DESENVOLVIMENTO DE UM CHATBOT FINANCEIRO COM RECUPERAÇÃO AUMENTADA POR GERAÇÃO (RAG)[1]

Pedro Vitor Melo Bitencourt[2]
Daniel Hasan Dalip[3]
Danilo Boechat Seufitelli[4]

Submetido em: 10/06/2026 | Aprovado em: 18/06/2026

RESUMO

O setor financeiro depende de informações precisas e verificáveis extraídas de documentos corporativos complexos. Embora os Modelos de Linguagem de Grande Porte (LLMs) apresentem capacidades avançadas de linguagem natural, sua aplicação direta em finanças ainda enfrenta desafios relacionados a alucinações, imprecisões factuais e dificuldades em lidar com informações específicas e frequentemente atualizadas. Este trabalho avalia experimentalmente diferentes estratégias de recuperação em sistemas de Recuperação Aumentada por Geração (*Retrieval-Augmented Generation* — RAG) aplicados à análise financeira corporativa. O *pipeline* experimental utiliza documentos financeiros públicos de empresas brasileiras e compara *Similarity Search*, *Maximum Marginal Relevance* (MMR), *MultiQuery Retrieval* e *Parent Document Retrieval*. Para a avaliação, foi construído um *dataset* especializado de perguntas e respostas financeiras, avaliado com métricas do RAGAS, incluindo *Faithfulness*, *Context Precision*, *Context Recall* e *Answer Relevancy*. Os resultados indicam que *MultiQuery Retrieval* apresenta melhor desempenho em tarefas que exigem maior cobertura contextual, enquanto *Parent Document Retrieval* é mais eficaz em cenários que demandam maior precisão contextual e fundamentação factual.

Palavras-chave: Análise Financeira Corporativa. Recuperação Aumentada por Geração. Modelos de Linguagem de Grande Porte. Estratégias de Recuperação. RAGAS.

ABSTRACT

[1] Artigo científico elaborado para o trabalho de conclusão do curso de Engenharia de Computação do campus Nova Gameleira do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG.)

[2] Graduando em Engenharia de Computação no campus Nova Gameleira do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG.). pedromelobitencourt@gmail.com

[3] Orientador do trabalho. Docente do Departamento de Computação do campus Nova Gameleira do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG.). hasan@cefetmg.br

[4] Coorientador do trabalho. Docente do Departamento de Computação do campus Nova Gameleira do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG.). danilimster@gmail.com

The financial sector relies on accurate and verifiable information extracted from complex corporate documents. Although Large Language Models (LLMs) provide advanced natural language capabilities, their direct application in finance remains challenging due to hallucinations, factual inaccuracies, and difficulties in handling domain-specific and frequently updated information. This work experimentally evaluates different retrieval strategies in Retrieval-Augmented Generation (RAG) systems applied to corporate financial analysis. The experimental pipeline uses public financial documents from Brazilian companies and compares Similarity Search, Maximum Marginal Relevance (MMR), MultiQuery Retrieval, and Parent Document Retrieval. A domain-specific dataset of financial questions and answers was built and evaluated using RAGAS metrics, including *Faithfulness*, *Context Precision*, *Context Recall*, and *Answer Relevancy*. The results indicate that MultiQuery Retrieval performs better in tasks requiring broader contextual coverage, while Parent Document Retrieval is more effective in scenarios requiring higher contextual precision and factual grounding.

Keywords: Corporate Financial Analysis. Retrieval-Augmented Generation. Large Language Models. Retrieval Strategies. RAGAS.

1 INTRODUÇÃO

O uso de Grandes Modelos de Linguagem (*Large Language Models* — LLMs) para tarefas financeiras tem atraído considerável interesse da comunidade acadêmica e da indústria financeira por conta de sua capacidade de compreender e gerar linguagem natural de forma avançada. No entanto, utilizar LLMs no setor financeiro requer lidar com alguns desafios, tais como o risco de alucinações; a produção de conteúdo impreciso ou sem suporte factual; e a dificuldade de acessar informações atualizadas, específicas ou presentes apenas em documentos corporativos extensos. Além disso, LLMs genéricos podem falhar por falta de contexto, pela presença de terminologias especializadas e pela necessidade de interpretar informações contábeis, operacionais e estratégicas dispersas em relatórios, demonstrações financeiras e comunicados ao mercado.(Zhang *et al.*, 2023).

Uma estratégia emergente para reduzir esses problemas é a Recuperação Aumentada por Geração (do inglês *Retrieval-Augmented Generation*, RAG). Ou seja, o modelo de linguagem é complementado por um mecanismo de busca responsável por recuperar informações externas relevantes antes da geração da resposta. Ao fundamentar as respostas do LLM em documentos factuais recuperados de uma base de conhecimento, reduz-se o risco de respostas alucinadas e aumenta-se a possibilidade de produzir respostas contextualmente relevantes e verificáveis em fontes oficiais (Lewis *et al.*, 2020). Essa característica é particularmente relevante no setor financeiro, pois respostas incorretas sobre indicadores, resultados, riscos ou estratégias corporativas podem comprometer a interpretação dos dados e a tomada de decisão (Zhao *et al.*, 2024).

Neste contexto, este trabalho investiga o uso de arquiteturas RAG aplicadas à análise financeira corporativa, com foco na tarefa de perguntas e respostas sobre documentos públicos de empresas brasileiras. Esses documentos são disponibilizados principalmente por empresas listadas em bolsa de valores, por meio de páginas de Relações com Investidores (RI), em conformidade com exigências regulatórias e de transparência do mercado financeiro. Os documentos considerados incluem demonstrações financeiras, relatórios anuais, formulários de referência, *releases* de resultados, relatórios operacionais, relatórios de produção e vendas e comunicados corporativos. A base documental contempla empresas

de diferentes setores econômicos, incluindo instituições financeiras, mineração, petróleo e varejo, tais como Itaú, Vale, Petrobras e Magazine Luiza. A escolha dessas empresas busca aumentar a diversidade dos documentos analisados, incluindo diferentes terminologias, indicadores financeiros, estruturas documentais e contextos operacionais presentes no mercado corporativo brasileiro.

Embora sistemas RAG sejam promissores para aplicações financeiras, seu desempenho depende fortemente da qualidade da etapa de recuperação de informação. Em um *pipeline* RAG, a resposta gerada pelo modelo de linguagem é condicionada aos contextos recuperados (Gao *et al.*, 2024). Ou seja, mesmo quando o modelo gerador possui elevada capacidade linguística, a resposta final pode se tornar imprecisa, incompleta ou pouco fundamentada caso os documentos recuperados sejam incompletos, irrelevantes ou semanticamente inadequados. Dessa forma, a escolha da estratégia de recuperação é um componente central para a qualidade final do sistema.

O objetivo deste trabalho é avaliar experimentalmente como diferentes estratégias de recuperação em sistemas RAG afetam a qualidade do contexto recuperado e das respostas em análises financeiras corporativas. Serão comparadas estratégias representativas, como a busca por similaridade semântica (*Similarity Search*), *Maximum Marginal Relevance* (MMR), *MultiQuery Retrieval* e *Parent Document Retrieval*. Tais estratégias foram escolhidas por oferecerem abordagens distintas para recuperação de contexto (similaridade direta, diversificação, reformulação de consulta e preservação de contexto documental). Especificamente, o trabalho responde às seguintes perguntas de pesquisa (RQs):

> ***RQ1*** *A escolha da estratégia de recuperação influencia o desempenho de um sistema RAG no domínio financeiro corporativo?*
>
> ***RQ2*** *O desempenho de cada estratégia depende do perfil da pergunta (contexto amplo vs. precisão factual)?*

O framework RAGAS (Retrieval-Augmented Generation Assessment) (Es *et al.*, 2024) é utilizado no processo de avaliação. Ele mede simultaneamente a qualidade da recuperação de contexto e das respostas geradas. Em síntese, este trabalho apresenta uma análise empírica de estratégias RAG aplicadas à análise financeira corporativa e destaca como diferentes métodos de recuperação afetam a qualidade dos contextos recuperados e das respostas produzidas.

O restante deste trabalho está organizado da seguinte forma. A Seção 2 discute a fundamentação teórica do problema e o panorama atual da literatura. A Seção 3 mostra o processo metodológico utilizado para a construção da base documental e execução dos experimentos. Os resultados obtidos são apresentados na Seção 4, e as conclusões, limitações e possibilidades de trabalhos futuros são apresentadas na Seção 5.

2 FUNDAMENTAÇÃO TEÓRICA E TRABALHOS RELACIONADOS

Esta seção discute os principais conceitos teóricos relacionados ao desenvolvimento de sistemas de perguntas e respostas baseados em *Retrieval-Augmented Generation* aplicados ao domínio financeiro. Inicialmente, discutem-se os desafios e aplicações de Grandes Modelos de Linguagem (LLMs) em finanças na Seção 2.1. Em seguida, são apresentados os fundamentos da arquitetura RAG na Seção 2.2. Posteriormente, são discutidas as técnicas de fragmentação documental (*chunking*) na Seção 2.3, os conceitos de embeddings

e representação semântica na Seção 2.4, as estratégias de recuperação de informação na Seção 2.5 e os métodos de avaliação automática utilizando o framework RAGAS na Seção 2.6. Por fim, são apresentados os trabalhos relacionados na Seção 2.7. Também são discutidos trabalhos relacionados e abordagens recentes utilizadas em sistemas RAG especializados.

2.1 Grandes Modelos de Linguagem no Setor Financeiro

O setor financeiro se destaca devido ao alto volume de informações complexas, especializadas e frequentemente atualizadas. Relatórios corporativos, demonstrações financeiras, *releases* de resultados, formulários regulatórios e documentos de governança representam apenas parte do conjunto de informações utilizadas por analistas, investidores e instituições financeiras para tomada de decisão (Zhao *et al.*, 2024). Tradicionalmente, analisar esses documentos requer um considerável esforço manual e conhecimento especializado. Com o avanço dos Grandes Modelos de Linguagem, surgiram novas possibilidades de automatização de tarefas relacionadas à interpretação de documentos financeiros (Setty *et al.*, 2024), extração de informações (Gao *et al.*, 2024), análise textual (Zhang *et al.*, 2023) e sistemas de perguntas e respostas (Setty *et al.*, 2024; Gao *et al.*, 2024).

Modelos recentes como GPT,[5] Gemini[6] e DeepSeek,[7] demonstram elevada capacidade de compreensão de linguagem natural, raciocínio contextual e geração textual (Zhao *et al.*, 2024). Essa capacidade abre espaço para algumas aplicações, como análise automatizada de relatórios financeiros, sistemas de suporte a investidores, sumarização documental, análise de sentimento financeiro e *chatbots* especializados em finanças (Zhao *et al.*, 2024; Zhang *et al.*, 2023).

Entretanto, LLMs genéricos enfrentam limitações para o domínio financeiro. Um dos principais problemas é a geração de informações incorretas ou não fundamentadas, fenômeno conhecido como *hallucination*. Em aplicações financeiras, respostas incorretas podem causar interpretações equivocadas sobre resultados corporativos, riscos, estratégias empresariais ou indicadores financeiros. (Zhao *et al.*, 2024) Além disso, modelos treinados de forma genérica têm dificuldades quanto à terminologia especializada do setor. Documentos corporativos possuem tabelas, siglas, indicadores contábeis, informações regulatórias e linguagem técnica pouco presentes nos *corpora* generalistas utilizados no treinamento de LLMs (Yepes *et al.*, 2024).

Outro desafio envolve a atualização temporal das informações. Modelos generativos tradicionais possuem conhecimento limitado ao período utilizado durante o treinamento. Isso representa uma limitação crítica em finanças, onde resultados trimestrais, indicadores e eventos corporativos mudam constantemente (Gao *et al.*, 2024). Pesquisas recentes têm demonstrado que o fornecimento de contexto externo especializado pode reduzir significativamente esses problemas. Em vez de depender exclusivamente do conhecimento interno do modelo, abordagens baseadas em recuperação de informação permitem que respostas sejam fundamentadas em documentos reais e atualizados (Lewis *et al.*, 2020). Nesse contexto, sistemas *Retrieval-Augmented Generation* surgem como uma alternativa promissora para integrar recuperação semântica de documentos financeiros com geração textual baseada em LLMs.

[5] Chat GPT: <https://chatgpt.com/>
[6] Gemini: <https://gemini.google.com/>
[7] DeepSeek: <https://www.deepseek.com/>

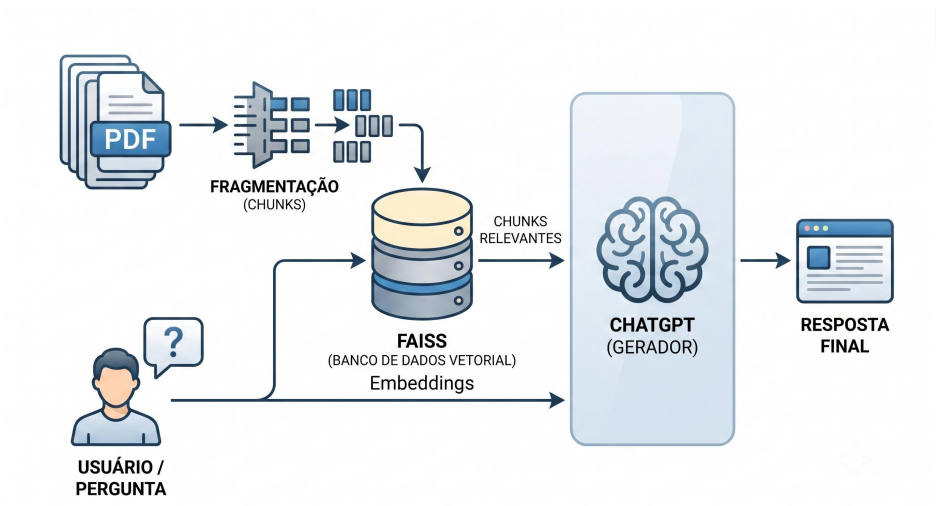
2.2 Recuperação Aumentada por Geração (RAG)

RAG é uma arquitetura que combina mecanismos de recuperação de documentos com modelos generativos de linguagem. A proposta é recuperar informações relevantes a partir de uma base documental externa e utilizá-las como contexto para gerar respostas pelo modelo de linguagem (Lewis *et al.*, 2020). No pipeline tradicional de RAG, ao receber uma pergunta do usuário, o sistema inicia uma etapa de recuperação semântica utilizando embeddings vetoriais. Em seguida, os documentos (ou trechos) recuperados são inseridos no prompt enviado ao LLM, que gera uma resposta fundamentada nas informações recuperadas (Gao *et al.*, 2024). Essa abordagem apresenta vantagens importantes em relação ao uso isolado de modelos generativos, tais como: (i) reduzir alucinações; (ii) utilizar informações atualizadas; (iii) rastrear a origem das respostas; (iv) possuir maior confiabilidade factual; e (v) adaptar a domínios especializados.

De forma geral, o pipeline RAG envolve etapas de carregamento documental, fragmentação textual, geração de embeddings, indexação vetorial, recuperação semântica, construção de contexto e geração da resposta pelo modelo de linguagem. No domínio financeiro, esses sistemas são particularmente relevantes porque permitem que modelos generativos utilizem documentos externos como fonte de evidência durante a geração das respostas. Essa fonte externa é relevante em cenários que exigem precisão factual, rastreabilidade e atualização constante das informações, como relatórios corporativos, comunicados ao mercado e demonstrações financeiras.

A Figura 1 mostra o fluxo de processamento adotado neste trabalho para responder perguntas financeiras a partir de documentos corporativos. Os documentos brutos são transformados em vetores, armazenados em um repositório vetorial e consultados por similaridade; os resultados recuperados são então combinados com a consulta do usuário para formar um contexto que alimenta o LLM, o qual gera a resposta final.

Figura 1 – Arquitetura do Sistema de Geração Aumentada por Recuperação.

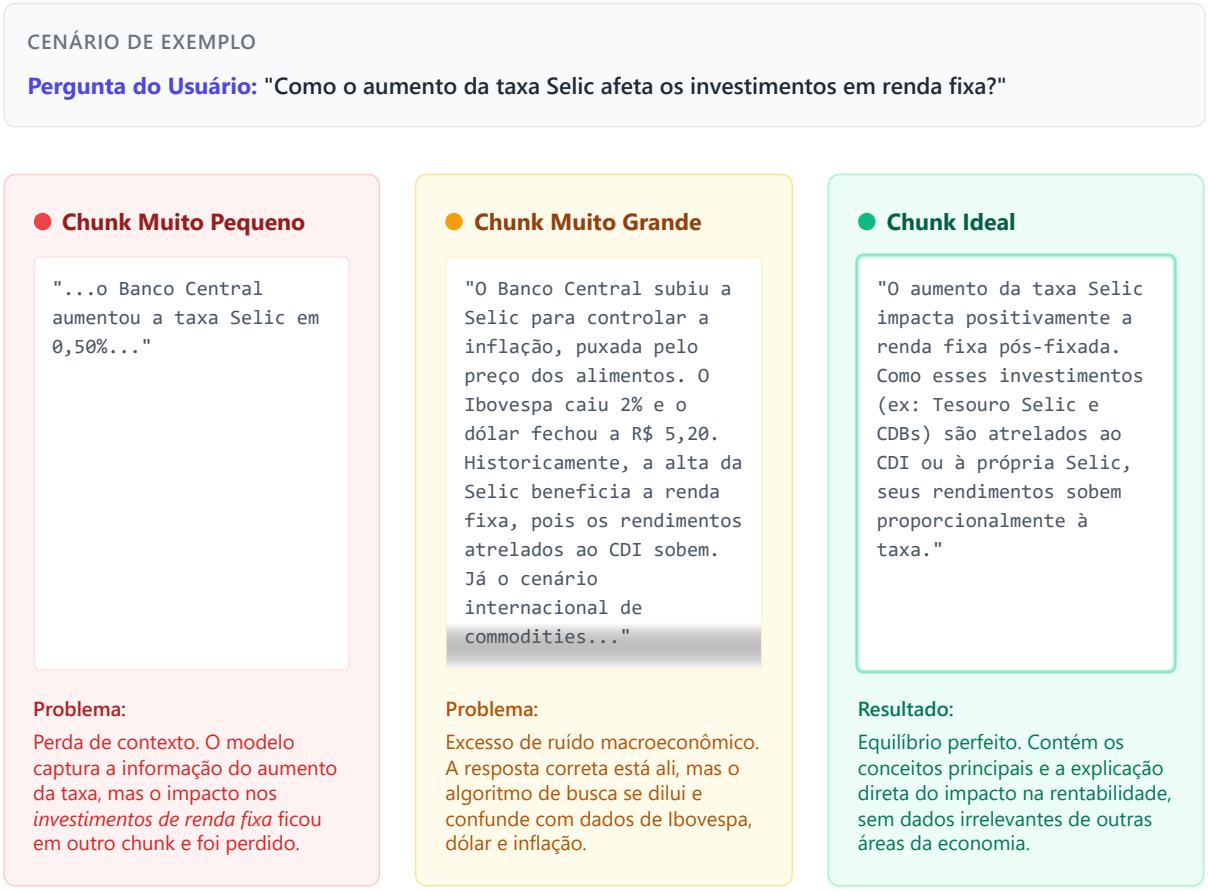


Fonte: Elaborado pelo próprio autor (2026).

2.3 Fragmentação de Documentos (*Chunking*) para Recuperação

Documentos financeiros frequentemente possuem muitas páginas e podem exceder o limite de contexto dos modelos de linguagem. Por esse motivo, é necessário dividi-los em partes menores, chamadas *chunks*, que podem ser indexadas e recuperadas individualmente durante o processo de busca semântica (Gao *et al.*, 2024; Yepes *et al.*, 2024). A estratégia de fragmentação influencia diretamente a recuperação de contexto e a qualidade das respostas geradas em sistemas RAG. *Chunks* muito pequenos podem remover informações importantes para a interpretação da pergunta, enquanto *chunks* muito grandes podem introduzir ruído e reduzir a precisão da recuperação (Yepes *et al.*, 2024). A Figura 2 ilustra este dilema conceitual em um cenário aplicado ao mercado financeiro.

Figura 2 – Ilustração do impacto do tamanho do *chunk* na recuperação de informações. Um *chunk* subdimensionado perde o contexto relacional exigido pela pergunta, enquanto um *chunk* superdimensionado introduz ruído macroeconômico, diluindo a informação relevante e prejudicando a precisão do modelo RAG.



Fonte: Elaborado pelo próprio autor (2026).

Neste trabalho, a fragmentação textual é utilizada como etapa de pré-processamento para viabilizar a indexação e recuperação dos documentos financeiros. A análise experimental principal foca nas estratégias de recuperação aplicadas aos documentos fragmentados, e não em uma comparação sistemática entre diferentes configurações de *chunking*. Essa delimitação permite isolar melhor o efeito de estratégias como *Similarity Search*, MMR, *MultiQuery Retrieval* e *Parent Document Retrieval* sobre a qualidade dos contextos recuperados e das respostas geradas.

2.4 Embeddings e Representação Semântica

Embeddings são representações vetoriais contínuas utilizadas para representar semanticamente textos em espaços multidimensionais. Em sistemas RAG, eles desempenham papel central na etapa de recuperação, pois permitem comparar perguntas e documentos por similaridade semântica, possibilitando a seleção de trechos potencialmente relevantes para compor o contexto enviado ao modelo gerador (Gao *et al.*, 2024). No domínio financeiro, a qualidade dessa representação é particularmente importante devido à presença de terminologia especializada, siglas, indicadores contábeis, informações temporais e linguagem corporativa específica (Setty *et al.*, 2024).

Neste trabalho, embeddings são utilizados para representar perguntas e documentos financeiros durante a etapa de recuperação. Embora diferentes configurações de embeddings possam influenciar o desempenho do pipeline, a análise experimental principal concentra-se na comparação entre estratégias de recuperação e modelos de linguagem.

Assim, os embeddings desempenham o papel de mecanismo de representação semântica subjacente ao processo de recuperação, enquanto as análises experimentais buscam investigar como diferentes estratégias de busca e diferentes modelos geradores afetam a qualidade dos contextos recuperados e das respostas produzidas.

2.5 Estratégias de Recuperação de Informação

A etapa de recuperação de contexto é um dos componentes centrais em sistemas RAG, pois a qualidade dos documentos recuperados influencia diretamente a infalibilidade da resposta final gerada pelo modelo. Dessa forma, quando os contextos recuperados são incompletos, irrelevantes ou pouco relacionados à pergunta, a resposta produzida pode se tornar imprecisa, incompleta ou pouco fundamentada (Gao *et al.*, 2024; Setty *et al.*, 2024; Barnett *et al.*, 2024).

Nesse sentido, diferentes estratégias de recuperação apresentam comportamentos distintos quanto à precisão, diversidade contextual e cobertura semântica. Neste trabalho, são investigadas as seguintes abordagens de recuperação: busca por similaridade semântica (*Similarity Search*), *Maximum Marginal Relevance* (MMR), *MultiQuery Retrieval* e *Parent Document Retrieval*.

Essas estratégias foram selecionadas por representarem diferentes formas de lidar com o problema de recuperação de contexto em sistemas RAG. Enquanto a busca por similaridade simples prioriza os trechos mais próximos semanticamente da consulta, MMR busca introduzir diversidade entre os documentos recuperados. Adicionalmente, Multi-Query Retrieval expande a pergunta original em múltiplas reformulações, aumentando a cobertura semântica da busca. Em contrapartida, Parent Document Retrieval procura preservar contexto mais amplo ao recuperar documentos maiores associados aos fragmentos inicialmente encontrados. Essas diferenças tornam possível investigar como distintas estratégias afetam métricas tais quais precisão contextual, cobertura da recuperação, relevância da resposta e fidelidade documental (Gao *et al.*, 2024; Setty *et al.*, 2024).

Similarity Search Representa uma das abordagens mais tradicionais em sistemas RAG. Nessa estratégia, os embeddings vetoriais da pergunta são comparados com os embeddings dos documentos indexados com medidas de similaridade, tal como a similaridade do cosseno. Os documentos semanticamente mais próximos da consulta são então recuperados e utilizados como contexto para o modelo de linguagem. Por sua simplicidade,

essa abordagem é frequentemente utilizada como *baseline* experimental para comparação com métodos mais avançados (Gao *et al.*, 2024; Setty *et al.*, 2024).

Maximum Marginal Relevance o MMR busca equilibrar relevância e diversidade entre os documentos recuperados. Diferentemente da busca por similaridade simples, que seleciona apenas os trechos mais próximos da pergunta, o MMR também considera a redundância entre os documentos selecionados, reduzindo a recuperação de trechos muito semelhantes (Carbonell; Goldstein, 1998). Essa característica pode ser útil em documentos financeiros extensos, nos quais informações similares podem aparecer em diferentes seções, embora o aumento da diversidade também possa introduzir contextos menos diretamente relacionados à pergunta.

MultiQuery Retrieval Utiliza um modelo de linguagem para gerar múltiplas reformulações da pergunta original. Cada reformulação é utilizada como uma consulta independente durante a recuperação vetorial, e os resultados são posteriormente agregados. Essa estratégia busca reduzir limitações causadas por ambiguidades linguísticas ou diferenças de terminologia entre a pergunta do usuário e os documentos indexados (Gao *et al.*, 2024). No contexto financeiro, essa abordagem pode ser útil devido à variedade de termos utilizados para representar indicadores, operações, riscos e resultados corporativos (Setty *et al.*, 2024).

Parent Document Retrieval Utiliza uma abordagem hierárquica de recuperação. Inicialmente, pequenos *chunks* semanticamente relevantes são identificados durante a busca vetorial; em seguida, documentos (ou trechos) maiores associados a esses *chunks* são retornados como contexto final. Essa estratégia visa preservar maior integridade contextual e reduzir problemas causados pela fragmentação excessiva dos documentos (Gao *et al.*, 2024). Em relatórios financeiros extensos, essa preservação de contexto pode ser importante, pois informações relevantes podem estar distribuídas entre tabelas, notas explicativas e seções correlacionadas (Yepes *et al.*, 2024).

2.5.1 Síntese das Estratégias de Recuperação

As estratégias apresentadas diferem principalmente na forma como equilibram relevância, diversidade contextual, cobertura semântica e preservação de contexto. Enquanto Similarity Search prioriza simplicidade e proximidade semântica direta entre consulta e documentos, MMR busca reduzir redundâncias entre os trechos recuperados. MultiQuery Retrieval amplia a cobertura da busca por meio de reformulações automáticas da consulta, ao passo que Parent Document Retrieval procura preservar maior integridade contextual ao recuperar trechos documentais mais amplos.

Essas diferenças tornam as estratégias adequadas para cenários distintos de recuperação. Em aplicações financeiras, por exemplo, perguntas que exigem integração de múltiplas evidências podem se beneficiar de maior cobertura semântica, enquanto perguntas dependentes de contexto documental mais amplo podem exigir maior preservação contextual. Dessa forma, a escolha da estratégia de recuperação pode influenciar diretamente a qualidade do contexto recuperado e, consequentemente, a qualidade das respostas geradas pelo sistema RAG.

2.6 Avaliação Automática com o Framework RAGAS

A avaliação de sistemas RAG envolve múltiplas dimensões, incluindo a qualidade da recuperação de contexto, a relevância da resposta e a fidelidade factual em relação

aos documentos recuperados. Diferentemente de tarefas tradicionais de geração textual, avaliar sistemas RAG requer analisar simultaneamente se os documentos corretos foram recuperados e se a resposta gerada está adequadamente fundamentada nesses documentos (Gao *et al.*, 2024; Es *et al.*, 2024).

Métricas tradicionais utilizadas em tarefas de tradução automática ou sumarização, como BLEU e ROUGE, baseiam-se principalmente na sobreposição textual entre uma resposta gerada e uma referência esperada. Embora sejam úteis em determinados cenários, essas métricas não capturam adequadamente aspectos centrais de sistemas RAG, como a relevância dos contextos recuperados, a fundamentação factual da resposta e a presença de informações não suportadas pelos documentos (Es *et al.*, 2024). Nesse contexto, frameworks especializados como o RAGAS (*Retrieval-Augmented Generation Assessment*) foram propostos para avaliação automática de sistemas RAG.

Neste trabalho, utiliza-se o framework RAGAS para a avaliação quantitativa dos experimentos realizados. As principais métricas consideradas são:

- **Faithfulness**: verifica se a resposta está fundamentada nos documentos recuperados;
- **Context Precision**: avalia a relevância dos documentos recuperados;
- **Context Recall**: mede se os contextos necessários foram efetivamente recuperados;
- **Answer Relevancy**: analisa se a resposta responde adequadamente à pergunta.

Esses indicadores auxiliam a compreender os trade-offs entre diferentes estratégias de recuperação. Por exemplo, valores mais altos de *Context Recall* podem indicar maior cobertura contextual, mas não necessariamente implicam maior *Faithfulness*, caso o contexto recuperado também introduza informações irrelevantes. Dessa forma, a avaliação considera tanto o desempenho quantitativo das configurações avaliadas quanto a interpretação qualitativa do comportamento das estratégias em diferentes tipos de perguntas financeiras.

2.7 Trabalhos Relacionados

A literatura sobre *Retrieval-Augmented Generation* (RAG) tem evoluído de aplicações generalistas para sistemas especializados em domínios que exigem precisão factual, rastreabilidade documental e atualização frequente das informações. Lewis et al. (Lewis *et al.*, 2020) apresentam uma das formulações iniciais de RAG, combinando modelos generativos com mecanismos de recuperação documental para tarefas intensivas em conhecimento. Posteriormente, Gao et al. (Gao *et al.*, 2024) sistematizam avanços recentes na área, incluindo estratégias de recuperação, geração, avaliação e otimização de pipelines RAG.

No domínio financeiro, o uso de LLMs apresenta desafios específicos, como terminologia contábil, presença de tabelas, múltiplos períodos fiscais, indicadores semelhantes e alto custo associado a respostas factualmente incorretas. Zhang et al. (Zhang *et al.*, 2023) investigam o uso de modelos aumentados por recuperação em tarefas financeiras, enquanto Setty et al. (Setty *et al.*, 2024) analisam estratégias para melhorar a recuperação em modelos de perguntas e respostas sobre documentos financeiros. De forma complementar, Yepes et al. (Yepes *et al.*, 2024) discutem a fragmentação de relatórios financeiros para

RAG, destacando que a forma como documentos são segmentados influencia a recuperação de informações relevantes.

A etapa de recuperação é um dos componentes mais determinantes em sistemas RAG, pois define quais evidências serão fornecidas ao modelo gerador. Estratégias simples, como busca por similaridade vetorial, são frequentemente utilizadas como *baseline*, enquanto técnicas como *Maximum Marginal Relevance* (MMR) (Carbonell; Goldstein, 1998) buscam equilibrar relevância e diversidade entre os documentos recuperados. Outras abordagens, como *MultiQuery Retrieval*, exploram reformulações automáticas da consulta para ampliar a cobertura semântica.

Além da recuperação, a avaliação de sistemas RAG exige considerar simultaneamente a qualidade dos contextos recuperados e a qualidade da resposta gerada. Es et al. (Es *et al.*, 2024) propõem o framework RAGAS, utilizado para avaliar dimensões como *Faithfulness*, *Context Precision*, *Context Recall* e *Answer Relevancy*. Roychowdhury et al. (Roychowdhury *et al.*, 2024) investigam métricas RAGAS em um domínio especializado, enquanto Barnett et al. (Barnett *et al.*, 2024) identificam falhas recorrentes em sistemas RAG, como recuperação de contexto irrelevante, uso inadequado dos trechos recuperados e respostas não suficientemente fundamentadas.

Diferentemente dos trabalhos anteriores, este estudo concentra-se na avaliação experimental de estratégias de recuperação aplicadas a documentos financeiros corporativos de empresas brasileiras. A análise não se limita ao desempenho médio geral das configurações avaliadas, mas também investiga o comportamento das estratégias por tipo de pergunta financeira, incluindo perguntas numéricas, explicativas, comparativas, estratégicas e multi-contextuais. Dessa forma, o trabalho busca contribuir para a compreensão dos trade-offs entre cobertura contextual, precisão da recuperação e fidelidade documental em sistemas RAG aplicados à análise financeira corporativa.

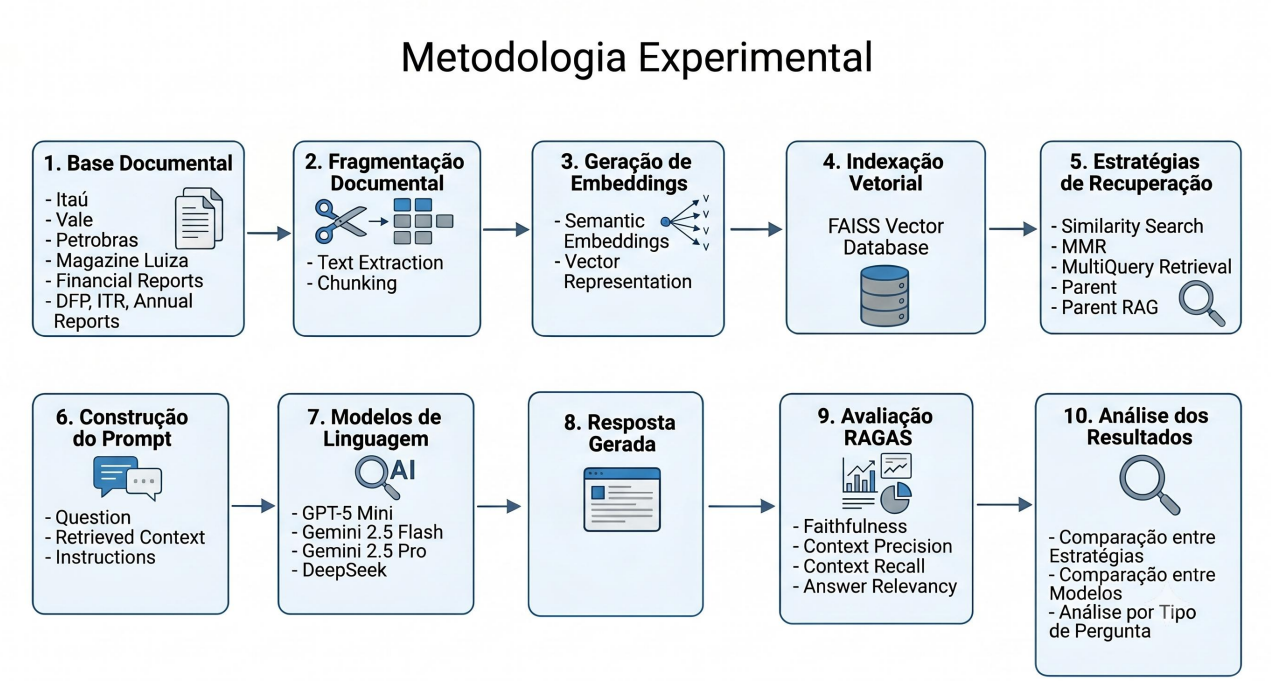
3 METODOLOGIA

Esta seção apresenta a metodologia utilizada para avaliar estratégias de recuperação em sistemas RAG aplicados à análise financeira corporativa. A metodologia está organizada em três partes principais. Primeiro, descrevemos a arquitetura geral do sistema e a base documental utilizada, incluindo a forma como os documentos foram obtidos e representados vetorialmente (Seções 3.1 e 3.2). Em seguida, detalhamos as etapas técnicas do pipeline: fragmentação documental, indexação vetorial, estratégias de recuperação avaliadas e construção dos prompts (Seções 3.3 a 3.5). Por fim, apresentamos a configuração experimental completa, incluindo a construção do dataset de avaliação, a execução dos experimentos e o processo de avaliação automatizada com o framework RAGAS (Seção 3.7).

3.1 Arquitetura do Sistema

A partir de uma pergunta em linguagem natural, o sistema recupera trechos relevantes em documentos financeiros, constrói um contexto a partir das informações recuperadas e utiliza um modelo de linguagem para gerar a resposta final. A Figura 1, apresentada na Seção 2, demonstra a arquitetura geral do pipeline experimental utilizado neste trabalho. O fluxo envolve etapas de preparação documental, representação vetorial dos documentos, recuperação semântica de contexto e geração de respostas fundamentadas em evidências documentais.

Figura 3 – Visão geral da metodologia experimental utilizada no trabalho.



Fonte: Elaborado pelo próprio autor (2026).

Inicialmente, os documentos financeiros são coletados e preparados para indexação. Em seguida, os conteúdos são fragmentados, representados por embeddings e armazenados em um índice vetorial. Durante a execução, a pergunta do usuário é utilizada para recuperar os fragmentos mais relevantes, que são incorporados ao prompt enviado ao modelo de linguagem. A resposta final é então gerada com base nos contextos recuperados.

Além disso, para garantir reprodutibilidade e eficiência computacional, o sistema utiliza mecanismos de persistência e cache para embeddings, respostas geradas e métricas calculadas. Essa infraestrutura reduz chamadas redundantes aos modelos e acelera a execução dos experimentos sem interferir nas comparações realizadas entre estratégias de recuperação e modelos de linguagem.

3.2 Base Documental e Representação Vetorial

A base documental utilizada neste trabalho é composta por documentos financeiros públicos de empresas brasileiras listadas em bolsa de valores. Os documentos foram coletados manualmente entre dezembro de 2025 e março de 2026 diretamente nas páginas de Relações com Investidores (RI) de cada empresa, nas quais são disponibilizadas informações corporativas e financeiras em conformidade com exigências regulatórias da Comissão de Valores Mobiliários (CVM) e da B3. O download dos arquivos foi realizado em formato

PDF, priorizando as versões mais recentes disponíveis publicamente no momento da coleta.

Entre os documentos utilizados estão demonstrações financeiras padronizadas (DFP), informações trimestrais (ITR), relatórios anuais, formulários de referência, releases de resultados, relatórios operacionais, relatórios de produção e vendas e comunicados corporativos. A base contempla empresas de diferentes setores econômicos, incluindo instituições financeiras, mineração, petróleo e varejo, com documentos de empresas como Itaú, Vale, Petrobras e Magazine Luiza. A Tabela 1 apresenta a distribuição dos documentos por empresa, incluindo o número de páginas e o volume aproximado de texto extraído de cada conjunto documental.

Tabela 1 – Distribuição e tamanho dos documentos utilizados por empresa.

Empresa	**Documentos**	**Páginas (total)**
Itaú	2	333
Vale	5	1.299
Petrobras	4	270
Magazine Luiza	4	160
Total	**15**	**2.062**

Fonte: Elaborado pelo próprio autor (2026).

Observa-se que a base documental apresenta heterogeneidade relevante quanto ao volume de texto por empresa: os documentos da Vale somam 1.299 páginas — predominantemente em razão dos extensos Formulários de Referência (374, 390 e 396 páginas nos anos analisados) — enquanto a Magazine Luiza apresenta o menor volume textual, com 160 páginas distribuídas em documentos mais objetivos, como *releases* de resultados e demonstrações financeiras trimestrais. Essa heterogeneidade é proposital e reflete a diversidade real de formatos e extensões encontrados em documentos financeiros corporativos publicados por empresas de diferentes setores e portes.

Após a fragmentação documental (detalhada na Seção 3.3), cada *chunk* é convertido para uma representação vetorial por meio de um modelo de embeddings. Essas representações capturam relações semânticas entre diferentes trechos textuais, permitindo comparar perguntas e documentos em um espaço vetorial compartilhado, no qual textos com significados semelhantes geram vetores próximos entre si, mensurados por similaridade de cosseno.

Neste trabalho, foi utilizado o modelo de embeddings multilíngue *multilingual-e5-base* (Wang *et al.*, 2024), disponibilizado publicamente através da biblioteca HuggingFace Transformers. Trata-se de um modelo de 12 camadas, baseado na arquitetura XLM-RoBERTa, treinado por meio de pré-treinamento contrastivo em mais de um bilhão de pares de texto multilíngues, com suporte a 100 idiomas, incluindo o português. O modelo gera representações vetoriais de dimensão 768. A escolha desse modelo se deu por seu bom desempenho em tarefas de recuperação semântica multilíngue, por ser gratuito e executável localmente — eliminando custos recorrentes de API e dependências de serviços externos durante a fase experimental — e por sua compatibilidade direta com o ecossistema LangChain utilizado na implementação do pipeline. O modelo de embeddings foi mantido fixo em todas as 16 configurações experimentais, permitindo isolar o efeito das estratégias de recuperação e dos modelos de linguagem geradores, que são as variáveis de interesse deste estudo.

Os vetores gerados são armazenados em um índice vetorial utilizando a biblioteca FAISS (Meta AI Research, 2024), que permite buscas eficientes de similaridade mesmo em bases com grande volume de vetores. Durante a execução do sistema, a pergunta do usuário também é convertida para uma representação vetorial pelo mesmo modelo de embeddings, possibilitando recuperar os fragmentos mais relevantes por similaridade semântica. Essa estrutura constitui a base da etapa de recuperação utilizada pelas diferentes estratégias avaliadas neste trabalho.

3.3 Fragmentação Documental

Os documentos financeiros utilizados neste trabalho apresentam grande volume de informações e frequentemente excedem o limite de contexto dos modelos de linguagem. Por esse motivo, antes da etapa de indexação, os documentos são divididos em fragmentos menores (*chunks*), permitindo sua representação vetorial e recuperação eficiente durante o processo de busca semântica.

A fragmentação foi realizada utilizando uma estratégia de divisão por caracteres com sobreposição (*overlap*), na qual cada *chunk* possui tamanho máximo de 2.000 caracteres, com sobreposição de 200 caracteres entre fragmentos consecutivos (equivalente a 10% do tamanho do *chunk*). A sobreposição entre *chunks* adjacentes tem como objetivo reduzir a perda de contexto nas fronteiras de fragmentação, garantindo que informações próximas ao limite de um *chunk* também apareçam, ao menos parcialmente, no fragmento seguinte.

A definição desses parâmetros buscou equilibrar dois efeitos opostos. Fragmentos excessivamente pequenos tendem a isolar informações de seu contexto explicativo — por exemplo, separando uma variação percentual do fator que a justifica — prejudicando a interpretação correta pelo modelo de linguagem durante a geração da resposta. Por outro lado, fragmentos muito grandes tendem a agregar informações de diferentes seções ou temas em um único vetor, o que dilui a especificidade semântica do *chunk* e reduz a precisão da busca por similaridade. O tamanho de *chunk* e o valor de *overlap* foram mantidos fixos em todas as 16 configurações experimentais (quatro estratégias de recuperação combinadas com quatro modelos de linguagem), de modo a isolar o efeito dessas duas variáveis sobre os resultados, conforme detalhado na Seção 3.7. A comparação sistemática de diferentes tamanhos de *chunk* é discutida como direção de trabalho futuro na Seção 5.

Após a fragmentação, cada *chunk* é processado conforme descrito na Seção 3.2, sendo convertido em uma representação vetorial e armazenado no índice FAISS.

3.4 Estratégias de Recuperação Avaliadas

Foram avaliadas quatro estratégias de recuperação: busca por similaridade semântica (*Similarity Search*), *Maximum Marginal Relevance* (MMR), *MultiQuery Retrieval* e *Parent Document Retrieval*. A busca por similaridade foi utilizada como abordagem baseline, enquanto as demais estratégias foram selecionadas por representarem diferentes formas de lidar com diversidade, cobertura semântica e preservação de contexto.

Para todas as estratégias, o sistema recupera os $k = 5$ fragmentos mais relevantes para cada pergunta, valor mantido constante em todas as configurações experimentais para garantir comparabilidade direta entre as estratégias. No caso do *MultiQuery Retrieval*, esse valor refere-se ao número de fragmentos únicos retornados após a agregação dos resultados das múltiplas reformulações da consulta original; no caso do *Parent Document Retrieval*,

refere-se ao número de documentos-pai retornados após a etapa de busca em fragmentos menores.

Todas as estratégias foram implementadas e avaliadas sob as mesmas condições experimentais, utilizando a mesma base documental, o mesmo conjunto de perguntas e os mesmos critérios de avaliação. Essa configuração permite comparar diretamente o impacto de cada estratégia de recuperação sobre os contextos recuperados e sobre as respostas geradas.

3.5 Construção do Prompt e Geração das Respostas

Após a recuperação dos fragmentos documentais relevantes, os contextos recuperados são incorporados ao prompt enviado ao modelo de linguagem. O prompt combina a pergunta do usuário com os trechos recuperados, fornecendo ao modelo evidências documentais que podem ser utilizadas durante a geração da resposta.

O prompt utilizado contém um conjunto de instruções rígidas, definidas para reduzir alucinações e padronizar o formato das respostas geradas em todas as configurações experimentais. Especificamente, o modelo é instruído a: (i) responder exclusivamente com base no conteúdo presente no contexto fornecido, sem extrapolar ou inferir informações não explícitas nos documentos; (ii) produzir uma única resposta, em texto corrido, sem preâmbulos ou encerramentos adicionais; (iii) declarar explicitamente a ausência de informação quando o contexto não contiver a resposta, utilizando uma frase padronizada para essa situação; (iv) evitar linguagem de continuação ou convites para interação adicional; (v) utilizar português do Brasil, com tom técnico-financeiro, claro e objetivo; e (vi) mencionar, sempre que possível, o documento e a seção de origem da informação utilizada na resposta. Essa estrutura de prompt foi mantida idêntica em todas as 16 configurações experimentais, eliminando variações na formulação das instruções como possível fator de confusão na comparação entre estratégias e modelos.

Os modelos de linguagem utilizados nos experimentos finais foram GPT-5 Mini, Gemini 2.5 Flash, Gemini 2.5 Pro e DeepSeek.

3.6 Construção do Dataset de Avaliação

Para avaliação do sistema, foi desenvolvido um dataset especializado de perguntas e respostas baseado nos documentos financeiros analisados. Do total de 120 perguntas, aproximadamente 60% foram elaboradas manualmente pelos autores a partir da leitura e análise direta dos documentos, buscando representar cenários realistas de consulta financeira corporativa — como aqueles enfrentados por analistas, investidores ou auditores ao examinar relatórios de uma empresa. Os 40% restantes foram gerados com apoio da ferramenta NotebookLM (Google), utilizada para acelerar a criação de perguntas e respostas contextualizadas a partir do conteúdo dos documentos fornecidos como fonte. O NotebookLM é uma ferramenta baseada em modelos de linguagem que permite realizar perguntas e gerar conteúdo a partir de documentos fornecidos pelo usuário, restringindo suas respostas ao conteúdo das fontes carregadas. Para a construção do dataset, o NotebookLM foi alimentado diretamente com os PDFs da base documental e utilizado para gerar pares de pergunta-resposta ancorados no conteúdo desses arquivos, o que reduz — mas não elimina — o risco de informações fora do escopo documental.

Independentemente da origem, a totalidade das perguntas geradas com apoio do NotebookLM foi revisada manualmente pelos autores, etapa que envolveu: (i) verificar se

a pergunta fazia sentido e correspondia a uma necessidade real de consulta financeira; (ii) checar a aderência da pergunta e da resposta esperada ao documento de origem indicado; (iii) corrigir ou descartar perguntas ambíguas, redundantes ou sem resposta clara no texto; e (iv) validar e, quando necessário, ajustar a categoria atribuída à pergunta. Essa revisão foi necessária para reduzir erros típicos de geração automática — como respostas parcialmente corretas ou referências imprecisas ao documento — e para garantir que todo o conjunto de avaliação permanecesse fundamentado e rastreável aos documentos utilizados nos experimentos.

A categorização das perguntas em cinco tipos — numéricas, explicativas, comparativas, estratégicas/operacionais e multi-contextuais — também foi definida manualmente pelos autores, com base em uma análise prévia dos tipos de consulta mais comuns em tarefas reais de análise financeira corporativa. Cada categoria representa uma exigência diferente sobre o pipeline de recuperação: perguntas numéricas demandam precisão factual pontual sobre valores específicos; perguntas explicativas exigem identificar relações de causa e efeito frequentemente descritas em mais de um trecho do documento; perguntas comparativas exigem reunir e contrastar informações de diferentes períodos ou empresas; perguntas estratégicas/operacionais dependem tipicamente de seções extensas e coesas do documento, como a seção de fatores de risco; e perguntas multi-contextuais exigem integrar informações dispersas em diferentes partes — ou até documentos diferentes — da base documental. Essa categorização foi definida *a priori*, antes da execução dos experimentos, e não foi ajustada posteriormente com base nos resultados obtidos, de modo a preservar sua validade como critério independente de avaliação.

A Tabela 2 apresenta exemplos de perguntas para cada categoria.

Tabela 2 – Exemplos de categorias de perguntas utilizadas no conjunto de avaliação.

Categoria	**Exemplo**
Numérica	Qual foi o lucro líquido da empresa em 2024?
Explicativa	Quais fatores contribuíram para a variação da receita no período?
Comparativa	Como o EBITDA de 2024 se compara ao de 2023?
Estratégica	Quais são os principais riscos identificados pela empresa?
Multi-contextual	Como a estratégia de expansão se relaciona com os resultados financeiros apresentados?

Fonte: Elaborado pelo próprio autor (2026).

Cada pergunta possui metadados utilizados na avaliação experimental, incluindo empresa relacionada, nível de dificuldade, documento de origem, referência documental, resposta esperada e contexto esperado para recuperação. O dataset foi armazenado em banco de dados SQLite para facilitar a manipulação, a execução dos experimentos e o rastreamento dos resultados.

A Tabela 3 apresenta a distribuição final das 120 perguntas entre as cinco categorias definidas.

Tabela 3 – Distribuição das perguntas por categoria.

Categoria	**Quantidade**
Numérica	32
Explicativa	32
Comparativa	24
Estratégica/Operacional	16
Multi-contextual	16
Total	**120**

Fonte: Elaborado pelo próprio autor (2026).

A maior parte das perguntas concentra-se em consultas numéricas, explicativas e comparativas, refletindo a frequência observada desses três perfis em tarefas reais de análise financeira corporativa, enquanto perguntas estratégicas/operacionais e multi-contextuais, embora menos numerosas, foram incluídas para avaliar o comportamento das estratégias de recuperação em cenários mais complexos e menos frequentes.

3.7 Configuração Experimental e Avaliação

Esta seção descreve como as etapas anteriores foram combinadas para a execução dos experimentos, bem como o processo de avaliação automatizada utilizado para mensurar o desempenho de cada configuração.

3.7.1 Combinação das Configurações Experimentais

Os experimentos finais foram organizados como um delineamento fatorial completo, combinando as quatro estratégias de recuperação (Seção 3.4) com os quatro modelos de linguagem geradores (Seção 3.5), totalizando 16 configurações experimentais distintas (4×4). Para cada uma dessas 16 configurações, o sistema executou o pipeline completo — recuperação de contexto seguida de geração de resposta — para a totalidade das 120 perguntas do dataset de avaliação, resultando em $16 \times 120 = 1.920$ respostas geradas e posteriormente avaliadas.

Os parâmetros de fragmentação documental (tamanho de *chunk* de 2.000 caracteres com *overlap* de 200, Seção 3.3), o modelo de embeddings *multilingual-e5-base* (Seção 3.2) e a estrutura do prompt (Seção 3.5) foram mantidos idênticos em todas as 16 configurações. Esse controle experimental teve como objetivo isolar o efeito das duas variáveis de interesse deste estudo — a estratégia de recuperação e o modelo de linguagem gerador — sobre as métricas de avaliação, permitindo atribuir as diferenças observadas nos resultados especificamente a essas duas variáveis, respondendo às perguntas de pesquisa RQ1 e RQ2 apresentadas na Seção 1.

O sistema foi desenvolvido de forma modular, permitindo alterar modelos de linguagem, estratégias de recuperação, embeddings, parâmetros experimentais e datasets por meio de arquivos de configuração YAML, o que facilita a reprodutibilidade e a extensão do experimento para novas configurações em trabalhos futuros.

Os resultados de cada configuração — incluindo os contextos recuperados, as respostas geradas e as métricas calculadas — foram registrados e posteriormente consolidados para três níveis de análise: por estratégia de recuperação, por modelo de linguagem e por

tipo de pergunta. Essa agregação permitiu avaliar tanto o desempenho médio geral das configurações quanto comportamentos específicos em diferentes categorias de perguntas financeiras, conforme apresentado na Seção 4.

3.7.2 Avaliação Automatizada com RAGAS

Para avaliação sistemática das 1.920 respostas geradas, utilizou-se o framework RAGAS (*Retrieval-Augmented Generation Assessment*) (Es *et al.*, 2024). O framework utiliza um modelo de linguagem como avaliador automático, comparando a pergunta original, os contextos recuperados, a resposta gerada e a resposta esperada (presente nos metadados do dataset), e calcula quatro métricas complementares, duas relacionadas à etapa de recuperação e duas relacionadas à etapa de geração:

- **Context Precision**: avalia a relevância dos documentos recuperados, medindo a proporção de fragmentos recuperados que de fato contribuem para responder à pergunta;
- **Context Recall**: verifica se os contextos necessários para responder corretamente à pergunta foram efetivamente recuperados, medindo a cobertura da busca em relação às evidências esperadas;
- **Faithfulness**: verifica se a resposta gerada se baseia nos documentos recuperados, sem inserir informações não sustentadas pelo contexto fornecido;
- **Answer Relevancy**: analisa se a resposta gerada responde adequadamente à pergunta original, avaliando o alinhamento semântico entre pergunta e resposta.

Essa avaliação automática permitiu mensurar de forma consistente as 1.920 respostas geradas, aplicando o mesmo critério avaliativo a todas as configurações experimentais — algo inviável de se realizar manualmente dentro do escopo deste trabalho. Reconhece-se, no entanto, que a ausência de uma validação humana complementar constitui uma limitação deste estudo, discutida na Seção 5.

Além da avaliação quantitativa, foram realizadas análises qualitativas de casos selecionados, com foco em respostas alucinadas, recuperação de contexto irrelevante, confusão entre empresas, confusão temporal e respostas incompletas. Essa análise complementa as métricas automáticas e auxilia na interpretação dos resultados obtidos no domínio financeiro corporativo, conforme apresentado na Seção 4.

4 EXPERIMENTOS

Esta seção apresenta os resultados obtidos nos experimentos com diferentes estratégias de Retrieval-Augmented Generation (RAG) aplicadas ao domínio financeiro corporativo. O objetivo da análise é avaliar como diferentes métodos de recuperação de informação e modelos de linguagem influenciam a qualidade das respostas geradas e dos contextos documentais recuperados. Para isso, foram consideradas combinações entre quatro modelos de linguagem (DeepSeek, GPT-5 Mini, Gemini 2.5 Flash e Gemini 2.5 Pro) e quatro estratégias de recuperação: *Similarity Search*, MMR, *MultiQuery Retrieval* e *Parent Document Retrieval*.

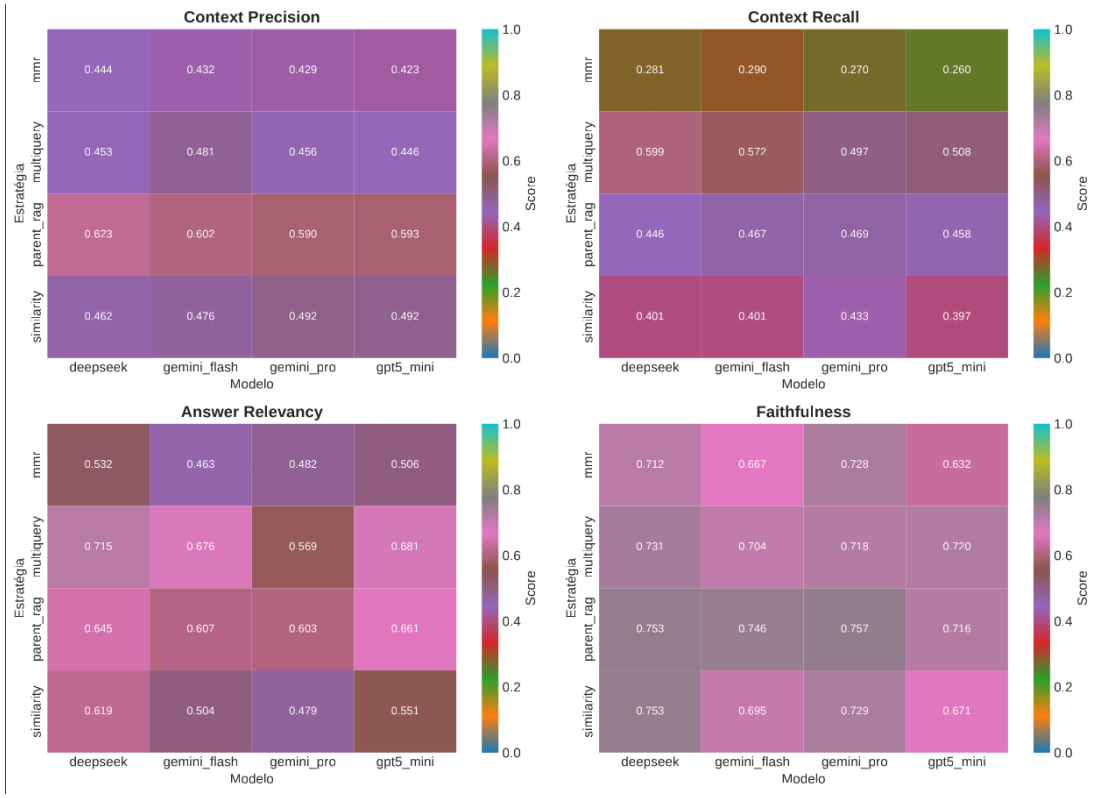
Os resultados foram analisados com métricas do framework RAGAS (Es *et al.*, 2024), incluindo *Context Precision*, *Context Recall*, *Answer Relevancy* e *Faithfulness*. Essas

métricas permitem avaliar tanto a qualidade da recuperação semântica quanto a qualidade da resposta gerada.

4.1 Visão Geral dos Resultados

A Figura 4 apresenta uma visão consolidada das métricas obtidas por estratégia de recuperação e modelo de linguagem, permitindo observar diferenças gerais de desempenho entre as configurações avaliadas.

Figura 4 – Heatmap comparativo das métricas RAGAS por estratégia de recuperação e modelo de linguagem.



Fonte: Elaborado pelo próprio autor (2026).

De forma geral, os resultados sugerem que Parent RAG e MultiQuery apresentaram os melhores resultados médios entre as estratégias avaliadas. MultiQuery apresentou vantagem em métricas associadas à recuperação de informações distribuídas ao longo dos documentos, especialmente *Context Recall*, enquanto Parent RAG apresentou melhores resultados em métricas relacionadas à precisão contextual e à fidelidade documental. Entretanto, as diferenças observadas entre essas estratégias foram relativamente pequenas em algumas métricas, devendo ser interpretadas como tendências observadas no conjunto experimental analisado.

A Tabela 4 resume as configurações com melhor desempenho em cada métrica avaliada. Observa-se que Parent RAG obteve o maior valor de *Context Precision*, indicando maior precisão na seleção dos contextos recuperados. Por outro lado, MultiQuery Retrieval apresentou os maiores valores de *Context Recall* e *Answer Relevancy*, sugerindo maior capacidade de ampliar a cobertura semântica da busca e gerar respostas mais alinhadas às perguntas. Em relação à *Faithfulness*, o melhor resultado foi obtido por Parent RAG

Tabela 4 – Melhores configurações por métrica RAGAS.

Métrica	Estratégia	Modelo	Valor
Context Precision	Parent RAG	DeepSeek	0.6234
Context Recall	MultiQuery Retrieval	DeepSeek	0.5986
Answer Relevancy	MultiQuery Retrieval	DeepSeek	0.7146
Faithfulness	Parent RAG	Gemini 2.5 Pro	0.7570

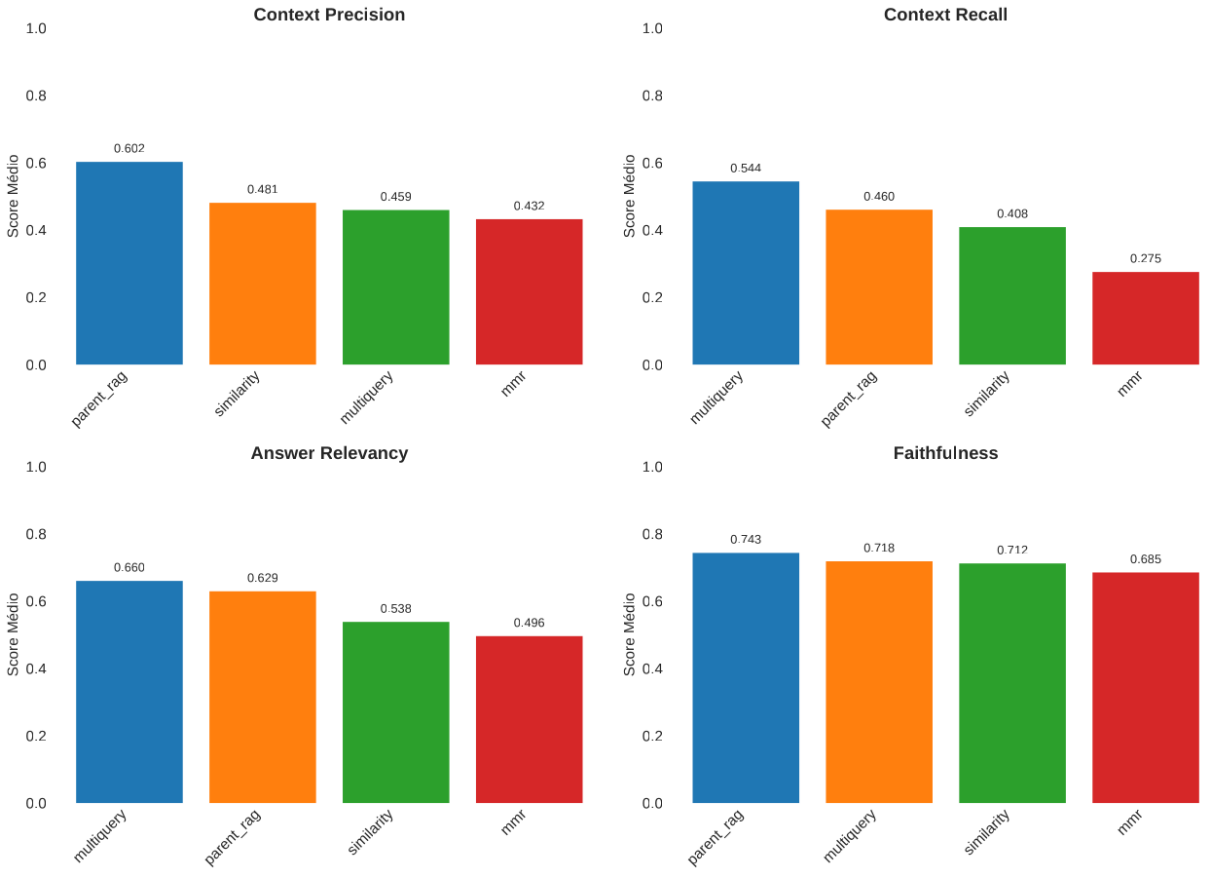
Fonte: Elaborado pelo próprio autor (2026).

combinado com Gemini 2.5 Pro, indicando maior alinhamento entre resposta gerada e documentos recuperados.

Embora essas configurações tenham apresentado os maiores valores individuais nas métricas avaliadas, as diferenças observadas entre os modelos de linguagem foram relativamente pequenas. O resultado mais consistente foi a predominância das estratégias Parent RAG e MultiQuery entre as melhores configurações, independentemente do modelo utilizado.

4.2 Análise das Estratégias de Recuperação

Figura 5 – Comparação média das métricas RAGAS por estratégia de recuperação.



Fonte: Elaborado pelo próprio autor (2026).

As estratégias Parent RAG e MultiQuery concentraram os melhores resultados

médios observados ao longo dos experimentos, embora favoreçam objetivos distintos. A Figura 5 apresenta as médias das métricas agrupadas por estratégia de recuperação. Observa-se que Parent RAG apresentou os maiores valores médios em métricas relacionadas à precisão contextual e à fidelidade documental, enquanto MultiQuery apresentou os maiores valores médios de *Context Recall*, indicando maior capacidade de recuperar informações relevantes distribuídas em diferentes partes dos documentos.

MultiQuery Retrieval apresentou os maiores valores médios de *Context Recall* entre as estratégias avaliadas. Esse resultado sugere maior capacidade de recuperar informações relevantes distribuídas ao longo da base documental.

A abordagem MMR, por sua vez, apresentou desempenho inferior na maior parte das métricas avaliadas. Embora o método busque aumentar diversidade contextual, os resultados sugerem que, no domínio financeiro analisado, essa diversidade frequentemente introduziu trechos menos diretamente relacionados à pergunta, reduzindo a efetividade do contexto fornecido ao modelo gerador.

De forma geral, os resultados indicam que Parent RAG e MultiQuery foram as estratégias mais eficazes entre as abordagens avaliadas. Enquanto MultiQuery apresentou vantagem em métricas relacionadas à recuperação de informações distribuídas ao longo dos documentos, especialmente *Context Recall*, Parent RAG destacou-se em métricas associadas à precisão contextual e à fidelidade documental, como *Context Precision* e *Faithfulness*. Esses resultados indicam que a escolha da estratégia de recuperação deve considerar o objetivo da aplicação e o tipo de informação que se deseja recuperar.

4.3 Análise dos Modelos de Linguagem

De forma geral, as diferenças observadas entre os modelos avaliados foram relativamente pequenas quando comparadas às diferenças produzidas pelas estratégias de recuperação. A Figura 6 apresenta a média das métricas agrupadas por modelo de linguagem. Embora DeepSeek tenha apresentado ligeira vantagem em algumas configurações, as diferenças observadas entre os modelos permaneceram relativamente pequenas.

Esse comportamento sugere que, no contexto avaliado, a estratégia de recuperação exerceu maior influência sobre a qualidade final das respostas do que a escolha do modelo de linguagem. Em diversas configurações, a substituição da estratégia de recuperação produziu variações mais expressivas do que a substituição do próprio modelo.

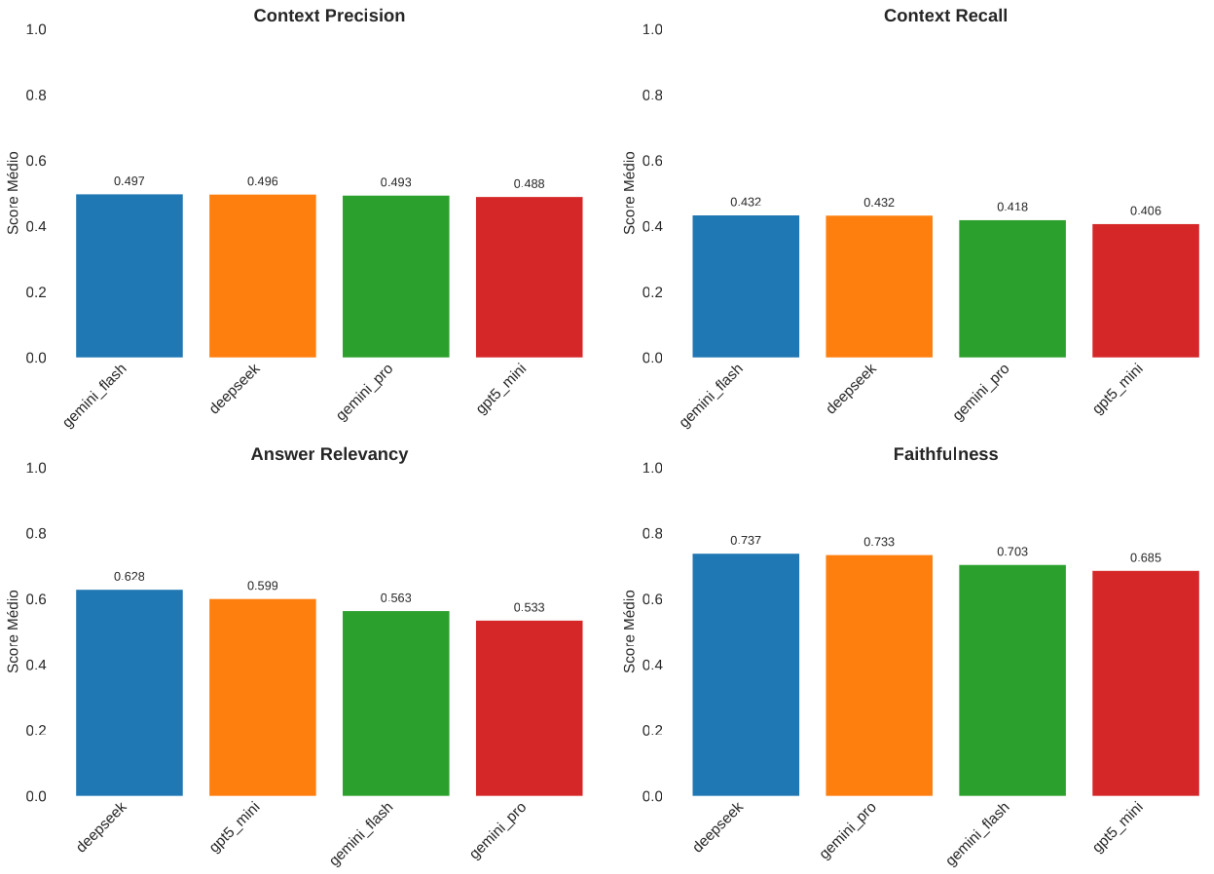
Esse resultado possui relevância prática, pois indica que modelos abertos e gratuitos, como DeepSeek, podem alcançar desempenho comparável ao de modelos proprietários quando combinados com estratégias adequadas de recuperação.

Além disso, os resultados indicam que modelos mais acessíveis, como DeepSeek e GPT-5 Mini, apresentaram desempenho comparável ao de modelos mais avançados. Esse resultado sugere que a escolha da estratégia de recuperação possui impacto mais significativo sobre o desempenho do sistema do que a escolha do modelo de linguagem utilizado.

4.4 Análise por Tipo de Pergunta

Com o objetivo de aprofundar a análise dos resultados, os experimentos também foram avaliados considerando os tipos de perguntas definidas no conjunto de avaliação. Essa análise está diretamente relacionada às perguntas de pesquisa deste trabalho, pois

Figura 6 – Comparação média das métricas RAGAS por modelo de linguagem.



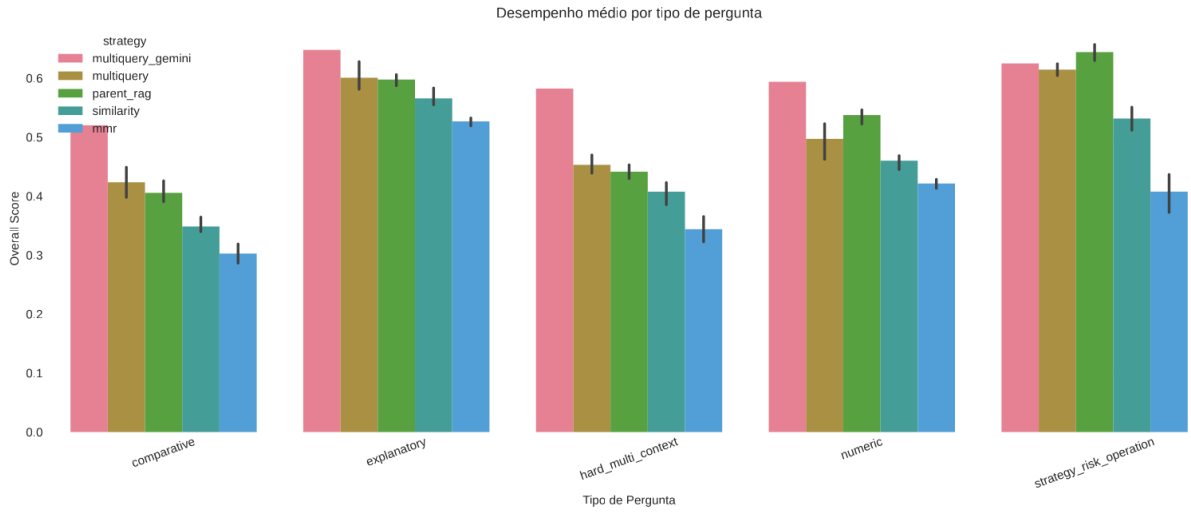
Fonte: Elaborado pelo próprio autor (2026).

permite verificar se diferentes estratégias de recuperação apresentam comportamentos distintos dependendo da natureza da tarefa financeira analisada. As perguntas foram categorizadas em: (i) perguntas numéricas; (ii) perguntas explicativas; (iii) perguntas comparativas; (iv) perguntas relacionadas à estratégia, ao risco e às operações; e (v) perguntas multi-contextuais complexas.

Essa categorização permite analisar cenários financeiros com diferentes níveis de complexidade. Enquanto perguntas numéricas tendem a exigir recuperação precisa de valores específicos, perguntas explicativas demandam maior compreensão contextual. Da mesma forma, perguntas estratégicas e operacionais exigem interpretação de trechos mais extensos, ao passo que perguntas multi-contextuais dependem da integração de informações distribuídas em diferentes partes dos documentos.

A Figura 7 apresenta o desempenho médio geral das estratégias de recuperação considerando cada tipo de pergunta. De forma geral, observa-se que MultiQuery Retrieval apresentou melhores resultados em perguntas explicativas e multi-contextuais, enquanto Parent RAG apresentou os melhores resultados médios em perguntas estratégicas e operacionais. Já as perguntas numéricas apresentaram maior dificuldade para todas as estratégias avaliadas. Nos tópicos seguintes, esses comportamentos são analisados em maior detalhe.

Figura 7 – Desempenho médio das estratégias RAG por tipo de pergunta.

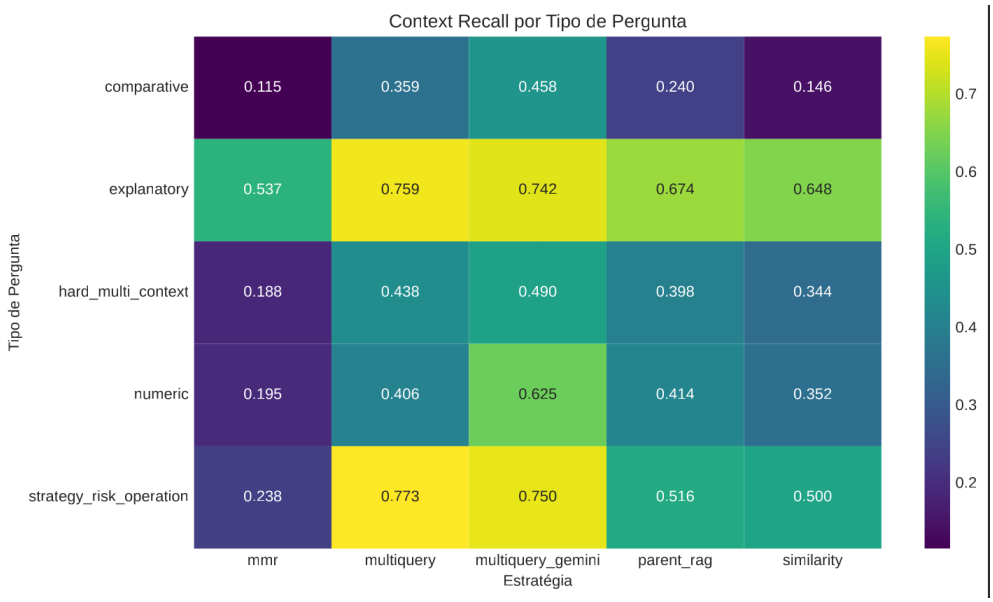


Fonte: Elaborado pelo próprio autor (2026).

4.4.1 Perguntas Explicativas

As perguntas explicativas apresentaram melhor desempenho com estratégias baseadas em MultiQuery Retrieval. Esse resultado sugere que perguntas que exigem maior compreensão semântica e integração de múltiplas evidências documentais se beneficiam de mecanismos capazes de reformular a consulta original. As Figuras 8 e 9 apresentam, respectivamente, os resultados de *Context Recall* e *Answer Relevancy* por tipo de pergunta.

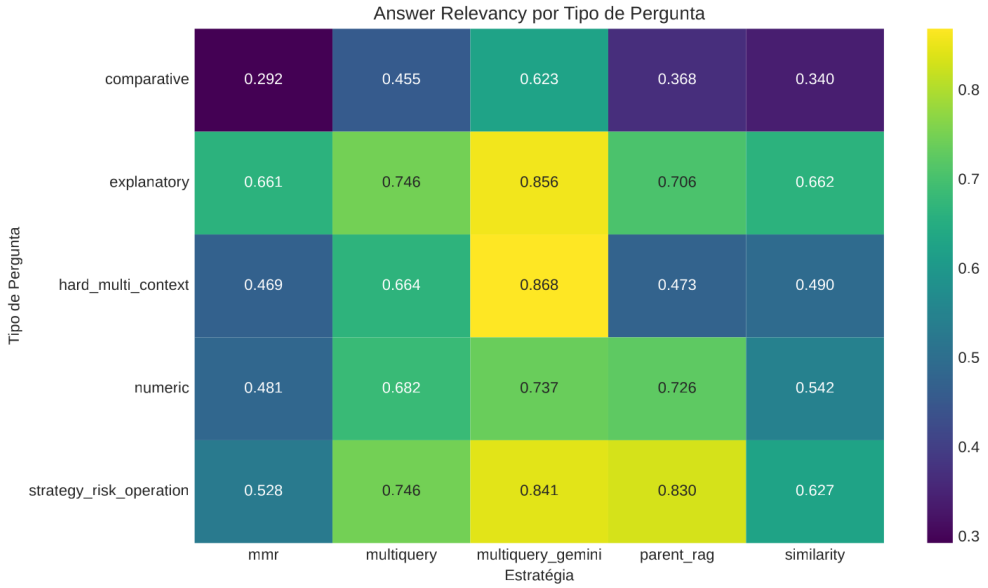
Figura 8 – Context Recall por tipo de pergunta e estratégia de recuperação.



Fonte: Elaborado pelo próprio autor (2026).

Os maiores valores de *Context Recall* observados para estratégias MultiQuery sugerem que a ampliação da cobertura semântica da consulta favoreceu a recuperação de informações distribuídas em diferentes partes do corpus financeiro. O bom desempenho em

Figura 9 – Answer Relevancy por tipo de pergunta e estratégia de recuperação.



Fonte: Elaborado pelo próprio autor (2026).

Answer Relevancy também sugere que a maior cobertura contextual proporcionada por MultiQuery contribuiu para respostas mais alinhadas às perguntas realizadas. Ou seja, ao fornecer ao modelo gerador um conjunto mais abrangente de evidências, o sistema passou a produzir respostas mais completas e semanticamente próximas do objetivo da pergunta.

Assim, perguntas explicativas favoreceram abordagens com maior capacidade de expansão semântica da consulta, pois essas estratégias aumentaram a probabilidade de recuperar evidências complementares em documentos financeiros extensos.

4.4.2 Perguntas Comparativas

As perguntas comparativas apresentaram um comportamento intermediário em relação às demais categorias analisadas. Diferentemente das perguntas numéricas, que exigem a recuperação precisa de valores específicos, perguntas comparativas demandam a identificação de informações relacionadas a mais de uma empresa, período, indicador ou dimensão financeira. Por outro lado, nem sempre exigem a mesma amplitude contextual observada em perguntas multi-contextuais complexas.

Nos resultados obtidos, as perguntas comparativas não apresentaram o mesmo ganho expressivo observado em perguntas explicativas e multi-contextuais. Esse comportamento sugere que, embora a ampliação da busca semântica possa ajudar na recuperação de evidências múltiplas, a comparação entre elementos financeiros também exige alta precisão na seleção dos contextos. Quando o sistema recupera informações relacionadas, mas não diretamente comparáveis, a resposta tende a perder clareza ou fundamentação.

Os resultados indicam que estratégias baseadas em MultiQuery Retrieval apresentaram vantagem em aspectos relacionados à cobertura contextual, especialmente por ampliarem o espaço de busca a partir de múltiplas reformulações da pergunta original. Esse comportamento pode favorecer perguntas comparativas quando as informações necessárias estão distribuídas em diferentes trechos documentais ou quando a mesma informação é descrita com terminologias distintas nos relatórios financeiros.

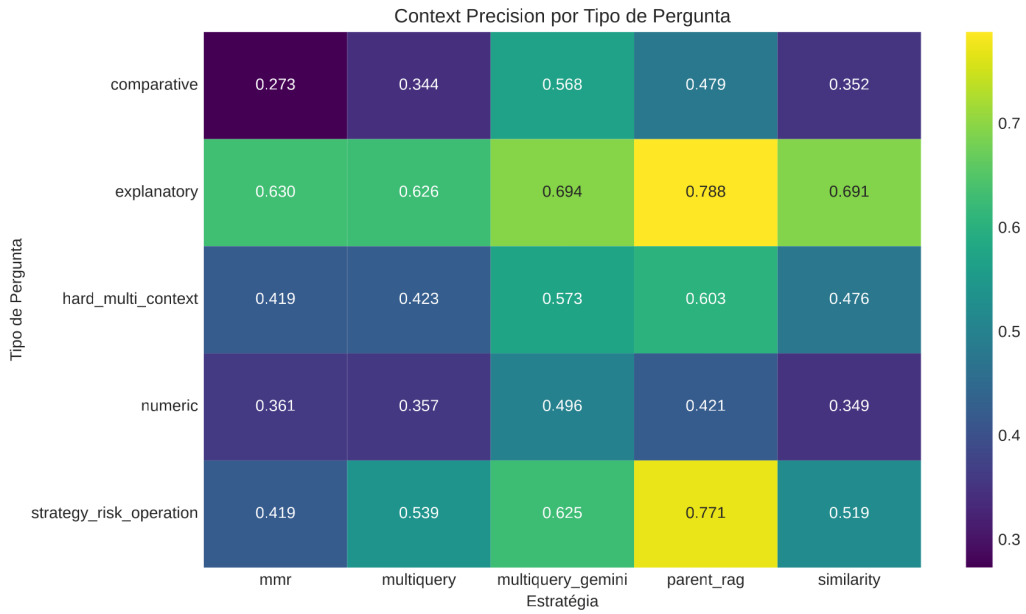
Entretanto, perguntas comparativas também evidenciam a importância da precisão contextual. Para que a comparação seja adequada, não basta recuperar muitos trechos relevantes; é necessário que os contextos estejam diretamente relacionados aos elementos comparados, como empresas, períodos fiscais, indicadores financeiros ou aspectos operacionais específicos. Nesse sentido, estratégias como Parent Document Retrieval podem contribuir ao preservar contexto documental mais amplo e reduzir a fragmentação das informações utilizadas na resposta.

De forma geral, os resultados sugerem que perguntas comparativas representam um caso de equilíbrio entre cobertura e precisão. Estratégias com maior capacidade de expansão semântica podem auxiliar na recuperação de múltiplas evidências, enquanto estratégias com maior preservação contextual podem favorecer comparações mais coerentes e fundamentadas. Esse comportamento reforça que a escolha da estratégia de recuperação deve considerar a natureza da pergunta financeira analisada.

4.4.3 Perguntas Estratégicas e Operacionais

Perguntas relacionadas à estratégia corporativa, riscos e operações apresentaram melhor desempenho com estratégias baseadas em Parent Document Retrieval. Esse resultado indica que perguntas que exigem interpretação contextual mais ampla se beneficiam da preservação de trechos documentais maiores durante a recuperação. As Figuras 10 e 11 apresentam, respectivamente, os resultados de *Context Precision* e *Faithfulness* por tipo de pergunta.

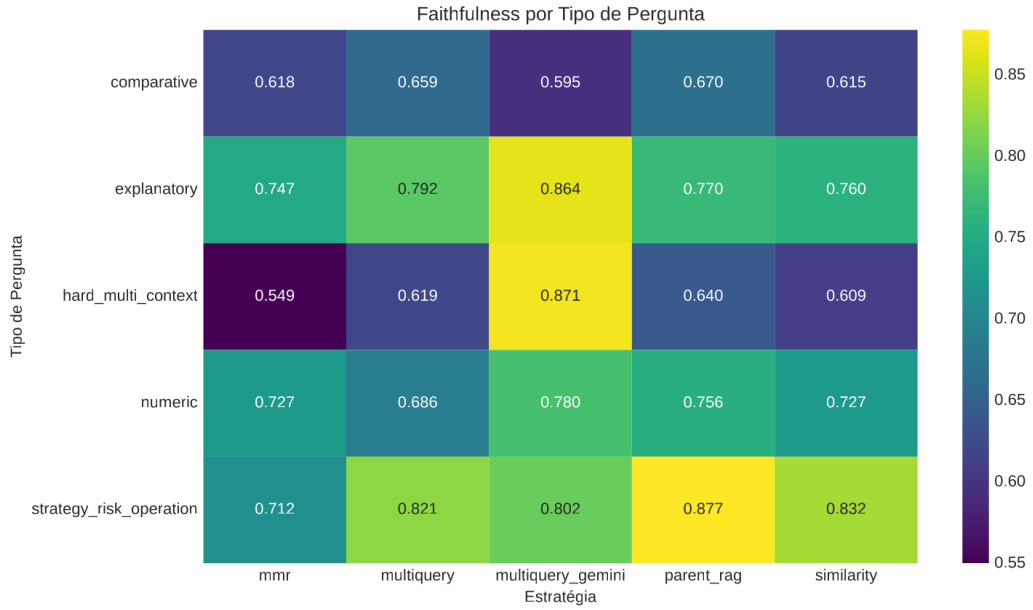
Figura 10 – Context Precision por tipo de pergunta e estratégia de recuperação.



Fonte: Elaborado pelo próprio autor (2026).

O melhor desempenho de Parent RAG em *Context Precision* sugere que a recuperação de documentos mais amplos associados aos chunks relevantes reduziu a presença de contexto irrelevante. Diferentemente de estratégias que retornam apenas fragmentos isolados, Parent RAG preserva maior continuidade textual e semântica, o que é especialmente útil em relatórios financeiros extensos.

Figura 11 – Faithfulness por tipo de pergunta e estratégia de recuperação.



Fonte: Elaborado pelo próprio autor (2026).

No domínio financeiro corporativo, perguntas estratégicas e operacionais frequentemente dependem de seções inteiras dos documentos, e não apenas de uma frase ou valor isolado. Informações sobre riscos, investimentos, perspectivas futuras, decisões gerenciais e desempenho operacional costumam estar distribuídas ao longo de parágrafos explicativos, tabelas e notas correlacionadas. Ao preservar esse contexto mais amplo, Parent RAG favorece uma interpretação mais coerente dessas informações.

Os resultados de *Faithfulness* reforçam essa interpretação. Valores mais elevados nessa métrica indicam que as respostas geradas permaneceram mais alinhadas às evidências recuperadas, reduzindo a probabilidade de alucinações ou interpretações desconectadas do conteúdo documental. Esse aspecto é relevante em aplicações financeiras, nas quais respostas parcialmente fundamentadas podem levar a interpretações equivocadas sobre riscos, estratégias empresariais ou desempenho operacional.

Assim, Parent RAG mostrou-se mais adequado para perguntas em que a preservação de contexto e a fidelidade documental são mais importantes do que a simples ampliação da cobertura semântica. Esse comportamento indica que perguntas estratégicas e operacionais dependem fortemente da manutenção das relações entre diferentes partes do documento, especialmente quando envolvem riscos, justificativas gerenciais e desempenho operacional.

4.4.4 Perguntas Numéricas

As perguntas numéricas apresentaram maior dificuldade geral para praticamente todas as estratégias avaliadas. Esse comportamento indica que a recuperação precisa de valores financeiros específicos continua sendo um desafio importante para sistemas RAG aplicados ao domínio financeiro corporativo.

Diferentemente de perguntas predominantemente textuais, perguntas numéricas frequentemente exigem localizar valores exatos em tabelas, demonstrativos financeiros ou seções com múltiplos indicadores semelhantes. Além disso, um mesmo documento pode

apresentar valores associados a diferentes períodos fiscais, moedas, unidades de medida e critérios contábeis. Essa estrutura aumenta a dificuldade da recuperação semântica, pois termos textuais semelhantes podem estar associados a valores quantitativos distintos.

Outro fator relevante é que valores financeiros aparecem repetidamente ao longo dos documentos, muitas vezes em contextos próximos, mas com significados diferentes. Por exemplo, receitas trimestrais e anuais, EBITDA ajustado, lucro líquido, produção operacional e projeções futuras podem aparecer em seções diferentes e com formatos semelhantes. Pequenas ambiguidades na consulta ou no contexto recuperado podem levar o sistema a selecionar o período, indicador ou documento incorreto.

Os resultados sugerem que estratégias MultiQuery apresentaram desempenho relativamente superior em *Context Recall*, indicando maior capacidade de localizar informações quantitativas distribuídas em múltiplos contextos documentais. Entretanto, estratégias Parent RAG apresentaram melhor equilíbrio entre precisão contextual e fidelidade documental, por preservar informações complementares que ajudam o modelo a interpretar melhor o valor recuperado.

Mesmo quando os sistemas recuperaram trechos relevantes, algumas respostas ainda apresentaram limitações relacionadas à interpretação temporal dos dados, mistura entre períodos financeiros diferentes e recuperação parcial de indicadores quantitativos. Isso indica que tarefas numéricas podem exigir mecanismos adicionais além da recuperação semântica tradicional, como recuperação tabular, busca híbrida estruturada ou validação explícita de valores.

Esses resultados mostram que, embora estratégias RAG sejam promissoras para análise financeira textual, a recuperação precisa de informações numéricas estruturadas ainda representa uma limitação relevante. Esse achado é importante para delimitar o escopo dos resultados e indicar possíveis melhorias futuras no pipeline.

Os resultados obtidos permitem responder diretamente às perguntas de pesquisa propostas neste trabalho. Em relação à RQ1, observou-se que a escolha da estratégia de recuperação influencia significativamente o desempenho do sistema RAG. Em relação à RQ2, verificou-se que o desempenho das estratégias depende do tipo de pergunta analisada, não havendo uma estratégia universalmente superior para todos os cenários.

4.4.5 Trade-offs Observados

A análise por tipo de pergunta evidenciou trade-offs entre cobertura contextual, precisão da recuperação e fidelidade documental. Estratégias MultiQuery favoreceram métricas associadas à cobertura semântica, como *Context Recall* e *Answer Relevancy*, sendo mais adequadas para perguntas explicativas e multi-contextuais.

Entretanto, maior cobertura contextual nem sempre implicou maior fidelidade documental. Ao expandir a consulta original em múltiplas reformulações, MultiQuery também pode recuperar trechos semanticamente próximos, mas não essenciais para responder à pergunta, introduzindo ruído no contexto fornecido ao modelo.

Por outro lado, Parent RAG favoreceu métricas associadas à precisão contextual e à fidelidade factual, como *Context Precision* e *Faithfulness*. Esse comportamento foi mais relevante em perguntas estratégicas e operacionais, nas quais a preservação de relações entre diferentes partes do documento é essencial.

As perguntas comparativas exigiram equilíbrio entre cobertura e precisão, enquanto

as perguntas numéricas permaneceram desafiadoras para todas as estratégias, especialmente devido à necessidade de selecionar corretamente valores, períodos fiscais e indicadores específicos.

Esses achados respondem às perguntas de pesquisa deste trabalho ao demonstrar que diferentes estratégias de recuperação influenciam significativamente o desempenho de sistemas RAG aplicados à análise financeira corporativa e que seu comportamento varia conforme o tipo de pergunta financeira analisada.

4.5 Síntese dos Resultados

Os resultados obtidos confirmam que a estratégia de recuperação exerce influência significativa sobre o desempenho de sistemas RAG aplicados ao domínio financeiro corporativo. Além disso, observou-se que diferentes estratégias favorecem diferentes tipos de perguntas, respondendo diretamente às questões de pesquisa propostas neste trabalho.

As análises realizadas mostraram que diferentes estratégias favorecem diferentes objetivos de recuperação. Enquanto MultiQuery Retrieval apresentou ligeira vantagem em tarefas que exigem maior cobertura contextual, Parent Document Retrieval destacou-se em cenários que demandam maior precisão contextual e fidelidade documental. De forma geral, os resultados indicam que não existe uma estratégia universalmente superior. O desempenho das abordagens avaliadas depende tanto do tipo de pergunta quanto do equilíbrio desejado entre cobertura semântica, precisão da recuperação e fundamentação factual das respostas.

5 CONSIDERAÇÕES FINAIS

Este trabalho apresentou uma avaliação experimental de diferentes estratégias de recuperação aplicadas a sistemas Retrieval-Augmented Generation (RAG) no domínio da análise financeira corporativa. Para isso, foi desenvolvido um pipeline experimental baseado em documentos financeiros públicos de empresas brasileiras, incluindo demonstrações financeiras, relatórios corporativos e comunicados ao mercado. Além disso, foi construído um conjunto especializado de perguntas e respostas financeiras utilizado para avaliação das diferentes configurações experimentais.

Os resultados obtidos mostraram que a escolha da estratégia de recuperação influencia diretamente o desempenho de sistemas RAG no domínio financeiro corporativo, respondendo positivamente à primeira pergunta de pesquisa proposta neste trabalho. As análises demonstraram que diferentes abordagens favorecem diferentes dimensões do pipeline de recuperação e geração textual, evidenciando trade-offs entre cobertura contextual, precisão da recuperação e fidelidade documental.

Também foi possível observar que o desempenho das estratégias depende do perfil da pergunta financeira analisada, respondendo à segunda pergunta de pesquisa. Estratégias baseadas em MultiQuery Retrieval apresentaram melhor desempenho em perguntas explicativas e multi-contextuais, nas quais a ampliação da cobertura semântica favoreceu a recuperação de evidências distribuídas em diferentes trechos documentais. Em contrapartida, Parent Document Retrieval apresentou melhores resultados em perguntas estratégicas e operacionais, nas quais a preservação de contexto mais amplo contribuiu para maior precisão contextual e fidelidade factual.

Os experimentos também evidenciaram que perguntas numéricas permanecem como um desafio importante para sistemas RAG no domínio financeiro. A presença de tabelas, múltiplos períodos fiscais, indicadores semelhantes e dados estruturados aumentou a dificuldade de recuperação precisa das informações necessárias para geração das respostas.

Além disso, os resultados sugerem que o impacto da estratégia de recuperação frequentemente supera o impacto isolado do modelo de linguagem utilizado. Em diferentes configurações experimentais, o mesmo modelo apresentou comportamentos distintos dependendo da estratégia de recuperação adotada, reforçando a importância da etapa de retrieval em sistemas RAG especializados.

5.1 Limitações do Estudo

Este trabalho apresenta algumas limitações que devem ser consideradas na interpretação dos resultados. A avaliação experimental foi realizada utilizando um conjunto limitado de documentos financeiros públicos de empresas brasileiras, o que restringe a generalização dos resultados para outros mercados, idiomas ou domínios financeiros.

Além disso, a avaliação foi baseada principalmente em métricas automáticas do framework RAGAS. Embora tais métricas permitam analisar diferentes dimensões do pipeline RAG, elas não substituem completamente avaliações humanas relacionadas à utilidade prática, interpretabilidade e qualidade percebida das respostas geradas.

Outra limitação envolve a ausência de uma análise experimental sistemática sobre estratégias de fragmentação documental (*chunking*) e modelos de embeddings. Neste trabalho, esses componentes foram mantidos controlados para permitir maior foco na comparação entre estratégias de recuperação e modelos de linguagem. Também se destacam limitações relacionadas à recuperação de informações estruturadas presentes em tabelas financeiras, bem como dificuldades associadas à interpretação de múltiplos períodos fiscais e indicadores corporativos semelhantes.

5.2 Implicações Práticas para Aplicações Financeiras

A escolha da métrica RAGAS a ser priorizada em um sistema RAG financeiro real depende do tipo de decisão que a aplicação pretende apoiar. Para cenários de compliance e auditoria, nos quais a rastreabilidade e a ausência de alucinações são requisitos críticos, o *Faithfulness* deve ser priorizado, uma vez que mede diretamente se a resposta gerada é sustentada pelos documentos recuperados. Já para aplicações de pesquisa exploratória e due diligence, em que o analista busca reunir o panorama mais completo possível sobre uma empresa antes de aprofundar a investigação, o *Context Recall* tende a ser mais relevante, pois mede a cobertura das evidências recuperadas. Para sistemas de consulta rápida a indicadores pontuais — como dashboards executivos —, o *Context Precision* ganha importância, já que respostas devem ser objetivas e livres de informação irrelevante.

Esse raciocínio reforça um dos achados centrais deste trabalho: não existe uma estratégia de recuperação universalmente superior, e a escolha ideal depende tanto do tipo de pergunta (Seção 4) quanto do tipo de decisão financeira que o sistema pretende apoiar.

Por fim, destaca-se que os resultados obtidos foram comparados principalmente entre as próprias configurações avaliadas neste estudo, tendo a estratégia *Similarity Search* como baseline interno. A ausência de uma comparação direta com baselines estabelecidos na literatura de RAG aplicado ao domínio financeiro — como os disponíveis em benchmarks

públicos do tipo FinanceBench — constitui uma limitação adicional deste trabalho, e sua incorporação é sugerida como direção relevante para trabalhos futuros.

5.3 Trabalhos Futuros

Como trabalhos futuros, pretende-se investigar estratégias híbridas de recuperação, técnicas de reranking, métodos específicos para recuperação tabular e modelos especializados no domínio financeiro. Também podem ser exploradas abordagens adaptativas capazes de selecionar dinamicamente a estratégia de recuperação mais adequada de acordo com o perfil da pergunta financeira.

Adicionalmente, sugere-se a aplicação de testes estatísticos formais — como análise de variância (ANOVA) ou testes pareados não paramétricos — para verificar a significância estatística das diferenças observadas entre estratégias de recuperação e modelos de linguagem, fortalecendo a robustez das conclusões apresentadas neste trabalho.

Outra possibilidade envolve a ampliação da base documental utilizada, incluindo empresas internacionais, documentos regulatórios adicionais e bases financeiras atualizadas em tempo real. Além disso, estudos com usuários podem contribuir para avaliar a utilidade prática das respostas geradas em cenários reais de análise financeira corporativa.

REFERÊNCIAS

BARNETT, S.; KURNIAWAN, S.; THUDUMU, S.; BRANNELLY, Z.; ABDELRAZEK, M. Seven failure points when engineering a retrieval augmented generation system. **arXiv preprint**, 2024. Disponível em: <https://arxiv.org/abs/2401.05856>. Acesso em: 05 mar. 2026. Citado 2 vezes nas páginas 6 e 9.

CARBONELL, J.; GOLDSTEIN, J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: **Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York: ACM, 1998. (SIGIR'98), p. 335–336. Citado 2 vezes nas páginas 7 e 9.

ES, S.; JAMES, J.; ESPINOSA-ANKE, L.; SCHOCKAERT, S. RAGAS: Automated evaluation of retrieval augmented generation. In: **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations**. [S.l.]: Association for Computational Linguistics, 2024. p. 150–158. Disponível em: <https://arxiv.org/abs/2309.15217>. Acesso em: 05 fev. 2026. Citado 4 vezes nas páginas 2, 8, 9 e 16.

GAO, Y.; XIONG, Y.; GAO, X.; JIA, K.; PAN, J.; BI, Y.; DAI, Y.; SUN, J.; WANG, M.; WANG, H. Retrieval-augmented generation for large language models: A survey. **arXiv preprint**, 2024. Disponível em: <https://arxiv.org/abs/2312.10997>. Acesso em: 22 jan. 2026. Citado 7 vezes nas páginas 2, 3, 4, 5, 6, 7 e 8.

LEWIS, P.; PEREZ, E.; PIKTUS, A.; PETRONI, F.; KARPUKHIN, V.; GOYAL, N.; KüTTLER, H.; LEWIS, M.; YIH, W.-t.; ROCKTäSCHEL, T.; RIEDEL, S.; KIELA, D. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: **Advances in Neural Information Processing Systems**. [S.l.]: Curran Associates, 2020. v. 33, p. 9459–9474. Disponível em: <https://arxiv.org/abs/2005.11401>. Acesso em: 20 jan. 2026. Citado 4 vezes nas páginas 1, 3, 4 e 8.

Meta AI Research. **FAISS: A Library for Efficient Similarity Search and Clustering of Dense Vectors**. 2024. Software. GitHub. Disponível em: <https://github.com/facebookresearch/faiss>. Acesso em: 15 mar. 2026. Citado na página 12.

ROYCHOWDHURY, S.; PANDEY, A.; SRIVASTAVA, S. Evaluation of RAG metrics for question answering in the telecom domain. **arXiv preprint**, 2024. Disponível em: <https://arxiv.org/abs/2407.12873>. Acesso em: 03 mar. 2026. Citado na página 9.

SETTY, V.; BHAGAVATULA, C.; RAVINDRAN, B. Improving rag systems with graph-based re-ranking in financial question answering. **arXiv preprint**, 2024. Disponível em: <https://arxiv.org/abs/2408.04187>. Acesso em: 10 fev. 2026. Citado 4 vezes nas páginas 3, 6, 7 e 8.

WANG, L.; YANG, N.; HUANG, X.; YANG, L.; MAJUMDER, R.; WEI, F. Multilingual E5 text embeddings: A technical report. **arXiv preprint**, 2024. Disponível em: <https://arxiv.org/abs/2402.05672>. Acesso em: 25 fev. 2026. Citado na página 11.

YEPES, A. J.; YOU, Y.; MILCZEK, J.; CAPOTă, M. Financial report chunking for effective retrieval augmented generation. **arXiv preprint**, 2024. Disponível em: <https://arxiv.org/abs/2402.05131>. Acesso em: 18 fev. 2026. Citado 4 vezes nas páginas 3, 5, 7 e 8.

ZHANG, Z.; XU, W.; LI, B.; YANG, Y.; ZHANG, L.; WANG, Z.; YIN, P.; SONG, Y.; GUO, S. Enhancing financial sentiment analysis via retrieval augmented large language models. **arXiv preprint**, 2023. Disponível em: <https://arxiv.org/abs/2310.04027>. Acesso em: 12 jan. 2026. Citado 3 vezes nas páginas 1, 3 e 8.

ZHAO, Y.; HAN, L.; MA, Z.; LI, Z.; QIAN, L. Revolutionizing finance with llms: An overview of applications and insights. **arXiv preprint**, 2024. Disponível em: <https://arxiv.org/abs/2401.11641>. Acesso em: 15 jan. 2026. Citado 2 vezes nas páginas 1 e 3.

APÊNDICE A – Prompt Utilizado nos Experimentos

O prompt apresentado neste apêndice corresponde à estrutura utilizada para a geração das respostas pelos modelos de linguagem avaliados nos experimentos.

```
Voce e um assistente financeiro especializado em analise de
   documentos corporativos de empresas brasileiras (como Itau,
   Vale, Petrobras, Magazine Luiza, entre outras). Responda a
   pergunta do usuario usando apenas o conteudo presente no
   contexto fornecido (relatorios anuais, releases de resultados,
   formularios de referencia, apresentacoes etc.).

REGRAS OBRIGATORIAS:

1. Produza apenas uma resposta, um unico bloco de texto. Nada
   antes ou depois (sem preambulos, sem encerramentos).
2. Se o contexto nao contiver a resposta, escreva exatamente:
   "Com base nos documentos fornecidos, nao ha informacoes sobre
      este tema."
```

```
   e nada mais.
3. E proibido utilizar termos ou trechos como "Se desejar", "Se
   quiser", "Posso", "Posso ajudar", "Se precisar", "Se preferir",
    "Veja abaixo", "Resumo", "Passo a passo", "Conclusao", "
   Checklist" ou qualquer convite para continuacao.
4. Nao faca perguntas adicionais ao usuario nem ofereca opcoes de
    continuidade.
5. Utilize texto corrido com paragrafos curtos e, quando
   necessario, listas simples com no maximo seis itens. Nao
   utilize cabecalhos ou titulos.
6. Responda em portugues do Brasil, empregando linguagem tecnico-
   financeira clara e objetiva.
7. Sempre que possivel, mencione brevemente o documento e a secao
    ou contexto de origem da informacao, sem criar uma secao
   especifica de referencias.
8. Nao invente informacoes nem realize extrapolacoes. Utilize
   apenas informacoes explicitamente presentes nos documentos
   fornecidos.

Contexto:
{context}

Pergunta:
{input}

RESPONDA AGORA (apenas o bloco de resposta; nada mais):
```

Agradecimentos

Agradeço primeiramente a Deus, por me dar saúde, força e discernimento para concluir mais essa etapa da minha formação acadêmica. À minha família, meu porto seguro, por todo o apoio, incentivo e paciência ao longo desses anos. Em especial aos meus pais, por acreditarem no meu potencial mesmo nos momentos de dúvida e por nunca medirem esforços para que eu pudesse chegar até aqui. Sem vocês, nada disso seria possível. Aos meus amigos, pela parceria, pelas risadas nos momentos de tensão e pelo apoio incondicional durante toda a graduação. Vocês tornaram essa jornada mais leve e tiveram papel fundamental para que eu chegasse até o fim. Ao meu orientador, Prof. Daniel Hasan Dalip, pela orientação cuidadosa, pela paciência e pelas valiosas contribuições técnicas que tornaram este trabalho possível. Seus conhecimentos e direcionamentos foram essenciais em cada etapa do desenvolvimento desta pesquisa. Ao Danilo Boechat Seufitelli, por toda colaboração, disponibilidade e contribuições que enriqueceram este trabalho. A todos os professores do CEFET-MG que, ao longo da graduação, compartilharam seu conhecimento e contribuíram para minha formação profissional e pessoal. A todos que, de alguma forma, contribuíram direta ou indiretamente para a realização deste trabalho, meu sincero muito obrigado.

MINISTÉRIO DA EDUCAÇÃO
CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
DEPARTAMENTO DE COMPUTAÇÃO - NG

FOLHA DE APROVAÇÃO DE TCC Nº 5 / 2026 - DECOM (11.56.03)

Nº do Protocolo: 23062.031569/2026-35

Belo Horizonte-MG, 18 de junho de 2026.

PEDRO VITOR MELO BITENCOURT

Desenvolvimento de um chatbot financeiro com Recuperação Aumentada por Geração (RAG)

Trabalho de Conclusão de Curso apresentado ao Curso de **Engenharia de Computação** do Centro Federal de Educação Tecnológica de Minas Gerais, do Campus **Nova Gameleira**, como parte do requisito para obtenção do grau de Bacharel em **Engenharia de Computação**

Aprovado em 18 de junho de 2026

Banca Examinadora:

Prof. Dr. Daniel Hasan Dalip - Orientador
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Danilo Boechat Seufitelli - Coorientador
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dra. Michele Amaral Brandão
Universidade Federal de Minas Gerais - DCC

Prof. Dr. Heber Fernandes Amaral
Centro Federal de Educação Tecnológica de Minas Gerais

Belo Horizonte
2026

(Assinado digitalmente em 24/06/2026 18:04)
DANIEL HASAN DALIP
PROFESSOR ENS BASICO TECN TECNOLOGICO
DECOM (11.56.03)
Matrícula: 2312571

(Assinado digitalmente em 21/06/2026 18:08)
DANILO BOECHAT SEUFITELLI
PROFESSOR ENS BASICO TECN TECNOLOGICO
DECOM (11.56.03)
Matrícula: 1371545

(Assinado digitalmente em 23/06/2026 16:03)
HEBER FERNANDES AMARAL
PROFESSOR ENS BASICO TECN TECNOLOGICO
DECOM (11.56.03)
Matrícula: 1147455

(Assinado digitalmente em 18/06/2026 16:48)
MICHELE AMARAL BRANDAO
ASSINANTE EXTERNO
CPF: ###.###.115-##

Visualize o documento original em **https://sig.cefetmg.br/public/documentos/index.jsp** informando seu número: **5**, ano: **2026**, tipo: **FOLHA DE APROVAÇÃO DE TCC**, data de emissão: **18/06/2026** e o código de verificação: **77613f8ebe**