Evaluation of RAG Retrieval Strategies with Quantized Qwen3 Models for University Documents

Lucas S. de Oliveira[1], **César Guarienti**[1], **Leandro D. de Jesus** [1], **Anderson C. S. de Oliveira**[1], **Claudia A. Martins**[1]

[1]Instituto de Computação – Universidade Federal de Mato Grosso (UFMT)
78060-900 – Cuiabá – MT – Brazil

lucassteffensoliveria@hotmail.com, {cesar,claudia}@ic.ufmt.br

anderson.oliveira@ufmt.br, prm.leandro86@gmail.com

Abstract. *This study compared five retrieval strategies for local Retrieval-Augmented Generation (RAG) systems using three quantized Qwen3 models, 37 institutional documents, 50 questions, and three repetitions per configuration. Reranking achieved the highest aggregate medians for Context Precision and Answer Correctness. Qwen3 8B with reranking obtained the best individual result. Although naive and hybrid had lower latency, reranking provided the best balance between quality and latency in the evaluated corpus.*

Resumo. *Este estudo comparou cinco estratégias de recuperação para sistemas locais de Geração Aumentada por Recuperação (RAG), utilizando três modelos Qwen3 quantizados, 37 documentos institucionais, 50 perguntas e três repetições por configuração. A estratégia reranking obteve as maiores medianas agregadas de Context Precision e Answer Correctness. O Qwen3 8B com reranking alcançou o melhor resultado individual. Embora naive e hybrid apresentassem menor latência, o reranking proporcionou o melhor equilíbrio entre qualidade e latência no corpus avaliado.*

1. Introdução

A consulta a documentos institucionais, como editais, portarias, resoluções, leis, formulários e instruções normativas, pode exigir a localização de informações frequentemente distribuídas em diferentes fontes oficiais. Em muitos casos, esses documentos não se encontram centralizados, o que dificulta sua identificação, comparação e utilização [Van Opijnen and Santos 2017]. Além disso, parte desse material está sujeita a alterações ao longo do tempo, por meio de retificações, atualizações e revogações [Oliveira et al. 2026, Van Opijnen and Santos 2017].

Em processos administrativos e acadêmicos, essa atividade pode tornar-se particularmente complexa em razão do elevado volume documental, da linguagem normativa empregada e da necessidade de obtenção de respostas consistentes e alinhadas ao conteúdo oficial disponível [Wojtasik et al. 2025]. Nesse contexto, sistemas baseados em modelos de linguagem podem contribuir para facilitar a interação com bases documentais, desde que as respostas geradas estejam efetivamente fundamentadas nos documentos utilizados como fonte de informação [Lewis et al. 2020, Wojtasik et al. 2025, Kulkarni et al. 2026].

A técnica de *Retrieval-Augmented Generation* (RAG) proposta por [Lewis et al. 2020] surge como uma alternativa relevante para esse tipo de aplicação, pois

combina a capacidade linguística dos *Large Language Models* (LLM) com mecanismos de recuperação de informações em bases de conhecimento. Entretanto, o desempenho de um sistema RAG não depende apenas do modelo gerador, mas também da qualidade dos segmentos recuperados, assim, diferentes abordagens de recuperação podem produzir variações significativas nas respostas geradas para os usuários.

Este estudo comparou cinco estratégias de recuperação *naive, hybrid, HyDE, reranking e adaptive* aplicadas a três LLMs da família *Qwen3* locais quantizados. O experimento utiliza um corpus composto por documentos relacionados ao concurso docente de uma universidade federal brasileira e um conjunto de 50 perguntas e respostas definidas pelos autores. A comparação considera métricas de recuperação, qualidade das respostas, eficiência operacional e comportamento de abstenção, buscando identificar como diferentes estratégias de recuperação influenciam o desempenho de sistemas RAG executados em infraestrutura computacional local.

2. Fundamentação Teórica

2.1. Large Language Models (LLMs)

Os *Large Language Models* (LLMs) são modelos treinados em grandes volumes de dados para processar e gerar linguagem natural [Shanahan 2024]. No contexto de perguntas e respostas, sua utilidade está na capacidade de processar perguntas em linguagem natural e produzir respostas textuais para o usuário [Jeong et al. 2024]. Contudo, seu uso isolado pode produzir informações inexatas resultantes de alucinações [Ji et al. 2023]. Por essa razão, em aplicações que exigem maior fidelidade ao conteúdo das fontes, os LLMs podem ser integrados a mecanismos de recuperação de informação, como ocorre em sistemas RAG [Gao et al. 2024, Lewis et al. 2020].

2.2. Componentes e Estratégias de Recuperação em Sistemas RAG

De acordo com [Gao et al. 2024], o RAG pode ser compreendido como uma técnica de inteligência artificial que aperfeiçoa os LLMs ao adicionar um mecanismo de recuperação de trechos relevantes em bases de conhecimento externas aos pesos do modelo. Em sua forma básica, o *Naive RAG* caracteriza-se por uma arquitetura linear composta por três etapas: indexação da base documental, recuperação de informações relevantes e geração da resposta.

Ainda segundo [Gao et al. 2024], cada etapa possui atribuições específicas. A indexação é responsável pela preparação da base documental, abrangendo processos como conversão de arquivos, limpeza, remoção de duplicatas, segmentação, transformação dos trechos em vetores por meio de modelos de *embedding* e seu armazenamento em um banco de dados vetorial. Na etapa de recuperação, a consulta do usuário é submetida ao mesmo modelo de *embedding*, permitindo o cálculo de similaridade entre a consulta e os trechos armazenados. Por fim, na geração, os segmentos recuperados são incorporados ao *prompt* enviado ao LLM, que produz a resposta a ser apresentada ao usuário.

A técnica *Naive RAG* evoluiu dando origem a novas abordagens [Gao et al. 2024]. Dentre elas, destacam-se a combinação de métodos de busca (busca híbrida) [Bruch et al. 2023], a geração de documentos hipotéticos para promover maior alinhamento semântico entre a consulta e a base de conhecimento (*HyDE*) [Gao et al. 2023], a

reclassificação dos trechos recuperados para aumentar a precisão dos resultados (*reranking*) [Nogueira and Cho 2019] e a implementação de mecanismos de roteamento adaptativo, capazes de ajustar dinamicamente a estratégia empregada conforme a complexidade da solicitação do usuário [Jeong et al. 2024].

2.2.1. Estratégias de Recuperação

Na recuperação híbrida, abordagens semânticas e lexicais são combinadas de modo que ambas se complementem e contribuam para a melhoria dos resultados de recuperação. A busca semântica avalia a similaridade de significado entre a consulta do usuário e os documentos armazenados na base de dados, enquanto a busca lexical procura correspondências exatas entre os termos da consulta e os documentos. Os resultados obtidos por essas abordagens são posteriormente combinados e reclassificados por meio de funções de fusão [Bruch et al. 2023].

Proposta por [Gao et al. 2023], a estratégia *Hypothetical Document Embeddings* (HyDE) utiliza a entrada do usuário para gerar previamente um documento hipotético por meio de um LLM. Esse documento busca responder à consulta com base apenas no conhecimento paramétrico do modelo, podendo, portanto, conter alucinações ou informações imprecisas. Ainda assim, espera-se que o texto gerado enriqueça a consulta original com contexto adicional, aproximando-a semanticamente dos documentos presentes na base de dados. Em seguida, o documento hipotético é transformado em uma representação vetorial por meio de um modelo de *embedding*, sendo essa representação utilizada para realizar a etapa de recuperação.

Outra abordagem amplamente empregada é o *reranking*, que consiste em realizar uma busca inicial mais abrangente, recuperando um conjunto maior de segmentos candidatos para, posteriormente, aplicar um método de reordenação sobre os resultados obtidos [Nogueira and Cho 2019]. De acordo com [Gao et al. 2024], essa reordenação pode ser realizada por meio de métricas pré-definidas ou de modelos capazes de avaliar a relevância dos segmentos recuperados em relação à consulta do usuário. Nesse contexto, [Nogueira and Cho 2019] utilizaram o modelo BERT para calcular escores de relevância e reordenar os resultados obtidos na etapa inicial de recuperação.

Por fim, a recuperação adaptativa parte do princípio de que consultas com diferentes níveis de complexidade não devem, necessariamente, ser submetidas a uma única estratégia fixa. Abordagens estáticas podem gerar custos computacionais desnecessários para consultas simples ou desempenho insuficiente para consultas mais complexas. Dessa forma, a consulta do usuário é previamente analisada e classificada, sendo posteriormente direcionada para a estratégia de recuperação mais adequada de acordo com seu nível de complexidade [Jeong et al. 2024].

3. Metodologia

Este estudo avaliou o desempenho de diferentes estratégias de recuperação em sistemas de Recuperação Aumentada por Geração (RAG) utilizando modelos de linguagem locais quantizados. A metodologia foi estruturada em quatro etapas: construção da base de dados, implementação das estratégias de recuperação, execução experimental e avaliação

dos resultados por métricas de recuperação, qualidade da resposta, eficiência operacional e comportamento de abstenção.

3.1. Base de Dados

O corpus utilizado foi composto por 37 documentos relacionados a um concurso docente de uma universidade federal brasileira, incluindo edital, retificações, anexos, instruções normativas, portarias, decretos, leis, resoluções e formulários. Os documentos, disponibilizados nos formatos PDF, DOCX e TXT, passaram por um processo de extração textual e limpeza, com remoção de cabeçalhos repetidos e URLs. Em seguida, os textos foram segmentados em blocos de até 512 tokens com sobreposição (*overlap*) de 64 tokens, buscando reduzir perdas de contexto entre segmentos consecutivos. Os segmentos resultantes foram armazenados no banco vetorial *Qdrant*.

Para a avaliação foi construído um conjunto de 50 perguntas e respectivas respostas esperadas. Destas, 46 possuíam resposta presente no corpus e 4 foram deliberadamente formuladas fora do escopo documental, permitindo avaliar a capacidade do sistema de se abster de responder quando a informação não estivesse disponível.

3.2. Estratégias de Recuperação

Foram avaliadas cinco estratégias de recuperação: *naive*, *hybrid*, *HyDE*, *reranking* e *adaptive*. Em todas elas, a busca inicial recuperou até 20 segmentos candidatos do banco vetorial. Posteriormente, uma política de seleção reduziu esse conjunto para no máximo cinco segmentos enviados ao modelo de linguagem, preservando a ordem de relevância e limitando a seleção a dois segmentos por documento, com o objetivo de aumentar a diversidade das fontes recuperadas.

Como linha de base do experimento, a estratégia *naive* utiliza exclusivamente busca vetorial densa. A consulta do usuário é convertida em um *embedding* pelo modelo BGE-M3 e utilizada diretamente para recuperação dos segmentos mais relevantes no *Qdrant*. Os resultados obtidos por essa abordagem servem como referência para comparação com as demais estratégias.

Buscando combinar diferentes formas de recuperação, a estratégia *hybrid* integra busca semântica e busca lexical. O modelo BGE-M3 gera representações densas e esparsas da consulta, permitindo a execução simultânea dos dois mecanismos de busca. Os resultados recuperados são posteriormente combinados pela técnica *Reciprocal Rank Fusion* (RRF), utilizando $k = 60$, e submetidos à política padrão de seleção dos contextos. O RRF foi escolhido por sua simplicidade e eficiência na combinação de rankings oriundos de diferentes mecanismos de recuperação [Cormack et al. 2009].

Diferentemente das abordagens anteriores, a estratégia *HyDE* realiza uma expansão da consulta antes da etapa de recuperação. Inicialmente, o modelo de linguagem gera um documento hipotético relacionado à pergunta do usuário por meio de um prompt orientado para textos administrativos genéricos, evitando a criação de datas ou valores inexistentes. Esse documento é então transformado em vetor pelo modelo BGE-M3 e utilizado como consulta para recuperação dos segmentos relevantes no banco vetorial.

Após a recuperação inicial, a estratégia *reranking* acrescenta uma etapa de reordenação dos resultados. Os 20 segmentos candidatos obtidos na busca vetorial são

avaliados pelo modelo *bge-reranker-v2-m3*, que atribui escores de relevância para cada par consulta-segmento. Com base nesses escores, os resultados são reclassificados antes da seleção final dos contextos enviados ao modelo de linguagem.

Por fim, a estratégia *adaptive* emprega um mecanismo de roteamento baseado em LLM para selecionar dinamicamente a estratégia de recuperação mais adequada para cada consulta. O roteador classifica a pergunta e escolhe uma entre as estratégias *naive*, *hybrid*, *HyDE* ou *reranking*. Perguntas simples tendem a seguir a rota *naive*. Consultas contendo datas, valores ou referências específicas são direcionadas para *hybrid*. Perguntas que exigem síntese ou envolvem múltiplos critérios tendem a utilizar *reranking* e consultas curtas ou ambíguas são encaminhadas para *HyDE*. Caso o roteamento não produza uma saída válida, a estratégia *naive* é utilizada como alternativa padrão. Como a decisão é realizada pelo próprio modelo de linguagem, a latência observada nessa estratégia inclui também o custo computacional da etapa de roteamento.

3.3. Experimento

O experimento avaliou a combinação das cinco estratégias de recuperação com três modelos de linguagem da família Qwen3 (*Qwen3 1.7B*, *Qwen3 4B-Instruct-2507* e *Qwen3 8B*), todos quantizados no formato *Q4_K_M*. Cada configuração foi executada em três repetições utilizando o mesmo conjunto de 50 perguntas.

Os modelos foram disponibilizados por meio do *Ollama* e configurados com parâmetros fixos (*temperature*=0.0, *seed*=42 e *top_k*=1), reduzindo a variabilidade estocástica da geração e favorecendo a comparabilidade dos resultados.

Todas as estratégias utilizaram o modelo de *embedding* BGE-M3, com vetores de 1.024 dimensões. Este modelo foi escolhido por oferecer suporte multilíngue e permitir simultaneamente busca lexical e semântica [Chen et al. 2024]. Na estratégia de reordenamento foi utilizado o modelo multilíngue *bge-reranker-v2-m3* como *cross-encoder*.

Os experimentos foram realizados em um computador equipado com processador Intel i5-10400, 32 GB de memória RAM e GPU Nvidia RTX 3060 com 12 GB de VRAM. O framework *Ollama* foi utilizado para execução dos modelos de linguagem, enquanto o *Qdrant* foi empregado como banco de dados vetorial para armazenamento e recuperação dos segmentos documentais.

Antes da execução das avaliações, todos os modelos foram previamente carregados em memória e executados uma vez. Esse procedimento teve como objetivo evitar que o custo de inicialização influenciasse as medições de desempenho.

3.4. Métricas

A avaliação foi conduzida considerando quatro dimensões complementares: recuperação da informação, qualidade da resposta, eficiência operacional e comportamento de abstenção.

Para avaliar a recuperação da informação foram utilizadas as métricas *Recall@5*, *nDCG@5*, *Context Precision* e *Context Recall*. O *Recall@5* mede a proporção de documentos relevantes recuperados entre os cinco primeiros resultados, enquanto o *nDCG@5* avalia a qualidade do ranqueamento considerando a posição dos documentos relevantes. As métricas *Context Precision* e *Context Recall*, implementadas no framework RAGAS

[Es et al. 2024, RAGAS 2026], medem respectivamente a relevância dos segmentos recuperados e a capacidade do sistema em recuperar todos os documentos necessários para responder à consulta.

A qualidade das respostas foi mensurada por meio das métricas *Faithfulness*, *Answer Relevancy* e *Answer Correctness*, também disponibilizadas pelo framework RAGAS. Essas métricas seguem a abordagem *LLM-as-Judge*, utilizando o modelo *Gemini 2.5 Flash Lite* como avaliador semântico e o modelo *gemini-embedding-001* para as operações de similaridade exigidas pelo framework.

Como medida de eficiência operacional foi considerada a latência total do sistema, definida como o tempo transcorrido entre o envio da pergunta e a geração da resposta final.

A capacidade de abstenção foi analisada utilizando as quatro perguntas sem resposta presente no corpus. Considerou-se uma abstenção correta quando o sistema retornava exatamente a mensagem padronizada: "Informação não encontrada no contexto fornecido.".

Para a análise estatística dos resultados, adotou-se a mediana como medida principal de tendência central, reduzindo a influência de valores extremos. As taxas de abstenção foram apresentadas como proporções de ocorrência.

3.5. Disponibilidade de dados e código

Os códigos-fonte, documentos do corpus, conjunto de perguntas do benchmark, respostas de referência, *prompts*, configurações experimentais e resultados brutos e consolidados estão disponíveis publicamente no *repositório*.

4. Resultados

O experimento produziu 2.250 respostas, distribuídas entre 15 combinações de modelos e estratégias de recuperação. Para assegurar comparabilidade entre as configurações, as métricas de qualidade da resposta, recuperação e latência foram calculadas sobre as 46 perguntas com documento de referência, totalizando 138 observações por configuração após as três repetições. As quatro perguntas sem documento de referência foram reservadas exclusivamente para a análise de abstenção.

4.1. Qualidade e recuperação

A Tabela 1 apresenta as medianas de *Answer Correctness* para cada combinação de modelo e estratégia de recuperação. Observa-se que a estratégia *reranking* obteve os melhores resultados nos três modelos avaliados, alcançando medianas de $0{,}569$, $0{,}591$ e $0{,}721$ para os modelos Qwen3 1.7B, 4B e 8B, respectivamente. O melhor desempenho do experimento foi registrado na combinação entre Qwen3 8B e *reranking*. Em contraste, a estratégia *HyDE* apresentou os menores valores de correção na maior parte das configurações, enquanto *hybrid* e *adaptive* ocuparam posições intermediárias.

Ao analisar conjuntamente o efeito do modelo e da estratégia de recuperação, verifica-se que os ganhos obtidos pela melhoria da recuperação foram comparáveis ou até superiores aos decorrentes do aumento do número de parâmetros do modelo. Como exemplo, o Qwen3 4B utilizando *reranking* ($0{,}591$) superou o desempenho do Qwen3 8B com as estratégias *naive* ($0{,}553$) e *adaptive* ($0{,}563$). Além disso, o ganho observado entre

naive e *reranking* no modelo Qwen3 8B foi de aproximadamente 30%. Esses resultados reforçam a importância da etapa de recuperação em sistemas RAG, indicando que a seleção de contextos relevantes pode exercer impacto tão significativo quanto a utilização de modelos geradores maiores.

Tabela 1. Mediana de *Answer Correctness* por modelo e estratégia de recuperação.

Estratégia	Qwen3 1.7B	Qwen3 4B	Qwen3 8B
Naive	0,533	0,471	0,553
Hybrid	0,540	0,534	0,597
Reranking	**0,569**	**0,591**	**0,721**
HyDE	0,362	0,445	0,400
Adaptive	0,523	0,506	0,563

A Tabela 2 apresenta as medianas agregadas das métricas de recuperação e qualidade da resposta para cada estratégia avaliada. Observa-se que a estratégia *reranking* obteve os melhores resultados em *Context Precision* (1,000) e *Answer Correctness* (0,598), superando as demais abordagens. Esse resultado sugere que a etapa de reordenação dos segmentos recuperados contribuiu para aumentar a relevância dos contextos fornecidos ao modelo gerador, refletindo em respostas mais corretas. A estratégia *hybrid* apresentou desempenho intermediário, com ganhos em relação à abordagem *naive*, enquanto o *HyDE* registrou os menores valores em *Context Precision*, *Answer Relevancy*, *Answer Correctness* e *nDCG@5*, indicando menor efetividade na recuperação de informações relevantes para o corpus analisado.

Também se observa que as métricas *Faithfulness*, *Context Recall* e *Recall@5* atingiram mediana igual a 1,000 para todas as estratégias. Esse comportamento indica que, na maior parte das consultas, os documentos relevantes foram recuperados e utilizados na geração das respostas, reduzindo o poder discriminatório dessas métricas nesse experimento. Dessa forma, as diferenças entre as estratégias tornaram-se mais evidentes em métricas relacionadas à qualidade do ranqueamento e da resposta final, particularmente *Context Precision* e *Answer Correctness*. Em conjunto, os resultados sugerem que o principal benefício do *reranking* não está em recuperar uma quantidade maior de documentos relevantes, mas em selecionar e ordenar de forma mais eficaz os contextos que serão utilizados pelo modelo de linguagem.

Tabela 2. Medianas agregadas das métricas por estratégia.

Métrica	*Naive*	*Hybrid*	*Reranking*	*HyDE*	*Adaptive*
Faithfulness	1,000	1,000	1,000	1,000	1,000
Context Precision	0,694	0,700	**1,000**	0,333	0,700
Context Recall	1,000	1,000	1,000	1,000	1,000
Answer Relevancy	0,837	0,876	0,876	0,831	0,874
Answer Correctness	0,515	0,547	**0,598**	0,400	0,542
Recall@5	1,000	1,000	1,000	1,000	1,000
nDCG@5	0,631	0,631	0,631	0,500	0,631

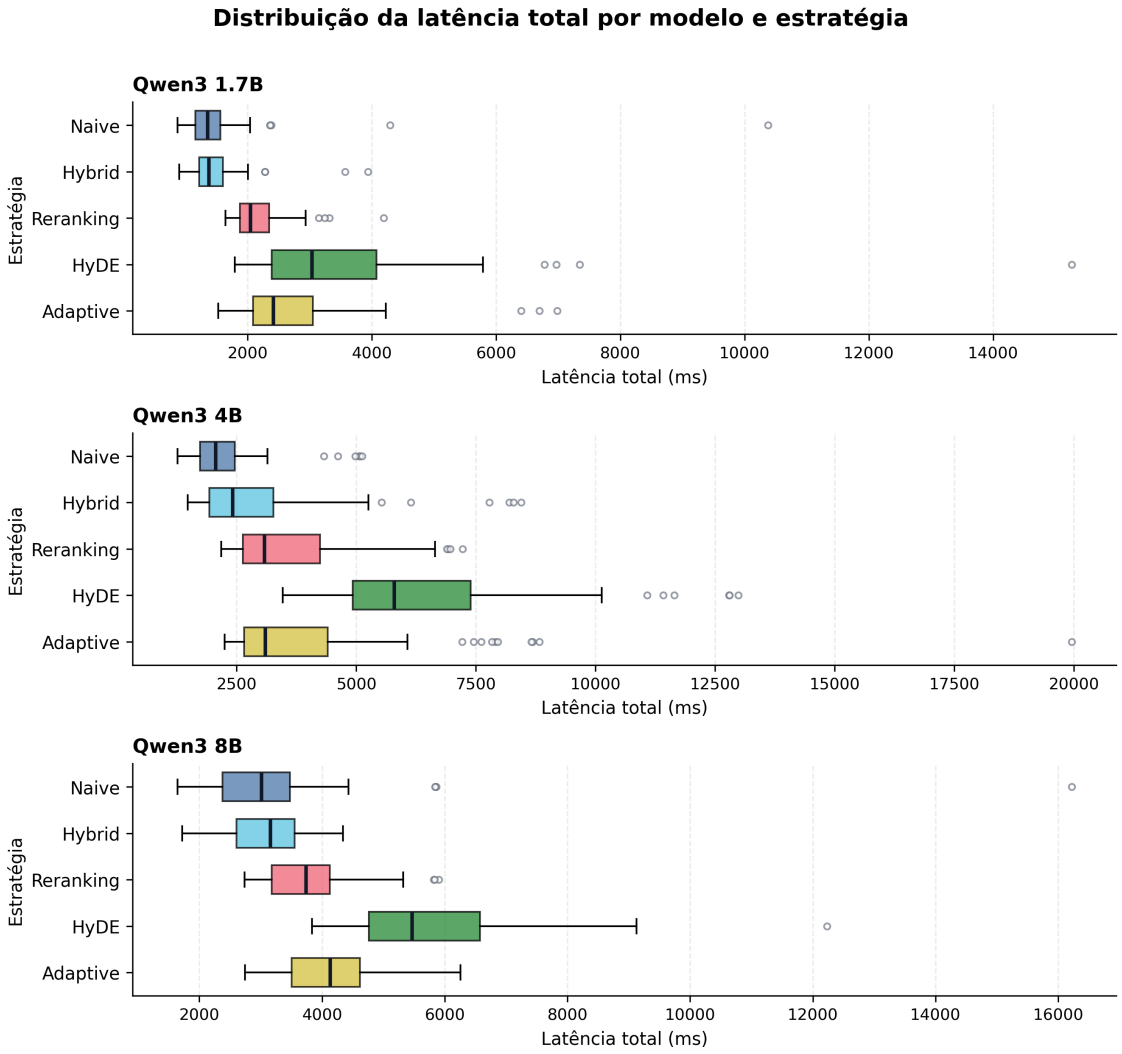


Figura 1. Distribuição da latência total por modelo e estratégia de recuperação.

4.2. Latência, trade-off e abstenção

A Figura 1 apresenta a distribuição da latência total para cada combinação de modelo e estratégia de recuperação. Observa-se que as estratégias *naive* e *hybrid* apresentaram, de forma consistente, os menores tempos de resposta nos três modelos avaliados. O aumento do porte do modelo resultou em um crescimento gradual da latência para todas as estratégias, porém mantendo a mesma ordem relativa de desempenho. A estratégia *reranking* ocupou uma posição intermediária, apresentando latência superior à das abordagens *naive* e *hybrid*, mas inferior à observada nas estratégias *HyDE* e *adaptive*. Esse comportamento era esperado devido ao custo adicional da etapa de reordenação dos segmentos recuperados.

A estratégia *HyDE* apresentou as maiores medianas de latência e também a maior dispersão dos tempos de resposta, evidenciada pela amplitude das caixas e pela presença de diversos valores extremos. Esse resultado decorre da necessidade de uma etapa adicional de geração para construção do documento hipotético antes da recuperação dos contextos. Em contraste, as estratégias *naive* e *hybrid* apresentaram distribuições mais con-

centradas, indicando maior previsibilidade operacional. Embora o *reranking* imponha um custo computacional adicional, sua distribuição permaneceu relativamente estável quando comparada ao *HyDE*, sugerindo um aumento moderado de latência em troca dos ganhos observados nas métricas de qualidade da resposta. Dessa forma, os resultados indicam que o principal trade-off do experimento ocorreu entre a maior qualidade proporcionada pelo *reranking* e a menor latência observada nas estratégias *naive* e *hybrid*.

A Tabela 3 apresenta o comportamento dos modelos diante de perguntas sem resposta disponível no corpus e de perguntas respondíveis. Observa-se que os modelos Qwen3 4B e Qwen3 8B apresentaram maior capacidade de recusa correta para consultas fora do escopo, abstendo-se adequadamente em 46 e 45 dos 60 casos avaliados, respectivamente. Em contraste, o modelo Qwen3 1.7B recusou corretamente apenas 15 das 60 perguntas sem resposta no corpus, indicando maior propensão a gerar respostas mesmo quando a informação necessária não estava disponível.

Tabela 3. Abstenção por modelo.

Modelo	Recusa correta fora do escopo	Recusa indevida em perguntas respondíveis
Qwen3 1.7B	15 / 60	6 / 690
Qwen3 4B	46 / 60	114 / 690
Qwen3 8B	45 / 60	75 / 690

Por outro lado, essa maior capacidade de abstenção foi acompanhada por um aumento das recusas indevidas em perguntas respondíveis. O Qwen3 1.7B recusou apenas 6 das 690 perguntas com resposta disponível, enquanto os modelos Qwen3 4B e Qwen3 8B recusaram 114 e 75 perguntas, respectivamente. Esse resultado sugere que os modelos maiores adotaram uma postura mais conservadora, reduzindo respostas para consultas fora do escopo, mas aumentando o risco de negar respostas que poderiam ser respondidas a partir do contexto recuperado.

Dessa forma, observa-se um trade-off entre capacidade de abstenção e cobertura das respostas. Embora esses resultados contribuam para caracterizar o comportamento dos modelos em cenários de ausência de informação, o número reduzido de perguntas fora do escopo não permite generalização do comportamento de abstenção.

5. Conclusão

Este estudo comparou cinco estratégias de recuperação em sistemas RAG locais aplicados a documentos institucionais. Entre as configurações avaliadas, o *reranking* apresentou os melhores resultados nas métricas com maior poder discriminatório, alcançando medianas agregadas de $1{,}000$ em *Context Precision* e de $0{,}598$ em *Answer Correctness*. O melhor desempenho individual foi obtido pelo Qwen3 8B com *reranking*, com mediana de $0{,}721$ em *Answer Correctness*. Como *Faithfulness*, *Context Recall* e *Recall@5* atingiram medianas máximas em todas as estratégias, os resultados indicam que as diferenças se concentraram na seleção e na ordenação dos contextos, e não na recuperação de uma quantidade maior de documentos relevantes.

As estratégias *naive* e *hybrid* apresentaram as menores latências, enquanto o *reranking* ocupou posição intermediária e proporcionou maior correção das respostas. Assim, para o corpus analisado, o *reranking* apresentou o melhor compromisso descritivo

entre qualidade e latência. A análise de abstenção, por sua vez, mostrou que os modelos Qwen3 4B e Qwen3 8B recusaram corretamente mais perguntas fora do escopo, mas também produziram mais recusas indevidas, evidenciando a necessidade de calibrar cobertura e conservadorismo.

6. Declaração de Uso de IA Generativa

Em conformidade com o art. 9º da Portaria CNPq nº 2.664, de 6 de março de 2026, declara-se o uso do Codex, da OpenAI, como ferramenta de Inteligência Artificial Generativa durante a elaboração deste artigo. A ferramenta foi utilizada para auxiliar na geração e revisão de código, bem como na revisão linguística e estrutural do texto.

Referências

Bruch, S., Gai, S., and Ingber, A. (2023). An analysis of fusion functions for hybrid retrieval. *ACM Trans. Inf. Syst.*, 42(1).

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. (2024). M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Cormack, G. V., Clarke, C. L. A., and Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Es, S., James, J., Espinosa Anke, L., and Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. In Aletras, N. and De Clercq, O., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Gao, L., Ma, X., Lin, J., and Callan, J. (2023). Precise zero-shot dense retrieval without relevance labels. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. (2024). Retrieval-augmented generation for large language models: A survey.

Jeong, S., Baek, J., Cho, S., Hwang, S. J., and Park, J. (2024). Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7036–7050, Mexico City, Mexico. Association for Computational Linguistics.

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

Kulkarni, A., Jhaveri, P. R., Shrestha, P., Han, Y. T., Amini, R., and Madahian, B. (2026). Long-context long-form question answering for legal domain. In Matusevych, Y., Eryiğit, G., and Aletras, N., editors, *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 728–751, Rabat, Morocco. Association for Computational Linguistics.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Nogueira, R. and Cho, K. (2019). Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

Oliveira, V. T. d., Silva, D. O. d., Souza, M. d. A., Lima, M. R., Oliveira, S. S. T. d., and Rosa, T. C. (2026). Retrieval-augmented generation and knowledge graphs in Portuguese-language legal documents. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 1–10, Salvador, Brazil. Association for Computational Linguistics.

RAGAS (2026). List of available metrics. Available in: `https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/`. Accessed: 2026-05-20.

Shanahan, M. (2024). Talking about large language models. *Commun. ACM*, 67(2):68–79.

Van Opijnen, M. and Santos, C. (2017). On the concept of relevance in legal information retrieval. *Artif. Intell. Law*, 25(1):65–87.

Wojtasik, K., Berdowski, A., Okulska, I., and Piasecki, M. (2025). Polichat: Retrieval augmented generation on university documents and regulations. In Paszynski, M., Barnard, A. S., and Zhang, Y. J., editors, *Computational Science – ICCS 2025 Workshops*, pages 273–288, Cham. Springer Nature Switzerland.