KG-RAG: A Hybrid Retrieval Architecture for Institutional Normative Documents

Célia Y. S. Pereira[1], Gabriele S. Araújo[2], Marcelino S. da Silva[1], Sandio M. dos Santos[1], Richard C. da S. Rêgo[1], Fábio M. F. Lobato[1,2]

[1] Universidade Federal do Oeste do Pará, Santarém, PA, Brasil
celia.pereira@discente.ufopa.edu.br
[2] Universidade de São Paulo, São Carlos, SP, Brasil
fabio.lobato@icmc.usp.br

Abstract. Higher education institutions deal with complex hierarchical normative documents, a scenario where traditional Retrieval-Augmented Generation (RAG) systems suffer from context fragmentation and hallucinations. Furthermore, public data management requires strictly local processing to ensure technological sovereignty. To address these limitations, this paper proposes and evaluates the Knowledge Graph RAG (KG-RAG) architecture, a hybrid pipeline applied to the documentary ecosystem of the Federal University of Western Pará (UFOPA). The model integrates an institutional ontology, Contextual Chunking, and query decomposition to isolate administrative boundaries before response generation, processing a corpus of 11,047 chunks. An empirical evaluation using the RAGAS framework demonstrated that the approach outperformed a standard hybrid baseline, achieving gains of 30.1% in Faithfulness and 35.2% in Context Recall. The results provide evidence that retrieval guided by logical relationships mitigates context loss in deep normatives, ensuring informational accuracy and privacy in closed government environments.

CCS Concepts: • **Computing methodologies** → **Natural language generation**.

Keywords: Chatbot, Data Privacy, Knowledge Graphs, Large Language Models, Retrieval-Augmented Generation

1. INTRODUÇÃO

Instituições de ensino superior e órgãos públicos lidam diariamente com uma vasta gama de normativas [Oliveira et al. 2025]. Nesse contexto, garantir acesso ágil e preciso a essas informações supera a barreira da "transparência passiva" e promove a transparência institucional, que é a base para o entendimento claro da informação pelo usuário [Cappelli et al. 2023]. Além disso, contribui para a otimização dos processos administrativos. No entanto, a recuperação de informações em documentos institucionais complexos desafia sistemas automatizados de suporte à decisão, mesmo em interfaces conversacionais aplicadas a ambientes organizacionais práticos [Seabra et al. 2025]. A Geração Aumentada por Recuperação (RAG) [Lewis et al. 2020] padrão (Naive RAG) baseia-se na fragmentação e na recuperação com vetores densos. Para melhorar a precisão, adota-se o RAG híbrido, que integra vetores densos (semânticos) e esparsos (léxicos, como o BM25). Contudo, ambas as abordagens enfrentam limitações em documentos hierárquicos, pois a fragmentação arbitrária rompe a organização interna e as relações lógicas originais [Xu et al. 2024]. Essa perda de rastreabilidade nas fronteiras administrativas favorece alucinações, nas quais os sistemas misturam regras setoriais [Zhu et al. 2025]. Ademais, a dependência de *Application Programming Interfaces* (APIs) externas de Inteligência Artificial (IA) expõe o ecossistema a vulnerabilidades, exigindo processamento local para garantir a soberania tecnológica dos dados públicos [Marcacini et al. 2025].

Para mitigar o problema dos *chunks* isolados e aprimorar a rastreabilidade, este trabalho propõe

e avalia a arquitetura *Knowledge Graph Retrieval-Augmented Generation* (KG-RAG)[1], um *pipeline* híbrido projetado para assegurar a precisão na recuperação de normativas complexas. A solução é aplicada e validada ao ecossistema documental da Universidade Federal do Oeste do Pará (UFOPA). Considerando que os Grafos de Conhecimento fornecem dados estruturados e conhecimento factual que tornam o raciocínio do sistema mais inteligente e preciso [Fensel et al. 2020], o modelo proposto diferencia-se do RAG padrão por meio de três inovações técnicas: I) a modelagem de uma Ontologia Institucional dedicada a estabelecer limites entre Pró-Reitorias, regulamentos e cursos; II) a aplicação de uma técnica de enriquecimento de metadados pré-indexação (*Contextual Chunking*); III) a decomposição de consultas (*Query Decomposition*). A extração de subgrafos visa equilibrar a eficácia da busca com a capacidade de raciocínio do modelo [Li et al. 2025].

Além da contribuição técnica na engenharia de recuperação, o trabalho apresenta experimentos para avaliar quantitativamente a exatidão das informações geradas. A eficácia da arquitetura proposta é mensurada por meio do *framework* RAGAS [Es et al. 2024], visando atestar seu desempenho em comparação a um modelo base (*baseline*) vetorial tradicional. O intuito dessa avaliação experimental é garantir precisão nas respostas e atestar que a execução estritamente local atende às exigências de privacidade e segurança, sem depender de APIs proprietárias. A condução dos experimentos empíricos e a avaliação da arquitetura são norteadas pelas seguintes Perguntas de Pesquisa (PPs). **PP-1:** Qual é o impacto do *Contextual Chunking* e da decomposição de consultas na exatidão das informações recuperadas? **PP-2:** É possível manter a qualidade das respostas operando o sistema em uma infraestrutura estritamente local para proteger os dados institucionais?

O restante deste artigo está organizado como segue. Trabalhos relacionados são descritos na Seção 2. A Seção 3 detalha os materiais e métodos do estudo. A Seção 4 apresenta os resultados, suas avaliações e responde às perguntas de pesquisa. Por fim, a Seção 5 expõe as considerações finais e pesquisas futuras.

2. TRABALHOS RELACIONADOS

0 estudo de [Oliveira et al. 2025] propôs e avaliou uma arquitetura híbrida denominada J-KGRAG, que integra um grafo de conhecimento a um modelo RAG para reduzir o problema de recuperação de normas independentes. Os autores concluíram que a expansão baseada em grafo apresentou os melhores resultados, superando o *baseline* puramente vetorial, com ganhos significativos na correção factual (+75%) e na precisão geral das respostas (+16%). Contudo, não exploraram modelos de peso aberto na infraestrutura local para garantir a soberania dos dados públicos, dependendo de APIs proprietárias (GPT-4o), sem avaliar o enriquecimento contextual pré-indexação.

No trabalho de [Xavier and da Silva Soares 2025] foi proposto um *pipeline* de recuperação híbrida, livre de treinamento, que combina grafos de conhecimento e representações vetoriais densas para otimizar a extração de evidências em sistemas de perguntas e respostas. Os autores concluíram que o método híbrido apresentou os melhores resultados, superando, de forma consistente, os desempenhos isolados em métricas como Recall@10, MRR e nDCG@10. Porém, o trabalho se restringiu a bases de conhecimento gerais e focou na fase de recuperação, não contemplando a fase de geração de respostas e sem explorar o potencial dos *Large Language Models* (LLMs) para a decomposição de perguntas.

A pesquisa de [Zhu et al. 2025] apresenta o *framework* KG2RAG, que utiliza grafos de conhecimento para guiar a expansão de recuperação e organizar o contexto, com o objetivo de mitigar o problema de fragmentos isolados e melhorar a coerência das respostas geradas. Eles chegaram à conclusão de que o modelo apresentou vantagens consistentes em relação aos métodos RAG tradicionais, melhorando simultaneamente a precisão da recuperação e a qualidade da resposta final ao preservar as relações factuais entre os textos. Para avaliar o impacto prático dessa arquitetura, o desempenho do sistema proposto será comparado ao de um modelo *baseline* de RAG híbrido, baseado exclusivamente na

[1]Repositório do projeto: `https://github.com/celiaysp/KG_RAG.git`

combinação de recuperações vetoriais densas e esparsas.

3. MATERIAIS E MÉTODOS

Para solucionar as limitações de fragmentação e de perda de contexto em documentos institucionais, este estudo propõe a arquitetura KG-RAG. O modelo consiste em um *pipeline* de recuperação híbrida livre de treinamento, estruturado para operar estritamente em ambiente local, mitigando a dependência de APIs externas e assegurando a privacidade dos dados públicos. A engenharia da solução é orquestrada por meio do *framework* LangChain, utilizando o servidor Ollama para a inferência local, e baseia sua capacidade de raciocínio lógico no conhecimento documental injetado dinamicamente no *prompt*. A visão geral dessa arquitetura é ilustrada na Figura 1, cujo *pipeline* implementa o pré-processamento multimodal e a modelagem ontológica como base estrutural, seguidos pela decomposição de consultas para orquestrar a busca híbrida, culminando no reordenamento neuronal e na geração controlada de respostas.

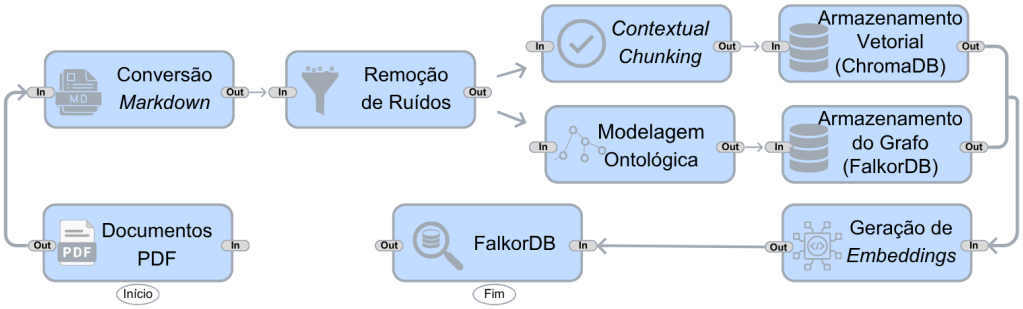

Figura. 1. Fluxo de ingestão de dados, modelagem ontológica e indexação multimodal.

3.1 Pré-processamento Multimodal e Modelagem Ontológica

A base de conhecimento processada no experimento é constituída por um *corpus* documental não estruturado no formato *Portable Document Format* (PDF). Para este estudo, foram processados 33 documentos administrativos da UFOPA, incluindo 2 Relatórios de Gestão e 31 Boletins de Serviço referentes aos anos de 2024 a 2026. Para a divisão estrutural do *corpus*, empregou-se a estratégia RecursiveCharacterTextSplitter,configurada com um tamanho de fragmento (*chunk size*) de 1.500 caracteres e uma sobreposição (*chunk overlap*) de 300 caracteres, utilizando quebras hierárquicas do formato *Markdown* como delimitadores principais. Esse processo de segmentação resultou em 11.047 fragmentos (*chunks*) aptos à etapa de indexação. Para preservar a estrutura do documento, incluindo tabelas e a hierarquia das seções, os arquivos foram inicialmente convertidos para o formato *Markdown* por meio de um processo de extração multimodal utilizando a biblioteca pymupdf4llm[2] (versão 1.27.2.3). Em seguida, aplicou-se um processo de limpeza baseado em expressões regulares para remover informações repetitivas e elementos sem valor semântico, como cabeçalhos e marcas de paginação, gerando um texto mais consistente.

Para solucionar o problema dos blocos isolados, implementou-se a técnica de enriquecimento de metadados denominada *Contextual Chunking*. Nessa abordagem, cada fragmento foi processado pelo modelo Gemma 3 (12 bilhões de parâmetros), antes de ser indexado. O modelo gerou um cabeçalho hierárquico conciso, com até 15 palavras, que resume o contexto administrativo do trecho e mantém a conexão com a fonte original (e.g., prefixando o fragmento com a tag [Contexto: Regulamento de Bolsas de Assistência Estudantil da PROGES]), facilitando a rastreabilidade das informações. De forma paralela, o texto limpo foi usado na construção do grafo de conhecimento por meio da biblioteca *GraphRAG SDK*[3], integrada ao banco de dados em grafo FalkorDB[4]. A arquitetura se destaca pela

[2] https://github.com/pymupdf/PyMuPDF
[3] https://github.com/FalkorDB/GraphRAG-SDK
[4] https://www.falkordb.com/

utilização de uma ontologia institucional específica, desenvolvida para representar de forma explícita os diferentes componentes da universidade, incluindo Pró-Reitorias, Regulamentos, Cursos, Matrizes curriculares e Bolsas/Auxílios. Por meio de relações direcionadas, como Gerencia, Regula e Oferece, o sistema conseguiu mapear conexões e dependências regulatórias entre esses elementos antes de realizar qualquer etapa de geração de resposta. A Figura 1 ilustra o *pipeline* arquitetural completo de pré-processamento e de indexação híbrida.

3.2 Decomposição de Consultas e Orquestração Híbrida

Na fase de recuperação, visando à precisão e à redução de ruídos, o sistema realizou um tratamento prévio da consulta, como mostra a Figura 2. Nesta etapa, o modelo de linguagem foi instruído via *prompt* a decompor a pergunta do usuário em quatro subperguntas independentes, isolando entidades-chave, como nomes próprios, cargos e siglas, para alimentar os motores de busca.

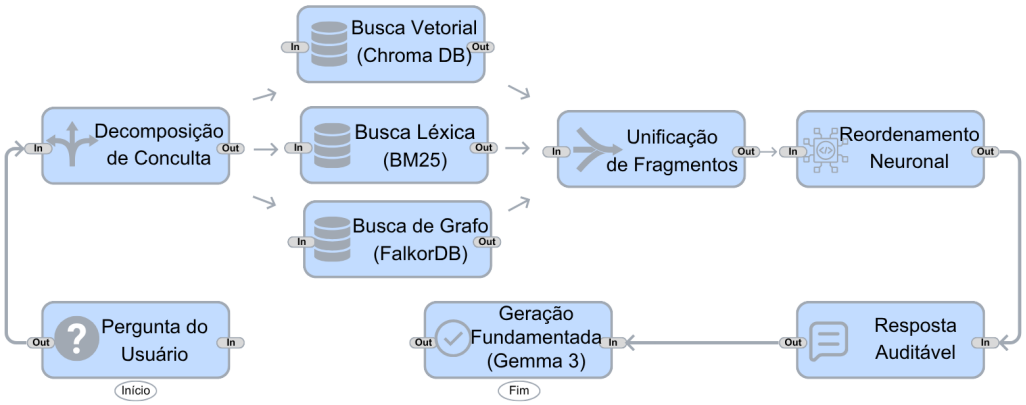

Figura. 2. Arquitetura de inferência híbrida com decomposição de consultas e reordenamento neuronal.

Essas consultas expandidas alimentaram simultaneamente um motor de busca híbrido, procedimento justificado pela eficiência de bancos de dados multimodais em coordenar buscas léxicas, vetoriais e estruturais baseadas em grafos [Xavier and da Silva Soares 2025]. A busca combinada foi estruturada nos seguintes passos:

Recuperação Semântica: Executada no banco de dados vetorial ChromaDB, que captura a similaridade de intenção por meio do modelo de *embeddings* nomic-embed-text-v2-moe. A vetorização opera em sua dimensionalidade padrão de 768 dimensões. A escolha justificou-se por sua arquitetura de pesos abertos,que oferece bom desempenho e baixo custo computacional local. Essa eficiência viabiliza a vetorização de textos na infraestrutura local, preservando a privacidade dos dados. Além disso, a abordagem densa, aliada ao *chunking* adequado, tem eficácia comprovada em textos longos e normativos no cenário brasileiro [Justino et al. 2025];

Recuperação Léxica: Utilizando o algoritmo *Best Matching* 25 (BM25), instanciado com seus hiperparâmetros padrão empíricos ($K_1 = 1.5$ e b $= 0.75$) para a busca de palavras exatas. A inclusão do BM25 é essencial, pois compensa uma limitação inerente aos modelos vetoriais densos, que costumam ter dificuldade com termos muito específicos [Robertson and Zaragoza 2009]. No caso dos documentos institucionais, o algoritmo atua como um mecanismo direto para localizar números de portarias e siglas, informações precisas frequentemente perdidas quando o sistema busca apenas pelo sentido geral;

Recuperação Estrutural: A biblioteca GraphRAG-SDK (versão 1.3.0) foi utilizada para recuperar o contexto das conexões mapeadas na ontologia. A heurística de extração no grafo baseia-se no resgate de relacionamentos diretos de primeira ordem (*1-hop*) a partir das entidades identificadas na consulta. Essa busca auxilia na resolução de uma limitação dos vetores, que encontram assuntos semelhantes, mas não percebem a ordem hierárquica nem os vínculos institucionais. O grafo mapeia

o caminho lógico entre regulamentos e departamentos, amarrando as regras organizacionais ao texto [Edge et al. 2024].

3.3 Reordenamento Neuronal e Geração Controlada

Após a busca paralela, os resultados dos três índices foram associados diretamente e a duplicação foi eliminada por correspondência exata. Para refinar esse volume, foi aplicada uma etapa de reordenamento (*reranking*) *zero-shot* (sem *fine-tuning* adicional no domínio) utilizando o modelo *Cross-Encoder* BAAI/bge-reranker-v2-m3. Diferentemente do *Reciprocal Rank Fusion* (RRF), focado na combinação matemática de posições, o *Cross-Encoder* processou a pergunta e o texto simultaneamente. Essa atenção cruzada avalia a aderência semântica real, garantindo maior precisão contextual. Embora apresente maior custo computacional, seu uso foi justificado por atuar apenas na triagem final, restringindo o contexto aos 12 fragmentos mais relevantes. Esse limite (top_k=12) foi definido empiricamente para equilibrar a retenção de contexto útil com a capacidade de inferência local.

Na fase de geração, esse contexto filtrado alimentou o modelo de linguagem conforme as regras definidas no *System Prompt*. A principal instrução foi o princípio de Isolamento de Entidades, que impede a IA de misturar normas ou regras de diferentes setores da universidade. Para mitigar o risco de alucinações, o modelo foi forçado a construir a resposta exclusivamente a partir dos dados recuperados, ignorando os conhecimentos paramétricos prévios. O *prompt* possui restrições explícitas e absolutas, instruções diretas, como "Responda somente com informações presentes no contexto", "Nunca invente, suponha, complete ou extrapole informações ausentes". Essa abordagem ancorou a geração do texto de forma rigorosa aos fragmentos fornecidos. Além disso, o *pipeline* exigiu que a resposta final cite a fonte e a página exata do documento utilizado, garantindo a rastreabilidade da informação. Todo o ambiente de desenvolvimento foi estruturado em Python, utilizando o *framework* LangChain para orquestrar as etapas do KG-RAG e o servidor Ollama para gerenciar o modelo de linguagem Gemma 3 (12B). Os experimentos foram executados em um servidor institucional equipado com uma GPU NVIDIA RTX A5500 (24 GB de VRAM). A infraestrutura de armazenamento seguiu os requisitos de privacidade com a adoção do banco de dados ChromaDB para armazenar os vetores, por ser uma solução leve e capaz de rodar totalmente *offline* [5], e do FalkorDB para armazenar o grafo de conhecimento, garantindo o processamento rápido das conexões institucionais no próprio servidor.

3.4 Avaliação Experimental

Para comprovar a eficiência do KG-RAG em relação ao RAG híbrido, foram conduzidos testes com o *framework* automatizado RAGAS [Es et al. 2024]. A grande vantagem dessa abordagem é a capacidade de auditar o sistema em partes, isolando o desempenho da etapa de recuperação e da etapa de geração. Isso traz transparência e torna o experimento totalmente reprodutível [Knollmeyer et al. 2026]. As perguntas empregadas na avaliação foram elaboradas manualmente para espelhar dúvidas reais da instituição. O banco e os testes contêm 35 consultas, distribuídas igualmente entre 5 frentes: i) busca de entidades específicas; ii) atribuição de competências; iii) listagem estrutural; iv) relações hierárquicas; e v) dados quantitativos. Com diferentes níveis de complexidade, o modelo foi desafiado desde a busca por siglas até o cruzamento de regras. A reprodutibilidade no *framework* RAGAS foi garantida por meio de uma temperatura nula.

O desempenho foi mensurado por quatro métricas fundamentais (com escores de 0 a 1) [Es et al. 2024; Knollmeyer et al. 2026]: **Fidelidade:** calculada como a razão entre as afirmações geradas que são suportadas pelo contexto e o total de afirmações da resposta (avaliadas de forma autônoma pelo LLM juiz). **Relevância da Resposta:** obtida pela similaridade de cosseno entre a pergunta original e as consultas reversas geradas a partir da resposta. **Precisão de Contexto:** Utiliza a métrica *Mean Average Precision* (MAP) para avaliar a proporção de fragmentos úteis priorizados nas primeiras posições do ranqueamento. **Revocação de Contexto:** Calcula a proporção de sentenças da resposta ideais (*ground truth*) efetivamente recuperadas da base de dados.

[5] `https://www.trychroma.com/`

4. RESULTADOS E DISCUSSÕES

A etapa inicial de preparação do *corpus* por meio de expressões regulares resultou em uma redução de 3,93% em relação ao volume original, passando de 4.004.553 para 3.847.084 *tokens*. Como exemplo prático, cabeçalhos ruidosos (como "Ir para o conteúdo") foram completamente suprimidos da indexação. O conjunto de perguntas empregado foi elaborado para refletir as demandas reais de recuperação informacional da universidade. Seguindo as diretrizes da literatura sobre a avaliação de sistemas baseados em LLMs, o banco de testes foi estruturado em diferentes níveis de complexidade. Além de perguntas com respostas diretas, foram inseridas consultas nas quais o modelo é forçado a cruzar normativas e relacionar dados dispersos. A Tabela I apresenta uma amostra representativa das perguntas empregadas.

Tabela I. Amostra das consultas de validação utilizadas na avaliação, categorizadas por intenção de busca.

Categoria / Desafio	**Exemplo de Pergunta**
Busca de Entidade Específica	Qual pró-reitoria cuida das bolsas de assistência estudantil?
Atribuição de Competência	Quem aprova o calendário acadêmico da UFOPA?
Listagem Estrutural	Quais são os pró-reitores da Ufopa e suas pró-reitorias?
Relação Hierárquica (*Multi-hop*)	Qual é a pró-reitoria de pesquisa? Quais são as suas direções?
Dado Quantitativo	Qual o total de recursos extraorçamentários captados em 2025?

A análise dos resultados quantitativos, detalhados na Tabela II e ilustrados graficamente na Figura 3, permite responder diretamente à PP-1: a integração do *Contextual Chunking* e da ontologia institucional eleva significativamente a precisão do sistema. A métrica de Fidelidade, que atesta se o modelo construiu a resposta exclusivamente com fatos presentes nos documentos, subiu de 0,73 no modelo *baseline* para 0,95 no KG-RAG, um aumento de 30,1%. Esse resultado evidencia que o grafo de conhecimento atua como um delimitador lógico, restringindo as inferências do modelo e mitigando alucinações. Em instituições públicas, essa confiabilidade é essencial.

Tabela II. Comparativo das métricas de avaliação (RAGAS) entre a abordagem híbrida padrão e o artefato KG-RAG.

Métrica	**RAG Híbrido (Padrão)**	**KG-RAG (Proposto)**	**Variação**
Fidelidade	0,73	0,95	+30,1%
Relevância da Resposta	0,51	0,62	+21,5%
Precisão de Contexto	1,00	1,00	0,0%
Revocação de Contexto	0,71	0,96	+35,2%

A principal melhoria em relação ao RAG híbrido reflete-se na Revocação de Contexto, que apresentou um avanço expressivo. O desafio da consulta *"Quem aprova o calendário acadêmico da UFOPA?"* exemplifica bem esse cenário. O modelo *baseline* recuperou normativas semelhantes, mas misturou as atribuições e forneceu informações incompletas. O KG-RAG, por sua vez, utilizou as relações de dependência mapeadas na ontologia para isolar o setor responsável, resgatando a totalidade do contexto necessário e solucionando a interferência comum em buscas complexas.

Na análise qualitativa, evidenciou-se que o *Contextual Chunking* mitigou a perda semântica comum no processamento de tabelas. Fragmentos resultantes de quebras de página que contêm apenas dados numéricos e linhas isoladas passaram a preservar seu significado por meio da inserção do metadado gerado pelo LLM. Por exemplo, uma linha que estava solta foi enriquecida com o cabeçalho "[Contexto: Tabela de bolsas de assistência estudantil concedidas pela PROGES em 2024] - Auxílio Moradia: 150', garantindo o vínculo com o regulamento original". Somado a isso, o algoritmo léxico BM25 demonstrou sua importância ao compensar a dificuldade dos *embeddings* de lidar com correspondências exatas, garantindo o resgate assertivo de siglas importantes como PROGES e PROEN.

Por fim, a PP-2 trata da viabilidade técnica de operar essa arquitetura com segurança. Os testes confirmaram o processamento estritamente local. Ao atingir uma fidelidade de 0,95 com um modelo de pesos abertos do Gemma 3 (12B) em um servidor institucional, o sistema mitiga a necessidade de

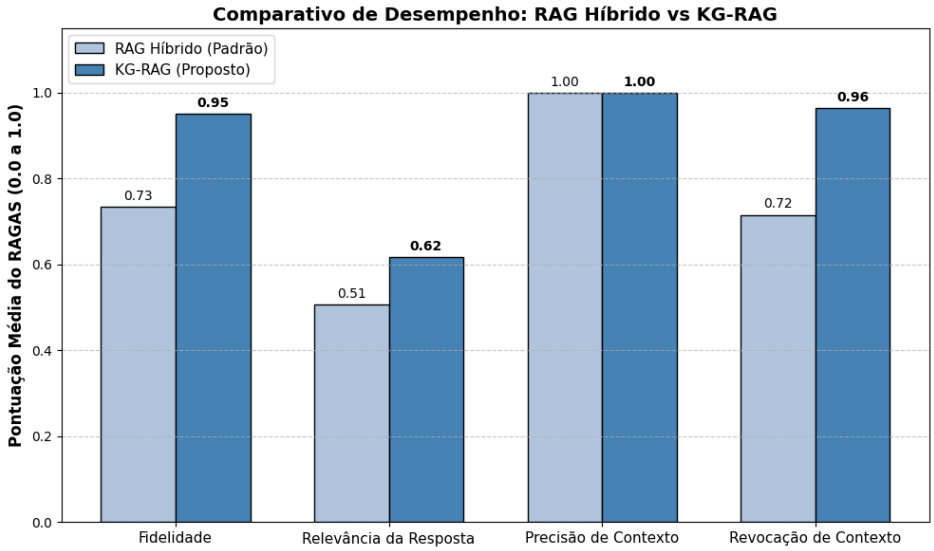

Figura. 3. Comparativo de desempenho das métricas do RAGAS entre o RAG Híbrido e o KG-RAG.

enviar documentos sensíveis para APIs de terceiros. A arquitetura demonstra que é possível processar normativas hierárquicas com precisão, assegurando a soberania tecnológica e a privacidade necessárias à gestão de dados públicos.

5. CONCLUSÃO

Este trabalho propôs e avaliou a arquitetura KG-RAG, um pipeline híbrido de recuperação de informação projetado para mitigar a fragmentação e a perda de contexto em documentos normativos complexos. Para superar as limitações das buscas puramente textuais, a solução uniu a flexibilidade da busca semântica à precisão estrutural de uma ontologia institucional. Ao aplicar o modelo no ecossistema administrativo da UFOPA, a avaliação evidenciou que é possível alcançar precisão informacional ao operar o sistema de forma estritamente local, preservando a privacidade dos dados públicos e a soberania tecnológica da instituição. Delimitada à amostra avaliada, cujo tamanho de 35 questões configura uma limitação metodológica sujeita a vieses, a integração da busca em grafo com o *Contextual Chunking* apresentou desempenho favorável em relação ao *baseline*, registrando crescimento de 30,1% na Fidelidade e de 35,2% na Revocação de Contexto. Adicionalmente, as limitações operacionais do *pipeline* exigem um esforço computacional considerável na indexação do grafo e elevam o tempo de inferência local.

Apesar dos resultados promissores, a adoção do *pipeline* proposto apresenta limitações operacionais que precisam ser apontadas. A construção automatizada do grafo exige um esforço computacional considerável na etapa de indexação dos documentos, especialmente na extração estruturada de entidades e de relações. Adicionalmente, a orquestração paralela dos múltiplos canais de busca aumenta o tempo de resposta da inferência. Essas características apontam a necessidade de refinamentos de desempenho antes que seja feita a expansão do modelo. Como melhorias para trabalhos futuros, planeja-se conduzir um estudo de ablação para avaliar o impacto individual de cada componente do modelo. Uma análise formal de sensibilidade de hiperparâmetros, especificamente quanto ao limite de fragmentos (top_k), além da expansão da ontologia com novas normativas institucionais. O objetivo final deste experimento é implementar a arquitetura do modelo em uma interface conversacional completa e otimizada. Com isso, busca-se transformar o sistema em um instrumento prático que democratize o acesso transparente à informação pública, com rapidez, precisão e segurança institucional.

Agradecimentos e Declaração de uso de IA generativa
O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Brasil – Código de Financiamento 001, do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) – DT–303031/2023–9 e da Financiadora de Estudos e Projetos (FINEP) (ProAmazonia – 2373/24).

Declara-se que os modelos de IA generativa *Gemini* 3 e *Grammarly Pro* foram empregados como ferramentas de suporte ao longo da pesquisa, sendo utilizados para revisão gramatical do texto, e o programa *Antigravity* para auxiliar na revisão de código. A autoria e a responsabilidade integral pelo conteúdo final, incluindo a verificação de plágio e as correções, são da primeira pessoa autora.

REFERÊNCIAS

Cappelli, C., Oliveira, R., and Nunes, V. Linguagem simples como pilar da transparência. *Humanidades & Inovação* 10 (9): 32–45, 2023.

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., and Larson, J. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Es, S., James, J., Anke, L. E., and 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*. pp. 150–158, 2024.

Fensel, D., Şimşek, U., Angele, K., Huaman, E., Kärle, E., Panasiuk, O., Toma, I., Umbrich, J., and Wahler, A. Introduction: what is a knowledge graph? In *Knowledge graphs: Methodology, tools and selected use cases*. Springer, pp. 1–10, 2020.

Justino, A. F., Junior, A. F. L. J., and Lobato, F. M. F. A comparative study of bert models for semantic retrieval of brazilian legal precedents. In *Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)*. SBC, pp. 65–72, 2025.

Knollmeyer, S., Caymazer, O., Koval, L., Akmal, M. U., Asif, S., Mathias, S. G., and Grossmann, D. Evaluating retrieval augmented generation: A comprehensive review of evaluation dimensions, question types, and application. *SN Computer Science* 7 (5): 576, 2026.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* vol. 33, pp. 9459–9474, 2020.

Li, M., Miao, S., and Li, P. Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation. In *International Conference on Learning Representations*. Vol. 2025. pp. 6061–6089, 2025.

Marcacini, R. M., Valverde-Rebaza, J. C., Turine, M. A., Santos, B. N., Levcovitz, S., and Rezende, S. O. Llm4gov: A privacy-preserving approach to teacher-student fine-tuning of distilled llms for the public sector. In *Latin American Symposium on Digital Government (LASDiGov)*. SBC, pp. 273–284, 2025.

Oliveira, V. T., Lima, M. R., Teles, S., and Dias, E. S. J-kgrag: A hybrid retrieval-augmented generation architecture for legal norm understanding with knowledge graphs. In *Congresso Latino-Americano de Software Livre e Tecnologias Abertas (Latinoware)*. SBC, pp. 202–209, 2025.

Robertson, S. and Zaragoza, H. *The probabilistic relevance framework: BM25 and beyond*. Vol. 4. Now Publishers Inc, 2009.

Seabra, A., Cavalcante, C., Medeiros, G., Nepomuceno, J., Sallenave, V., Ruberg, N., and Lifschitz, S. Pergunte ao bi: Uma interface conversacional para apoio à tomada de decisão usando modelos de linguagem. In *Simpósio Brasileiro de Banco de Dados (SBBD)*. SBC, pp. 142–147, 2025.

Xavier, O. C. and da Silva Soares, A. Training-free hybrid evidence retrieval for question answering: Dynamic fusion of knowledge-graph triples and dense text embeddings. In *Simpósio Brasileiro de Banco de Dados (SBBD)*. SBC, pp. 644–657, 2025.

Xu, Z., Cruz, M. J., Guevara, M., Wang, T., Deshpande, M., Wang, X., and Li, Z. Retrieval-augmented generation with knowledge graphs for customer service question answering. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*. pp. 2905–2909, 2024.

Zhu, X., Xie, Y., Liu, Y., Li, Y., and Hu, W. Knowledge graph-guided retrieval augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. pp. 8912–8924, 2025.