



INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS ANGICAL
CURSO ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

VINÍCIUS GABRIEL DE SOUSA SOARES

**GRAPHRAG APLICADO A UMA INTELIGÊNCIA ARTIFICIAL CONVERSACIONAL
PARA REDUÇÃO DE ALUCINAÇÕES**

ANGICAL
2026

VINÍCIUS GABRIEL DE SOUSA SOARES

GRAPHRAG APLICADO A UMA INTELIGÊNCIA ARTIFICIAL CONVERSACIONAL
PARA REDUÇÃO DE ALUCINAÇÕES

Trabalho de Conclusão de Curso (artigo científico) apresentado como exigência parcial para obtenção do diploma do Curso de Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Angical.

Orientador: Prof. Me. Charleno Queiroz Pires.

GRAPHRAG APLICADO A UMA INTELIGÊNCIA ARTIFICIAL CONVERSACIONAL PARA REDUÇÃO DE ALUCINAÇÕES

Vinícius Gabriel de Sousa Soares¹

Prof. Me. Charleno Queiroz Pires²

RESUMO

Os modelos de linguagem de grande escala impulsionaram o desenvolvimento de sistemas de inteligência artificial conversacional capazes de responder consultas em linguagem natural. Entretanto, esses modelos apresentam limitações relacionadas à geração de informações incorretas ou não fundamentadas nas fontes utilizadas, fenômeno conhecido como alucinação. Entre as abordagens propostas para reduzir esse problema, destaca-se a Geração Aumentada por Recuperação, que utiliza informações externas durante o processo de geração de respostas. Contudo, a recuperação tradicional baseada principalmente em similaridade vetorial apresenta limitações na representação explícita das relações semânticas entre entidades, dificultando consultas que dependem de conhecimentos interconectados. Nesse contexto, este trabalho teve como objetivo desenvolver e avaliar um agente conversacional baseado em Geração Aumentada por Recuperação com Grafos, comparando seu desempenho com um agente baseado em recuperação tradicional quanto à aderência às informações recuperadas e à ocorrência de respostas não fundamentadas. A pesquisa caracteriza-se como aplicada e exploratória, desenvolvida por meio de um estudo comparativo entre duas arquiteturas de recuperação de conhecimento. Foram desenvolvidos dois agentes utilizando o Google Agent Development Kit, empregando o mesmo modelo de linguagem e a mesma fonte documental, sendo o agente de recuperação tradicional submetido a uma limitação de indexação decorrente da cota da interface de programação de aplicações utilizada. A avaliação foi realizada por meio de perguntas padronizadas, analisando a fidelidade das respostas ao conteúdo documental e o reconhecimento da ausência de informações. Os resultados indicaram maior controle contextual e rastreabilidade no agente baseado em grafos.

Palavras-chave: inteligência artificial; graphrag; modelos de linguagem de grande escala; neo4j; sistemas conversacionais.

ABSTRACT

Large language models have driven the development of conversational artificial intelligence systems capable of answering queries in natural language. However, these models still present limitations related to the generation of incorrect or unsupported information from the sources used, a phenomenon known as hallucination. Among the approaches proposed to mitigate this problem, Retrieval-Augmented Generation stands out by incorporating external information during response generation. Nevertheless, traditional retrieval based mainly on vector similarity presents limitations

¹Vinícius Gabriel de Sousa Soares – Instituto Federal de Educação, Ciência e Tecnologia do Piauí – Campus Angical. E-mail: viniciulaika98@gmail.com

²Prof. Me. Charleno Queiroz Pires – Instituto Federal de Educação, Ciência e Tecnologia do Piauí – Campus Angical. E-mail: charlenopires@ifpi.edu.br

in the explicit representation of semantic relationships between entities, which may affect queries that require reasoning over interconnected information. In this context, this study aimed to develop and evaluate a conversational agent based on Graph Retrieval-Augmented Generation, comparing its performance with an agent based on traditional retrieval regarding adherence to retrieved information and the occurrence of unsupported responses. The research is characterized as applied and exploratory and was developed through a comparative study between two knowledge retrieval architectures. Two agents were developed using the Google Agent Development Kit, employing the same language model and the same documentary source, while the traditional retrieval agent was subject to an indexing limitation due to the quota restrictions of the application programming interface used. The evaluation was conducted through standardized questions, analyzing the fidelity of responses to the document content and the ability to recognize missing information. The results indicated greater contextual control and traceability in the graph-based agent.

Keywords: artificial intelligence; graphrag; large language models; neo4j; conversational systems.

Data de aprovação: 28/07/2026

1 INTRODUÇÃO

A Inteligência Artificial Generativa (IAG) consolidou-se como uma das principais tecnologias emergentes da Computação contemporânea, impulsionada principalmente pelos avanços dos Modelos de Linguagem de Grande Porte (*Large Language Models* – LLMs). Esses modelos são capazes de compreender, interpretar e gerar textos em linguagem natural com elevado grau de fluência e coerência, sendo amplamente empregados em agentes conversacionais, assistentes virtuais, sistemas de apoio à decisão, mecanismos de busca inteligentes e ferramentas de automação textual (Bommasani et al., 2021; OpenAI, 2023; Google Team, 2023; Touvron et al., 2023).

Apesar do desempenho expressivo desses modelos em diversas tarefas de Processamento de Linguagem Natural, estudos recentes apontam limitações relacionadas à confiabilidade factual das respostas produzidas. Conforme discutem Ji et al. (2023) e Huang et al. (2025), os LLMs estão sujeitos ao fenômeno das alucinações, caracterizado pela geração de conteúdos plausíveis do ponto de vista linguístico, porém incorretos, inconsistentes ou não fundamentados em evidências verificáveis. Esse problema representa um desafio significativo para aplicações em domínios que exigem elevado grau de precisão, como educação, saúde, setor jurídico, administração pública e ambientes corporativos.

Com o objetivo de reduzir essa limitação, arquiteturas *Retrieval-Augmented Generation* (RAG) foram propostas para integrar fontes externas de conhecimento ao processo de geração textual. Nessa abordagem, antes de elaborar uma resposta, o modelo recupera documentos considerados relevantes para a consulta do usuário e utiliza essas informações como contexto adicional durante a geração da resposta (Lewis et al., 2020; Gao et al., 2023). Dessa forma, busca-se reduzir a dependência exclusiva

do conhecimento paramétrico armazenado durante o treinamento do modelo. Entretanto, embora o RAG tradicional apresente ganhos relevantes de desempenho, sua estratégia de recuperação baseada predominantemente em similaridade vetorial pode apresentar limitações na representação explícita das relações semânticas existentes entre diferentes entidades e conceitos.

Diante desse contexto, esta pesquisa tem como objetivo geral desenvolver e avaliar um agente conversacional baseado em GraphRAG, comparando seu desempenho com um agente fundamentado na abordagem tradicional de Retrieval-Augmented Generation (RAG), analisando a aderência das respostas às informações disponíveis na base de conhecimento e a ocorrência de respostas não sustentadas pelas evidências recuperadas. Para alcançar esse objetivo, foram definidos como objetivos específicos a construção de uma base de conhecimento estruturada em grafo a partir da extração de entidades e relacionamentos presentes nos documentos utilizados; o desenvolvimento de dois agentes conversacionais utilizando o Google Agent Development Kit (ADK), sendo um baseado em RAG e outro em GraphRAG; a realização de experimentos comparativos utilizando o mesmo conjunto de dados e modelo de linguagem; e a análise dos resultados obtidos considerando critérios de fidelidade ao conteúdo documental, rastreabilidade e reconhecimento da ausência de informações.

A justificativa desta pesquisa está relacionada à necessidade de desenvolver sistemas de inteligência artificial conversacional mais confiáveis, visto que modelos de linguagem podem produzir respostas plausíveis, porém não fundamentadas nas informações disponíveis. Nesse sentido, a investigação de arquiteturas que associam recuperação de informações e estruturas de conhecimento, como o GraphRAG, torna-se relevante devido à capacidade dessas abordagens de organizar informações por meio de entidades e relacionamentos, favorecendo maior contextualização, rastreabilidade e exploração de conexões semânticas durante o processo de geração de respostas (Peng et al., 2024). Dessa forma, este trabalho busca contribuir para a análise experimental das diferenças entre RAG tradicional e GraphRAG em cenários de recuperação de conhecimento para agentes conversacionais.

2 MODELOS DE LINGUAGEM DE GRANDE ESCALA

Os Modelos de Linguagem de Grande Escala (*Large Language Models* – LLMs) representam um dos principais avanços recentes da Inteligência Artificial Generativa e do Processamento de Linguagem Natural. Baseados em arquiteturas de redes neurais profundas treinadas sobre grandes volumes de dados textuais, esses modelos são capazes de compreender, interpretar e gerar linguagem natural com elevado grau de coerência e fluidez, possibilitando sua aplicação em agentes conversacionais, sistemas de recomendação, mecanismos de busca inteligentes, assistentes virtuais e diversas outras aplicações baseadas em linguagem natural.

Segundo Bommasani et al. (2021), os LLMs integram um conjunto mais amplo de modelos fundamentais (*foundation models*), treinados em larga escala e posteriormente adaptados para diferentes tarefas. Essa característica permite que um único modelo execute atividades distintas, como tradução automática, sumarização, classificação de textos, resposta a perguntas e geração de conteúdo.

Modelos dessa natureza apresentam elevada capacidade de generalização por meio da aprendizagem em contexto (*in-context learning*), conforme sintetizado na citação a seguir:

Modelos de linguagem são aprendizes few-shot, capazes de realizar tarefas com pouco ou nenhum treinamento específico para a tarefa, baseando-se apenas em alguns exemplos fornecidos no prompt (Brown et al., 2020, p. 2, tradução nossa).

A arquitetura Transformer, proposta por Vaswani et al. (2017), consolidou-se como a base estrutural dos LLMs contemporâneos ao introduzir o mecanismo de auto-atenção (*self-attention*), permitindo o processamento paralelo de sequências textuais e ampliando significativamente a capacidade de representação contextual das palavras. Aliada ao aumento da capacidade computacional e à disponibilidade de grandes volumes de dados para treinamento, essa arquitetura possibilitou o desenvolvimento de modelos cada vez maiores e mais sofisticados.

Entretanto, apesar do elevado desempenho observado em diversas tarefas, os LLMs operam de forma probabilística, realizando a predição do próximo token com base nos padrões estatísticos aprendidos durante o treinamento. Dessa forma, esses modelos não possuem mecanismos intrínsecos para verificar a veracidade das informações produzidas nem para distinguir, de forma determinística, conteúdos verdadeiros de informações incorretas ou inexistentes.

Essa característica constitui uma limitação importante em aplicações institucionais, organizacionais e científicas, nas quais a precisão das informações representa um requisito essencial. Conforme destacam Ji et al. (2023), a ausência de mecanismos de validação factual está diretamente relacionada ao fenômeno das alucinações, considerado atualmente um dos principais desafios para a utilização prática de modelos de linguagem em ambientes críticos.

Além das limitações relacionadas à confiabilidade das respostas, Bommasani et al. (2021) destacam que o uso de modelos fundamentais também impõe desafios relacionados à transparência, à segurança e ao uso responsável da Inteligência Artificial. Nesse contexto, cresce o interesse por arquiteturas capazes de complementar o conhecimento paramétrico dos LLMs por meio de fontes externas confiáveis, motivando o desenvolvimento de abordagens como o Retrieval-Augmented Generation

(RAG) e, posteriormente, o GraphRAG.

2.1 RETRIEVAL-AUGMENTED GENERATION (RAG)

A arquitetura *Retrieval-Augmented Generation* (RAG) foi proposta por Lewis et al. (2020) como uma estratégia para reduzir a dependência exclusiva do conhecimento paramétrico armazenado pelos Modelos de Linguagem de Grande Porte. Nessa abordagem, o modelo passa a consultar uma base documental externa durante o processo de geração da resposta, permitindo utilizar informações atualizadas e específicas de determinado domínio.

A arquitetura RAG pode ser sintetizada da seguinte forma:

Modelos de Geração Aumentada por Recuperação (RAG) combinam um gerador sequence-to-sequence pré-treinado com um índice de vetores densos de documentos, permitindo que o modelo recupere e incorpore conhecimento externo relevante durante a geração (Lewis et al., 2020, p. 9459, tradução nossa).

O funcionamento do RAG pode ser dividido em três etapas principais. Inicialmente, a consulta realizada pelo usuário é convertida em um vetor de representação (*embedding*), utilizado para recuperar os documentos semanticamente mais próximos em uma base vetorial. Em seguida, os documentos recuperados são incorporados ao *prompt* enviado ao modelo de linguagem. Finalmente, o LLM gera a resposta considerando simultaneamente o conhecimento recuperado e o conhecimento adquirido durante seu treinamento.

A principal contribuição dessa arquitetura consiste em reduzir a necessidade de memorizar grandes volumes de informações durante o treinamento do modelo, permitindo que o conhecimento permaneça armazenado externamente e possa ser atualizado de maneira independente. Além disso, o RAG possibilita fundamentar as respostas em documentos específicos, reduzindo parcialmente a ocorrência de informações desatualizadas e ampliando a confiabilidade das respostas produzidas.

Entretanto, apesar dessas vantagens, o mecanismo de recuperação do RAG tradicional baseia-se predominantemente em similaridade vetorial calculada a partir de *embeddings* densos. Embora esse processo apresente elevada eficiência para localizar documentos semanticamente semelhantes, ele não representa explicitamente as relações existentes entre entidades, eventos ou conceitos presentes na base documental.

Como consequência, consultas que exigem raciocínio relacional, integração de informações distribuídas em diferentes documentos ou inferências envolvendo múltiplas conexões semânticas (*multi-hop reasoning*) podem apresentar limitações. Nes-

nas situações, o sistema tende a recuperar documentos semelhantes do ponto de vista semântico, mas nem sempre consegue estabelecer adequadamente as relações necessárias para produzir respostas completas e contextualizadas.

Essa limitação tem motivado o desenvolvimento de novas arquiteturas capazes de integrar mecanismos de recuperação vetorial com estruturas semânticas explícitas, como os grafos de conhecimento. Entre essas abordagens destaca-se o GraphRAG, que amplia o processo de recuperação ao incorporar relações estruturadas entre entidades, favorecendo maior contextualização, interpretabilidade e redução da ocorrência de alucinações (Edge et al., 2024; Peng et al., 2024; Han et al., 2025).

2.2 ALUCINAÇÕES EM SISTEMAS DE INTELIGÊNCIA ARTIFICIAL

As alucinações constituem uma das principais limitações dos Modelos de Linguagem de Grande Porte (LLMs), sendo consideradas um dos maiores desafios para a utilização segura da Inteligência Artificial Generativa em aplicações que exigem elevada confiabilidade das informações.

Ji et al. (2023, p. 3, tradução nossa) definem alucinação como “conteúdo gerado que não faz sentido ou é infiel ao conteúdo de origem fornecido, embora pareça fluente e plausível”.

Esse fenômeno caracteriza-se pela geração de informações incorretas, inexistentes ou incompatíveis com as evidências disponíveis, embora apresentadas de forma linguisticamente coerente. Em razão dessa característica, as respostas produzidas podem aparentar elevado grau de credibilidade, dificultando que usuários identifiquem imediatamente erros factuais ou inconsistências.

As alucinações podem manifestar-se de diferentes maneiras, incluindo fabricação de fatos, criação de referências bibliográficas inexistentes, associação incorreta entre entidades, respostas contraditórias e inferências incompatíveis com o contexto recuperado. Essas características demonstram que, apesar da capacidade dos modelos de linguagem em produzir textos coerentes, ainda existem limitações relacionadas à verificação factual das informações geradas.

Sob o ponto de vista técnico, esse comportamento decorre da natureza probabilística dos LLMs, cujo objetivo consiste em prever o próximo token mais provável com base nos padrões estatísticos aprendidos durante o treinamento, e não em validar a veracidade das informações produzidas. Consequentemente, quando o conhecimento necessário não está adequadamente representado em seus parâmetros ou quando o contexto fornecido é insuficiente, o modelo pode produzir respostas semanticamente coerentes, porém factualmente incorretas.

Esse problema torna-se ainda mais crítico em aplicações institucionais, acadêmicas, jurídicas, médicas e organizacionais, nas quais decisões podem ser tomadas com base nas respostas fornecidas pelos sistemas conversacionais. Dessa forma,

pesquisas recentes têm direcionado esforços para o desenvolvimento de mecanismos capazes de reduzir a ocorrência de alucinações por meio da incorporação de conhecimento externo, recuperação de documentos e utilização de estruturas semânticas, como ocorre nas arquiteturas Retrieval-Augmented Generation (RAG) e GraphRAG (Lewis et al., 2020; Edge et al., 2024).

2.3 GRAFOS DE CONHECIMENTO

Grafos de conhecimento (*Knowledge Graphs*) consistem em estruturas semânticas destinadas à representação organizada de entidades e dos relacionamentos existentes entre elas. Diferentemente das bases documentais tradicionais, nas quais as informações permanecem predominantemente em formato textual, os grafos estruturam o conhecimento na forma de nós (entidades) conectados por arestas (relacionamentos), permitindo representar explicitamente as conexões existentes entre diferentes conceitos. Segundo Ehrlinger e WöB (2016), um grafo de conhecimento pode ser compreendido como uma representação formal do conhecimento baseada em entidades, relacionamentos e semântica, favorecendo a integração e a reutilização das informações em diferentes aplicações. Complementando essa definição, Hogan et al. (2021) destacam que os grafos de conhecimento possibilitam organizar informações de maneira estruturada, facilitando consultas, inferências e a descoberta de relações semânticas entre os elementos de um domínio.

A literatura destaca a função semântica dos grafos de conhecimento da seguinte forma:

Grafos de conhecimento fornecem representações semânticas estruturadas capazes de conectar entidades e relacionamentos, possibilitando o raciocínio contextual, a integração de conhecimento e a descoberta semântica (Hogan et al., 2021, tradução nossa).

Essa forma de representação favorece consultas complexas e mecanismos de inferência que exploram não apenas informações isoladas, mas também os relacionamentos existentes entre elas. Dessa forma, torna-se possível realizar raciocínio sobre múltiplas entidades, identificar conexões indiretas e ampliar a contextualização das informações utilizadas durante processos de recuperação de conhecimento. Em comparação com bases documentais convencionais, os grafos de conhecimento permitem representar explicitamente as relações semânticas entre conceitos, tornando a recuperação das informações mais estruturada e contextualizada (Ehrlinger; WÖB, 2016; Hogan et al., 2021).

Além da organização estruturada do conhecimento, os grafos favorecem a interoperabilidade entre diferentes fontes de dados, permitindo conectar entidades pertencentes

centes a múltiplos domínios e enriquecer o contexto utilizado durante a recuperação das informações. Essa característica tem contribuído para a crescente adoção de grafos de conhecimento em aplicações de Inteligência Artificial, mecanismos de busca semântica e sistemas de apoio à decisão (Hogan et al., 2021).

Embora a construção de grafos de conhecimento exija etapas adicionais, como extração de entidades, identificação de relacionamentos, resolução de ambiguidades e atualização contínua da base de conhecimento, essa abordagem tem sido amplamente utilizada em aplicações que demandam elevado grau de organização semântica e recuperação contextualizada das informações (Ehrlinger; Wöß, 2016; Hogan et al., 2021).

2.4 GRAPHRAG

O *Graph Retrieval-Augmented Generation* (GraphRAG) representa uma evolução das arquiteturas *Retrieval-Augmented Generation* ao incorporar grafos de conhecimento ao processo de recuperação das informações utilizadas pelos Modelos de Linguagem de Grande Porte.

A arquitetura foi proposta por Edge et al. (2024), que demonstraram como a integração entre LLMs e grafos de conhecimento possibilita recuperar informações considerando explicitamente os relacionamentos existentes entre entidades, favorecendo respostas mais contextualizadas e fundamentadas.

A abordagem GraphRAG pode ser sintetizada da seguinte forma:

O GraphRAG combina os pontos fortes de grandes modelos de linguagem com grafos de conhecimento estruturados, possibilitando o raciocínio de múltiplos saltos (*multi-hop reasoning*), a recuperação de informações com contexto mais rico e respostas mais explicáveis (Edge et al., 2024, tradução nossa).

Diferentemente do RAG tradicional, cuja recuperação ocorre predominantemente por similaridade vetorial entre *embeddings*, o GraphRAG utiliza estruturas relacionais para representar o conhecimento presente na base documental. Assim, além de recuperar documentos semanticamente relevantes, o sistema pode explorar conexões explícitas entre entidades, permitindo percorrer diferentes caminhos dentro do grafo para responder consultas mais complexas.

Essa capacidade favorece mecanismos de raciocínio em múltiplos saltos (*multi-hop reasoning*), nos quais a resposta depende da combinação de diversas relações existentes entre entidades distribuídas em diferentes documentos. Como consequência, o GraphRAG amplia a contextualização das respostas, melhora a recuperação de conhecimento e reduz a dependência exclusiva da similaridade textual utilizada pelas

arquiteturas RAG convencionais (Peng et al., 2024).

Além disso, a utilização de grafos proporciona maior rastreabilidade das evidências utilizadas durante a geração das respostas, característica importante para aplicações que exigem transparência e auditabilidade. Em vez de recuperar apenas trechos de documentos semanticamente semelhantes, o sistema pode identificar quais entidades e relacionamentos participaram do processo de recuperação, aumentando a interpretabilidade do resultado obtido.

Diversos trabalhos recentes reforçam o potencial dessa arquitetura. Peng et al. (2024) apresentam um levantamento abrangente das abordagens GraphRAG e destacam sua capacidade de integrar recuperação semântica e raciocínio baseado em grafos. Guo et al. (2024) demonstram que a incorporação de grafos de conhecimento melhora a recuperação de informações estruturadas e favorece processos de geração aumentada. Han et al. (2025) evidenciam que a utilização de agentes associados a grafos amplia a eficiência da recuperação e do raciocínio em sistemas baseados em LLMs. De forma complementar, Zhang et al. (2026) exploram arquiteturas multiagentes baseadas em GraphRAG para raciocínio sobre conhecimento complexo, enquanto Papageorgiou et al. (2025) demonstram que a integração entre GraphRAG, recuperação vetorial e sistemas multiagentes produz respostas mais confiáveis e contextualizadas em aplicações de governo eletrônico.

Embora apresente benefícios expressivos relacionados à recuperação semântica, ao raciocínio contextual e à explicabilidade, a implementação do GraphRAG envolve maior complexidade arquitetural, exigindo etapas adicionais de extração de entidades, construção do grafo de conhecimento, armazenamento em bancos de dados orientados a grafos e manutenção contínua da estrutura relacional.

Em virtude dessas características, o GraphRAG tem se consolidado como uma das abordagens mais promissoras para aumentar a confiabilidade de agentes conversacionais baseados em Inteligência Artificial, especialmente em aplicações que demandam elevada precisão factual, rastreabilidade das evidências utilizadas e mitigação da ocorrência de alucinações (Peng et al., 2024; Han et al., 2025).

2.5 NEO4J E BANCOS DE DADOS EM GRAFOS

Os bancos de dados orientados a grafos constituem uma abordagem de armazenamento de informações baseada na representação explícita de entidades e dos relacionamentos existentes entre elas. Diferentemente dos bancos de dados relacionais tradicionais, que organizam os dados predominantemente em tabelas conectadas por chaves, os bancos de grafos utilizam nós (*nodes*) para representar entidades e arestas (*edges*) para representar os relacionamentos entre essas entidades, formando estruturas altamente conectadas. Essa forma de organização favorece a representação de conhecimento complexo e tem sido amplamente empregada na construção de grafos

de conhecimento utilizados por aplicações de Inteligência Artificial (Ehrlinger; Wöß, 2016; Hogan et al., 2021).

Segundo Ehrlinger e Wöß (2016), um grafo de conhecimento pode ser compreendido como uma representação estruturada de entidades e seus relacionamentos, organizada de modo a possibilitar o compartilhamento, a integração e a reutilização do conhecimento. Complementando essa definição, Hogan et al. (2021) destacam que os grafos de conhecimento representam informações de forma semântica, permitindo que sistemas computacionais realizem consultas, inferências e navegação contextual a partir das conexões existentes entre diferentes elementos de um domínio.

Grafos de conhecimento representam informações por meio de entidades e relações semânticas, possibilitando a integração, a representação e a exploração de conhecimento estruturado em diferentes domínios (Hogan et al., 2021, tradução nossa).

Essa forma de representação torna os bancos de dados orientados a grafos particularmente adequados para aplicações que exigem recuperação contextualizada de informações, raciocínio relacional e navegação por múltiplas conexões entre entidades. Diferentemente da recuperação baseada exclusivamente em similaridade textual, a utilização de grafos permite explorar explicitamente a estrutura semântica existente na base de conhecimento, favorecendo respostas mais contextualizadas e fundamentadas.

Nesse contexto, o Neo4j destaca-se como um dos principais Sistemas Gerenciadores de Banco de Dados Orientados a Grafos (SGBDG), sendo amplamente utilizado em aplicações de Inteligência Artificial, sistemas de recomendação, análise de redes e grafos de conhecimento. Sua estrutura fundamenta-se no modelo *Property Graph*, no qual nós e relacionamentos podem armazenar propriedades, possibilitando consultas complexas por meio da linguagem Cypher.

No contexto das arquiteturas GraphRAG, o Neo4j desempenha papel fundamental ao armazenar o grafo de conhecimento construído a partir da extração automática de entidades e relacionamentos presentes na base documental. Dessa forma, durante a etapa de recuperação das informações, o sistema pode consultar não apenas documentos semanticamente semelhantes, mas também percorrer conexões existentes entre diferentes entidades, favorecendo mecanismos de raciocínio relacional e recuperação contextualizada. Essa capacidade torna os bancos de grafos uma infraestrutura adequada para aplicações que demandam recuperação de conhecimento baseada em relações semânticas explícitas (Hogan et al., 2021).

Neste trabalho, o Neo4j foi utilizado como camada de persistência do grafo de conhecimento gerado automaticamente a partir dos documentos utilizados na base

experimental. Após a extração das entidades e de seus respectivos relacionamentos, essas informações foram armazenadas no banco de dados orientado a grafos, permitindo que o agente baseado em GraphRAG recuperasse conhecimento considerando tanto as entidades quanto suas conexões semânticas. Essa estratégia diferencia-se da arquitetura RAG tradicional, cuja recuperação ocorre predominantemente por similaridade vetorial entre documentos.

2.6 GOOGLE AGENT DEVELOPMENT KIT (GOOGLE ADK)

O desenvolvimento de agentes inteligentes modernos envolve a integração entre Modelos de Linguagem de Grande Porte, mecanismos de recuperação de informações, ferramentas externas e componentes responsáveis pela execução de tarefas específicas. Nesse contexto, frameworks para construção de agentes fornecem mecanismos destinados à orquestração do fluxo de execução, permitindo organizar processos de interpretação das solicitações dos usuários, recuperação de informações e geração de respostas.

O Google Agent Development Kit (Google ADK) é um framework desenvolvido pelo Google para facilitar a implementação de agentes baseados em modelos generativos, oferecendo recursos para definição de agentes, ferramentas, memória, fluxos de execução e integração com modelos da família Gemini.

No presente trabalho, o Google ADK foi utilizado como infraestrutura para implementação dos dois agentes conversacionais desenvolvidos: um baseado na arquitetura Retrieval-Augmented Generation (RAG) e outro fundamentado em GraphRAG. Em ambos os casos, o framework foi responsável pela orquestração das interações entre o usuário, o modelo Gemini e os mecanismos de recuperação das informações, garantindo que ambos os agentes utilizassem a mesma infraestrutura computacional, diferindo apenas na estratégia de recuperação do conhecimento.

A arquitetura desenvolvida integra o Google ADK ao modelo Gemini, à biblioteca LangChain, ao banco de dados vetorial utilizado pelo agente RAG e ao banco de dados orientado a grafos Neo4j empregado pelo agente GraphRAG. Essa organização permitiu realizar uma comparação controlada entre as duas arquiteturas, reduzindo interferências decorrentes da utilização de tecnologias distintas.

A utilização de agentes baseados em modelos de linguagem está diretamente relacionada aos avanços dos LLMs, que apresentam elevada capacidade de compreensão e geração de linguagem natural, mas ainda enfrentam limitações relacionadas à confiabilidade factual das respostas produzidas (Bommasani et al., 2021).

Grandes modelos de linguagem demonstram capacidades notáveis na compreensão e geração de linguagem; no entanto, ainda podem produzir informações factualmente incorretas ou sem embasamento, tornando a integração

de conhecimento externo essencial para aumentar a confiabilidade (Ji et al., 2023, tradução nossa).

Nesse contexto, o Google ADK desempenha, nesta pesquisa, o papel de camada de orquestração dos agentes conversacionais, coordenando a comunicação entre o usuário, o modelo Gemini e os mecanismos de recuperação de conhecimento. Enquanto o agente RAG recupera informações por similaridade vetorial, o agente GraphRAG consulta o grafo de conhecimento armazenado no Neo4j, permitindo que a comparação entre ambas as arquiteturas seja realizada mantendo-se constantes os demais componentes da solução desenvolvida.

2.7 TRABALHOS RELACIONADOS

O avanço das arquiteturas baseadas em *Retrieval-Augmented Generation* (RAG) e, mais recentemente, em *Graph Retrieval-Augmented Generation* (GraphRAG), tem impulsionado o desenvolvimento de pesquisas voltadas à melhoria da recuperação de informações, da fundamentação das respostas e da redução de alucinações em sistemas baseados em Modelos de Linguagem de Grande Porte (LLMs). Diversos estudos recentes demonstram que a incorporação de grafos de conhecimento possibilita representar explicitamente relações semânticas entre entidades, favorecendo mecanismos de raciocínio mais complexos e aumentando a confiabilidade das respostas geradas.

O trabalho seminal de Edge et al. (2024) apresentou a arquitetura GraphRAG, propondo a integração entre modelos de linguagem e grafos de conhecimento construídos automaticamente a partir de documentos textuais. Os autores demonstram que a utilização de comunidades semânticas, extração de entidades e recuperação baseada em relacionamentos permite produzir respostas mais contextualizadas, especialmente em consultas que exigem raciocínio envolvendo múltiplas entidades (*multi-hop reasoning*). Esse trabalho consolidou o GraphRAG como uma evolução das arquiteturas RAG tradicionais e serviu como base para pesquisas posteriores.

Na sequência, Papageorgiou et al. (2025) propuseram uma arquitetura híbrida baseada em múltiplos agentes para aplicações em governo eletrônico. A solução integra recuperação vetorial por *embeddings*, recuperação baseada em GraphRAG e busca em fontes externas na Web, sendo composta por agentes especializados responsáveis pela recuperação, validação e geração das respostas. Segundo os autores, essa combinação aumenta a rastreabilidade das evidências utilizadas, melhora a contextualização das respostas e reduz a propagação de informações incorretas em ambientes institucionais.

Para validar a arquitetura proposta, Papageorgiou et al. (2025) utilizaram documentos públicos da plataforma *European Commission Press Corner*, a partir dos quais

foi construído um grafo de conhecimento empregado durante o processo de recuperação das informações. Os resultados obtidos demonstraram que a integração entre GraphRAG, recuperação vetorial e pesquisa na Web produziu respostas mais completas e confiáveis quando comparada ao uso isolado da recuperação textual baseada em similaridade.

Além dos trabalhos aplicados, pesquisas recentes de revisão bibliográfica também evidenciam o crescimento dessa área. Peng et al. (2024) realizaram um levantamento abrangente sobre GraphRAG, identificando diferentes estratégias para construção automática de grafos de conhecimento, recuperação baseada em entidades e mecanismos de raciocínio relacional. Os autores destacam que a utilização de estruturas semânticas representa uma evolução significativa em relação ao RAG tradicional, principalmente em tarefas que exigem recuperação contextualizada de informações.

De forma semelhante, Han et al. (2025) apresentam uma revisão sobre arquiteturas Agentic GraphRAG, demonstrando que a integração entre múltiplos agentes inteligentes e grafos de conhecimento amplia a capacidade de planejamento, recuperação de informações e tomada de decisão por sistemas conversacionais. Segundo os autores, essa abordagem representa uma tendência promissora para aplicações que exigem elevado grau de confiabilidade e interpretabilidade.

Outro trabalho relevante é apresentado por Coimbra (2025), que investigou a utilização de grafos de conhecimento associados a modelos de linguagem em aplicações conversacionais. O autor destaca que a representação explícita das relações entre entidades contribui para ampliar a qualidade da recuperação das informações e reduzir inconsistências presentes em respostas produzidas exclusivamente por modelos generativos, reforçando o potencial da utilização de grafos em arquiteturas baseadas em LLMs.

Embora essas pesquisas utilizem GraphRAG ou arquiteturas semanticamente semelhantes, existem diferenças importantes em relação ao presente trabalho. O estudo de Edge et al. (2024) concentra-se na proposição da arquitetura GraphRAG; Papageorgiou et al. (2025) desenvolvem uma solução híbrida baseada em múltiplos agentes para aplicações de governo eletrônico; Han et al. (2025) investigam arquiteturas agentivas; e Coimbra (2025) explora a aplicação de grafos de conhecimento em sistemas conversacionais.

Em contrapartida, esta pesquisa possui como objetivo principal analisar a contribuição do GraphRAG para a redução de alucinações em agentes conversacionais por meio de um estudo comparativo exploratório entre duas arquiteturas distintas: uma baseada em Retrieval-Augmented Generation tradicional e outra fundamentada em GraphRAG. Para garantir a comparabilidade entre os resultados, ambos os agentes foram desenvolvidos utilizando a mesma base documental, o mesmo modelo de linguagem (Gemini), a mesma linguagem de programação (Python), o framework Google

Agent Development Kit (Google ADK) e a biblioteca LangChain, diferenciando-se exclusivamente pelo mecanismo de recuperação das informações.

Dessa forma, esta pesquisa contribui para a literatura ao avaliar, em um mesmo ambiente experimental, o impacto da utilização de grafos de conhecimento na recuperação contextual das informações, na coerência das respostas e na mitigação de alucinações, fornecendo evidências sobre as vantagens e limitações da arquitetura GraphRAG quando comparada ao RAG tradicional.

3 METODOLOGIA

Esta pesquisa caracteriza-se, quanto à sua natureza, como aplicada, por buscar desenvolver uma solução computacional voltada à redução de alucinações em agentes conversacionais baseados em Inteligência Artificial. Segundo Gil (2002), a pesquisa aplicada possui como finalidade a geração de conhecimentos voltados à solução de problemas específicos, envolvendo aplicação prática dos resultados obtidos.

Quanto aos objetivos, a pesquisa possui caráter exploratório. De acordo com Gil (2002), pesquisas exploratórias têm como finalidade proporcionar maior familiaridade com um problema, tornando-o mais explícito e possibilitando o aprimoramento de ideias ou a formulação de hipóteses. Nesse contexto, busca-se compreender como a utilização da arquitetura GraphRAG pode contribuir para a redução de alucinações em sistemas conversacionais baseados em Modelos de Linguagem de Grande Porte.

Em relação à abordagem do problema, a pesquisa adota métodos qualitativos e quantitativos. A abordagem qualitativa consistiu na análise das respostas produzidas pelos agentes conversacionais, observando aspectos como fidelidade ao documento utilizado, coerência contextual, fundamentação das respostas nas informações recuperadas e ocorrência de conteúdos não suportados pela base documental. Já a abordagem quantitativa baseou-se na mensuração dos resultados obtidos durante os testes, considerando indicadores como número de respostas corretas, respostas parcialmente corretas, respostas incorretas, ocorrência de alucinações e capacidade do agente em reconhecer a ausência de informações na base de conhecimento.

Quanto aos procedimentos técnicos, o trabalho foi desenvolvido por meio de pesquisa bibliográfica e de um estudo comparativo exploratório. A pesquisa bibliográfica forneceu a fundamentação teórica acerca dos Modelos de Linguagem de Grande Porte (LLMs), Retrieval-Augmented Generation (RAG), GraphRAG, grafos de conhecimento, bancos de dados orientados a grafos e alucinações em Inteligência Artificial. Conforme Gil (2002), a pesquisa bibliográfica é desenvolvida a partir de material já elaborado, constituído principalmente por livros e artigos científicos. Complementarmente, realizou-se um estudo comparativo exploratório entre duas arquiteturas de recuperação de informações, RAG tradicional e GraphRAG, buscando analisar suas diferenças quanto à recuperação do conhecimento e à confiabilidade das respostas

produzidas.

Nesse contexto, esta pesquisa compreendeu o desenvolvimento de dois agentes conversacionais e a comparação de seus desempenhos utilizando o mesmo conjunto de dados, o mesmo modelo de linguagem e critérios padronizados de avaliação. Conforme Gil (2002), pesquisas exploratórias possibilitam ampliar o conhecimento sobre determinado problema por meio da investigação sistemática de seus fenômenos, permitindo maior compreensão sobre o objeto estudado. Dessa forma, o desenvolvimento dos protótipos e sua avaliação comparativa permitiram analisar as contribuições das arquiteturas RAG e GraphRAG para a redução de alucinações em agentes conversacionais.

Inicialmente, foi realizado um levantamento bibliográfico acerca dos Modelos de Linguagem de Grande Porte (LLMs), do fenômeno das alucinações em Inteligência Artificial, das arquiteturas Retrieval-Augmented Generation (RAG), dos grafos de conhecimento e da abordagem GraphRAG. A busca pelos trabalhos científicos foi realizada em bases de dados e mecanismos de pesquisa acadêmica, como Google Scholar, ACM Digital Library, IEEE Xplore, SpringerLink e arXiv, priorizando artigos publicados em periódicos e conferências de relevância entre os anos de 2020 e 2026. Como principais referências, destacam-se os trabalhos de Lewis et al. (2020), Hogan et al. (2021), Ji et al. (2023), Edge et al. (2024), Papageorgiou et al. (2025), Coimbra (2025) e os levantamentos recentes sobre GraphRAG publicados na literatura científica. Essa etapa possibilitou compreender os conceitos necessários para o desenvolvimento dos protótipos e para definição dos critérios utilizados na comparação entre as arquiteturas.

Após a fundamentação teórica, foram desenvolvidos dois agentes conversacionais utilizando o Google Agent Development Kit (Google ADK). O primeiro agente foi implementado utilizando a arquitetura Retrieval-Augmented Generation (RAG), enquanto o segundo foi desenvolvido utilizando a arquitetura GraphRAG. Ambos empregaram o mesmo modelo de linguagem (Gemini), a mesma base documental e o mesmo conjunto de perguntas de avaliação. A principal diferença entre os agentes esteve no mecanismo de representação e recuperação das informações. Entretanto, durante a construção do índice vetorial do agente RAG tradicional, a quantidade de trechos processados foi limitada a 20 chunks em razão das restrições de cota da API utilizada para geração dos embeddings, enquanto o agente GraphRAG utilizou um processo distinto de extração de entidades e relacionamentos para construção do grafo de conhecimento.

Para implementação dos protótipos foi utilizado um ambiente computacional local composto por um computador desktop equipado com processador AMD Ryzen 5 7600X, 32 GB de memória RAM DDR5 operando a 6000 MHz, unidade de armazenamento SSD NVMe de 1 TB e placa gráfica NVIDIA GeForce RTX 5070 com 12 GB de

memória dedicada. O desenvolvimento foi realizado utilizando o Visual Studio Code (VS Code) como ambiente de desenvolvimento integrado (IDE), empregando a linguagem Python, ambiente virtual (*venv*), Google Agent Development Kit (Google ADK), modelo Gemini, banco de dados Neo4j, biblioteca LangChain e demais dependências necessárias para processamento dos documentos, recuperação das informações e integração entre os componentes da arquitetura.

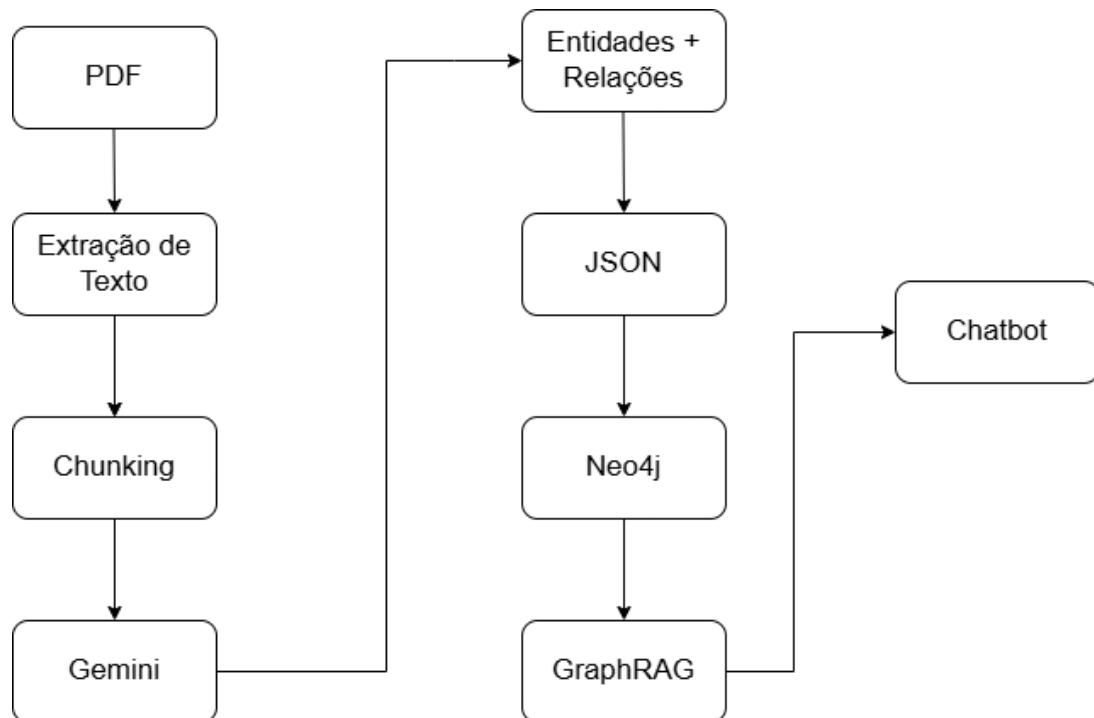
Para construção da base de conhecimento foi utilizado um documento em formato PDF contendo informações relacionadas ao GraphRAG. Inicialmente, o documento foi submetido ao processo de extração textual, seguido da segmentação em múltiplos trechos (*chunks*) utilizando sobreposição (*overlap*), preservando o contexto semântico entre segmentos consecutivos.

Na arquitetura GraphRAG, cada trecho foi processado pelo modelo Gemini para identificação automática das entidades presentes no texto e dos relacionamentos existentes entre elas. As entidades e relações extraídas foram armazenadas inicialmente em formato JSON e posteriormente inseridas em um banco de dados orientado a grafos Neo4j. Nesse banco, cada entidade foi representada como um nó e cada relacionamento como uma aresta, estruturados por meio da linguagem Cypher. Essa modelagem permitiu organizar semanticamente o conhecimento presente no documento e possibilitou consultas baseadas em relacionamentos entre entidades, conforme os princípios descritos por Edge et al. (2024).

Na arquitetura RAG tradicional, o mesmo documento foi submetido ao processo de segmentação textual e indexação vetorial, permitindo que a recuperação das informações ocorresse por similaridade semântica entre *embeddings*. Entretanto, em razão das limitações de cota da API utilizada na etapa de geração dos *embeddings*, a indexação do agente RAG tradicional foi limitada aos primeiros 20 *chunks* gerados a partir do documento. No agente GraphRAG, o processamento ocorreu por meio da extração de entidades e relacionamentos e posterior construção do grafo de conhecimento no Neo4j. Dessa forma, embora ambos os agentes tenham utilizado o mesmo documento como fonte de conhecimento, a quantidade de conteúdo efetivamente processada e a forma de representação das informações foram distintas, devendo essa limitação ser considerada na análise dos resultados.

A Figura 1 apresenta, de forma resumida, as etapas metodológicas desenvolvidas nesta pesquisa, desde a preparação da base documental até a comparação dos resultados obtidos pelos agentes conversacionais.

Figura 1 – Fluxograma das etapas de desenvolvimento e avaliação dos agentes conversacionais



Fonte: Elaborado pelo autor (2026).

Conforme ilustrado na Figura 1, o desenvolvimento da pesquisa compreendeu as etapas de levantamento bibliográfico, preparação da base documental, construção da base vetorial e do grafo de conhecimento, implementação dos agentes conversacionais e realização do estudo comparativo exploratório entre as arquiteturas RAG e GraphRAG.

Após a construção das duas arquiteturas, foram elaboradas dez perguntas padronizadas, distribuídas em três categorias: perguntas cujas respostas estavam explicitamente presentes no documento utilizado como base de conhecimento, perguntas parcialmente relacionadas ao conteúdo disponível e perguntas cujas respostas não estavam presentes no documento. Essa organização permitiu avaliar tanto a capacidade de recuperação das informações quanto o comportamento dos agentes diante de situações em que não havia evidências suficientes para responder à consulta.

Os dois agentes foram submetidos exatamente ao mesmo conjunto de perguntas, utilizando o mesmo modelo de linguagem, a mesma base documental e os mesmos parâmetros de execução. Além disso, foi utilizado o mesmo *prompt* base para ambos os agentes, alterando-se exclusivamente o mecanismo de recuperação das informações. Enquanto o agente RAG recuperava trechos do documento por similaridade vetorial, o agente GraphRAG recuperava entidades e relacionamentos armazenados no grafo de conhecimento, garantindo que a comparação refletisse apenas as diferenças entre as arquiteturas de recuperação.

As respostas produzidas foram analisadas qualitativamente considerando critérios relacionados à recuperação correta das informações, coerência contextual, fundamentação nas evidências recuperadas e ocorrência de informações não suportadas pelo documento utilizado como base de conhecimento.

Como forma de documentação dos experimentos, foram registradas capturas de tela durante a execução dos testes em ambos os agentes conversacionais. Posteriormente, os resultados foram organizados em quadros comparativos e analisados de forma conjunta, permitindo identificar diferenças entre as arquiteturas quanto à recuperação das informações, à coerência das respostas e à ocorrência de alucinações.

Com base nessa análise, foram calculados indicadores quantitativos relacionados à ocorrência de respostas corretas, respostas parcialmente corretas, respostas incorretas, reconhecimento da ausência de informações e taxa de respostas não fundamentadas. Esses resultados permitiram comparar objetivamente o desempenho das arquiteturas RAG e GraphRAG quanto à confiabilidade das respostas produzidas e à capacidade de reduzir a geração de informações não fundamentadas na base documental.

O código-fonte desenvolvido para implementação dos agentes conversacionais, bem como os scripts utilizados na construção da base de conhecimento e na realização dos experimentos, está disponível em repositório público no GitHub, permitindo a reprodução dos experimentos e contribuindo para a transparência e reprodutibilidade da pesquisa.

Repositório do projeto:

<https://github.com/vnzin-rb/Agentes>

4 RESULTADOS E DISCUSSÃO

Nesta seção são apresentados e discutidos os resultados obtidos por meio do estudo comparativo exploratório realizado entre o agente conversacional baseado em *Retrieval-Augmented Generation* (RAG) tradicional e o agente fundamentado na arquitetura *Graph Retrieval-Augmented Generation* (GraphRAG). A avaliação buscou verificar, nas condições estabelecidas nesta pesquisa, em que medida a utilização de um grafo de conhecimento contribuiu para reduzir a ocorrência de respostas não fundamentadas e ampliar a aderência das respostas às informações presentes na base documental.

A análise foi fundamentada na arquitetura de recuperação aumentada proposta por Lewis et al. (2020), na abordagem GraphRAG apresentada por Edge et al. (2024), nas discussões de Ji et al. (2023) acerca das alucinações em Modelos de Linguagem de Grande Porte e nos estudos recentes que destacam a utilização de grafos de conhecimento como mecanismo de recuperação contextual e rastreável (Hogan et al.,

2021; Papageorgiou et al., 2025).

Inicialmente, apresenta-se a construção do grafo de conhecimento utilizado pelo agente GraphRAG. Em seguida, são comparadas as respostas produzidas pelos dois agentes diante de um conjunto padronizado de dez perguntas. Por fim, são apresentadas as métricas quantitativas e a discussão dos resultados à luz da literatura, considerando as vantagens e limitações observadas no estudo.

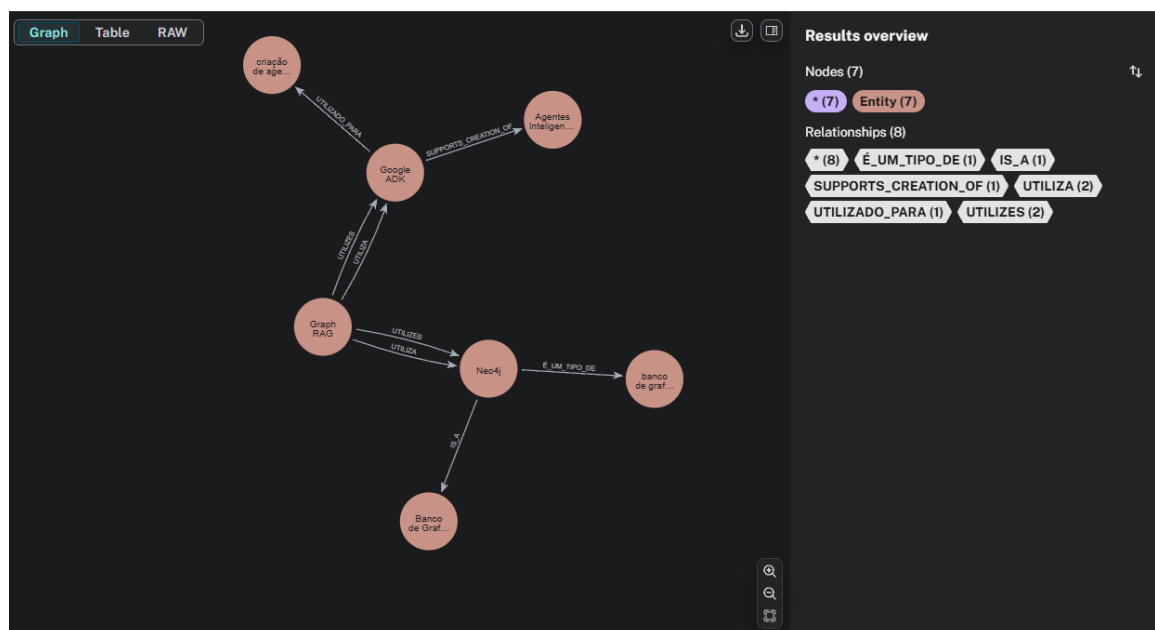
4.1 CONSTRUÇÃO DO GRAFO DE CONHECIMENTO

Após a extração das entidades e dos relacionamentos presentes no documento selecionado como base de conhecimento, realizou-se a construção automática do grafo utilizando o banco de dados orientado a grafos Neo4j. Nesse processo, cada entidade identificada foi representada por um nó, enquanto os relacionamentos extraídos foram armazenados como arestas responsáveis por conectar os respectivos elementos.

A utilização dessa estrutura permitiu representar explicitamente as conexões semânticas existentes entre os conceitos presentes no documento. Dessa forma, o agente baseado em GraphRAG passou a recuperar o contexto mediante consultas às entidades e aos relacionamentos armazenados no grafo, em vez de depender exclusivamente da similaridade vetorial entre os trechos textuais.

A Figura 2 apresenta parte do grafo de conhecimento gerado durante o desenvolvimento da pesquisa.

Figura 2 – Grafo de conhecimento construído automaticamente para utilização pelo agente GraphRAG



Fonte: Elaborado pelo autor (2026).

Conforme apresentado na Figura 2, o grafo contém entidades relevantes do domínio analisado, como *GraphRAG*, *Neo4j*, *Google ADK* e outros conceitos identificados no documento. As arestas representam os relacionamentos semânticos extraídos entre essas entidades, formando uma estrutura interconectada utilizada pelo agente GraphRAG durante o processo de recuperação das informações.

Essa organização possibilitou que o agente consultasse não apenas conceitos isolados, mas também as relações existentes entre eles. Tal característica está de acordo com Hogan et al. (2021), que destacam que grafos de conhecimento permitem representar explicitamente entidades e relacionamentos, favorecendo consultas contextuais e a exploração de conexões semânticas.

4.2 COMPARAÇÃO ENTRE O RAG TRADICIONAL E O GRAPHRAG

Para avaliar o comportamento das arquiteturas estudadas, foi realizado um estudo comparativo exploratório entre os dois agentes conversacionais desenvolvidos. O primeiro agente utilizou a arquitetura RAG tradicional, baseada na recuperação vetorial de trechos do documento, conforme os princípios apresentados por Lewis et al. (2020). O segundo agente foi implementado com a arquitetura GraphRAG, na qual as entidades e os relacionamentos extraídos do documento foram estruturados em um grafo de conhecimento armazenado no Neo4j, seguindo os fundamentos descritos por Edge et al. (2024).

Ambos os agentes utilizaram o mesmo documento como fonte de conhecimento, o mesmo modelo de linguagem, o mesmo conjunto de perguntas, os mesmos parâmetros gerais de execução e instruções equivalentes de geração. A diferença entre os agentes concentrou-se no mecanismo de recuperação do contexto. Enquanto o agente RAG utilizou a busca por similaridade vetorial, o agente GraphRAG realizou consultas às entidades e aos relacionamentos registrados no grafo de conhecimento.

Foram elaboradas dez perguntas padronizadas, distribuídas entre consultas cujas respostas estavam explicitamente presentes no documento, consultas que exigiam a associação entre conceitos e perguntas cujas respostas não estavam disponíveis na base de conhecimento. Essa organização permitiu avaliar a capacidade de recuperação das informações e o comportamento dos agentes diante da ausência de evidências documentais.

As respostas foram analisadas com base nos seguintes critérios: aderência ao conteúdo da base documental, correção da informação apresentada, coerência contextual, reconhecimento da ausência de informações e ocorrência de conteúdos não sustentados pelas evidências recuperadas. Neste trabalho, uma resposta foi classificada como alucinação quando apresentou afirmações factuais que não estavam fundamentadas no documento utilizado como fonte de conhecimento, ainda que essas afirmações pudessem ser verdadeiras em fontes externas.

O Quadro 1 apresenta a síntese dos resultados obtidos para as dez perguntas aplicadas aos dois agentes.

Quadro 1 – Comparação entre as respostas dos agentes RAG tradicional e GraphRAG

Pergunta	GraphRAG	RAG tradicional	Análise
O que é GraphRAG?	Respondeu com base nas informações recuperadas do grafo de conhecimento.	Apresentou uma resposta conceitualmente correta, acrescentando informações gerais não presentes no documento.	Ambos responderam corretamente. O GraphRAG manteve maior aderência ao conteúdo recuperado, enquanto o RAG complementou a resposta com conhecimento paramétrico do modelo.
Qual banco de dados é utilizado pelo GraphRAG neste projeto?	Respondeu corretamente que o banco utilizado foi o Neo4j.	Apresentou uma explicação genérica sobre diferentes bancos de dados utilizados em soluções GraphRAG.	O GraphRAG recuperou a informação específica do projeto. O RAG respondeu de forma coerente, porém não recuperou a informação solicitada diretamente da base documental.
Qual tecnologia é utilizada para a criação dos agentes inteligentes?	Respondeu corretamente que foi utilizado o Google ADK.	Não recuperou a informação esperada e sugeriu tecnologias como LangChain e Llamaindex.	O GraphRAG respondeu conforme o documento. O RAG apresentou informações plausíveis, porém incompatíveis com a implementação descrita no projeto.
Qual é a relação entre GraphRAG e Neo4j neste projeto?	Informou que o Neo4j foi utilizado para armazenar e consultar o grafo de conhecimento.	Explicou a relação conceitual entre GraphRAG e bancos de dados em grafos, sem se limitar à implementação descrita.	Ambos produziram respostas corretas em termos conceituais. Entretanto, o GraphRAG permaneceu aderente ao contexto do projeto, enquanto o RAG respondeu de forma mais genérica.
Quem é o autor do trabalho “Integração de RAG e GraphRAG”?	Respondeu corretamente: Vagner Marcos Coimbra.	Também respondeu corretamente: Vagner Marcos Coimbra.	Ambos os agentes recuperaram adequadamente a informação presente na base documental.
Quem criou o GraphRAG?	Informou que essa informação não estava disponível na base de conhecimento consultada.	Respondeu que o GraphRAG foi criado pela Microsoft Research.	A resposta do RAG é factualmente correta, porém foi produzida com conhecimento paramétrico do modelo e não a partir da base documental utilizada no experimento. O GraphRAG reconheceu corretamente a ausência dessa informação na fonte consultada.
Qual foi o ano de criação do GraphRAG?	Informou que a base de conhecimento não continha essa informação.	Respondeu 2024 com base no conhecimento do modelo.	Embora correta, a resposta do RAG não estava fundamentada nas informações recuperadas do documento. O GraphRAG limitou-se às evidências disponíveis na base.
Qual algoritmo específico o GraphRAG utiliza para gerar os grafos?	Informou que essa informação não estava registrada no grafo consultado.	Mencionou os algoritmos Leiden e Louvain, acrescentando informações técnicas não presentes no documento.	A resposta do RAG é majoritariamente correta, embora contenha informação adicional não fundamentada na base documental. O GraphRAG respondeu apenas com base nas evidências recuperadas.
Qual é a arquitetura interna do Google ADK utilizada neste projeto?	Informou que não havia dados suficientes para descrever essa arquitetura.	Indicou corretamente que o documento não apresentava essa informação.	Ambos os agentes reconheceram adequadamente a ausência de evidências na base documental.
Quantos usuários utilizaram o sistema GraphRAG no estudo?	Informou que a base não continha registros sobre testes com usuários.	Também informou que o documento não apresentava experimentos com usuários.	Ambos os agentes reconheceram corretamente que essa informação não estava disponível no documento, evitando apresentar dados inexistentes.

Fonte: Elaborado pelo autor (2026).

A comparação apresentada no Quadro 1 demonstra que ambos os agentes foram capazes de responder adequadamente a consultas diretamente sustentadas pelo

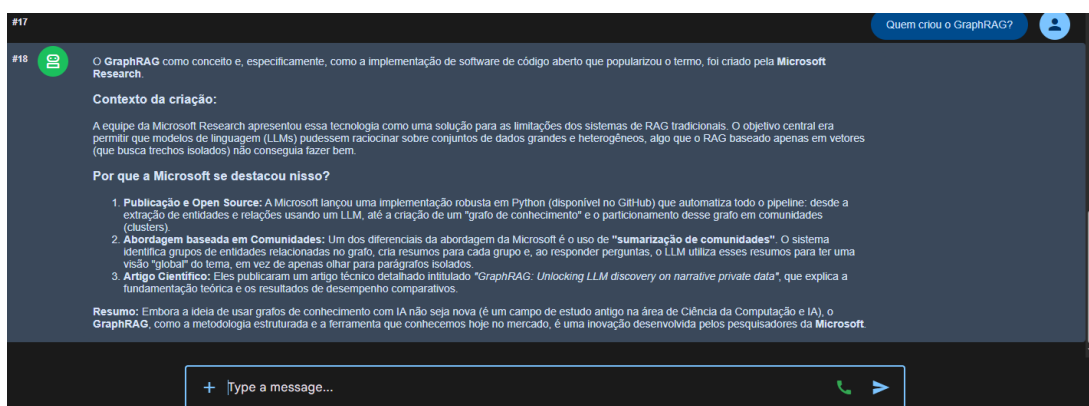
documento. Entretanto, diferenças mais expressivas foram observadas nas perguntas que exigiam relações específicas entre entidades e, principalmente, nas consultas cujas respostas não estavam disponíveis na base de conhecimento.

O agente GraphRAG apresentou maior tendência de restringir suas respostas às informações recuperadas do grafo. Quando não encontrou evidências suficientes, informou a indisponibilidade da informação. Em contrapartida, o agente RAG tradicional utilizou, em determinadas situações, o conhecimento paramétrico do modelo para complementar as respostas, mesmo quando os dados não estavam presentes no documento utilizado no estudo.

É importante destacar que uma informação externa à base documental não é necessariamente falsa. No entanto, considerando o objetivo desta pesquisa, respostas não fundamentadas no documento foram classificadas como alucinações ou extrapolações, pois não apresentaram aderência à fonte de conhecimento definida para o estudo. Esse critério permitiu avaliar a capacidade de cada agente de permanecer restrito às evidências recuperadas.

Após a análise das dez perguntas, foram selecionadas três situações representativas para demonstrar visualmente o comportamento dos agentes. As figuras apresentadas a seguir ilustram um caso em que o agente RAG utilizou conhecimento externo ao documento, um caso de reconhecimento da ausência de informação pelo GraphRAG e uma situação em que ambos os agentes responderam corretamente com base na documentação disponível.

Figura 3 – Resposta do agente RAG tradicional para a pergunta “Quem criou o GraphRAG?”



Fonte: Elaborado pelo autor (2026).

A Figura 3 apresenta uma situação em que o agente RAG tradicional forneceu uma informação que não estava presente na base documental utilizada. Embora a resposta seja factualmente correta segundo a literatura sobre GraphRAG, ela foi produzida a partir do conhecimento paramétrico do modelo de linguagem, e não das evidências recuperadas durante a consulta. Esse comportamento evidencia menor

aderência à base documental empregada no experimento, ainda que não caracterize necessariamente uma alucinação.

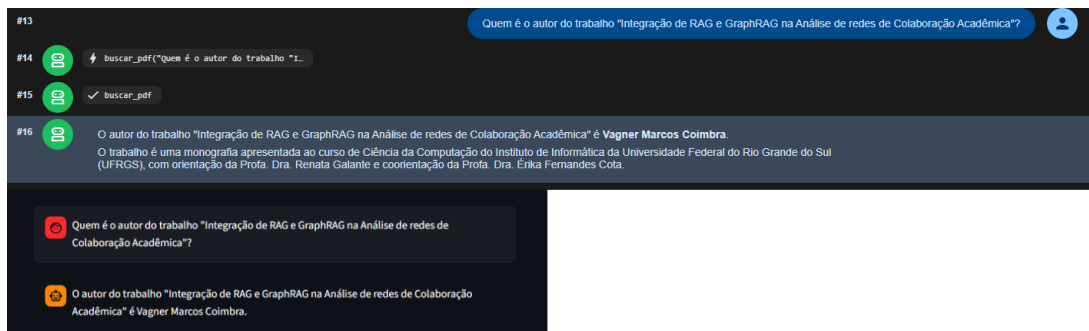
Figura 4 – Resposta do agente GraphRAG para a pergunta “Quem criou o GraphRAG?”



Fonte: Elaborado pelo autor (2026).

Em contraste, a Figura 4 demonstra que o agente GraphRAG restringiu sua resposta às informações disponíveis no grafo de conhecimento. Como não foram encontradas entidades ou relações suficientes para responder à pergunta, o agente reconheceu corretamente a ausência de evidências na base documental e evitou complementar a resposta com conhecimento externo.

Figura 5 – Comparação entre as respostas dos agentes para uma informação presente na base documental



Fonte: Elaborado pelo autor (2026).

A Figura 5 apresenta uma situação em que ambos os agentes produziram respostas corretas, uma vez que a informação solicitada estava explicitamente presente na base documental. Esse resultado demonstra que as duas abordagens são capazes de recuperar informações disponíveis na fonte, embora tenham apresentado comportamentos distintos quando as consultas envolveram informações ausentes ou exigiram maior aderência ao conteúdo documental.

4.3 MÉTRICAS DE AVALIAÇÃO DOS RESULTADOS

Além da análise qualitativa, foi realizada uma avaliação quantitativa das respostas produzidas para as dez perguntas padronizadas. As métricas consideradas buscaram analisar o grau de aderência das respostas à base documental utilizada, distinguindo respostas fundamentadas nas evidências recuperadas, respostas corretas obtidas por meio do conhecimento paramétrico do modelo, respostas incorretas e a capacidade de reconhecer corretamente a ausência de informações na base de conhecimento.

Para fins desta avaliação, foram consideradas fundamentadas na base documental as respostas que recuperaram corretamente informações presentes no documento utilizado ou reconheceram adequadamente que determinada informação não estava disponível na base de conhecimento. Também foi adotada uma categoria específica para respostas factualmente corretas, porém produzidas com conhecimento externo ao documento, caracterizando menor aderência às evidências recuperadas, embora não necessariamente configurando alucinação conforme a definição apresentada por Ji et al. (2023).

A Tabela 2 apresenta os resultados quantitativos obtidos pelos dois agentes.

Tabela 2 – Métricas de avaliação dos agentes conversacionais

Métrica	RAG tradicional	GraphRAG
Perguntas avaliadas	10	10
Respostas fundamentadas na base documental	5	10
Respostas corretas com conhecimento externo	3	0
Respostas incorretas	2	0
Reconhecimento correto da ausência de informação	3	5

Fonte: Elaborado pelo autor (2026).

Antes da análise dos resultados, destaca-se que as respostas corretas produzidas com conhecimento externo ao documento não foram classificadas como alucinações nesta pesquisa. Essas respostas foram tratadas como casos de menor aderência à base documental, uma vez que eram factualmente corretas, porém não estavam fundamentadas nas evidências recuperadas durante a consulta. Essa distinção permitiu separar a correção factual da fundamentação documental na avaliação dos agentes.

Os resultados apresentados na Tabela 2 evidenciam diferenças relevantes entre os comportamentos dos agentes avaliados. O agente GraphRAG apresentou todas as respostas fundamentadas nas evidências recuperadas da base documental, totali-

zando 10 respostas aderentes ao conteúdo disponível. Esse resultado demonstra que a utilização do grafo de conhecimento contribuiu para manter o agente mais restrito às informações recuperadas durante o processo de consulta.

Em contrapartida, o agente RAG tradicional apresentou 5 respostas fundamentadas na base documental, além de 3 respostas corretas obtidas por meio do conhecimento paramétrico do modelo e 2 respostas classificadas como incorretas. Embora algumas respostas geradas pelo RAG fossem factualmente verdadeiras, elas apresentaram menor aderência ao documento utilizado como fonte de conhecimento, pois incorporaram informações externas ao contexto recuperado.

Além disso, o GraphRAG apresentou maior capacidade de reconhecer a ausência de informações na base de conhecimento, identificando corretamente situações em que não havia evidências suficientes para responder às perguntas realizadas. Dessa forma, os resultados indicam que, nas condições avaliadas nesta pesquisa, a arquitetura GraphRAG proporcionou maior controle sobre a origem das informações utilizadas nas respostas e maior fidelidade ao contexto documental quando comparada ao RAG tradicional.

O agente RAG tradicional apresentou cinco respostas fundamentadas na base documental. Em três situações, produziu respostas factualmente corretas utilizando conhecimento paramétrico previamente adquirido pelo modelo de linguagem, embora essas informações não estivessem presentes no documento utilizado durante os experimentos. Além disso, foram observadas duas respostas incorretas relacionadas à recuperação de informações específicas da implementação desenvolvida neste trabalho.

Esse comportamento evidencia que a recuperação vetorial, embora forneça contexto adicional ao modelo de linguagem, não impede completamente a utilização de conhecimento paramétrico durante a geração das respostas. Conforme discutido por Ji et al. (2023), modelos de linguagem podem produzir respostas plausíveis e factualmente corretas sem que estejam necessariamente fundamentadas nas evidências fornecidas ao processo de recuperação.

Outro aspecto relevante foi o reconhecimento da ausência de informações. O agente RAG identificou corretamente três situações em que o documento não continha elementos suficientes para responder à pergunta, enquanto o agente GraphRAG apresentou esse comportamento em cinco situações. Esse resultado demonstra maior capacidade do GraphRAG em restringir suas respostas às informações efetivamente disponíveis na base de conhecimento utilizada no experimento.

Os resultados também se aproximam das discussões apresentadas por Edge et al. (2024), que destacam que a organização das informações por entidades, relacionamentos e comunidades semânticas amplia a contextualização e a rastreabilidade do processo de recuperação. Da mesma forma, Hogan et al. (2021) afirmam que os

grafos de conhecimento permitem representar explicitamente conexões entre entidades, favorecendo consultas fundamentadas em relações semânticas além da simples similaridade textual.

Papageorgiou et al. (2025) também identificaram benefícios relacionados à confiabilidade e à rastreabilidade ao combinar recuperação vetorial, grafos de conhecimento e agentes especializados. Embora a arquitetura apresentada pelos autores possua maior complexidade e seja direcionada a aplicações de governo eletrônico, seus resultados reforçam o potencial do GraphRAG para produzir respostas mais aderentes às evidências recuperadas.

Entretanto, os resultados desta pesquisa devem ser analisados considerando suas limitações. O estudo utilizou apenas um documento como base de conhecimento, um conjunto reduzido de dez perguntas e um único modelo de linguagem. Além disso, a classificação das respostas foi realizada pelo próprio autor, não sendo empregada uma rubrica de avaliação validada, múltiplos avaliadores independentes ou medidas de concordância entre avaliadores. Também não foram realizados testes estatísticos ou estimativas de intervalo de confiança, de modo que os resultados obtidos possuem caráter descritivo e exploratório. Assim, embora o GraphRAG tenha apresentado maior aderência à base documental nas condições avaliadas, esses resultados não permitem generalizações para outros domínios, bases documentais ou modelos de linguagem, sendo necessários estudos futuros com amostras maiores, múltiplas execuções e protocolos de avaliação mais robustos.

Dessa forma, os resultados obtidos indicam que, no cenário analisado, o agente baseado em GraphRAG apresentou maior aderência à base documental e maior capacidade de restringir suas respostas às evidências efetivamente recuperadas quando comparado ao agente RAG tradicional. A estruturação das informações em entidades e relacionamentos contribuiu para que o agente reconhecesse com maior frequência a ausência de evidências e evitasse complementar suas respostas com conhecimento externo ao documento utilizado no experimento.

4.4 PADRONIZAÇÃO DAS INSTRUÇÕES DOS AGENTES

Com o objetivo de ampliar a comparabilidade entre as arquiteturas avaliadas, os dois agentes conversacionais utilizaram a mesma instrução de sistema, o mesmo modelo de linguagem (Gemini), o mesmo documento de origem e o mesmo conjunto de perguntas. As principais diferenças estiveram nos mecanismos de representação e recuperação do conhecimento: o agente RAG utilizou indexação vetorial e recuperação por similaridade semântica, enquanto o agente GraphRAG utilizou entidades e relacionamentos armazenados em um grafo de conhecimento. Ressalta-se, entretanto, que o agente RAG apresentou uma limitação operacional durante a etapa de indexação, sendo processados 20 chunks devido às restrições de cota da API de embeddings

utilizada.

O Quadro 2 apresenta a instrução de sistema utilizada pelos dois agentes durante a realização dos experimentos.

Quadro 2 – Instrução de sistema utilizada pelos agentes conversacionais

Instrução de sistema
Você é um assistente de Inteligência Artificial especializado em responder perguntas utilizando exclusivamente o contexto recuperado durante a consulta. Responda apenas com base nas informações fornecidas pelo contexto recuperado. Não utilize conhecimento externo ou informações adquiridas durante o treinamento do modelo. Caso o contexto não contenha informações suficientes para responder à pergunta, informe explicitamente que a informação não está disponível na base de conhecimento. Não faça suposições, não complete respostas utilizando conhecimento próprio e não invente informações. Sempre priorize respostas objetivas, coerentes e fundamentadas exclusivamente nas evidências recuperadas.

Fonte: Elaborado pelo autor (2026).

O Quadro 2 evidencia que ambos os agentes utilizaram exatamente a mesma instrução de sistema durante a execução dos experimentos. Dessa forma, controlaram-se as variáveis relacionadas ao comportamento do modelo de linguagem, mantendo-se constantes o modelo utilizado, a base documental, os parâmetros de geração e o conjunto de perguntas. Dessa forma, foram mantidos constantes o modelo generativo, a instrução de sistema e o conjunto de perguntas. A comparação concentrou-se principalmente nas diferenças entre os mecanismos de recuperação utilizados por cada arquitetura. Contudo, deve-se considerar a limitação operacional existente no agente RAG tradicional quanto à quantidade de chunks indexados, aspecto que pode ter influenciado parte dos resultados observados.

5 CONSIDERAÇÕES FINAIS

Este estudo teve como objetivo desenvolver e avaliar uma aplicação de inteligência artificial conversacional utilizando a abordagem GraphRAG, analisando sua capacidade de reduzir respostas inconsistentes e alucinações em comparação com um agente baseado exclusivamente na técnica tradicional de Retrieval-Augmented Generation (RAG). A implementação realizada possibilitou a construção de um agente capaz de utilizar um grafo de conhecimento estruturado a partir das informações extraídas de documentos, permitindo a recuperação de dados considerando entidades e

relações estabelecidas entre os elementos presentes na base de conhecimento.

Os resultados obtidos durante os experimentos demonstraram que o comportamento dos agentes apresentou diferenças significativas quanto à utilização das informações recuperadas. O agente baseado em RAG tradicional apresentou bom desempenho em perguntas diretamente relacionadas ao conteúdo documental, porém, em determinados cenários, utilizou informações provenientes do conhecimento prévio do modelo de linguagem, gerando respostas que ultrapassavam os dados disponíveis na base consultada. Esse comportamento evidenciou uma das principais limitações dos sistemas conversacionais baseados apenas em recuperação vetorial.

Em contrapartida, o agente desenvolvido com GraphRAG apresentou maior controle sobre o contexto utilizado para geração das respostas, uma vez que a recuperação das informações foi realizada a partir do grafo de conhecimento construído. Durante os testes comparativos, observou-se que a abordagem foi capaz de identificar situações nas quais determinadas informações não estavam presentes na base, evitando complementar as respostas com conhecimento externo ao documento utilizado. Dessa forma, os resultados indicam que a utilização de estruturas baseadas em grafos pode contribuir para o desenvolvimento de sistemas conversacionais mais confiáveis, com maior aderência às evidências recuperadas e potencial para mitigar a ocorrência de respostas não fundamentadas.

Apesar dos resultados alcançados, algumas limitações foram identificadas durante o desenvolvimento da aplicação. A primeira está relacionada ao tamanho e à especificidade da base documental utilizada nos experimentos, sendo necessário avaliar a abordagem em conjuntos maiores de documentos e diferentes áreas de conhecimento. Adicionalmente, o agente RAG tradicional apresentou uma limitação operacional durante a etapa de indexação vetorial. Em razão das restrições de cota da API utilizada para geração dos embeddings, apenas os primeiros 20 chunks do documento foram incorporados ao índice vetorial. Essa restrição reduziu a quantidade de conteúdo disponível para recuperação pelo agente RAG e pode ter influenciado seu desempenho em determinadas consultas. Dessa forma, os resultados observados não podem ser atribuídos exclusivamente às diferenças arquiteturais entre RAG e GraphRAG, sendo recomendável que trabalhos futuros realizem novos experimentos com a indexação integral da base documental. Além disso, embora tenha sido realizada uma avaliação quantitativa baseada na classificação das respostas quanto à aderência à base documental, ao uso de conhecimento externo, à ocorrência de respostas incorretas e ao reconhecimento da ausência de informação, não foram empregadas métricas automatizadas e amplamente consolidadas na literatura para avaliação de sistemas de recuperação e geração de informações, como RAGAs (Retrieval-Augmented Generation Assessment), ROUGE (Recall-Oriented Understudy for Gisting Evaluation) e BLEU (Bilingual Evaluation Understudy). Também não foram realizados testes esta-

tísticos, múltiplas execuções ou avaliações independentes por diferentes avaliadores, de modo que os resultados obtidos possuem caráter exploratório e descritivo. Outra limitação identificada está relacionada à disponibilidade de estudos científicos amplamente consolidados sobre inteligência artificial conversacional aplicada a técnicas como GraphRAG, uma vez que esse campo de pesquisa apresenta desenvolvimento recente. Durante o levantamento bibliográfico, observou-se a dificuldade em encontrar trabalhos com alto nível de impacto e ampla validação experimental, principalmente quando relacionados especificamente à utilização de grafos de conhecimento em agentes conversacionais.

Entretanto, os resultados desta pesquisa devem ser analisados considerando suas limitações. O estudo utilizou apenas um documento como base de conhecimento, um conjunto reduzido de dez perguntas e um único modelo de linguagem. Além disso, a classificação das respostas foi realizada pelo próprio autor, não sendo empregada uma rubrica de avaliação validada, múltiplos avaliadores independentes ou medidas de concordância entre avaliadores. Também não foram realizados testes estatísticos ou estimativas de intervalo de confiança, de modo que os resultados obtidos possuem caráter descritivo e exploratório. Assim, embora o GraphRAG tenha apresentado maior aderência à base documental nas condições avaliadas, esses resultados não permitem generalizações para outros domínios, bases documentais ou modelos de linguagem, sendo necessários estudos futuros com amostras maiores, múltiplas execuções e protocolos de avaliação mais robustos.

Embora o aumento da quantidade de *chunks* disponíveis ao agente RAG pudesse ampliar sua cobertura documental e potencialmente melhorar a recuperação de informações, tal alteração não eliminaria as diferenças estruturais entre as arquiteturas. O RAG tradicional continuaria fundamentado predominantemente em similaridade vetorial, enquanto o GraphRAG manteria a capacidade de representar e percorrer relações explícitas entre entidades, característica particularmente relevante para consultas que exigem associação entre múltiplas informações. Entretanto, novos experimentos com a indexação integral do documento no agente RAG são necessários para avaliar quantitativamente o impacto dessa limitação.

Como trabalhos futuros, sugere-se a realização de experimentos utilizando bases documentais mais extensas, bem como a adoção de métricas automatizadas e amplamente consolidadas para avaliação de sistemas RAG, como RAGAs, ROUGE e BLEU, de modo a complementar a análise qualitativa realizada neste estudo. Também recomenda-se a comparação do GraphRAG com outras arquiteturas de recuperação e geração de informações, incluindo abordagens híbridas que integrem recuperação vetorial e grafos de conhecimento, visando aprimorar a precisão, a confiabilidade e a capacidade de raciocínio dos agentes conversacionais. Além disso, espera-se que o avanço das pesquisas em inteligência artificial conversacional amplie a disponibili-

dade de estudos sobre GraphRAG e técnicas relacionadas, possibilitando comparações mais abrangentes e fundamentadas, uma vez que essa área ainda se encontra em estágio de consolidação na literatura científica.

Dessa forma, conclui-se que a aplicação do GraphRAG apresenta potencial como uma alternativa para melhorar a confiabilidade de sistemas de inteligência artificial conversacional, especialmente em cenários nos quais a geração de respostas precisa manter maior aderência às informações disponíveis na base de conhecimento. Nas condições avaliadas neste estudo, o GraphRAG demonstrou maior capacidade de restringir suas respostas às evidências recuperadas e de reconhecer corretamente a ausência de informações, apresentando indícios de que essa abordagem pode contribuir para a mitigação de respostas não fundamentadas e, conseqüentemente, para a redução da ocorrência de alucinações em agentes conversacionais.

REFERÊNCIAS

BOMMASANI, Rishi et al. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*, 2021. Disponível em: <https://arxiv.org/abs/2108.07258>. Acesso em: 23 jul. 2026.

BROWN, Tom B. et al. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, v. 33, p. 1877-1901, 2020. Disponível em: <https://arxiv.org/abs/2005.14165>. Acesso em: 23 jul. 2026.

COIMBRA, Vagner Marcos. *Integração de RAG e GraphRAG para Sistemas Baseados em Modelos de Linguagem*. 2025. Monografia (Graduação em Ciência da Computação) – Universidade Federal de Uberlândia, Uberlândia, 2025.

EDGE, Darren et al. From Local to Global: A GraphRAG Approach to Query-Focused Summarization. *arXiv preprint arXiv:2404.16130*, 2024. Disponível em: <https://arxiv.org/abs/2404.16130>. Acesso em: 23 jul. 2026.

EHRLINGER, Lisa; WÖß, Wolfram. Towards a Definition of Knowledge Graphs. In: SEMANTICS POSTERS AND DEMOS TRACK, 2016, Leipzig. *Proceedings [...]* CEUR Workshop Proceedings, v. 1695, 2016. Disponível em: <http://ceur-ws.org/Vol-1695/paper4.pdf>. Acesso em: 23 jul. 2026.

GAO, Yunfan et al. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997*, 2023. Disponível em: <https://arxiv.org/abs/2312.10997>. Acesso em: 23 jul. 2026.

GIL, Antonio Carlos. *Como elaborar projetos de pesquisa*. 4. ed. São Paulo: Atlas, 2002.

GOOGLE TEAM. Gemini: A Family of Highly Capable Multimodal Models. *arXiv*

preprint *arXiv:2312.11805*, 2023. Disponível em: <https://arxiv.org/abs/2312.11805>. Acesso em: 23 jul. 2026.

GUO, Zirui et al. KRAGEN: Knowledge-Graph Augmented Generation for Information Extraction. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2024. *Proceedings [...]* EMNLP Industry Track, 2024. Disponível em: <https://aclanthology.org/2024.emnlp-industry.86/>. Acesso em: 23 jul. 2026.

HAN, Xiang et al. GEAR: Graph-Enhanced Agent for Retrieval-Augmented Generation. In: FINDINGS OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: ACL 2025, 2025. *Findings [...]* ACL, 2025. Disponível em: <https://aclanthology.org/2025.findings-acl.624.pdf>. Acesso em: 23 jul. 2026.

HOGAN, Aidan et al. Knowledge Graphs. *ACM Computing Surveys*, v. 54, n. 4, p. 1-37, 2021. DOI: 10.1145/3447772. Disponível em: <https://doi.org/10.1145/3447772>. Acesso em: 23 jul. 2026.

HUANG, Lei et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Computing Surveys*, 2025. Disponível em: <https://arxiv.org/abs/2311.05232>. Acesso em: 23 jul. 2026.

Jl, Ziwei et al. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, v. 55, n. 12, p. 1-38, 2023. DOI: 10.1145/3571730. Disponível em: <https://doi.org/10.1145/3571730>. Acesso em: 23 jul. 2026.

LEWIS, Patrick et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, v. 33, p. 9459-9474, 2020. Disponível em: <https://arxiv.org/abs/2005.11401>. Acesso em: 23 jul. 2026.

OPENAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023. Disponível em: <https://arxiv.org/abs/2303.08774>. Acesso em: 23 jul. 2026.

PAPAGEORGIOU, George et al. Hybrid Multi-Agent GraphRAG for E-Government: Towards a Trustworthy AI Assistant. *Applied Sciences*, v. 15, n. 11, p. 6315, 2025. DOI: 10.3390/app15116315. Disponível em: <https://www.mdpi.com/10.3390/app15116315>. Acesso em: 23 jul. 2026.

PENG, Bo et al. Graph Retrieval-Augmented Generation: A Survey. *arXiv preprint arXiv:2408.08921*, 2024. Disponível em: <https://arxiv.org/abs/2408.08921>. Acesso em: 23 jul. 2026.

TOUVRON, Hugo et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*, 2023. Disponível em: <https://arxiv.org/abs/2307.09288>. Acesso em: 23 jul. 2026.

VASWANI, Ashish et al. Attention Is All You Need. *Advances in Neural Information Processing Systems*, v. 30, 2017. Disponível em: <https://arxiv.org/abs/1706.03762>.

Acesso em: 23 jul. 2026.

ZHANG, Y. et al. Multi-Agent GraphRAG for Complex Knowledge Reasoning. *Scientific Reports*, v. 16, 2026. Disponível em: <https://www.nature.com/articles/s41598-026-47145-x>. Acesso em: 23 jul. 2026.