baixo custo computacional, reduzido consumo de memória e possibilidade de execução em ambientes locais, características alinhadas às restrições de infraestrutura comuns a órgãos públicos. Modelos dessa categoria representam uma alternativa intermediária entre abordagens puramente extrativas e o uso de modelos de grande porte hospedados em nuvem, permitindo avaliar ganhos potenciais de fluidez e síntese textual sem comprometer requisitos de segurança da informação.

Tabela 4 - Características do Modelo Qwen3-0.6B

Característica	Valor
Modelo	Qwen3-0.6B
Fabricante	*Alibaba Cloud (Alibaba Group*)
Tipo	*Causal Language Model*
Número total de parâmetros	0,6 bilhão
Comprimento máximo de contexto	32.768 tokens
Licença	Apache 2.0

Fonte**:** elaboração do autor, com base na documentação técnica do Qwen3 (QWEN TEAM, 2025).

Nesse cenário, o experimento S4 busca, analisar se a adoção de um modelo generativo leve é tecnicamente viável e adequada ao domínio normativo,

considerando aspectos como aderência ao contexto, risco de alucinação e compatibilidade com arquiteturas leves e institucionais.

2.2.3.2.2 Cenário S5: Geração com modelo externo de referência

O experimento S5 foi estruturado com o objetivo de utilizar um modelo de linguagem externo como referência comparativa, empregando o GPT-4o mini (Tabela 5) como componente de geração em uma arquitetura RAG. Diferentemente dos cenários anteriores, este experimento não tem como finalidade propor uma solução aplicável ao contexto institucional, mas sim estabelecer um ponto de comparação para análise do desempenho relativo de modelos leves executados localmente.

O GPT-4o mini foi selecionado por representar uma classe de modelos generativos de pequeno porte projetados para oferecer bom custo-benefício, baixa latência e qualidade de resposta, sendo indicado para aplicações que envolvem passagem de grandes volumes de contexto e arquiteturas de RAG, conforme descrito em sua documentação oficial (OpenAI, 2024).

Tabela 5 – Características do GPT-4o mini

Característica	Valor
Modelo	GPT-4o mini
Fabricante	*OpenAI*
Tipo	*Large Language Model (LLM)* proprietário
Número total de parâmetros	Não divulgado pelo fornecedor
Comprimento máximo de contexto	Até 128.000 *tokens*
Estágio de treinamento	Pré-treinamento e pós-treinamento
Licença	Proprietária

Fonte**:** elaboração do autor, com base na documentação técnica do GPT-4o mini (OPENAI, 2024).

O cenário S5 busca, portanto, servir como base comparativa, auxiliando na análise da viabilidade técnica de arquiteturas leves frente a soluções baseadas em modelos externos, sem desconsiderar restrições institucionais como custo, segurança da informação e dependência de serviços em nuvem.

2.2.3.2.3 Cenário S6: Geração com modelo local quantizado (Qwen3-0.6B em Q8 via Llama.cpp)

O experimento S6 foi concebido para avaliar o impacto da execução de um modelo de linguagem leve em ambiente local utilizando técnicas de quantização, especificamente o Qwen3-0.6B quantizado em Q8 e executado por meio do *framework* Llama.cpp. Esse cenário mantém a configuração lógica do experimento S4 — geração condicionada exclusivamente aos trechos normativos recuperados via BM25 — diferenciando-se apenas pela forma de execução do modelo.

O Llama.cpp é uma implementação otimizada para inferência local de modelos de linguagem, permitindo sua execução sem dependência de GPU dedicada ou infraestrutura em nuvem. A quantização em Q8 reduz a precisão dos pesos para 8 *bits*, diminuindo o consumo de memória e o custo computacional, com impacto controlado sobre a qualidade das respostas. Essa abordagem viabiliza o uso de modelos generativos em hardware convencional, adequado a ambientes institucionais com restrições de infraestrutura.

Nesse experimento, o uso do Qwen3-0.6B em Q8 permite investigar se os ganhos de eficiência proporcionados pela quantização preservam níveis aceitáveis de coerência, aderência ao contexto e fidelidade ao conteúdo normativo.

2.2.3.2.4 Cenário S7: Geração com modelo local quantizado de maior porte (Qwen3-1.7B em Q8 via Llama.cpp)

O experimento S7 amplia a análise do cenário anterior ao empregar um modelo de linguagem local de maior porte, o Qwen3-1.7B (Tabela 6), também quantizado em Q8 e executado via Llama.cpp. A estrutura do experimento permanece inalterada em relação aos cenários S4 e S6, com recuperação baseada em BM25 e geração restrita ao contexto normativo recuperado.

Tabela 6 – Características do modelo Qwen3-1.7B

Característica	Valor
Modelo	Qwen3-1.7B
Fabricante	*Alibaba Cloud (Alibaba Group)*
Tipo	*Causal Language Model*
Número total de parâmetros	1,7 bilhões
Comprimento máximo de contexto	32.768 *tokens*
Licença	*Apache License 2.0*

Fonte**:** elaboração do autor, com base na documentação técnica do Qwen3 (QWEN TEAM, 2025).

A adoção de um modelo com maior número de parâmetros permite avaliar o equilíbrio entre capacidade expressiva e custo computacional em ambientes locais, sendo a quantização em Q8 um mecanismo essencial para viabilizar sua execução. O cenário S7 analisa se o aumento de capacidade do modelo resulta em ganhos perceptíveis de qualidade em relação ao cenário S6 e se tais ganhos justificam o impacto adicional em latência e consumo de recursos.

2.2.4 Análise crítica do desempenho dos modelos

A avaliação dos experimentos foi conduzida com base no framework Ragas (Retrieval-Augmented Generation Assessment), uma biblioteca de código aberto voltada à avaliação sistemática de aplicações baseadas em modelos de linguagem, especialmente arquiteturas de geração aumentada por recuperação. O Ragas foi selecionado por oferecer um conjunto de métricas estruturadas que permitem comparar diferentes configurações de sistemas RAG de forma consistente, superando abordagens baseadas exclusivamente em métricas lexicais ou em avaliações manuais de baixa escalabilidade (Ragas, 2025).

2.2.4.1 Avaliação por modelos de linguagem (*LLM judges*)

Um elemento central do framework Ragas é o uso de modelos de linguagem como avaliadores (*LLM judges*) para operacionalizar métricas de natureza semântica. Nesse paradigma, um modelo avaliador é empregado para analisar consultas, trechos recuperados, respostas geradas e respostas de referência, estimando relações de relevância, cobertura e aderência ao contexto. Essa estratégia abstrai a complexidade da avaliação manual e viabiliza a mensuração de aspectos críticos em sistemas RAG, como relevância contextual, similaridade semântica e fidelidade da resposta ao conteúdo recuperado (Ragas, 2025).

2.2.4.2 Métricas de qualidade da recuperação

Nos cenários de recuperação (S1 a S3), a avaliação concentrou-se em três métricas principais: precisão do contexto (*context precision*), revocação do

contexto (*context recall*) e similaridade semântica da resposta (*answer semantic similarity).*

A métrica precisão do contexto avalia a capacidade do mecanismo de busca em priorizar trechos relevantes nas primeiras posições do ranqueamento. É calculada como a média ponderada da precisão ao longo das posições do conjunto recuperado, conforme a Equação 1:

$$\frac{\sum_{k=1}^{K}(Precisão@k \times v_k)}{(Número\ total\ de\ itens\ relevantes\ nos\ top\ K\ resultados)} \quad (1),$$

Onde v_k indica se o item na posição k é relevante (1) ou não (0)

A métrica revocação do contexto mensura o quanto das informações relevantes esperadas foi efetivamente recuperado pelo sistema, com foco na minimização da omissão de evidências essenciais. No Ragas, essa métrica é operacionalizada a partir da decomposição da resposta de referência em afirmações (*claims*), avaliando-se quantas dessas afirmações podem ser sustentadas pelos trechos recuperados, conforme a Equação 2:

$$\frac{\text{N° de afirmações da resposta de referência suportadas pelo contexto recuperado}}{\text{Número total de afirmações na resposta de referência}} \quad (2)$$

A métrica similaridade semântica da resposta foi utilizada em todos os cenários experimentais para estimar o grau de alinhamento semântico entre a resposta gerada e a resposta de referência do questionário. Essa métrica baseia-se na comparação entre representações vetoriais (embeddings) da resposta e da referência, produzidas por modelos semânticos. Valores mais elevados indicam maior equivalência de significado, independentemente de variações na redação textual.

2.2.4.3 Métricas de qualidade da geração

Nos cenários de geração (S4 a S7), além da similaridade semântica da resposta, foi empregada a métrica fidelidade (*faithfulness*), destinada a avaliar a consistência factual da resposta em relação ao contexto recuperado. Essa métrica é particularmente relevante para identificar o risco de alucinação, fenômeno no qual o

modelo gera informações plausíveis porém não fundamentadas nas evidências fornecidas.

A fidelidade é calculada a partir da identificação das afirmações presentes na resposta gerada e da verificação de seu suporte nos trechos recuperados, conforme a Equação 3:

$$\frac{\text{Número de afirmações na resposta suportadas pelo contexto recuperado}}{\text{Número total de afirmações na resposta}} \quad (3)$$

As métricas de precisão do contexto, revocação do contexto, similaridade da resposta e fidelidade são expressas em valores normalizados no intervalo entre 0 e 1, nos quais valores mais próximos de 1 indicam melhor desempenho.

2.2.4.4 Métricas de desempenho computacional

Além das métricas de qualidade, foram coletadas métricas de desempenho computacional para caracterizar o custo associado à execução de cada cenário experimental. Essas métricas são particularmente relevantes para avaliar a viabilidade de implantação da arquitetura em ambientes institucionais com restrições de infraestrutura.

A latência média, expressa em segundos, representa o tempo médio necessário para produzir uma resposta a cada consulta, considerando as etapas de recuperação e, quando aplicável, geração.

A vazão, medida em consultas por segundo (questões por segundo – qps), indica a capacidade de processamento do sistema, ou seja, quantas consultas podem ser atendidas por unidade de tempo.

O consumo de memória do índice, expresso em *megabytes* (MB), corresponde à quantidade de memória ocupada pelas estruturas de indexação mantidas em memória principal durante a execução dos experimentos. A análise desse indicador permite avaliar o impacto das diferentes estratégias sobre os requisitos de armazenamento e a viabilidade de adoção em ambientes com recursos limitados.

2.2.4.5 Síntese do protocolo de avaliação

A Tabela 7 sintetiza as métricas aplicadas a cada bloco de experimentos:

Tabela 7 - Métricas de avaliação por bloco experimental

Bloco	Cenários	Métricas de qualidade	Métricas computacionais
Recuperação	S1, S2, S3	Precisão do contexto, revocação do contexto, similaridade da resposta	Latência, vazão, memória do índice
Geração	S4, S5, S6, S7	Similaridade da resposta, fidelidade	Latência, vazão

Fonte: elaborado pelo autor.

A combinação dessas métricas permitiu uma análise multidimensional do desempenho dos cenários, contemplando tanto a qualidade das respostas produzidas quanto a viabilidade operacional de cada configuração no contexto institucional do CBMDF.

2.2.5 Desenvolvimento do protótipo experimental

Como parte do delineamento metodológico, foi desenvolvido um protótipo funcional com o objetivo de viabilizar a execução controlada dos experimentos definidos neste estudo. O protótipo foi concebido como instrumento de apoio à experimentação, permitindo configurar combinações de modelos de linguagem, métodos de recuperação e parâmetros de geração, bem como registrar as respostas e os contextos normativos recuperados.

O desenvolvimento do protótipo buscou atender a três requisitos metodológicos principais: (i) permitir o chaveamento explícito entre diferentes arquiteturas RAG, (ii) assegurar a rastreabilidade das evidências utilizadas na resposta e (iii) possibilitar a repetição dos experimentos sob condições equivalentes. A descrição detalhada da interface implementada e de seu uso nos cenários experimentais é apresentada no Capítulo de Resultados e Discussão.

2.2.6 Aspectos éticos da pesquisa

A pesquisa não envolve dados pessoais, sensíveis ou sigilosos. Os documentos analisados são de acesso público, provenientes de bases institucionais oficiais do CBMDF.

Durante a redação, ferramentas de inteligência artificial generativa foram utilizadas de forma auxiliar, restritas à revisão linguística e formatação, sem interferência nas decisões analíticas ou metodológicas. Além destes fins, utilizou-se modelos de inteligência artificial específicos para correção e revisão de código. O uso dessas ferramentas respeitou as normas de citação, as diretrizes de integridade científica e as políticas institucionais do CBMDF.

2.3 Resultados e discussão

2.3.1 Avaliação de Métodos de Recuperação de Informação (S1-S3)

Esta seção apresenta a análise comparativa dos três métodos de recuperação de informação avaliados: BM25 (S1), busca vetorial com FAISS (S2) e abordagem híbrida combinando BM25 e FAISS (S3). Em todos os cenários, a geração das respostas foi conduzida de forma estritamente extrativa, permitindo isolar o impacto das estratégias de recuperação sobre os resultados obtidos.

2.3.1.1 Métricas de Qualidade

A Tabela 8 apresenta os valores médios das métricas de precisão, revocação e similaridade da resposta para os três cenários de recuperação.

Tabela 8 - Métricas de qualidade para os métodos de recuperação (S1 ao S3)

Cenário	Método de recuperação	Precisão do contexto	Revocação do contexto	Similaridade da resposta
S1	BM25	0,685	0,776	0,856
S2	FAISS (vetorial)	0,539	0,648	0,863
S3	Híbrido	0,657	0,755	0,860

Fonte: o autor.

O cenário S1 apresentou os melhores resultados tanto em precisão quanto em revocação do contexto, indicando maior capacidade de recuperar trechos

normativos diretamente relevantes às consultas. Esse desempenho pode ser atribuído à natureza lexical do algoritmo, particularmente adequada ao domínio jurídico-normativo, no qual a correspondência exata de termos técnicos, expressões legais e dispositivos normativos desempenha papel central.

O cenário S2 apresentou valores inferiores de precisão e revocação, sugerindo que a recuperação puramente semântica, embora eficiente do ponto de vista computacional, tende a recuperar trechos semanticamente relacionados que nem sempre correspondem ao dispositivo normativo mais pertinente à consulta. Esse comportamento é consistente com o domínio avaliado, no qual a precisão terminológica é frequentemente mais relevante do que a similaridade conceitual ampla.

O cenário S3 apresentou desempenho intermediário, com valores próximos aos do BM25 em revocação, porém ligeiramente inferiores em precisão. Esse resultado indica que, para o corpus normativo avaliado e sob a estratégia de agregação adotada, a combinação entre recuperação lexical e semântica não produziu ganhos significativos de qualidade em relação ao cenário S1, que emprega o modelo BM25 isoladamente.

2.3.1.2 Similaridade da Resposta

Os valores de similaridade da resposta foram elevados e bastante próximos em todos os cenários, variando entre 0,856 e 0,863. Essa consistência sugere que, quando a geração é estritamente extrativa, a qualidade semântica da resposta final é menos sensível ao método de recuperação empregado, desde que ao menos um trecho normativo relevante seja recuperado.

Esse resultado reforça a interpretação de que, nos cenários extrativos, a etapa de recuperação impacta principalmente a qualidade do contexto, enquanto a resposta final tende a manter elevada similaridade com a resposta-referência sempre que o contexto mínimo adequado está disponível.

2.3.1.3 Desempenho Computacional

A Tabela 9 evidencia os resultados em termos de performance computacional entre os experimentos S1, S2 e S3.

Tabela 9 - Métricas de desempenho computacional (S1 ao S3)

Cenário	Latência média (segundos)	Vazão (questões por segundo)	Memória do índice (MB)
S1	0,021	48,35	45,59
S2	0,009	110,48	17,16
S3	0,038	26,26	62,25

Fonte: o autor.

O cenário S2 (FAISS) apresentou a menor latência e a maior vazão, evidenciando sua eficiência computacional, além de menor consumo de memória. Tais características tornam a busca vetorial particularmente atraente para cenários com alto volume de consultas ou restrições severas de recursos computacionais.

O cenário S1 (BM25) apresentou desempenho intermediário, com maior consumo de memória, mas ainda mantendo latência e vazão compatíveis com aplicações institucionais de consulta normativa.

O cenário S3 (híbrido) apresentou maior latência e menor vazão, resultado esperado devido à execução sequencial de dois métodos de recuperação. Esses resultados indicam que, nas condições experimentais adotadas, a sobrecarga computacional do método híbrido não foi compensada por ganhos proporcionais de qualidade.

2.3.1.4 Discussão dos Métodos de Recuperação

Os resultados evidenciam relações distintas entre qualidade de recuperação e eficiência computacional. O modelo BM25 mostrou-se mais adequado ao domínio normativo avaliado, priorizando precisão e cobertura do contexto. A busca vetorial com FAISS destacou-se pela eficiência computacional, ainda que com menor aderência ao contexto normativo específico. Já a abordagem híbrida, embora conceitualmente promissora, não apresentou vantagem clara no corpus analisado, indicando que estratégias de fusão mais sofisticadas poderiam ser exploradas em trabalhos futuros.

2.3.2 Avaliação dos Modelos de Geração (S4 à S7)

Esta seção compara dois modelos de geração empregados na arquitetura RAG: o Qwen3-0.6B, executado localmente (S4), e o GPT-4o mini, utilizado como modelo externo de referência (S5). Em ambos os cenários, foi adotado o BM25 como método de recuperação, de modo a isolar o impacto do modelo gerador.

2.3.2.1 Métricas de Qualidade

A Tabela 10 descreve o resultado dos ensaios experimentais S4 a S7, destacando as métricas de similaridade de resposta e fidelidade, usadas para avaliar a fase de geração da arquitetura de RAG implementada.

Tabela 10 - Métricas de qualidade para os modelos de geração (S4 a S7)

Cenário	Modelo de geração	Similaridade da resposta	Fidelidade
S4	Qwen3-0.6B (local)	0,893	0,537
S5	GPT-4o mini (externo)	0,896	0,692
S6	Qwen3-0.6B quantizado (Q8) via Llama.cpp (local)	0,892	0,502
S7	Qwen3-1.7B quantizado (Q8) via Llama.cpp (local)	0,904	0,644

Fonte: o autor.

Os resultados indicam que o modelo GPT-4o mini (S5) apresentou o maior valor de fidelidade entre os cenários avaliados, sugerindo maior capacidade de manter aderência ao contexto normativo fornecido durante a geração da resposta. Esse comportamento é compatível com modelos treinados em larga escala e projetados para operar com janelas de contexto amplas.

Entre os modelos executados localmente sem quantização, o Qwen3-0.6B (S4) apresentou desempenho próximo ao do modelo externo em termos de similaridade da resposta, porém com valor inferior de fidelidade. Esse resultado indica que, embora o modelo seja capaz de produzir respostas semanticamente próximas às referências, apresenta maior propensão a extrapolações em relação ao conteúdo normativo recuperado.

Nos cenários que empregaram quantização em Q8, observam-se comportamentos distintos entre os modelos avaliados. O Qwen3-0.6B quantizado (S6)

apresentou redução na similaridade da resposta e, de forma mais acentuada, na fidelidade, indicando limitação na capacidade de manter aderência estrita ao contexto normativo. Esse resultado sugere que, em modelos de pequeno porte, a redução de precisão associada à quantização pode acentuar restrições já impostas pelo número reduzido de parâmetros.

Em contraste, o Qwen3-1.7B quantizado (S7) apresentou o maior valor de similaridade da resposta e fidelidade superior à observada em S4 e S6. Esse desempenho indica que o aumento da capacidade do modelo foi suficiente para preservar a qualidade da geração mesmo com a adoção de quantização, evidenciando a influência do porte do modelo sobre a aderência ao conteúdo recuperado.

De forma geral, os resultados indicam uma relação entre capacidade do modelo, forma de execução (local ou externa) e fidelidade ao contexto normativo. Os experimentos demonstram que o tamanho do modelo exerceu influência mais relevante sobre a qualidade e a aderência das respostas do que o nível de quantização empregado, uma vez que os cenários S6 e S7 utilizaram Q8 e apresentaram desempenhos distintos. Assim, modelos locais de maior porte tendem a manter maior fidelidade ao conteúdo recuperado, enquanto modelos menores demandam maior rigor arquitetural para reduzir o risco de geração desvinculada das evidências.

2.3.2.2 Desempenho Computacional

A Tabela 11 apresenta as métricas de desempenho computacional dos experimentos S4, S5, S6 e S7, considerando a latência média de resposta e a vazão (questões por segundo– qps). Essas métricas são particularmente relevantes nesta etapa, pois os cenários analisados incorporam modelos de geração com diferentes tamanhos, formas de execução (local ou remota) e níveis de otimização, impactando diretamente o desempenho global da arquitetura de RAG.

Tabela 11 - Métricas de desempenho computacional (S4 a S7)

Cenário	Latência média (s)	Vazão (qps)
S4	5,59	0,179

S5	1,79	0,436
S6	3,56	0,281
S7	6,92	0,145

Fonte: o autor.

Nos experimentos conduzidos em infraestrutura local (S4, S6 e S7), observa-se que a latência média tende a crescer conforme o aumento da complexidade do modelo de geração. O cenário S6, que utiliza o modelo Qwen3-0.6B quantizado (Q8) via Llama.cpp, apresentou redução moderada de latência em relação ao S4, indicando que a quantização contribui para ganhos de eficiência computacional. No entanto, essa otimização ocorre com impacto limitado na vazão, que permanece próxima aos valores observados no S4.

Por outro lado, o cenário S7, baseado no modelo Qwen3-1.7B quantizado (Q8), apresentou a maior latência média e a menor vazão entre os experimentos locais. Esse comportamento evidencia o custo computacional associado ao aumento do número de parâmetros, mesmo quando técnicas de quantização são aplicadas, reforçando a compensação entre capacidade de geração e desempenho em ambientes com recursos restritos.

O cenário S5, que utiliza o modelo GPT-4o mini por meio de API externa, apresentou desempenho computacional significativamente superior, com menor latência média e maior vazão. Esse resultado reflete o uso de infraestrutura altamente otimizada e escalável, não diretamente comparável aos cenários executados localmente, mas relevante como referência de desempenho máximo esperado em condições ideais.

De forma geral, os resultados indicam que, em ambientes locais, a quantização permite ganhos pontuais de eficiência, porém não elimina as limitações impostas pelo aumento da complexidade dos modelos. Esses achados reforçam a importância de considerar o equilíbrio entre qualidade da geração e viabilidade computacional na definição de uma arquitetura leve, especialmente no contexto institucional e operacional que orienta esta pesquisa.

2.3.2.3 Discussão dos Modelos de Geração

Os resultados não indicam uma oposição direta entre modelos locais e modelos hospedados em nuvem, mas sim diferentes compromissos entre qualidade, latência e custo computacional. O modelo Qwen3-0.6B apresentou desempenho próximo ao do GPT-4o mini, com diferenças percentuais moderadas nas métricas de qualidade de resposta, compensadas por menor latência e maior previsibilidade operacional.

Já o modelo Qwen3-1.7B superou o GPT-4o mini em algumas métricas específicas de geração de resposta, ao custo de aumento significativo de latência e maior demanda computacional. Esses resultados sugerem que modelos locais podem atingir níveis comparáveis de robustez, desde que dimensionados adequadamente ao contexto de uso.

2.3.3 Síntese dos Resultados Experimentais

Os resultados obtidos nos sete cenários experimentais permitem identificar padrões de desempenho e estabelecer critérios para orientar decisões arquiteturais no contexto de consulta normativa do CBMDF. Esta seção consolida os principais achados e apresenta uma análise comparativa das alternativas avaliadas.

2.3.3.1 Comparação entre métodos de recuperação

A análise dos cenários S1, S2 e S3 evidenciou que o método de recuperação lexical (BM25) apresentou desempenho superior ao método semântico (FAISS) nas métricas de precisão e revocação do contexto, contrariando a expectativa inicial de que abordagens baseadas em *embeddings* ofereceriam vantagens em domínios especializados.

Esse resultado pode ser atribuído às características específicas do corpus normativo do CBMDF, no qual a terminologia técnica, as referências a dispositivos legais e a estrutura formal dos documentos favorecem a correspondência lexical exata. Para ilustrar esse comportamento, considere-se a consulta "Qual portaria traz o regimento interno do CBMDF vigente?", presente no questionário de

referência. Nesse caso, os termos "portaria", "regimento interno" e "CBMDF" aparecem literalmente no documento relevante (Portaria nº 24/2020), favorecendo o método lexical. De forma similar, consultas como "Em qual norma do CBMDF está disposto o concurso de movimentações?" ou "Qual norma regula a realização de Inspeções de Saúde?" dependem de correspondência exata com a nomenclatura oficial dos atos normativos.

Por outro lado, consultas formuladas em linguagem mais descritiva, como "Como deve ser realizada a regulação médica pelo bombeiro militar em atendimentos a vítimas de violência doméstica?", poderiam teoricamente se beneficiar da recuperação semântica. Contudo, mesmo nesses casos, o BM25 demonstrou capacidade de identificar os trechos relevantes, sugerindo que a densidade terminológica do corpus normativo mantém a eficácia da correspondência lexical.

O método híbrido (S3), embora conceitualmente promissor, não produziu ganhos que compensassem o aumento de latência e complexidade. Esse achado sugere que, para o corpus avaliado, a simples agregação de resultados lexicais e semânticos não é suficiente para superar o desempenho do BM25 isolado. Estratégias mais sofisticadas de fusão, como ponderação dinâmica ou reordenação por modelo neural, poderiam ser investigadas em trabalhos futuros.

2.3.3.2 Comparação entre modelos de geração

A análise dos cenários S4 a S7 revelou um espectro de “balanças” entre qualidade, fidelidade e custo computacional. A Tabela 12 sintetiza os principais indicadores para apoiar a comparação:

Tabela 12 - Comparação consolidada dos modelos de geração

Critério	S4 (Qwen3-0.6B)	S5 (GPT-4o mini)	S6 (Qwen3-0.6B Q8)	S7 (Qwen3-1.7B Q8)
Similaridade da resposta	0,893	0,896	0,892	0,904
Fidelidade	0,537	0,692	0,502	0,644
Latência média (s)	5,59	1,79	3,56	6,92
Execução local	Sim	Não	Sim	Sim
Custo operacional	Baixo	Por uso	Baixo	Baixo

Fonte: elaborado pelo autor.

O modelo GPT-4o mini (S5) apresentou a maior fidelidade ao contexto recuperado, indicando menor propensão à geração de informações não fundamentadas. Contudo, sua utilização implica dependência de serviço externo, custos por utilização e potenciais restrições de segurança da informação para dados institucionais.

Entre os modelos locais, o Qwen3-1.7B quantizado (S7) alcançou o melhor equilíbrio entre qualidade e fidelidade, superando inclusive o GPT-4o mini em similaridade da resposta. Esse resultado demonstra que modelos de código aberto, adequadamente dimensionados, podem atingir desempenho competitivo em domínios especializados.

O cenário S6 (Qwen3-0.6B quantizado) apresentou a menor fidelidade entre todos os experimentos (0,502), indicando que a combinação de modelo pequeno com quantização leve compromete significativamente a capacidade de aderência ao contexto normativo. Esse resultado desaconselha a adoção dessa configuração para aplicações que exijam confiabilidade das respostas.

2.3.3.3 Critérios de decisão para o contexto institucional

Com base nos resultados experimentais, foram identificados critérios objetivos para orientar a escolha da configuração mais adequada ao CBMDF:

a) Se a prioridade for simplicidade e controle: a abordagem extrativa com BM25 (S1) oferece o melhor equilíbrio entre qualidade de recuperação, previsibilidade e baixo custo computacional. Nessa configuração, o sistema retorna diretamente os trechos normativos relevantes, sem risco de alucinação ou geração de informações não fundamentadas. Essa abordagem é particularmente indicada para consultas objetivas que demandam a identificação de dispositivos específicos, como prazos, composições de guarnição ou procedimentos operacionais padronizados.

b) Se for desejável síntese em linguagem natural com execução local: o modelo Qwen3-1.7B quantizado (S7) apresentou o melhor desempenho entre as alternativas locais, com fidelidade aceitável (0,644) e similaridade elevada (0,904). A latência de 6,92 segundos por consulta é compatível com uso interativo em contexto administrativo. Essa configuração pode ser indicada para consultas mais complexas,

nas quais a síntese de múltiplas informações em linguagem natural agregue valor à resposta.

c) Se houver disponibilidade orçamentária e autorização para uso de serviços externos: o GPT-4o mini (S5) oferece a maior fidelidade (0,692) e menor latência (1,79s), sendo indicado para cenários em que a qualidade da síntese é prioritária e as restrições de segurança da informação possam ser adequadamente tratadas.

d) Configurações não recomendadas: o cenário S6 (Qwen3-0.6B quantizado) não é recomendado para uso institucional devido à baixa fidelidade, que aumenta o risco de respostas desvinculadas do conteúdo normativo oficial.

2.3.3.4 Limitações e considerações metodológicas

O estudo apresenta limitações relacionadas ao corpus normativo analisado e ao questionário de referência adotado, que representa apenas uma amostra das possíveis consultas. Os experimentos foram conduzidos em infraestrutura local, sem uso de GPU dedicada, o que influenciou as métricas de latência observadas. Ainda assim, os resultados obtidos permitem avaliar o comportamento da arquitetura proposta e indicam caminhos para trabalhos futuros, como validação com usuários finais, integração com sistemas corporativos, automatização da atualização do corpus e avaliação sob uso concorrente.

2.3.4 Validação da arquitetura proposta

Esta seção apresenta a validação da arquitetura proposta por meio da análise das funcionalidades implementadas no protótipo experimental, avaliando sua adequação aos requisitos metodológicos definidos e sua aplicabilidade ao contexto institucional do CBMDF.

2.3.4.1 Verificação dos requisitos metodológicos

Conforme estabelecido na seção 2.2.5, o protótipo foi desenvolvido para atender a três requisitos principais: chaveamento entre configurações, rastreabilidade

das evidências e reprodutibilidade dos experimentos. A Tabela 13 apresenta a verificação do atendimento a cada requisito:

Tabela 13 - Verificação Dos Requisitos Metodológicos Do Protótipo

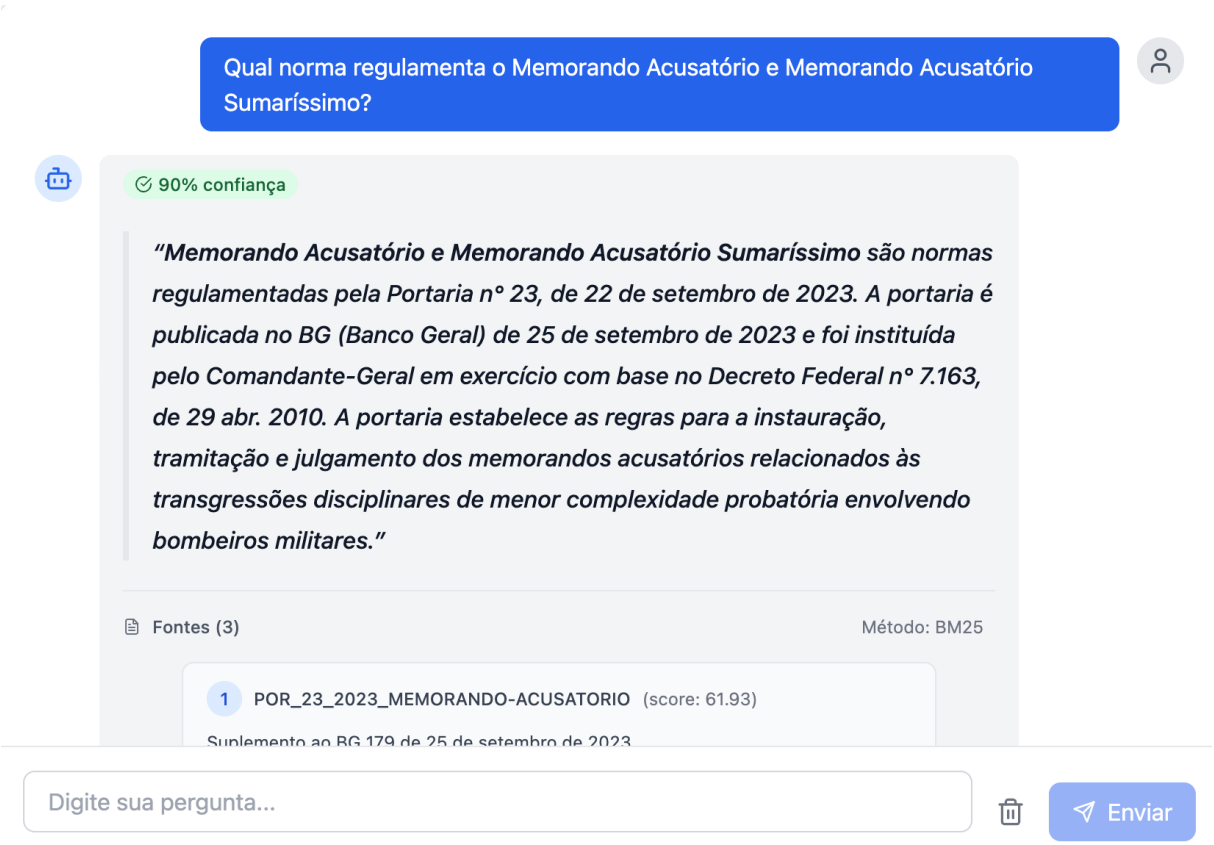
Requisito	Funcionalidade implementada	Status
Chaveamento entre configurações	Seleção de modelo e método de recuperação na interface de consulta	Atendido
Rastreabilidade das evidências	Exibição dos trechos recuperados com indicação de fonte e escore	Atendido
Reprodutibilidade dos experimentos	Registro automático de parâmetros, consultas e respostas	Atendido

Fonte: elaborado pelo autor.

2.3.4.2 Interface de interação por *chat* e inspeção de evidências

A interface de consulta, ilustrada na Figura 3, implementa o padrão de interação conversacional (*chat*), permitindo que o usuário formule perguntas em linguagem natural e receba respostas contextualizadas. Abaixo de cada resposta, o sistema exibe os trechos normativos recuperados, acompanhados da identificação do documento de origem e do escore de relevância atribuído pelo método de recuperação.

Figura 3 – Resposta do sistema a uma pergunta sobre tema institucional



Fonte: o autor.

Durante a execução dos experimentos, observou-se que as respostas nos cenários com geração (S4 a S7) apresentaram estrutura textual coerente e linguagem adequada ao contexto normativo, salvo pequenas imprecisões terminológicas, como a geração de "Banco Geral" em vez de "Boletim Geral (BG)". Esse tipo de "alucinação" pontual pode ser mitigado por meio de prompts mais restritivos ou pela incorporação de glossários institucionais. A exibição dos trechos recuperados mostrou-se útil para identificar esses casos, reforçando a relevância do mecanismo de rastreabilidade.

2.3.4.3 Interfaces de configuração e gerenciamento

Além da interface de consulta, o protótipo inclui funcionalidades administrativas para configuração do sistema e gerenciamento do corpus documental.

A interface de gerenciamento de modelos de linguagem, apresentada no Anexo A, permite cadastrar, editar e selecionar modelos, bem como definir parâmetros

como limite de tokens, temperatura e penalidade de repetição. Essa funcionalidade foi empregada na configuração dos cenários experimentais S4 a S7, assegurando a execução dos modelos sob parâmetros controlados.

A interface de gestão de documentos, apresentada no Anexo B, possibilita a inclusão e manutenção dos documentos que compõem o corpus normativo. O sistema exibe o catálogo dos 1.227 documentos indexados, com informações sobre tipo documental, status de processamento e data de inclusão, permitindo atualização controlada da base, ainda que com intervenção manual.

A Figura 4 ilustra o painel de seleção disponível na interface de chat, que permite ao usuário alternar entre modelos e métodos de recuperação antes de cada consulta. Essa funcionalidade foi essencial para a condução dos experimentos e, em uso institucional, pode ser restrita a perfis administrativos.

Figura 4 – Seção do *chat* de configuração do modelo e método de recuperação de informação.

Fonte: o autor.

2.3.4.4 Limitações e perspectivas de evolução

O protótipo apresenta limitações inerentes a uma prova de conceito. O mecanismo de autenticação não está integrado aos sistemas corporativos do CBMDF, a escalabilidade sob uso concorrente não foi avaliada e a atualização do corpus normativo ainda depende de intervenção manual.

O corpus utilizado é composto apenas por documentos normativos de acesso público, não envolvendo dados pessoais ou informações sensíveis. A ampliação da solução para documentos restritos exigiria integrações adicionais, como autenticação institucional, controle de acesso por perfis e adequação aos requisitos da Lei Geral de Proteção de Dados Pessoais (LGPD).

Como perspectivas de evolução, destacam-se a integração com sistemas corporativos, a automatização da atualização do corpus, a avaliação de desempenho sob carga concorrente, a validação de usabilidade com usuários finais e o aprofundamento de estratégias de recuperação híbrida no cenário S3. Os resultados obtidos indicam que a arquitetura proposta fornece base adequada para essas investigações e para a evolução da solução em ambiente institucional.

3 CONSIDERAÇÕES FINAIS

Este trabalho partiu de uma inquietação prática: É viável construir uma arquitetura de PLN para o CBMDF que, operando sob um determinado arranjo de técnicas, seja mais leve que modelos robustos e apresente desempenho satisfatório para o acesso às normas da Corporação? A resposta a essa pergunta exigiu percorrer um caminho que articulou fundamentos teóricos do Processamento de Linguagem Natural, experimentação técnica com diferentes arquiteturas e a construção de um protótipo funcional capaz de materializar as possibilidades investigadas.

Os resultados obtidos revelaram nuances que merecem reflexão. A superioridade do método de recuperação lexical (BM25) sobre a busca semântica, embora contraintuitiva à primeira vista, converge diretamente com os achados de Pedroso, Ladeira e Faleiros (2019). Ao avaliarem técnicas de recuperação da informação em coleções de documentos jurídicos, os autores concluíram não haver superioridade significativa dos modelos semânticos, elegendo o BM25 como o método

mais adequado devido ao seu excelente equilíbrio entre eficácia e simplicidade computacional.

No domínio normativo, em que a precisão terminológica é essencial e cada palavra carrega peso jurídico, abordagens clássicas mostraram-se não apenas viáveis, mas preferíveis. Esse achado reforça a importância de avaliar as tecnologias à luz do contexto específico de aplicação, resistindo à tentação de adotar soluções mais sofisticadas apenas por serem mais recentes.

Por outro lado, os experimentos com modelos de geração demonstraram que é possível obter respostas sintetizadas de qualidade aceitável utilizando modelos leves executados localmente, sem dependência de serviços externos. O desempenho do modelo Qwen3-1.7B quantizado, que superou o GPT-4o mini em similaridade de resposta, ilustra o potencial dos modelos de código aberto para aplicações institucionais que demandam autonomia tecnológica e controle sobre os dados processados. Essa constatação é particularmente relevante para organizações públicas como o CBMDF, que precisam equilibrar inovação tecnológica com requisitos de segurança da informação e independência de fornecedores externos.

A construção do protótipo representou a aplicação prática dos conceitos teóricos discutidos no trabalho. As interfaces desenvolvidas demonstraram a viabilidade de uma consulta normativa intuitiva, com respostas fundamentadas e rastreáveis, sem exigir conhecimentos técnicos especializados. A exibição dos trechos normativos recuperados permite ao usuário verificar a aderência da resposta aos documentos oficiais.

Este estudo apresenta limitações. O questionário de referência constitui uma amostra do conjunto possível de consultas, e a validação com usuários finais permanece como etapa futura. Além disso, a dinâmica do acervo normativo, marcada por atualizações frequentes, impõe desafios contínuos de manutenção do corpus.

Ainda assim, os resultados indicam que a adoção de arquiteturas leves de Processamento de Linguagem Natural para consulta normativa é tecnicamente viável no contexto do CBMDF. Os experimentos demonstram que soluções dessa natureza podem ser implementadas com recursos computacionais modestos, compatíveis com a realidade de órgãos públicos.

Espera-se que este trabalho contribua para as discussões sobre modernização tecnológica no CBMDF e que a arquitetura proposta sirva de base para investigações futuras voltadas à ampliação da eficiência no acesso à informação normativa.

REFERÊNCIAS

AUTORIDADE NACIONAL DE PROTEÇÃO DE DADOS. **Radar tecnológico n. 3: inteligência artificial generativa**. Brasília, DF: ANPD, 2024.

BAIGORRI, Pablo Federico. **Implementação de tecnologias de gestão do conhecimento para investigação de incêndios no Distrito Federal**. 2020. Trabalho de Conclusão de Curso (Curso de Aperfeiçoamento de Oficiais) - Corpo de Bombeiros Militar do Distrito Federal, Brasília, 2020.

CASELI, H. M.; NUNES, M. G. V. **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 3. ed. *[S.l.]*: BPLN, 2024.

CHOMSKY, Noam. **Syntactic Structures**. 's-Gravenhage: Mouton & Co., 1957.

CORPO DE BOMBEIROS MILITAR DO DISTRITO FEDERAL. Portaria nº 24, de 25 de novembro de 2020. Aprova o Regimento Interno do Corpo de Bombeiros Militar do Distrito Federal, revoga a Portaria n° 6, de 15 de abril de 2020 e dá outras providências. **Suplemento ao Boletim Geral nº 223**, de 1º de dezembro de 2020, Brasília, 2020.

CORPO DE BOMBEIROS MILITAR DO DISTRITO FEDERAL. **Planejamento Estratégico do CBMDF 2025–2030**. Brasília: CBMDF, 2024. Disponível em: https://www.cbm.df.gov.br/portarias-internas-do-cbmdf/portaria-de-13-de-janeiro-de-2025-planejamento-estrategico-do-cbmdf-2025-2030. Acesso em: 12 nov. 2025.

COSTA, José Alfredo Ferreira; DANTAS, Nielsen Castelo Damasceno. Análise comparativa de embeddings jurídicos aplicados a algoritmos de clustering. In: **CONGRESSO BRASILEIRO DE INTELIGÊNCIA COMPUTACIONAL**, XVI., 2023, Salvador, BA. Anais […]. Salvador, BA: SBIC, 2023. p. 1–9. DOI: 10.21528/CBIC2023-181.

GIL, Antonio Carlos. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2008.

HAKDAĞLI, Özlem. **Hybrid Question-Answering System: A FAISS and BM25 Approach for Extracting Information from Technical Document**. Orclever Proceedings of Research and Development, v. 5, n. 1, p. 226–237, 31 dez. 2024.

JURAFSKY, Daniel; MARTIN, James H. **Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition with language models**. 3. ed. Manuscrito online. Stanford University, 2026. Disponível em: https://web.stanford.edu/~jurafsky/slp3. Acesso em: 13 jan. 2026.

LIMA, João Alberto de Oliveira. **Unlocking legal knowledge with multi-layered embedding-based retrieval**. CoRR, 2024. Disponível em: https://doi.org/10.48550/arXiv.2411.07739. Acesso em: 12 nov. 2025.

MANNING, Christopher D.; SCHÜTZE, Hinrich. **Foundations of Statistical Natural Language Processing**. Cambridge: MIT Press, 1999.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Fundamentos de metodologia científica**. 6. ed. São Paulo: Atlas, 2007.

MICROSOFT. **RAG (Geração Aumentada de Recuperação) no Azure Databricks**. Disponível em: <https://learn.microsoft.com/pt-br/azure/databricks/generative-ai/retrieval-augmented-generation>. Acesso em: 19 dez. 2025.

MOREIRA, Viviane P. Recuperação de Informação. *In*: CASELI, H. M.; NUNES, M. G. V (Orgs.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 3. ed. *[S.l.]*: BPLN, 2024.

ONG, Ryan. **O que é o Faiss (Facebook AI Similarity Search)?** Disponível em: <https://www.datacamp.com/pt/blog/faiss-facebook-ai-similarity-search>. Acesso em: 17 dez. 2025.

OPENAI. **GPT-4o mini: inteligência com bom custo-benefício**. 2024**.** Disponível em: <https://openai.com/pt-BR/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>. Acesso em: 17 dez. 2025.

PEDROSO, Daniel; LADEIRA, Marcelo; FALEIROS, Thiago. **Does Semantic Search Performs Better than Lexical Search in the Task of Assisting Legal Opinion Writing?** *In*: IEEE, dez. 2019.

POLO, Felipe Maia *et al.* **LegalNLP - Natural Language Processing methods for the Brazilian Legal Language**. *In*: Sociedade Brasileira de Computação, 29 nov. 2021.

PRODANOV, Cleber Cristiano; FREITAS, Ernani César de. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico.** 2. ed. Novo Hamburgo: Feevale, 2013.

QWEN TEAM. **Qwen3 technical report**. arXiv, 2025. Disponível em: https://arxiv.org/abs/2505.09388. Acesso em: 13 jan. 2026.

RAGAS. **Core Concepts**. 2025. Disponível em: <https://docs.ragas.io/en/latest/concepts>. Acesso em: 23 nov. 2025.

RODRIGUES, Ramon Gouveia; MARCACINI, Ricardo Marcondes**. Aplicação de RAG para identificação de atos judiciais similares na Justiça Eleitoral**. 2024. DOI: 10.11606/003226631.

SAKIYAMA, Kenzo *et al.* PLN no Direito - Ementas: Geração Automática de Ementas Jurídicas. *In*: CASELI, H. M.; NUNES, M. G. V (Orgs.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português.** 3. ed. *[S.l.]*: BPLN, 2024.

SILVA, Danillo Alvin Mendes e. **A possibilidade do uso estratégico da inteligência**

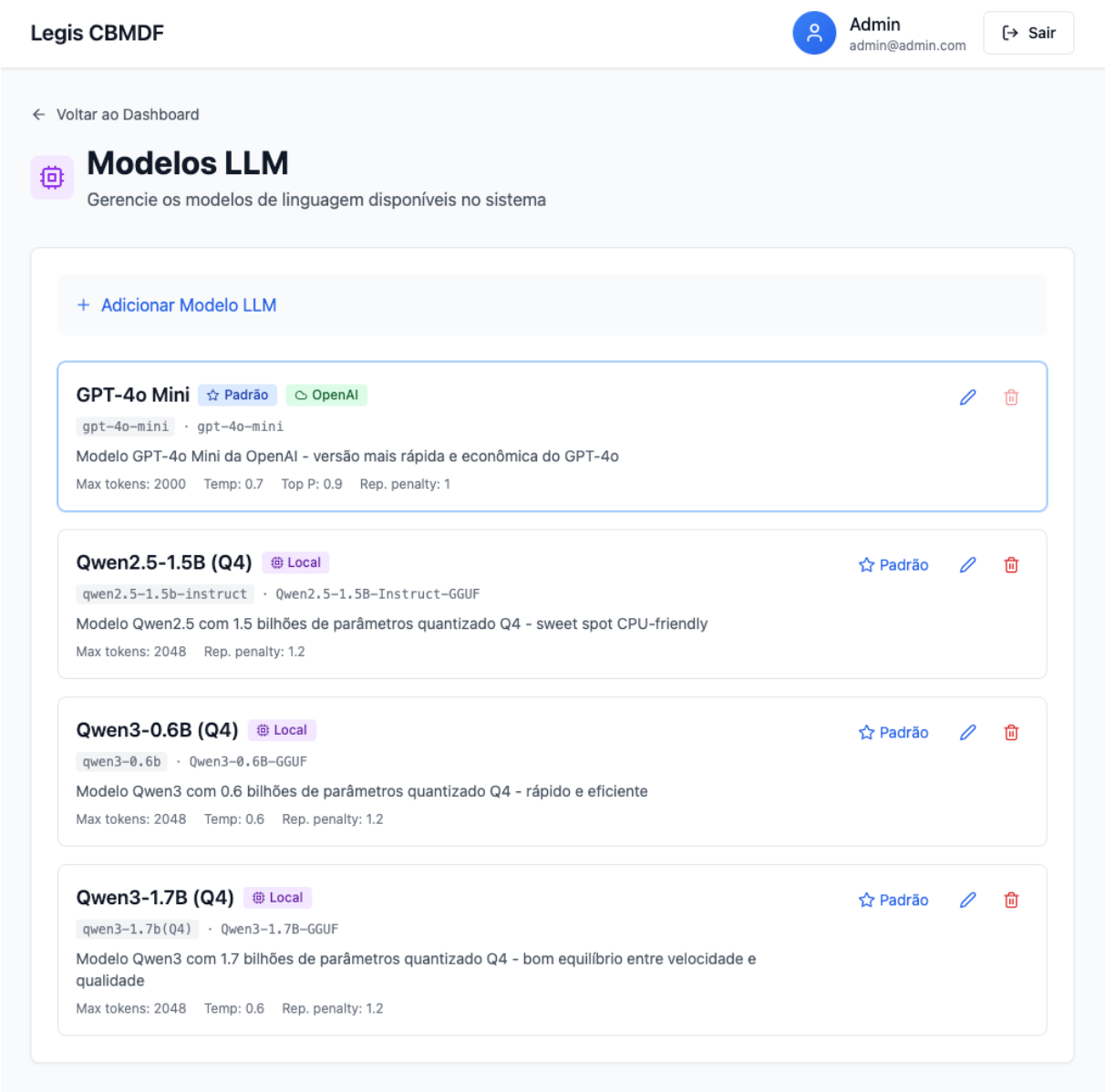
artificial como ferramenta de inovação na segurança contra incêndio e pânico: uma abordagem aplicada ao DESEG/CBMDF. 2025. Trabalho de Conclusão de Curso (Curso de Altos Estudos para Oficiais) – Corpo de Bombeiros Militar do Distrito Federal, Brasília, 2025.

SINHA, Debu. **Mastering Chunking Strategies for RAG: Best Practices & Code Examples**. Disponível em: <https://community.databricks.com/t5/technical-blog/the-ultimate-guide-to-chunking-strategies-for-rag-applications/ba-p/113089>. Acesso em: 18 dez. 2025.

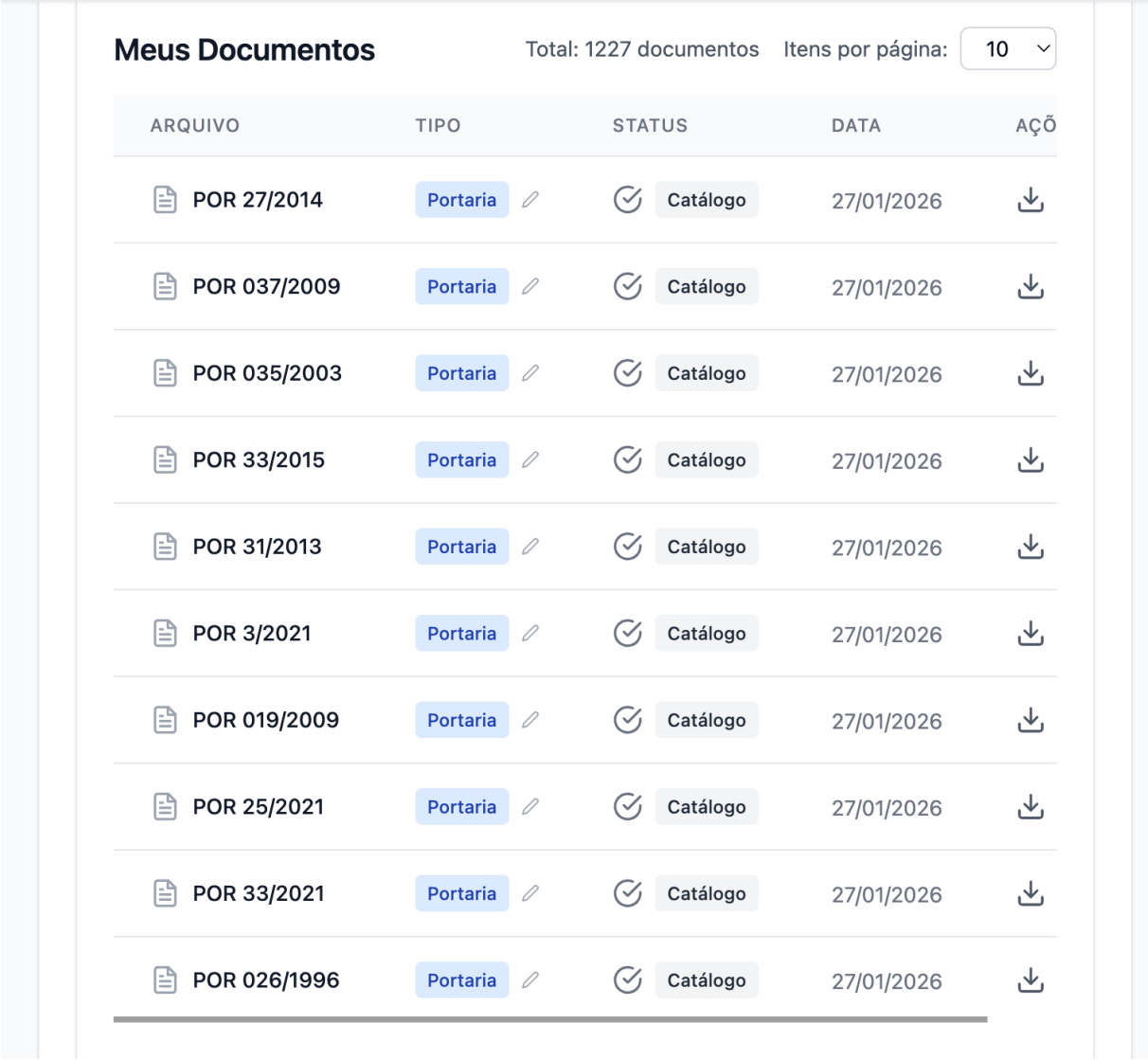
VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N.; KAISER, Lukasz; POLOSUKHIN, Illia. Attention Is All You Need. In: GUYON, Isabelle; VON LUXBURG, Ulrike; BENGIO, Samy; WALLACH, Hanna; FERGUS, Rob; VISHWANATHAN, S. V. N.; GARNETT, Roman (eds.). **Advances in Neural Information Processing Systems**, v. 30. Red Hook, NY: Curran Associates, Inc., 2017. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. Acesso em: 02 out. 2025.

WINASTWAN, Ruben. **Visualizing word embedding with PCA and t-SNE: creating an interactive visualization of word embedding in 2D or 3D**. *Towards Data Science*, [s. l.], 1 out. 2020. Disponível em: https://towardsdatascience.com/. Acesso em: 12 jan. 2026.

ANEXO A – TELA DO PROTÓTIPO PARA GERENCIAMENTO DE MODELOS DE LARGA LINGUAGEM (LLM)



ANEXO B – TELA DO PROTÓTIPO PARA GESTÃO DE DOCUMENTOS



GLOSSÁRIO

ALGORITMO: conjunto de instruções ou regras lógicas que um computador segue para resolver um problema ou executar uma tarefa.

AMBIENTE LOCAL: infraestrutura computacional instalada dentro da própria instituição (no caso, o CBMDF), sem depender de serviços de nuvem externos.

APRENDIZADO DE MÁQUINA (*MACHINE LEARNING*): ramo da Inteligência Artificial em que sistemas aprendem automaticamente a partir de exemplos, ajustando seu comportamento sem necessidade de programação explícita.

ARQUITETURA LEVE: estrutura de sistema projetada para ter baixo consumo de memória e processamento, podendo ser executada em equipamentos comuns, como servidores institucionais.

***CHUNKING* (SEGMENTAÇÃO TEXTUAL):** técnica que divide um texto em partes menores (*chunks*), geralmente com base em sentido ou estrutura, para facilitar o processamento por modelos de linguagem.

CORPUS: conjunto de textos ou documentos usados em um experimento de Processamento de Linguagem Natural. No estudo, corresponde ao conjunto de normas e regulamentos do CBMDF.

***EMBEDDING* (VETORIZAÇÃO SEMÂNTICA):** representação numérica (em forma de vetor) que traduz o significado de palavras, frases ou trechos de texto, permitindo medir semelhanças de sentido entre eles.

FAISS (*FACEBOOK AI SIMILARITY SEARCH*): biblioteca de código aberto desenvolvida pelo Facebook AI Research, usada para buscas rápidas em grandes volumes de vetores, localizando textos de significado semelhante.

***GOLDSET*:** conjunto de referência com pares de perguntas e respostas corretas, criado manualmente para avaliar o desempenho do sistema — funciona como um "gabarito" de validação.

LATÊNCIA: tempo médio que o sistema leva para retornar uma resposta após o recebimento de uma consulta.

LLM (*LARGE LANGUAGE MODEL*): modelo de linguagem de grande porte, treinado com grandes volumes de texto, capaz de compreender e gerar linguagem natural (ex.: *GPT*, *Gemini*, *LLaMA*).

PLN (*PROCESSAMENTO DE LINGUAGEM NATURAL / NATURAL LANGUAGE PROCESSING – NLP*): área da Inteligência Artificial que permite que computadores compreendam, processem e gerem linguagem humana.

PRECISÃO (*PRECISION*): proporção de respostas corretas entre todas as respostas retornadas pelo sistema.

REVOCAÇÃO (*RECALL*): proporção de respostas corretas que o sistema conseguiu encontrar em relação ao total de respostas esperadas no questionário de referência.

***SENTENCE-TRANSFORMERS*:** biblioteca de código aberto que gera *embeddings* de frases ou parágrafos, permitindo medir a similaridade semântica entre textos.

TOKENIZAÇÃO (*TOKENIZATION*): processo de dividir o texto em unidades básicas chamadas *tokens* (palavras, números ou símbolos), que são interpretadas pelos modelos de linguagem.

***TRANSFORMERS*:** arquitetura de redes neurais introduzida por Vaswani *et al*. (2017), que revolucionou o PLN ao permitir o processamento eficiente de longas sequências de texto.

VETORIZAÇÃO: processo de converter textos em vetores numéricos (*embeddings*), permitindo que o sistema avalie o conteúdo e estabeleça comparações de sentido.