INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA DE INTELIGÊNCIA ARTIFICIAL




Sistema Conversacional Especializado em Laudos de Honorários e Deontologia Médica com Recurso a Grafos de Conhecimento

RICARDO MIGUEL PEIXOTO FARIA
outubro de 2025

P.PORTO

Sistema Conversacional Especializado em Laudos de Honorários e Deontologia Médica com Recurso a Grafos de Conhecimento

Ricardo Miguel Peixoto Faria

Aluno nº: 1201405

Dissertação para obtenção do Grau de

Mestre em Engenharia de Inteligência Artificial

Orientador: Luiz Felipe Rocha de Faria, Professor Coordenador do Instituto Superior de Engenharia do Porto do Instituto Politécnico do Porto

Júri:

Presidente:

Paulo Sérgio dos Santos Matos, Professor Adjunto do Instituto Superior de Engenharia do Porto do Instituto Politécnico do Porto

Vogais:

Diogo Emanuel Pereira Martinho, Professor Adjunto do Instituto Superior de Engenharia do Porto do Instituto Politécnico do Porto

Luiz Felipe Rocha de Faria, Professor Coordenador do Instituto Superior de Engenharia do Porto do Instituto Politécnico do Porto

Porto, setembro 2025

Resumo

Este documento apresenta o desenvolvimento e a avaliação de um sistema conversacional especializado no domínio de laudos de honorários e deontologia médica, com base na *framework* LightRAG, que combina recuperação de informação com grafos de conhecimento para mitigar alucinações. A partir de um domínio complexo, normativo e sensível, procurou-se garantir respostas factualmente sustentadas em documentos institucionais. A arquitetura implementada alinha a pergunta do utilizador com passagens recuperadas do *corpus* e com entidades e relações do grafo, o que incentiva uma geração ancorada em evidências. A avaliação do sistema conversacional recorreu a métricas semânticas, onde se observou boa cobertura temática, de 84%, e elevada recuperação do contexto e de entidades, de 93% e 92% correspondente, mas com uma precisão de recuperação e utilização parcial do contexto reduzida, de 24% e 58% respetivamente, coerentes com a utilização de modelos locais de pequena dimensão para *embeddings* e geração e de um grafo pouco denso, composto por cinquenta (50) nós e cinco (5) relações, o que corresponde a um grau médio de 0.2 e uma densidade de 0.00408. Conclui-se que esta combinação é promissora para domínios críticos, mas a sua eficiência depende da qualidade do grafo, da seletividade do recuperador e da capacidade geradora. Propõe-se, como trabalho futuro, evoluir para modelos de *embeddings* e LLM de maior dimensão e curadoria contínua do grafo, o que visa maior precisão, melhor uso do contexto e menor probabilidade de alucinação.

Palavras-chave: *Conversation Artificial Intelligence System*, RAG, Grafos de Conhecimento, LightRAG, Laudos de Honorários Médicos, Deontologia Médica

Abstract

This document presents the development and evaluation of a specialized conversational system in the domain of medical fees and ethics reports, based on the LightRAG framework, which combines information retrieval with knowledge graphs to mitigate hallucinations. From a complex, normative, and sensitive domain, we sought to ensure factually supported responses in institutional documents. The implemented architecture aligns the user's question with passages retrieved from the corpus and with entities and relationships from the graph, which encourages evidence-based generation. The evaluation of the conversational system used semantic metrics, where good thematic coverage was observed, at 84%, and high context and entity retrieval, at 93% and 92% respectively, but with reduced retrieval accuracy and partial context utilization, of 24% and 58%, respectively, consistent with the use of small local models for embeddings and generation and a sparse graph, composed of fifty (50) nodes and five (5) relations, corresponding to an average degree of 0.2 and a density of 0.00408. We conclude that this combination is promising for critical domains, but its efficiency depends on the quality of the graph, the selectivity of the retriever, and the generative capacity. We propose, as future work, to evolve towards larger embedding and LLM models and continuous graph curation, which aims at greater accuracy, better use of context, and lower probability of hallucination.

Keywords: Conversation Artificial Intelligence System, RAG, Knowledge Graphs, LightRAG, Medical Fee Reports, Medical Deontology

Índice

Lista de Figuras

Lista de Tabelas

Acrónimos e Símbolos

Lista de Acrónimos

AI	*artificial intelligence*
AI Act	regulamento europeu da inteligência artificial
CAIS	*conversation artificial intelligence system*
EL	*entity linking*
GNN	*graph neural network*
HR	*human resources*
IR	*information retrieval*
LLM	*large language model*
NLP	*natural language processing*
RAG	*retrieval-augmented generation*
RGPD	regulamento geral sobre a proteção de dados

1 Introdução

Este documento representa uma dissertação desenvolvida para o Mestrado de Engenharia de Inteligência Artificial (MEIA) do Instituto Superior de Engenharia do Porto (ISEP). O documento contém as etapas e as informações relevantes que foram necessárias para a sua conclusão.

No presente Capítulo, Introdução, é apresentado o contexto da dissertação, a interpretação do problema a resolver, bem como os principais objetivos e contributos que a solução proporciona, são levantadas e discutidas questões éticas e legais que a solução enfrenta e, por último, a exposição do documento.

1.1 Enquadramento

Este projeto resulta da colaboração entre o autor e a DataCentric[1], em resposta a um projeto sugerido pela organização, que consiste na criação de um protótipo de um *conversation artificial intelligence system* (CAIS) especializado no domínio dos laudos de honorários e de deontologia médica.

A escolha deste tema para a dissertação deve-se à relevância da *artificial intelligence* (AI), em particular da área de *natural language processing* (NLP), que conta com um crescimento na investigação e no desenvolvimento tecnológico. O estudo (Mordor Intelligence, n.d.) indica que o mercado de CAISs deve crescer a uma taxa composta de crescimento anual de 24,32% para atingir 20,81 mil milhões de dólares no ano de 2029.

A complexidade dos protocolos que regulamentam os honorários e a deontologia médica, aliada à coexistência de diferentes regulamentos aplicáveis, leva frequentemente a dúvidas por parte dos profissionais de saúde. Estas dúvidas são habitualmente dirigidas ao departamento de *human resources* (HR), o que provoca um aumento nas tarefas destas equipas e, consequentemente, mais custos operacionais que poderiam ser reduzidos. A publicação (Nádia

[1] https://www.datacentric-tech.com/

Ventura, 2022) sobre a gestão de custos dos HR, indica que o tempo despendido no esclarecimento de dúvidas dos colaboradores é um custo extra que é necessário contabilizar.

Uma solução que tem sido cada vez mais adotada pelas organizações é a implementação de CAISs, que têm como objetivo assumir a responsabilidade destas tarefas que são realizadas pelos HR. O estudo (Android Standard Blog, n.d.) indica que os CAISs podem conversar com seres humanos 69% do tempo sem intervenção humana, o que reduz as solicitações de colaboradores em quase 70%.

Um dos principais desafios enfrentados pelos CAISs é o fenômeno da alucinação, o que faz estes sistemas produzirem informações factualmente incorretas ou inventadas, o que pode comprometer a confiabilidade das respostas geradas, especialmente em contextos críticos como o da saúde. O artigo (Z. Ji et al., 2022) destaca que até 30% das respostas gerados pelos CAISs podem conter erros factuais devido à alucinação.

1.2 Descrição do Problema

Este projeto foca no desenvolvimento de um CAIS, especializado no domínio dos laudos de honorários e de deontologia médica, que oferece um suporte eficiente e contínuo aos profissionais de saúde de uma instituição de saúde, como por exemplo, hospitais ou clínicas médicas.

O CAIS é capacitado para responder, de forma rápida e precisa, a questões colocadas por estes profissionais, o que permite ao esclarecimento sobre protocolos ou regulamentos a que estão sujeitos, nomeadamente tópicos relacionados com aspetos normativos, financeiros e práticos associados ao fornecimento de cuidados e à conduta profissional.

Este sistema contribui para uma melhoria da satisfação dos profissionais de saúde em relação à instituição para a qual prestam serviços devido ao acesso fácil e rápido que passarão a usufruir sobre o esclarecimento de dúvidas que tenham. Além disso, à medida que o CAIS comece a assumir este papel, permite que os colaboradores do departamento de HR destas instituições direcionem o seu tempo e esforço para atividades que possuem maior valor acrescentado.

1.2.1 Objetivos

O objetivo principal deste projeto consiste no desenvolvimento de um CAIS que esclareça questões em volta de protocolos e regulamentos de uma instituição e da Ordem dos Médicos de forma contextualizada e factualmente verdadeira.

Além disso, o projeto foi delineado com base em objetivos específicos que guiam o desenvolvimento do CAIS, o que proporciona uma abordagem sistemática e organizada. Os objetivos são:

- Revisão da Literatura: Realização de um levantamento de diversos tipos de CAISs, para destacar as suas características, definições técnicas, benefícios e limitações;
- Análise do Domínio: Compreensão do domínio dos laudos de honorários e de deontologia médica, que abrange diferentes protocolos e regulamentos aplicáveis;
- Desenvolvimento do Protótipo: Implementação do protótipo do CAIS onde é demonstrado o fluxo do sistema, os modelos e as técnicas utilizadas;
- Avaliação: Medição do desempenho do CAIS através de métricas, cuja avaliação asseguram a eficiência do sistema e geração de respostas factualmente verdadeiras e precisas;
- Documentação e Disseminação: Elaboração do documento da dissertação com os resultados obtidos e as metodologias documentadas.

Esta estrutura de objetivos oferece uma abordagem geral, na qual aborda as questões críticas para o desenvolvimento da solução.

1.2.2 Contributos

Neste projeto, propõe-se o contributo principal a utilização de técnicas para certificar que as respostas geradas pelo CAIS estão em conformidade com os protocolos e regulamentos, garantindo a contextualização, precisão e veracidade da informação.

1.3 Aspetos Éticos e Legais

A introdução do CAIS no contexto dos laudos de honorários e deontologia médica, levanta várias questões de ordem ética e legal. Estes tipos de sistemas podem trazer avanços em eficiência e rigor, mas é necessário analisar os impactos negativos, tanto para os profissionais, como para a instituição e os próprios utentes.

Neste subcapítulo é discutido os problemas éticos e legais que podem surgiu devido à utilização do CAIS nas instituições de saúde.

1.3.1 Regulamento Europeu da Inteligência Artificial

O Regulamento (UE) 2024/1689, conhecido como o Regulamento Europeu da Inteligência Artificial (AI Act) (Parlamento Europeu e do Conselho, 2024), surge como o primeiro enquadramento legal, aplicado a todos países membros da União Europeia, dedicado à AI. O objetivo deste regulamento é garantir que os sistemas de AI sejam desenvolvidos e utilizados em conformidade com os valores europeus, nomeadamente a centralidade no ser humano, a

confiança, a transparência e a proteção da saúde, segurança e direitos fundamentais, tais como a privacidade, a não discriminação e a defesa da democracia.

O AI Act impõe um conjunto de obrigações rigorosas aos sistemas implementados na área da saúde, o qual o CAIS proposto neste documento se enquadra, devido a este tipo de sistemas serem considerados como um sistema de risco elevado. O regulamento determina que todo o ciclo de vida do sistema, desde o desenho até à utilização, deve ser documentado, transparente e sujeito a mecanismos de avaliação e mitigação de riscos. Exige-se uma supervisão humana sobre o funcionamento do sistema, a rastreabilidade das decisões tomadas, bem como a possibilidade de auditoria por parte das entidades competentes.

O AI Act complementa, sem substituir, outros regulamentos europeus já existentes, o que reforça a necessidade de articulação entre a inovação tecnológica e a salvaguarda dos direitos dos cidadãos. Desta forma, a instituição responsável por um sistema de AI deve garantir que todos os requisitos legais, técnicos e éticos estão cumpridos, sob pena de responsabilidade civil e administrativa.

1.3.2 Regulamento Geral sobre a Proteção de Dados

O Regulamento (UE) 2016/67, conhecido como o Regulamento Geral sobre a Proteção de Dados (RGPD) (Parlamento Europeu e do Conselho, 2016), constitui a principal referência sobre a proteção de dados pessoais no espaço europeu e cobre especial importância os contextos onde se tratam dados de saúde, considerados de natureza sensível.

O RGPD estabelece princípios fundamentais como a legitimidade, lealdade, transparência, limitação das finalidades, minimização dos dados, exatidão e integridade. Exige que o tratamento dos dados relativos à saúde só possa ocorrer com o consentimento explícito do titular ou em situações específicas, como por exemplo, a prestação de cuidados ou de interesse público.

É reconhecido o direito dos titulares em não serem sujeitos, exceto em circunstâncias excecionais, a decisões que impactam significativamente as suas vidas de forma exclusivamente automática, garantindo o direito à intervenção humana, à explicação e à contestação destas decisões. Para além disto, o RGPD impõe a adoção de medidas técnicas e organizativas adequadas, como a anonimização, pseudonimização, controlo de acessos e registo de operações.

1.3.3 Perda de Postos de Trabalho

Como o CAIS é utilizado diretamente pelos profissionais de saúde, existe o risco da perda de postos de trabalho, pois elimina o intermediário do departamento de HR, que têm como uma das tarefas responder às questões colocadas por estes profissionais. O que pode resultar numa diminuição de volume de trabalho deste departamento, o que por consequência, elimina parte

dos postos de trabalho necessários. O relatório *Future of Jobs Report* (World Economic Forum, 2023) estima, globalmente, que até ao ano de 2027, 83 milhões de empregos serão eliminados e 69 milhões criados devido à AI, o que resulta numa redução líquida de 14 milhões postos de trabalho. Para mitigar este risco, o relatório recomenda requalificar os colaboradores para que estes executem outras funções de maior valor acrescentado.

1.3.4 Processos Jurídicos

Caso o CAIS forneça eventuais respostas factualmente incorretas ou imprecisas, existe o risco de os profissionais de saúde processarem a instituição de saúde que possui o sistema. Estas ações jurídicas podem ter impactos financeiros e na reputação da instituição, o que torna importante a garantia de certeza e a fiabilidade do CAIS. Para minimizar este risco, podem ser utilizadas técnicas de *grounding* para garantir respostas mais factualmente verdadeiras e precisas (Lee et al., 2024).

1.3.5 Sensibilidade da Informação

O CAIS apoia-se nos protocolos e regulamentos públicos ou os da própria instituição de saúde que possui o sistema, e a sua utilização será realizada pelos profissionais devidamente autorizados pela própria, pelo que assim, não existe risco em função da sensibilidade dos dados durante a utilização do CAIS.

Estes dados pertencem à instituição e estão sob o seu controlo, o que exclui, portanto, questões de privacidade ou de desvio de informação.

1.3.6 União Europeia em Comparação com os Estados Unidos da América

Os regulamentos e leis descritas neste subcapítulo são aplicadas somente à União Europeia, sendo que países fora desta união têm regulamentos e leis próprias, o que pode tornar o desenvolvimento do CAIS facilitado ou dificultado a nível ético e legal.

Segundo o relatório *AI Index* 2025 (Stanford University, 2025), os Estados Unidos da América é o país com o maior investimento privado em AI do mundo, no total de 109,1 mil milhões de dólares.

Na seguinte Tabela 1 - Comparação entre Estados Unidos da América e União Europeia, encontra-se sintetizadas características gerais da União Europeia e dos Estados Unidos da América, o que permite, deste modo, uma comparação facilitada e sucinta. Após a tabela, encontra-se uma discussão dos pontos levantados pela mesma.

Tabela 1 - Comparação entre Estados Unidos da América e União Europeia

Características	**Estados Unidos**	**União Europeia**
Abordagem Regulamentar	Ausência de lei federal geral para AI ou dados pessoais. Regulações por setor ou estado	Unificada e transversal aplicáveis a todos os Estados-Membros
Regulação de IA	Recomendações voluntárias; Leis emergentes a nível estadual e propostas no Congresso	Regulamento vinculativo que classifica sistemas de AI por risco, impõe obrigações para sistemas de risco elevado, e prevê fiscalização e sanções
Direitos dos Titulares dos Dados	Variam por jurisdição. Nem sempre há direito de acesso, retificação ou oposição a decisões automáticas	Direitos dos indivíduos. Acesso, retificação, eliminação, limitação do tratamento, portabilidade, oposição e não sujeição a decisões exclusivamente automáticas
Supervisão e Fiscalização	Agências setoriais ou estaduais	Autoridades nacionais independentes e Comissão Europeia com poderes de fiscalização, investigação e sanção
Sanções por Incumprimento	Tipicamente multas mais baixas e dependentes de legislação setorial/estadual	Multas elevadas de até 20 milhões de euros ou 4% do volume de negócios global (RGPD). AI Act prevê sanções adicionais
Avaliação de Impacto	Não obrigatório de forma geral. Algumas exigências em setores regulados	Obrigatório para sistemas de alto risco (AI Act) e para operações de tratamento de dados sensíveis (RGPD)
Transparência e Explicabilidade	Requisitos limitados e não uniformes, depende da legislação do setor	Obrigatório para sistemas de AI de risco elevado e para qualquer decisão automática significativa
Responsabilização	Responsabilidade variável conforme o setor; Foco em autorregulação	Obrigação de demonstrar conformidade com a lei e manter registos de operações
Foco Regulatório	Inovação e competitividade, com ênfase na autorregulação	Proteção dos direitos fundamentais, confiança pública e segurança jurídica

A União Europeia possui uma filosofia de regulação direcionada aos direitos humanos, enquadrada por legislações, como o RGPD e o AI Act, com uma separação entre inovação tecnológica e proteção dos direitos individuais. Esta abordagem consiste em requisitos para avaliação de impacto, transparência, rastreabilidade e responsabilidade institucional, particularmente para sistemas aplicados em setores críticos, como o da saúde. Existem órgãos públicos com autoridade para realizar auditorias, com a opção de aplicar sanções em caso de incumprimento.

Em contraste, os Estados Unidos da América não possui um quadro legal federal abrangente para a proteção de dados pessoais ou regulação de AI. A legislação é fragmentada e setorial, sendo que a proteção de dados depende de normas estaduais ou específicas por setor. As orientações federais, como o *Blueprint for an AI Bill of Rights (White House, 2022)*, apresentam-se sobretudo como recomendações e não como obrigações legais, o que faz ficar a cargo das próprias organizações a definição de grande parte dos mecanismos de controlo e transparência. A avaliação prévia de impacto, a obrigatoriedade de auditoria externa e a responsabilização institucional são menos exigentes do que na União Europeia, o que torna a inovação e a competitividade frequentemente privilegiadas em detrimento da proteção dos utilizadores.

1.4 Estrutura do Documento

O presente relatório encontra-se estruturado em seis Capítulos distintos.

O Capítulo 1, Introdução, apresenta uma síntese do projeto, tendo por base o contexto da proposta de solução, onde são dados a conhecer o tema de investigação, o problema que levou à proposta de elaboração do mesmo, os principais objetivos e contributos a concretizar na mesma e os aspetos éticos e legais que podem ser levantados durante o desenvolvimento e a utilização do CAIS.

No Capítulo 2, Estado da Arte, apresenta a metodologia de revisão e os trabalhos relacionados. São introduzidos os conceitos fundamentais, nomeadamente os *large language models* (LLM), o fenómeno da alucinação que estes modelos enfrentam, os grafos de conhecimento e o *grounding*. Discutem-se técnicas relevantes para garantir a factualidade, com destaque para o *entity linking* (EL) e o *retrieval-augmented generation* (RAG). Analisa-se ainda um conjunto de *frameworks* para a aplicação das técnicas, com foco no LightRAG e no GraphRAG.

No Capítulo 3, Construção do Grafo de Conhecimento, descrevem-se as ferramentas analíticas consideradas para a criação de um grafo de conhecimento. Detalha-se o grafo construído, a sua modelação, entidades, relações e são estudados os componentes principais do domínio. Por fim, é referido as possíveis utilidades deste grafo para o CAIS.

No Capítulo 4, Desenvolvimento do CAIS, documenta o CAIS implementado. Começa pela plataforma de execução de modelos e inclui os critérios de escolha e as suas parametrizações. Em seguida, descreve-se o fluxo operacional do sistema, a preparação e a ingestão das fontes

de conhecimento, os mecanismos de indexação e o ciclo pergunta-resposta. São apresentados os diferentes modos de resposta e as salvaguardas implementadas para a mitigação das alucinações.

No Capítulo 5, Avaliação, aborda os resultados obtidos pelo CAIS. Avalia-se o grafo de conhecimento por métricas clássicas e compara-se os grafos gerados pela *framework* LightRAG e pela ferramenta analítica InfraNodus. É explicada a possibilidade da manipulação manual do grafo do LightRAG e possibilidade de utilizar o grafo gerado pelo InfraNodus no CAIS. Procede-se à avaliação das respostas geradas pelo CAIS com recurso à *framework* Ragas, com detalhe das métricas consideradas, o conjunto de perguntas de teste e o protocolo experimental. Por fim, os resultados são discutidos e analisados.

No Capítulo 6, Conclusão, sumaria-se o impacto das estratégias de combate à alucinação e o papel do grafo de conhecimento na obtenção de respostas factualmente verdadeiras. Reflete-se sobre a implementação realizada, com destaque para as escolhas técnicas. São também descritas as limitações encontradas, tanto metodológicas como tecnológicas. O capítulo termina com propostas de trabalho futuro, quer ao nível da expansão do domínio, quer da melhoria contínua de modelos, métricas e mecanismos de verificação factual.

2 Estado da Arte

No presente Capítulo 2, Estado da Arte, é apresentado o estado da arte que sustenta o desenvolvimento do CAIS. É exposta a metodologia de pesquisa para a revisão da literatura e sintetizados os trabalhos relacionados. Nos conceitos base, é abordado os LLM, e são explorados os grafos de conhecimento e o *grounding* para combater o problema da alucinação que estes modelos enfrentam. Por fim, são discutidas as técnicas de RAG e de EL e das *frameworks* para a aplicação das técnicas mencionadas.

O objetivo é estabelecer os critérios técnicos para garantir contextualização, precisão e veracidade das respostas geradas pelo CAIS no domínio dos laudos de honorários e deontologia médica.

2.1 Metodologia

A metodologia corresponde às estratégias e aos procedimentos utilizados para alcançar os objetivos do estudo. Assim, são descritos os métodos usados para a seleção e a análise da literatura correspondente, descrevendo-se as perguntas de pesquisa, os termos de pesquisa e os critérios de inclusão e de exclusão adotados para assegurar a relevância e a qualidade das referências encontradas.

2.1.1 Questões de Investigação

A questão principal que guiou a criação das questões de investigação foi: "Qual é o atual estado da arte dos CAISs aplicados ao esclarecimento de dúvidas?". Para responder a esta questão, a literatura mais recente deve ser revista de acordo com três questões principais, representadas na Tabela 2 - Questões de Pesquisa.

Tabela 2 - Questões de Pesquisa

ID	Questão de Pesquisa
QP1	Quais são os principais tipos de fontes e estudos relacionados com os CAISs para o esclarecimento de dúvidas
QP2	Quais são os principais desafios e benefícios associados à utilização de CAISs para o esclarecimento de dúvidas
QP3	Quais são as principais funcionalidades e caraterísticas implementadas nos CAISs para o esclarecimento de dúvidas

Na primeira questão, os principais tipos de fontes de estudos são revistos para compreender quão diversos são estes estudos em termos de terem sido publicados numa revista, em conferências ou noutro canal de publicação. Além disso, esta questão pode também permitir compreender o impacto de cada publicação.

Na segunda questão, os principais benefícios e desafios identificados pelos autores de cada trabalho selecionado são revistos para compreender quais são as principais limitações e obstáculos atuais e as caraterísticas mais vantajosas a considerar para desenvolver com sucesso o CAIS.

Na terceira questão, as principais funcionalidades e caraterísticas são analisadas para identificar que tipo de tecnologias e técnicas é mais frequentemente consideradas no desenvolvimento de CAISs.

2.1.2 Termos de Pesquisa

As palavras-chave permitem selecionar artigos com temas que se alinham diretamente com o propósito do estudo. Através da utilização de base de dados de pesquisa, como por exemplo o IEEE, efetuaram-se pesquisas por intermédio da combinação de diversos termos, obtendo-se, assim, resultados pertinentes e variados.

A Tabela 3 - Consulta de Pesquisa compreende as palavras-chave alocadas por domínio e a pesquisa que foi realizada. Estas pesquisas foram elaboradas para incluir as variações linguísticas e sinónimos, de forma a melhor expandir os seus resultados.

Tabela 3 - Consulta de Pesquisa

Domínio	*Query* de Pesquisa
HR	(Human Resources OR HR OR Interview OR Business OR Informative) AND
CAISs	(Conversational System OR Conversational Agent OR Virtual Assistant OR Virtual Agent OR Chatbot)

2.1.3 Critérios de Elegibilidade

Os critérios de elegibilidade são utilizados para fazer a filtragem da literatura encontrada, o que influencia diretamente a seleção final da literatura pertinente aos objetivos do estudo. Estes critérios dão à pesquisa uma base em literatura de qualidade dentro do âmbito definido e exclui artigos irrelevantes e redundantes.

Na Tabela 4 - Critérios de Inclusão, são apresentados os critérios de inclusão, enquanto na Tabela 5 - Critérios de Exclusão são apresentados os critérios de exclusão utilizados neste estudo para definir a pertinência e adequabilidade dos artigos identificados, que conduziram a análise crítica da literatura.

Tabela 4 - Critérios de Inclusão

ID	Critério de Inclusão
CI1	A fonte deve focar na área de esclarecimento de dúvidas
CI2	A fonte explora CAISs ou equivalentes
CI3	A fonte identifica contributos, benefícios ou desafios

Tabela 5 - Critérios de Exclusão

ID	Critério de Exclusão
CE1	A fonte é anterior de 2015
CE2	A fonte não é em português ou inglês
CE3	A fonte está protegida por uma *paywall*

2.2 Trabalho Relacionados

A investigação do uso de CAISs em contextos de esclarecimentos tem vindo a existir progressos significativos. A seguir, apresenta-se os principais objetivos e contributos de cada trabalho analisado.

2.2.1 CrowdAssistant: A Virtual Buddy for Crowd Workers

No artigo (Abhinav et al., 2018) é explorado o desenvolvimento de um CAIS para os funcionários de plataformas de *crowdsourcing*. Este sistema tem como objetivo o auxílio na seleção de tarefas a pesquisar, na sugestão de taxas de pagamento, na recomendação de habilidades para desenvolvimento profissional e no fornecimento de orientação de carreira com a intenção de oferecer um suporte personalizado e contribuir para a experiência dos funcionários.

O desenvolvimento do CAIS foi projetado com NLP para interpretar as consultas dos funcionários e acionar os módulos específicos, como por exemplo, a seleção de tarefas. Algoritmos de aprendizagem supervisionada e processos de decisão de *markov* foram utilizados para prever as trajetórias de carreira e personalizar as recomendações com a utilização dos históricos dos utilizadores.

2.2.2 Conversational AI: Chatbots

No artigo (Meshram et al., 2021) é estudado o desenvolvimento dos CAISs, que se pontuam no tempo como sistemas baseados em regras até sistemas que empregam aprendizagem de máquina. É apresentado os setores onde estes sistemas podem ser usados, como na educação, saúde e negócios, e com a automatização das interações e a redução da dependência de interações humanas.

Os sistemas baseados em regras são apresentados como um método primordial, por meio de fluxos predefinidos para interações simples. CAISs mais desenvolvidos combinaram NLP com aprendizagem de máquina, que utiliza redes neuronais para aprendizagem contínua, o que torna possível, interações cada vez mais adaptáveis e personalizadas.

2.2.3 HR-Based Chatbot Using Deep Neural Network

No artigo (Rachana et al., 2022) é explorado o desenvolvimento de um CAIS com dois objetivos principais, sendo o primeiro a facilitação da comunicação entre os funcionários e o departamento de HR, através da automatização de respostas a perguntas frequentes sobre políticas de licenças, benefícios e cultura organizacional, e o segundo objetivo o auxílio na triagem de candidatos e na marcação das entrevistas destes. O principal contributo deste sistema é a redução da carga de trabalho do HR e a melhoria da experiência dos funcionários devido ao sistema estar constantemente disponível.

No desenvolvimento foram utilizadas redes neuronais profundas para o processamento das perguntas levantadas pelos funcionários e com isto a identificação de padrões nas consultas. Para a interpretação e categorização nas consultas, foram aplicadas técnicas de NLP, o que garantiu respostas mais precisas e relevantes para os funcionários.

2.2.4 Interview Bot for Improving Human Resource Management

No artigo (Suakanto et al., 2021) é explorado o desenvolvimento de um CAIS com o principal objetivo a automatização das entrevistas de emprego devido à substituição dos processos manuais tradicionais. O sistema permite a realização de entrevistas estruturadas, o armazenamento das respostas numa base de dados e a criação e disponibilização de relatórios analíticos para o departamento de HR. O principal contributo deste sistema é a padronização dos processos de recrutamento, a redução de custos e na possibilidade oferecida aos

candidatos de realizarem as entrevistas de forma conveniente, independentemente de restrições de tempo e localização.

No desenvolvimento foram aplicadas redes neuronais recorrentes para interpretar sequências textuais, o que permitiu a análise de respostas de maneira lógica e consistente, o que tornou possível o ajuste das perguntas com base nas suas respostas com base em técnicas de NLP.

2.2.5 RAG-Based LLM Chatbot Using Llama-2

No artigo (Vakayil et al., 2024) é explorado o desenvolvimento de um CAIS voltado para o apoio a vítimas de assédio sexual. O sistema é capaz de fornecer respostas empáticas e informações relevantes sobre leis e recursos locais, o que auxilia estas a encontrarem um suporte adequado. O principal contributo destaca-se pela sua abordagem sensível e confidencial, e no direcionamento das vítimas para ajuda profissional quando necessário.

No desenvolvimento foi aplicado o Llama-2, baseado em *transformers*, o que permite a geração de respostas contextualmente precisas e empáticas. A técnica de RAG foi aplicada para realizar a integração de informações atualizadas nas respostas, através de *embeddings* gerados pelo modelo *Bidirecional Encoder Representations from Transformers* para recuperação de informações relevantes o que permitiu ao sistema atingir uma precisão de 95%.

2.3 Conceitos Base

No presente subcapítulo, é realizado uma explicação dos LLMs e do conceito de alucinação que representa um problema enfrentado por estes modelos. Além disso, são expostos e discutidos dois conceitos base utilizados em diversas técnicas existentes para mitigar este problema.

2.3.1 LLM

Os LLMs são redes profundas treinadas em grandes *corpus*, capazes de reconhecer padrões, extrair informações, resumir, prever e gerar texto a partir da modelação da probabilidade de sequências linguísticas. À medida que escalam em dados e parâmetros, estes modelos passam a exibir capacidades emergentes não observadas em modelos menores, o que os torna solucionadores generalistas para uma gama ampla de tarefas de NLP (Zhao et al., 2025).

2.3.1.1 Arquitetura Geral do LLM

A arquitetura geral de um LLM pode ser dividia em seis (6) componentes principais, como ilustrada na Figura 1 - Componentes Principais do LLM, para garantir o funcionamento correto (Naveed et al., 2024).

Figura 1 - Componentes Principais do LLM

A tokenização é uma etapa onde o texto é segmentado em unidades mínimas chamadas de *tokens*, que funcionam como os blocos básicos de entrada para o modelo.

A codificação posicional consiste em mecanismos de codificação posicional para representar a posição das palavras no texto, sendo que estas codificações podem ser absolutas, relativas ou aprendidas.

O mecanismo de atenção permite ao modelo atribuir diferentes pesos aos *tokens* de acordo com a sua relevância no contexto, priorizando as partes mais informativas da sequência.

As funções de ativação determinam como as informações não-lineares são processadas nos LLMs.

A normalização de camadas é aplicada para estabilizar e acelerar o processo de treinamento. Em arquiteturas modernas de LLMs, a prática mais comum é a normalização pré-camada, aplicada antes do bloco de *multi-head attention*.

O pré-processamento dos dados envolve várias estratégias para melhorar a qualidade do *corpus*. Algumas destas técnicas são a filtragem baseada em heurísticas ou classificadores, a remoção de duplicados para evitar memorização excessiva e a exclusão de informações pessoais para a proteção dos dados.

2.3.1.2 Casos de Uso

Os LLMs podem ser utilizados para três (3) casos de uso distintos. Abordagens clássicas de NLP, como por exemplo, etiquetagem, extração de informação e geração. Abordagens de *information retrieval* (IR) onde os LLMs atuam como modelos de *ranking* ou aprimoram procuradores existentes. Por fim, abordagens de recomendação, seja como recomendadores *few-shot* ou como motores para enriquecer modelos tradicionais (Zhao et al., 2025).

2.3.2 Alucinação

Alucinação refere-se à produção de texto por parte do CAIS que é infiel ou desconectado de uma fonte de conhecimento, podendo ser incoerente, incorreto ou inventado. No contexto dos LLMs, a geração de conteúdo alucinado ocorre com frequência e afeta várias tarefas. Esta limitação é preocupante, sobretudo em aplicações críticas, como na área da saúde, onde erros factuais podem pôr em risco a segurança dos utilizadores (Z. Ji et al., 2022).

2.3.2.1 Alucinação Intrínseca e Extrínseca

A alucinação intrínseca ocorre quando a saída contradiz diretamente a fonte, por exemplo, um resumo gerado afirmar uma data ou facto que está incorreto face ao texto original. A alucinação extrínseca corresponde à inclusão de informação que não pode ser confirmada nem refutada pela fonte, sendo externa ao contexto (Z. Ji et al., 2022).

2.3.2.2 Métodos de Mitigação

Dentro de vários métodos possíveis para mitigar a alucinação, destacam-se a construção manual ou automática de *datasets* mais fiéis, o aumento da entrada com informação externa relevante, melhorias arquiteturais em *encoders* e *decoders*, e a introdução de mecanismos de controlo no processo de geração.

Métodos de estruturação da informação como grafos de conhecimento ou métodos de recuperação de informação como o RAG, também podem ser aplicados aos LLMs para combater este fenómeno.

Outro destaque são métodos de treino mais sofisticados, como *reinforcement learning*, *multi-task learning* ou técnicas de geração controlada, bem como o pós-processamento das saídas (Z. Ji et al., 2022).

2.3.2.3 Desafios Atuais e Futuro

Persistem vários desafios, como a avaliação fina de alucinações numéricas ou em textos longos, a distinção automática dos diferentes tipos de alucinação e a capacidade de ajustar o nível de alucinação conforme as exigências de cada contexto.

As tendências atuais apontam para o desenvolvimento de métricas mais robustas, integração de sistemas automáticos de verificação factual e a melhoria do raciocínio dos modelos para aumentar a fiabilidade e a transparência dos sistemas de NPL (Z. Ji et al., 2022).

2.3.3 Grafos de Conhecimento

Os grafos de conhecimento são uma representação estruturada da informação em formato de grafo. É composto por dois elementos, os nós que representam as entidades ou conceitos da vida real, e pelas arestas que conectam os nós entre eles e representam as suas relações ou associações (S. Ji et al., 2022).

Devido à informação estruturada e navegável, é possível fazer a utilização desta representação para combater a alucinação dos LLMs.

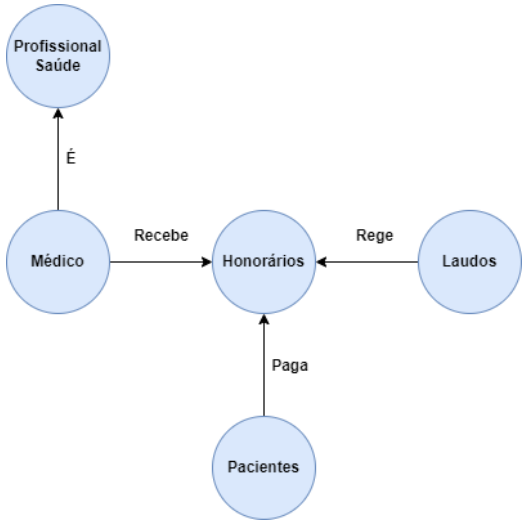


Figura 2 - Exemplo de Grafo de Conhecimento

Como ilustrado na Figura 2 - Exemplo de Grafo de Conhecimento, encontra-se um exemplo básico aplicado a um grafo, em que os nós representam conceitos, como as entidades, e estão interligados por arestas, que explicitam a sua ligação com os outros conceitos.

De acordo, um médico (entidade), é (associação) um profissional de saúde (entidade), e o médico recebe (relação) honorários (entidade) que são regidos (relação) pelos laudos (entidade) e pagos (relação) pelos pacientes (entidade).

2.3.3.1 Construção o Grafo de Conhecimento

A construção de um grafo de conhecimento pode ser composta por quatro etapas básicas. A aquisição de dados, o reconhecimento de entidades, a extração de relações e por fim, a criação do grafo de conhecimento (Bai et al., 2025).

Na aquisição de dados é realizada a recolha de dados textuais relevantes de diversas fontes, que podem incluir livros, artigos ou documentos específicos de um domínio.

No reconhecimento de entidades são identificadas as entidades presentes nas fontes adquiridas.

Na extração de relações é analisada a estrutura sintática e semântica das frases para extrair ligações significativas, o que se designa por extração de relações.

Por fim, na criação do grafo de conhecimento é criado a estrutura em grafo em que as entidades são nós e as relações são arestas. Este passo envolve a organização da informação extraída numa representação em grafo coerente.

2.3.3.2 Benefícios e Limitações dos Grafos de Conhecimento

Os benefícios desta representação encontram-se na estruturação do conhecimento, o que possibilita a realização de inferências e a organização dos fatos em estruturas interligadas, e desta forma, facilita o raciocínio lógico sobre dados complexos (Liang et al., 2024).

Todavia, a utilidade do grafo pode ficar comprometida caso a informação esteja incompleta ou seja vaga, devido às poucas ligações entre os nós que irão existir, e ainda devido ao facto da manutenção deste pode ser complexa ou morosa, caso a informação esteja em constante mudança (Pan et al., 2024).

2.3.4 Grounding

O *grounding* refere-se à capacidade do modelo de produzir respostas que estejam fundamentadas num determinado contexto ou fonte de informação externa, o que evita o modelo recorrer ao seu próprio conhecimento interno, que pode ser desatualizado ou irrelevante para o caso em questão (Lee et al., 2024).

2.3.4.1 Evolução do *Grounding*

Tradicionalmente, estudos como o (Weller et al., 2024) definem que um modelo está *grounded* se este gera respostas relevantes e corretas com base no contexto fornecido.

No entanto, o estudo (Lee et al., 2024) argumenta que esta definição é demasiado permissiva, pois permite que as respostas falhem ao omitir informações essenciais do contexto ou incluam informações que não foram explicitamente fornecidas.

Os autores propõem uma definição mais rigorosa, onde um modelo só está verdadeiramente *grounded* se utilizar toda a informação essencial presente no contexto externo fornecido e se limitar estritamente a este contexto, sem adicionar conhecimento extraído dos seus próprios parâmetros ou de fontes não explícitas.

2.3.4.2 Importância para os CAISs

A implementação eficaz do *grounding* é fundamental para a confiança, transparência e utilidade dos CAISs, principalmente em setores críticos como o da saúde. Quando um CAIS é capaz de fazer *grounding* rigorosamente num contexto dado, seja um documento ou um conjunto de políticas internas ou uma informação recém-obtida, aumenta significativamente a fiabilidade das respostas geradas.

Além disso, o *grounding* permite uma maior personalização, pois o CAIS pode adaptar-se a informações específicas, por exemplo, a diferentes instituições de saúde, integrando novos dados de forma eficiente e segura (Weller et al., 2024).

2.3.4.3 Benefícios e Limitações dos Grafos de Conhecimento

Os benefícios do *grounding* consistem no aumento da confiança dos utilizadores, ao reduzir a incidência de respostas inventadas ou alucinadas. O facto de cada afirmação poder ser rastreada até uma fonte concreta facilita a validação de informação e reduz os riscos associados à desinformação, o que torna os sistemas mais apropriados para domínios críticos, como a saúde. Além disso, o *grounding* promove a transparência e a auditabilidade dos modelos, o que facilita a explicação das decisões do sistema perante utilizadores ou entidades reguladoras. A utilização também favorece a atualização dinâmica dos CAISs, dado que o conhecimento relevante pode ser incorporado e atualizado de acordo com a evolução das fontes de conhecimento utilizadas sem depender exclusivamente do conhecimento armazenado nos parâmetros do modelo.

No entanto, o *grounding* apresenta algumas limitações, como na dificuldade em garantir que todas as partes relevantes do contexto externo são corretamente utilizadas pelo CAIS, especialmente quando o contexto é extenso, complexo ou heterogéneo. Os sistemas podem, por vezes, ignorar detalhes críticos, o que leva a respostas incompletas ou descontextualizadas. Outra limitação prende-se com a tendência dos modelos em recorrer ao conhecimento memorizado durante o pré-treino, mesmo quando existem informações contraditórias no contexto fornecido, o que pode comprometer a fidelidade da resposta. A dependência de fontes externas implica também desafios técnicos na integração eficiente e rápida dessas fontes, sobretudo em ambientes com informação dinâmica ou não estruturada. Por fim, há uma limitação inerente à avaliação automática do *grounding*, muitas métricas existentes não conseguem captar nuances subtis de fidelidade factual ou distinguir entre informação verdadeiramente fundamentada e respostas apenas plausíveis ou relevantes, mas não explicitamente suportadas pelo contexto (Lee et al., 2024; Weller et al., 2024).

2.4 Técnicas

No presente subcapítulo, são expostas e discutidas duas técnicas para a aplicação de grafos de conhecimento e de *grouding*, respetivamente, para o combate à alucinação dos LLMs. É realizada a descrição da técnica, o seu funcionamento e os benefícios e as limitações que as técnicas podem apresentar. Por fim, será realizada uma comparação entre as técnicas, exposta sobre a forma de tabela, seguida de uma discussão.

2.4.1 Entity Linking

O EL é uma técnica para associar menções de entidades num texto livre, que podem ser ambíguas e ter múltiplos significados, aos seus correspondentes exatos numa base de conhecimento estruturada.

Esta associação permite dotar os CAISs com uma compreensão semântica mais profunda e precisa do texto, o que facilita operações como desambiguação de nomes, enriquecimento de conhecimento, e respostas factualmente corretas a perguntas do utilizador (Shen et al., 2015).

2.4.1.1 Arquitetura Geral do EL

O EL trata-se de uma técnica que permite aos CAISs identificar e associar, de forma automática, menções ambíguas em texto livre às suas correspondentes entidades numa base de conhecimento estruturada.

Como ilustrado na Figura 3 - Arquitetura Geral do EL, o EL é um processo sequencial constituído por três (3) componentes, onde é iniciado pela introdução de um texto de entrada no componente de deteção de menções, seguido pelos componentes dedicados à geração e à desambiguação de candidatos, o que resulta na ligação final das entidades corretas ao texto de origem (Ganea & Hofmann, 2017).

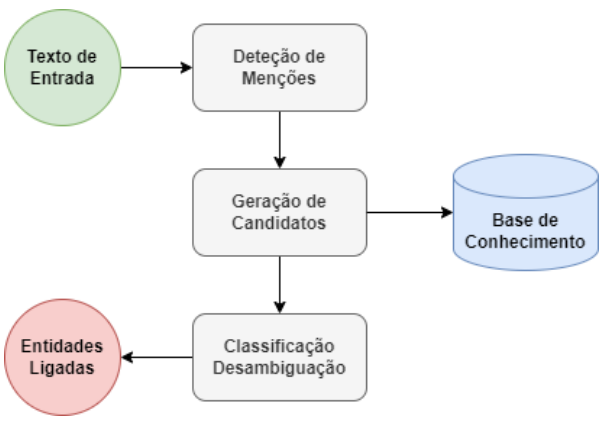


Figura 3 - Arquitetura Geral do EL

O primeiro componente consiste na deteção de menções, onde o sistema identifica no texto livre as expressões que potencialmente referem entidades, como nomes próprios, locais ou organizações. Por exemplo, num texto sobre saúde, menções como "Médico" ou "Hospital" são identificadas como possíveis referências a entidades relevantes.

A seguir, ocorre a geração de candidatos, componente na qual, para cada menção detetada, o sistema pesquisa numa base de conhecimento todas as entidades que possam corresponder àquela menção. Esta pesquisa baseia-se em dicionários de nomes, variantes ortográficas e outros metadados que ligam o texto a diferentes entidades possíveis.

No terceiro componente, realiza-se a classificação e a desambiguação dos candidatos. Neste componente, o sistema avalia, com base no contexto local e global do texto, qual das entidades candidatas é a mais provável para cada menção. Algoritmos de classificação analisam fatores como similaridade contextual, frequência e coocorrência de entidades.

Por fim, o resultado são entidades ligadas, ou seja, a atribuição final das menções do texto às entidades específicas da base de conhecimento. Este processo permite transformar texto não estruturado em informação semântica enriquecida, tornando possível a validação de factos, a agregação de conhecimento e o suporte a tarefas avançadas como pesquisa semântica e sistemas de resposta a perguntas (Kolitsas et al., 2018).

2.4.1.2 Benefícios e Limitações do EL

Os benefícios desta técnica consistem na capacidade de enriquecer bases de conhecimento automaticamente, melhorar a precisão de motores de busca semântica e suportar aplicações de pergunta-resposta mais inteligentes. Do ponto de vista prático, o EL permite resolver ambiguidades em grande escala, agregar e cruzar informações dispersas, e contribuir para a visão da *Web* Semântica, o que torna os dados mais estruturados e interligados (Shen et al., 2015). Com os avanços dos modelos neuronais, especialmente as abordagens ponta a ponta que combinam a deteção de menções e desambiguação num único modelo, tornou-se possível lidar melhor com contextos complexos e adaptar sistemas a domínios específicos, sem depender excessivamente de *features* desenhadas manualmente (Kolitsas et al., 2018).

No entanto, esta técnica depende da qualidade e cobertura da base de conhecimento utilizada, entidades ausentes ou mal representadas dificultam o *linking* correto, o que afetará negativamente a precisão do sistema. A dependência de grandes volumes de texto anotado pode ser um obstáculo para domínios especializados. Além disso, a introdução de relações artificiais durante processos de densificação pode adicionar ruído ao grafo, e existe um custo computacional não negligenciável associado ao treino e inferência em sistemas avançados de EL. Por fim, como mostra (Shen et al., 2015), é improvável que exista uma abordagem universalmente dominante, diferentes técnicas e *features* podem ter desempenhos muito variáveis conforme o tipo de dados, o domínio e os objetivos da aplicação (Ganea & Hofmann, 2017).

2.4.2 RAG

O RAG é uma técnica que permite aos modelos generativos terem a mesma capacidade de um sistema de *information retrieval* (IR). Esta técnica realiza uma modificação nos LLMs na parte das interações (Gao et al., 2023).

Nos modelos de NLP tradicionais, as respostas são geradas apenas com base em padrões e informações pré-aprendidas durante a fase de treino. No entanto, estes modelos são inerentemente limitados pelos dados com que foram treinados, o que conduz frequentemente a respostas que podem carecer de profundidade ou de conhecimentos específicos (Mousavi et al., 2025).

O RAG aborda esta limitação através da extração de dados de fontes externas conforme necessário durante o processo de geração da resposta. Quando é efetuada uma consulta, o CAIS começa por recuperar informações relevantes de um conjunto de dados ou de uma base de conhecimento, onde estas informações são utilizadas para orientar a geração da resposta (Gao et al., 2023).

2.4.2.1 Arquitetura Geral do RAG

O RAG trata-se de uma técnica que permite que os LLMs acedam dinamicamente a fontes externas de informação durante a geração de respostas (Hoang Tran, 2023).

Como ilustrado na Figura 4 - Arquitetura Geral do RAG, o RAG é um processo constituído por dois módulos principais, onde é iniciado por uma consulta inserida no modulo recuperador, seguido pelo modulo gerador que tem como saída uma resposta (Goku Mohandas & Philipp Moritz, 2023).

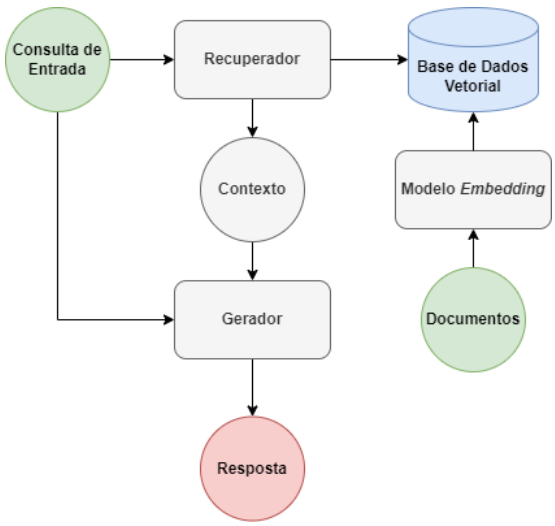


Figura 4 - Arquitetura Geral do RAG

O fluxo é iniciado com uma qualquer entrada à qual se pretenda que o CAIS responda, pode corresponder a uma pergunta ou a uma solicitação.

O recuperador, que é um modelo de *embedding*, converte a consulta de entrada num vetor e utiliza-a para pesquisar numa base de dados vetorial que contém vetores pré-computadores de contextos potenciais que foram gerados durante a fase de indexação, e assim, encontrar documentos e/ou informações relevantes que possam ser úteis para gerar uma resposta, de forma a responder à consulta inicial.

Os contextos que foram recuperados são passados para o gerador, que é um LLM. Estes contextos recuperados contêm a informação que o modelo necessita para gerar uma resposta.

O gerador utiliza a consulta de entrada e os contextos recuperados para gerar uma resposta baseada no seu conhecimento pré-existente e aumenta-a com detalhes específicos dos dados recuperados.

O fluxo é terminado com a geração da resposta que é agora enriquecida pelos dados externos recuperados no processo, tornando-a mais factualmente verdadeira e precisa.

2.4.2.2 Benefícios e Limitações do RAG

O benefício desta técnica consiste em fornecer conhecimento adicional aos LLMs, o que aumenta a qualidade das respostas geradas, tanto em termos de contexto, quanto de veracidade. Esta melhoria ocorre por meio da recuperação de informações relevantes de fontes externas, o que possibilita aos modelos acesso a conhecimentos fora do seu pré-treinamento.

No entanto, esta técnica também apresenta o risco de induzir os LLMs em erro, caso os textos recuperados contenham informação com ruído, imprecisos ou contraditórios em relação ao conhecimento já embutido no modelo. A presença destes problemas pode comprometer a coerência, a factualidade e a confiança nas respostas geradas (Xu et al., 2024).

2.4.3 Comparação entre Técnicas

As técnicas descritas neste subcapítulo encontram-se expostas na seguinte Tabela 6 - Comparação das Técnicas, onde são sintetizadas as suas definições, benefícios e limitações, o que permite, deste modo, uma comparação facilitada e sucinta. Após a tabela, encontra-se uma discussão dos pontos levantados pela mesma.

Tabela 6 - Comparação das Técnicas

Características	**RAG**	**EL**
Definição	Técnica que combina LLM com IR para responder às perguntas com referência a documentos específicos	Processo de identificar e ligar menções de entidades em texto livre às entidades correspondentes numa base de conhecimento estruturada
Benefícios	Fornece conhecimento aos LLM, o que aumenta o contexto e a veracidade das respostas	Reduz ambiguidade, aumenta precisão factual, permite integração com conhecimento externo, e suporta aplicações avançadas como pesquisa semântica e QA
Limitações	Pode induzir erros, devido a textos recuperados com ruído ou incorretos	Depende da cobertura e qualidade da base de conhecimento, pode falhar com entidades raras ou ausentes, custo computacional para grandes volumes de texto

O RAG é uma técnica que integra aos LLMs capacidades de IR para estes responderem a perguntas com utilização de referências diretas a documentos específicos. O seu principal benefício é fornecer conhecimento adicional aos LLMs, o que melhora o contexto e a veracidade das respostas geradas. No entanto, uma desvantagem do RAG é o potencial de induzir erros, pois os textos recuperados podem conter ruídos ou informações incorretas, o que afetará a precisão das respostas finais.

O EL é uma técnica de análise de texto que conecta texto não estruturado com bases de dados de conhecimento estruturadas, que identifica e liga menções ambíguas no texto a entidades concretas. O principal ponto forte desta abordagem está na redução da ambiguidade semântica e no aprimoramento do texto com metadados factuais para sistemas como pesquisa semântica, análise de redes de conhecimento ou sistemas de resposta a perguntas. Além disso, a associação de textos a entidades reconhecidas estabelece a base para a validação de factos e a interoperabilidade entre diferentes sistemas de informação. No entanto, está abordagem também pode criar problemas, como a dependência da cobertura e qualidade da base de conhecimento, e pode falhar em domínios com muitas entidades raras ou emergentes, existe também a ambiguidade em contextos textuais pouco informativos.

2.5 *Frameworks*

No presente subcapítulo, são expostas e discutidas duas *frameworks* consideradas para o desenvolvimento do CAIS especializado no domínio de laudos de honorários e deontologia médica. A *framework* é descrita, explicado o seu fluxo e os benefícios e as limitações que podem apresentar. Por fim, será realizada uma comparação entre as tecnologias, exposta sobre a forma de tabela, seguida de uma discussão.

2.5.1 LightRAG

O LightRAG é uma *framework* de RAG que integra estruturas de grafos de conhecimento no processo de indexação e recuperação textual, o que permite captar eficientemente as interdependências entre entidades e relações semânticas.

Utiliza um paradigma de recuperação de dois níveis que se estende desde informação detalhada sobre entidades até tópicos gerais, e possui um algoritmo de atualização incremental para adaptação rápida a novos dados. Esta abordagem gera respostas mais ricas em contexto, relevantes e coerentes, o que permite superar as limitações dos modelos RAG clássicos, baseados em representações planas de texto (Guo et al., 2025).

2.5.1.1 Fluxo do LightRAG

Como ilustrado na Figura 5 - Fluxo do LighRAG, o fluxo é divido entre duas etapas, a indexação e a consulta. Em ambas as etapas, o modelo de *embedding* e o LLM desempenham diferentes funções para gerar a resposta final ao utilizador.

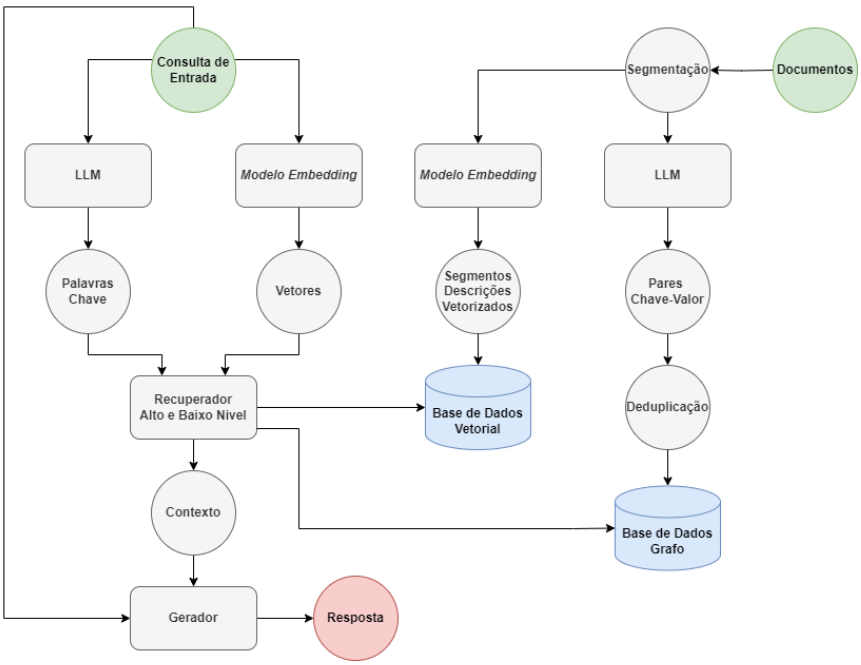


Figura 5 - Fluxo do LighRAG

O processo inicia-se pela segmentação dos documentos em pequenos blocos de texto, o que facilita a análise local e reduz a complexidade computacional. Com recurso a um LLM, são extraídas as entidades e as relações e são gerados pares chave-valor, onde a chave é um termo curto correspondente ao nome da entidade ou relação, e o valor é um resumo textual do contexto extraído do corpus. A deduplicação, com o auxílio do LLM e de heurísticas, funde as entidades e as relações equivalentes detetadas em diferentes segmentos, o que reduz redundâncias e mantem o grafo consistente. Estes elementos são guardados numa base de dados de grafo, onde fica armazenado a topologia e os metadados do grafo de conhecimento. Em paralelo, um modelo de *embeddings*, é usado para vetorizar os segmentos de texto e os pares chave-valor de entidades e relações, estes vetores são armazenados numa base de dados vetorial juntamente com referências cruzadas para os nós e arestas do grafo.

Ao receber novos documentos, estes são submetidos ao mesmo processo de extração de entidades e relações, o que origina um novo subgrafo. Este subgrafo é integrado no grafo existente por união dos conjuntos de nós e arestas, o que evita reconstruções totais. Assim, o sistema consegue adaptar rapidamente a novas informações, manter a integridade das ligações já estabelecidas e minimizar o custo computacional.

No momento da consulta, o LightRAG usa o LLM para extrair do pedido do utilizador as palavras-chave de dois tipos, as de baixo e alto nível. As palavras-chave de baixo nível são referentes a entidades ou detalhes específicos e as palavras-chave de alto nível são relativas a temas ou conceitos globais. Em seguida, o modelo de *embeddings* converte o pedido do utilizador em vetores para estes serem utilizados na pesquisa na base de dados vetorial para recuperar contextos similares. Em paralelo, o sistema pode explorar o grafo para expandir o contexto.

Estas duas vias, recuperação vetorial, que é de alto nível e exploração do grafo, que é de baixo nível, são complementares e fornecem um contexto rico e entre a visão global e o detalhe preciso.

Por fim, a informação recuperada é reunida e fornecida ao gerador, que é um LLM, que faz a fusão do contexto recuperado e os dados do grafo de conhecimento e produz a resposta final alinhada com o pedido do utilizador (Guo et al., 2025) (GeeksforGeeks, 2025b).

2.5.1.2 Benefícios e Limitações do LightRAG

O principal benefício do LightRAG é a sua capacidade de representar e recuperar relações complexas entre entidades através da integração de estruturas em grafo na indexação e recuperação de texto. Isto torna o sistema capaz de fornecer respostas mais contextuais, coerentes e personalizadas para consultas avançadas, o que transpõe a fragmentação que pode estar presente em técnicas convencionais baseadas apenas em representações planas de dados. Os dois paradigmas de recuperação, o de baixo e de alto nível, fornecem respostas elaboradas. Além disso, o algoritmo de atualização incremental é caracterizado pela sua capacidade de incorporar novas informações e manter o sistema atualizado, sem invocar reconstruções de índices (Guo et al., 2025).

Contudo, esta abordagem está dependente da qualidade da extração automática de entidades e relações, e erros ou incertezas neste ponto podem comprometer a riqueza semântica e a coerência das respostas. A gestão e a manutenção do grafo do conhecimento, especialmente em contextos dinâmicos e com enormes fluxos de dados, podem ser tecnicamente difíceis, o que exige medidas agressivas de deduplicação, atualização e controlo de inconsistências (GeeksforGeeks, 2025b).

2.5.2 GraphRAG

O GraphRAG é uma *framework* estruturada e holística para RAG que foi desenhada especificamente para trabalhar com dados nativamente em formato de grafo.

Esta abordagem facilita a adaptação a diferentes domínios e a integração de conhecimento relacional, que explora técnicas específicas de recuperação, organização e geração para maximizar a utilidade dos dados em grafo nas respostas geradas (Han et al., 2025).

2.5.2.1 Fluxo do GraphRAG

Como ilustrado na Figura 6 - Fluxo do GraphRAG, o fluxo é composto por cinco componentes principais, fonte de dados estruturados em grafos, o processador de consultas, o recuperador, o organizador e o gerador, o que permite captar a heterogeneidade e as relações complexas presentes nos dados.

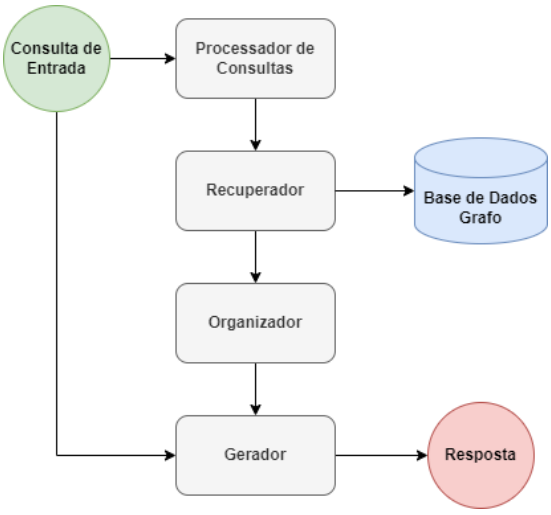


Figura 6 - Fluxo do GraphRAG

A fonte de dados constitui o núcleo do conhecimento, sob a forma de grafos de conhecimento, documentais, sociais, moleculares, entre outros. A sua diversidade exige processos próprios de construção e curadoria, adaptados ao domínio em causa. Este repositório pode evoluir dinamicamente, o que impacta os restantes módulos. É sobre esta base que todo o processo de recuperação e geração se desenvolve.

O processador de consultas converte a consulta do utilizador para uma forma adequada à estrutura de grafo dos dados, através de técnicas como o reconhecimento de entidades, a extração de relações e a estruturação de consultas. Este processamento permite o reconhecimento correto de entidades e relações dentro do grafo. Contém também mecanismos de decomposição e expansão de consultas, de modo a permitir o mapeamento entre perguntas complexas e a lógica do grafo. Assim, estabelece a ponte entre a linguagem natural e a estrutura relacional dos dados.

O recuperador tem a tarefa de procurar e recuperar, a partir do grafo, os elementos mais pertinentes com base na consulta já processada. Aplica técnicas como a travessia de grafos, a ligação de entidades, a correspondência relacional, os núcleos de grafos e as incorporações de grafos, que combina, opcionalmente, heurísticas e técnicas de aprendizagem profunda. Para tarefas complexas, o módulo utiliza, por defeito, técnicas iterativas e adaptativas e a integração neurossimbólica. O módulo permite que a recuperação pesquise eficazmente a semântica e a topologia do grafo.

O organizador desambigua o conteúdo recuperado, remove o ruído e a redundância através de processos de poda, reordenação e enriquecimento de grafos. Procura otimizar a informação para consumo do gerador, tornando-a semanticamente relevante e estruturalmente consistente. Utiliza, por exemplo, processos de classificação, sumarização e verbalização baseada na estrutura. O seu objetivo é fornecer um contexto limpo, focado e pronto para a fase de geração.

O gerador recebe a consulta e o conteúdo transformado e gera o resultado final, que pode ser estrutural ou textual. Utiliza modelos de linguagem avançados e codificadores específicos do domínio para manter as relações e a topologia. Adapta a geração às especificações da tarefa, de modo a garantir não só a factualidade, mas também a fidelidade à lógica dos grafos. Este módulo é necessário para transformar o conhecimento relacional em respostas exatas e informativas (Han et al., 2025) (GeeksforGeeks, 2025a).

2.5.2.2 Benefícios e Limitações do GraphRAG

O GraphRAG destaca-se pela sua generalidade e aptidão para ser utilizado em domínios em que os dados estão inerentemente organizados em estruturas de grafos, ou seja, grafos de conhecimento, grafos científicos, grafos sociais ou grafos de documentos. O ponto forte reside na capacidade de efetuar raciocínios *multi-hop*, inferências complexas e recuperação não só com base na semântica do texto, mas também com base na topologia e nas relações explícitas, o que facilita tarefas de nível superior que incluem a navegação, a agregação e a composição do conhecimento relacional. A sua arquitetura modular, permite que cada um dos elementos do *pipeline*, desde o processador de consultas ao organizador e gerador, seja adaptado aos requisitos específicos e às peculiaridades do tipo de grafo em análise, o que proporciona flexibilidade e eficiência (Han et al., 2025).

No entanto, esta flexibilidade pode criar problemas como a diversidade de formatos e padrões de relação em vários domínios dificulta a criação de soluções abrangentes e exige o desenvolvimento de técnicas de adaptação ou integração mútua. O desempenho de um sistema pode degradar-se se os algoritmos de pesquisa, a ligação de entidades ou a codificação dos grafos não forem corretamente adaptados ao ambiente em que são implementados, ou mesmo se a densidade e a complexidade dos grafos causarem redundância, ruído ou sobrecarga computacional. Além disso, a escalabilidade e a manutenção de grafos de grandes dimensões, nomeadamente em ambientes dinâmicos, necessitam de estratégias de atualização e limpeza para garantir a robustez e a fiabilidade do sistema (GeeksforGeeks, 2025a).

2.5.3 Comparação entre *Frameworks*

As *frameworks* descritas neste subcapítulo encontram-se expostas na seguinte Tabela 7 - Comparação das *Frameworks*, onde são sintetizadas diversas características, o que permite, deste modo, uma comparação facilitada e sucinta. Após a tabela, encontra-se uma discussão dos pontos levantados pela mesma.

Tabela 7 - Comparação das *Frameworks*

Características	**LightRAG**	**GraphRAG**
Estrutura do conhecimento	Integra grafos em indexação e recuperação; Relações explícitas	Dados nativamente em grafo; Múltiplos domínios
Mecanismo de recuperação	Paradigma dual de baixo e alto nível	Técnicas de busca em grafo, GNNs e grafos heterogéneos
Componente gerador	LLM, orientado por contexto extraído de grafos	LLM, informada por relações e estrutura dos grafos
Objetivo central	Recuperação eficiente, respostas contextuais e coesas, adaptação rápida	Integração de relações e dependências; Adaptação a domínios complexos
Gestão de atualização de dados	Algoritmo incremental para atualização dinâmica do índice	Foco em adaptabilidade a diferentes domínios e tipos de grafo
Eficiência computacional	Alta eficiência, redução de custos e resposta rápida	Variável conforme domínio e tipo de grafo
Benefícios principais	Síntese de informação múltiplas fonte, contextualização, eficiência	Captura de relações complexas e múltiplos domínios, suporte a raciocínio estruturado
Limitações identificadas	Dependente da qualidade da extração relacional e atualização de grafos	Desafios na heterogeneidade de formatos, adaptação e escalabilidade

Uma análise comparativa das *frameworks*, evidencia o de diversas soluções empregues para melhorar os CAISs com conhecimento externo de forma eficaz, eficiente e flexível. Cada uma das estruturas tem os seus próprios desafios no processo de integração entre a recuperação de informação e a geração de textos, o que reflete desenvolvimentos distintos na representação do conhecimento, nos mecanismos de recuperação e na otimização da geração.

O LightRAG é caracterizado pela inclusão explícita de estruturas em grafo nos passos de indexação e recuperação do texto. A abordagem é inspirada na compreensão das vulnerabilidades dos sistemas RAG tradicionais, que confiam nas representações planas de dados, o que limita a sua capacidade de capturar relações complexas entre entidades e manter coerência contextual em respostas para consultas sofisticadas. Ao adotar um paradigma de recuperação dupla, o LightRAG permite a seleção eficiente de informações, tanto em níveis

detalhados, quanto abstratos, por meio de uma arquitetura que integra grafos com incorporações vetoriais. A arquitetura permite a geração de respostas mais contextuais e coerentes, com menor fragmentação de informações. Além disso, o algoritmo de atualização incremental diferencia o LightRAG ao fornecer ajustes quase em tempo real para dados recém inseridos, o que é importante para situações dinâmicas ou fluxos de informação de alta frequência.

O GraphRAG é projetado para uma extensão da técnica de RAG onde as informações são intrinsecamente estruturadas em grafos, além dos limites apenas textual. A especificidade do GraphRAG está na necessidade de adaptar cada componente a diferentes formatos e domínios de grafos, tais como grafos de conhecimento, moléculas, redes sociais, documentos estruturados, entre outros. O módulo de recuperação, por exemplo, pode receber diversos algoritmos de busca em grafo diferentes e pode ser complementado por codificadores específicos, como as *graph neural network* (GNNs) para abrigar tanto a semântica quanto a topologia. A construção e a posterior geração de respostas têm em conta não só o material de texto recuperado, mas também as interligações estruturais, o que promove a capacidade de raciocínio *multi-hop* e a integração do conhecimento, que assenta em várias entidades correlacionadas. O GraphRAG enfatiza os desafios de heterogeneidade, interdependência e especificidade de domínio que fazem com que seja complicada a adoção de soluções universais. Por outro lado, ele propõe um modelo de especialização por domínio no qual as abordagens são personalizadas em função de cada categoria de grafo e tarefa.

3 Construção do Grafo de Conhecimento

No presente Capítulo 3, Construção do Grafo de Conhecimento, é feito um levantamento de ferramentas analíticas disponíveis para a criação de grafos de conhecimentos e, assim, realizar a construção deste com base em documentos normativos e deontológicos que regem a prática médica em Portugal, nomeadamente o Regulamento dos Laudos a Honorários (Ordem dos Médicos, n.d.) e o Regulamento de Deontologia Médica (Ordem dos Médicos, 2016).

O objetivo é estudar o domínio representado por estes documentos e explorar as relações entre os conceitos. Além disso, o grafo de conhecimento construído neste Capítulo poderá ser utilizado diretamente pelo CAIS desenvolvido, ou poderá servir para validar o grafo construído autonomamente pelo CAIS.

3.1 Ferramentas de Criação de Grafos de Conhecimento

No presente subcapítulo, são expostas e discutidas as potenciais escolhas de ferramentas analíticas para a construção do grafo de conhecimento. A exposição consiste numa visão geral das ferramentas e das suas principais funcionalidades, enquanto a discussão consiste em levantar os benefícios e as limitações que estas podem apresentar. Por fim, será realizada uma comparação entre as ferramentas, exposta sobre a forma de tabela, seguido de uma discussão.

3.1.1 InfraNodus

O InfraNodus [2] é uma ferramenta criada para o mapeamento, visualização, análise e compreensão de informações complexas por meio de grafos de conhecimento. O seu principal objetivo é ajudar os utilizadores a identificar padrões, conexões ocultas e lacunas em conjuntos de ideias, textos ou dados, o que permite uma abordagem mais sistemática para a organização e exploração do conhecimento.

[2] https://infranodus.com/

3.1.1.1 Funcionamento Base do InfraNodus

O funcionamento do InfraNodus baseia-se na conversão de textos ou conjuntos de dados em redes de conceitos interligados. Ao introduzir um texto, transcrição ou lista de ideias, a ferramenta identifica as palavras-chave mais importantes e constrói um grafo em que cada nó representa um conceito e cada aresta representa uma relação semântica entre eles. A ferramenta, de forma autónoma, destaca os tópicos centrais, mede a centralidade dos nós, sugere conexões e identifica as lacunas na informação, o que permite ao utilizador explorar visualmente as inter-relações do seu conteúdo.

3.1.1.2 Benefícios e Limitações do InfraNodus

Os principais benefícios do InfraNodus está na sua capacidade de transformar, de forma célere, informações não estruturadas em grafos de interpretação visual, o que ajuda no pensamento sistemático e na identificação de conceitos e as suas relações centrais ou negligenciadas. A ferramenta oferece análises automáticas que podem identificar conexões ocultas e sugerir novas questões ou caminhos de pesquisa, tornando-se útil para o *brainstorming*, a revisão de literatura, o estudo de textos complexos ou o planeamento estratégico.

No entanto, apresenta algumas limitações, o InfraNodus pode não captar nuances semânticas de maior profundidade ou relações contextuais subtis, devido à análise de coocorrência de palavras. Por fim, para obter resultados com maior precisão, é necessário que o utilizador possua nível de compressão sobre os textos e dados introduzidos, já que ruídos ou termos irrelevantes podem distorcer o grafo gerado. Para obter acesso à ferramenta é necessário fazer uma subscrição mensal paga, mas, possui uma versão *trial* de duas semanas que permite o acesso total.

3.1.2 Kumu

O Kumu[3] é uma ferramenta criada para o mapeamento, visualização, análise e compreensão de sistemas complexos por meio de mapas interativos de relações, semelhantes a grafos. O seu principal objetivo é ajudar os utilizadores a organizar as ideias, identificar dinâmicas e visualizar interconexões entre diferentes entidades, conceitos ou elementos dentro de projetos, redes sociais, sistemas ou ecossistemas.

3.1.2.1 Funcionamento Base do Kumu

O funcionamento do Kumu baseia-se na criação de mapas visuais onde os utilizadores podem adicionar nós que representam entidades, e conectá-los por arestas para representar a relação entre eles. O resultado são redes que apresentam uma estrutura de sistemas complexos. A ferramenta permite customizar a aparência dos mapas, inserir informações detalhadas sobre cada nó ou relação, e utilizar filtros ou categorias para explorar diferentes perspetivas.

[3] https://kumu.io/

3.1.2.2 Benefícios e Limitações do Kumu

Os principais benefícios do Kumu está na sua flexibilidade e facilidade de uso, tornando-se acessível para utilizadores sem experiência técnica em análise de redes. A ferramenta permite a construção de mapas complexos de forma visual e intuitiva, a integração de dados de diferentes fontes e a geração de relatórios visuais para o apoio de decisões estratégicas.

No entanto, apresenta algumas limitações, o seu destaque está na visualização e organização das conexões, não na análise de forma autónoma dos dados ou na deteção semântica profunda. Para projetos de maior dimensão ou com necessidade de análise quantitativa avançada, pode ser menos eficiente que ferramentas específicas de ciência de redes. Para obter acesso total à ferramenta é necessário fazer uma subscrição mensal paga, mas, possui uma versão gratuita que dá acesso às funcionalidades básicas da ferramenta.

3.1.3 Scapple

O Scapple[4] é uma ferramenta criada com o objetivo de facilitar o registo, a organização e a visualização de ideias. O seu principal objetivo é oferecer um espaço flexível para *brainstormings* e mapeamentos mentais, o que permite ao utilizador conectar pensamentos de forma não linear e construir relações entre as ideias de maneira espontânea.

3.1.3.1 Funcionamento Base do Scapple

O funcionamento do Scapple é simples e intuitivo, ao abrir um novo quadro, o utilizador pode criar notas soltas, que são de texto livre. As notas podem ser conectadas manualmente entre si, o resultado são redes visuais de ideias. Não há restrições sobre a forma como as notas são agrupadas ou relacionadas, o que possibilita liberdade total de organização, por exemplo organização em semelhança de grafo. O utilizador pode reorganizar, redimensionar e personalizar as notas conforme necessário.

3.1.3.2 Benefícios e Limitações do Scapple

O principal benefício do Scapple é a sua flexibilidade e simplicidade, tornando-se uma ferramenta para capturar ideias de forma célere sem precisar seguir regras rígidas de organização. Este conceito estimula a criatividade, já que o utilizador pode experimentar diferentes formas de ligar pensamentos e reorganizar tudo conforme novas ideias surgem. Além disso, por ser leve e direto, o Scapple é ideal para sessões de *brainstorming*, esquematização de argumentos ou estruturação inicial de textos acadêmicos e criativos.

[4] literatureandlatte.com/scapple/overview

No entanto, apresenta algumas limitações por não contar com recursos de análise automática, hierarquização avançada ou integração direta com bases de dados. A ferramenta restringe-se a ser um quadro de anotações visual, sem funções de análise semântica ou quantitativa. Projetos grandes podem se tornar difíceis de gerir visualmente, e a exportação para outros formatos é limitada. Para obter acesso à ferramenta é necessário fazer a compra de uma licença.

3.1.4 Comparação entre Ferramentas

As ferramentas descritas neste subcapítulo encontram-se expostas na seguinte Tabela 8 - Comparação das Ferramentas, onde são sintetizadas diversas características, o que permite, deste modo, uma comparação facilitada e sucinta. Após a tabela, encontra-se uma discussão dos pontos levantados pela mesma.

Tabela 8 - Comparação das Ferramentas

Características	**InfraNodus**	**Kumu**	**Scapple**
Objetivo Principal	Visualização e análise semântica de textos	Mapeamento e visualização de sistemas e relações	*Brainstorming* e mapeamento livre de ideias
Funcionamento base	Converte textos em grafos de conceitos; Análise semântica automática	Construção manual de mapas e relações; Organização visual	Criação livre de notas e conexões manuais em um quadro
Nível de automatização	Alta, deteta autonomamente conceitos, relações e o significado semântico	Baixa, visualização e categorização, sem análise semântica	Nenhuma
Visualização de grafos	Sim, interativa e automática	Sim, interativa, altamente customizável	Sim, manual, estilo quadro branco
Análise semântica	Sim	Não	Não
Plataforma	Web	Web	Desktop
Acesso	Subscrição paga mensal com versão *trial* disponível	Subscrição paga mensal com versão gratuita limitada disponível	Licença paga

As três ferramentas analisadas neste subcapítulo apresentam, cada uma delas, a sua própria abordagem de organização, visualização e exploração de informações, para atender a diferentes necessidades e casos de uso.

O InfraNodus destaca-se pelo seu alto nível de automatização na análise de texto, o que permite transformar rapidamente conteúdo não estruturado em grafos de conhecimento com perceções automáticas sobre conceitos, relações e possíveis lacunas. Estas capacidades tornam a ferramenta útil para casos de uso de extração de significado, deteção de padrões e geração de novas questões a partir de grandes quantidades de informação textual. A análise semântica automática facilita o mapeamento de campos de conhecimento ou a realização de *brainstorming* estruturado. Por outro lado, a dependência de algoritmos para determinar a relevância pode ignorar nuances contextuais.

Por outro lado, o Kumu segue uma abordagem livre e visual, voltada para o mapeamento manual de conceitos relacionados sobre sistemas ou estruturas organizacionais. Esta abordagem oferece à ferramenta capacidades de visualização interativa e customização, que permite moldar redes complexas de acordo com finalidades específicas, sem a necessidade de conhecimentos técnicos avançados. Os casos de uso são abrangentes, desde o desenho de projetos, a análise de *stakeholders* e a visualização de ecossistemas. Porém, a ausência de análise semântica automática e recursos quantitativos limita o seu potencial para análises mais profundas dos dados, o que faz que ele seja mais útil para organização e apresentação do que para extração de perceções automáticas.

Por fim, o Scapple oferece a maior liberdade criativa, que funciona como um quadro branco digital onde as notas e ideias podem ser colocadas e ligadas e relacionadas entre si sem restrições. Esta simplicidade é mais adequada para casos de uso como *brainstorming* rápido e desenvolvimento inicial de estrutura de projetos. No entanto, a falta de automatização ou análise estrutural torna o Scapple inadequado para sistemas de mapeamento com grandes quantidades de informação, o que limita a sua aplicação a contextos mais individuais e de organização inicial de ideias.

3.2 Grafo de Conhecimento

O grafo de conhecimento foi construído com recurso à ferramenta InfraNodus devido à identificação de entidades e à análise das relações semânticas entre estas de forma autónoma. O grafo exibe as principais áreas abordadas nos documentos analisados, organizando-se à volta de quatro componentes principais, sendo a saúde, pacientes, ética e honorários.

3.2.1 Estrutura e Componentes Principais do Grafo de Conhecimento

O componente ilustrado na seguinte Figura 7 - Grafo de Conhecimento – Saúde, foca nos aspetos normativos e práticos relacionados com o fornecimento de cuidados médicos.

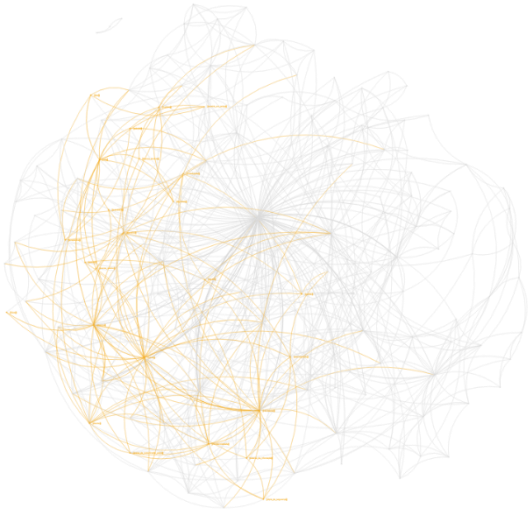


Figura 7 - Grafo de Conhecimento – Saúde

Este componente proporciona a compreensão das obrigações dos médicos em relação à proteção da saúde pública e ao cumprimento das melhores práticas, contendo, para esta finalidade, conceitos associados à qualidade dos serviços médicos prestados, às responsabilidades técnicas dos profissionais de saúde e à obrigação de atuação em emergências.

O componente ilustrado na seguinte Figura 8 - Grafo de Conhecimento – Pacientes, explora os direitos e deveres dos pacientes.

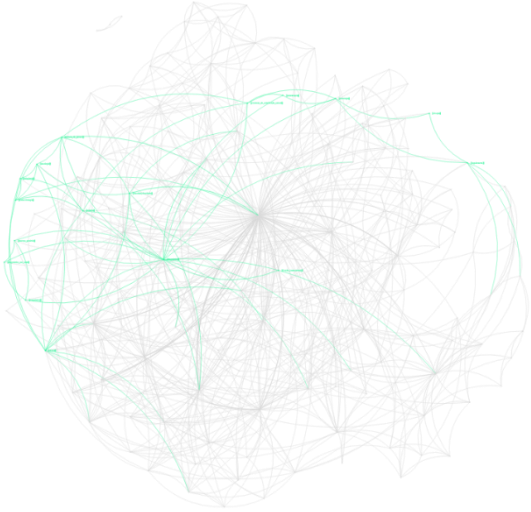


Figura 8 - Grafo de Conhecimento – Pacientes

Proporciona, assim, a compreensão das interações diretas entre médicos e pacientes, por parte dos profissionais de saúde, enfatizando a importância e o respeito da dignidade e autonomia

dos pacientes. Isto é possível devido ao destaque, pelo componente, de temas como o consentimento informado, a confidencialidade e o direito à livre escolha de médicos pelos pacientes.

O componente ilustrado na seguinte na Figura 9 - Grafo de Conhecimento – Ética, abrange as normas que regem a conduta profissional.

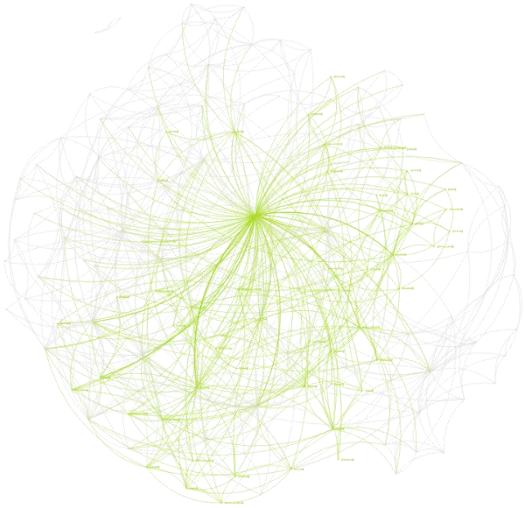


Figura 9 - Grafo de Conhecimento – Ética

Destaca o respeito pelos princípios da beneficência, não-maleficência, justiça e autonomia, o que permite o discernimento pelos dilemas éticos enfrentados pelos médicos, como a objeção de consciência, o sigilo profissional e o respeito pela vida humana.

O componente ilustrado na seguinte na Figura 10 - Grafo de Conhecimento - Honorários, organiza os princípios relacionados com a remuneração dos médicos.

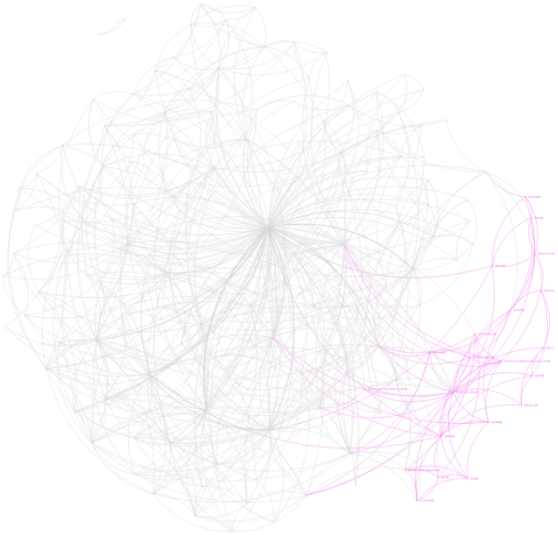


Figura 10 - Grafo de Conhecimento - Honorários

Foca-se, assim, em diretrizes sobre a fixação de valores, transparência na apresentação de contas e proibições relacionadas com práticas desleais e permite o entendimento de questões financeiras com a deontologia médica, o que garante a justa compensação pelo trabalho realizado.

Este grafo de conhecimento geral constitui a base para o desenvolvimento do CAIS, que poderá responder a questões relacionadas com as normas éticas e regulamentações sobre os honorários médicos. A estruturação das informações permite ao CAIS navegar entre as diversas áreas do grafo, o que torna a geração de respostas mais contextualizadas e factualmente verdadeiras.

3.2.2 Estudo do domínio e apoio ao CAIS

O grafo de conhecimento construído contribuiu para o estudo e para a análise do domínio dos laudos de honorários e deontologia médica. Através deste estudo, foi possível construir uma representação estruturada dos conceitos-chave e das suas inter-relações, o que permitiu uma melhor compreensão da rede de conhecimento.

O grafo de conhecimento construído pode ser integrado diretamente no CAIS, que permite que este utilize a estrutura do grafo como fonte de conhecimento, sem a necessidade de gerar um grafo autonomamente. Ao utilizar o grafo, o CAIS pode recorrer a uma representação já estruturada e validada do domínio, poupando tempo e recursos na construção do seu próprio grafo. Isto facilita a recuperação de informações precisas, uma vez que o grafo oferece uma base sólida de conceitos interligados, o que proporciona ao CAIS uma fonte externa confiável para enriquecer as suas respostas e garantir maior precisão nas interações com o utilizador.

Caso o grafo de conhecimento utilizado pelo CAIS seja gerado autonomamente pelo próprio, o grafo criado pelo InfraNodus pode ser utilizado para validações, correções ou melhorias. A comparação entre os dois grafos pode revelar divergências, lacunas de informação ou até mesmo relações incorretas que o sistema tenha extraído erroneamente.

4 Desenvolvimento do CAIS

No presente Capítulo 4, Desenvolvimento do CAIS, é descrito o desenvolvimento do CAIS. É exposta a seleção e parametrização de modelos via Ollama, seguido do fluxo do sistema baseado na *framework* LightRAG, e apresentados os diversos modos de resposta presentes na *framework*. Por fim, é documentado as parametrizações anti-alucinação e expostos os desafios e as limitações.

O objetivo é tornar o CAIS reproduzível. Para isso, são estabelecidos os critérios técnicos para a escolha dos modelos, a definição de práticas de ingestão e *grounding* que preservam contexto e proveniência, e é justificado o modo de resposta mais adequado por cenário, bem como medidas de mitigação de alucinações.

4.1 Ollama

O Ollama[5] é um servidor local de modelos LLM e de modelos de *embeddings*. Ele faz a gestão do descarregamento e a execução dos modelos, expõe uma API HTTP simples em *localhost* e responde a chamadas de geração de texto e de cálculo de *embeddings*.

A execução local contrasta com os serviços em nuvem, como a API da OpenAI. A nuvem oferece, em regra, modelos maiores e mais capazes, mas implica custos variáveis, dependência de conexão, limite de chamadas e questões de confidencialidade dos dados (Sen, 2013). Ao correr os modelos numa máquina local, a privacidade é garantida, as chamadas ao modelo são ilimitadas e gratuitas e existe uma maior reprodutibilidade experimental (Liu et al., 2021).

[5] https://ollama.org

4.1.1 Escolha e Parametrização dos Modelos

Quanto aos modelos escolhidos para o projeto, adotou-se um LLM compacto, o Gemma 2B[6], que oferece um equilíbrio entre requisitos de memória, latência e qualidade de resposta numa máquina sem GPU dedicada. Para os *embeddings*, optou-se por um modelo generalista de setecentos e sessenta e oito (768) dimensões, o nomic-embed-text[7].

Inicialmente, é necessário fazer o *pull* dos modelos, seguindo-se do arranque do serviço local que, por sua vez, permite que o CAIS se comunique com o Ollama através de requisições HTTP. Antes de realizar alguma chamada aos modelos, é recomendado realizar o pré-aquecimento destes através de uma curta chamada para os modelos iniciarem rapidamente e estarem prontos para as chamadas do CAIS.

4.1.2 Benefícios e Limitações

Os principais benefícios no uso de modelos locais concentram-se no controlo e na privacidade, pois todo o processamento ocorre num ambiente local, sem a necessidade de enviar dados para servidores externos. Isto também permite um funcionamento totalmente *offline*, o que é útil para cenários onde a conexão com a nuvem é limitada ou indesejada (Liu et al., 2021).

Por outro lado, os modelos locais de menor dimensão apresentam algumas limitações em comparação com os LLMs de grande escala, como os disponíveis na nuvem. A qualidade de raciocínio pode ser inferior, especialmente para tarefas mais complexas que exigem um maior poder computacional. A latência também tende a aumentar à medida que os documentos se tornam mais extensos ou quando o *hardware* utilizado é limitado, o que impacta o desempenho do CAIS. Além disso, o modelo local não possui capacidade de navegação na *web*, assim, qualquer referência externa que o modelo devolva é gerada com base no conhecimento do próprio modelo ou das fontes externais locais (Liu et al., 2021; Sen, 2013).

4.2 Fluxo do CAIS

O Fluxo do CAIS segue um *pipeline* com base na *framework* LightRAG, como ilustrado na Figura 11 - Fluxo do CAIS, que separa as operações de preparação e ingestão, que são executadas quando a base de conhecimento é alterada, das operações de consulta, que são executadas a cada pergunta do utilizador. Esta separação é essencial para manter latência baixa durante o uso normal, concentrando o custo computacional na fase de preparação.

[6] https://ollama.com/library/gemma2
[7] https://ollama.com/library/nomic-embed-text

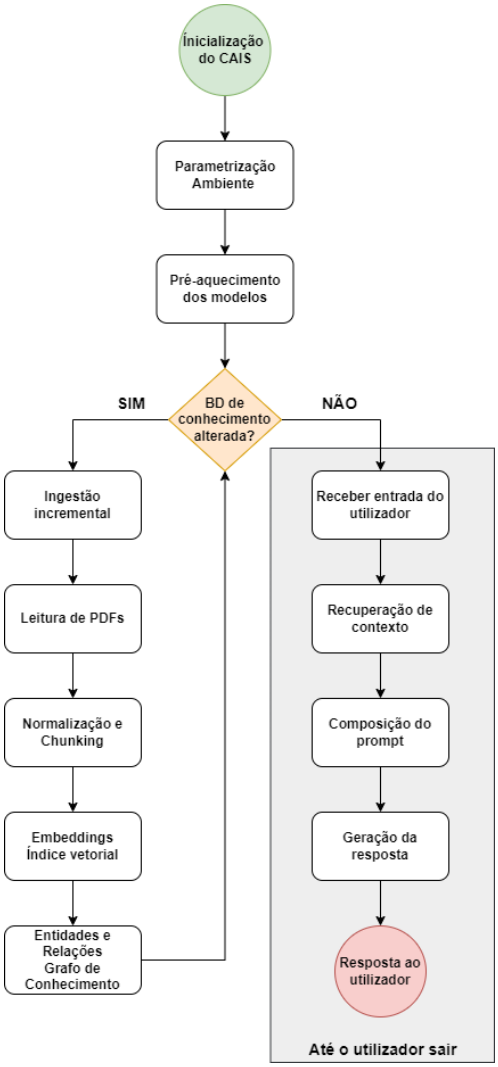


Figura 11 - Fluxo do CAIS

Quando o CAIS é iniciado, o Ollama carrega os modelos necessários. Na primeira chamada de cada modelo é realizado o pré-aquecimento com uma chamada curta antes de iniciar a ingestão ou responder à primeira pergunta do utilizador. Em paralelo, são definidos os parâmetros operacionais.

4.2.1 Preparação e Ingestão

A ingestão é incremental e desencadeada apenas quando entram novos documentos ou quando algum é atualizado. A inserção é feita documento a documento, o que facilita a observabilidade para caso algum documento apresente um problema, o reprocessamento seletivo e a recuperação de erros. O sistema suporta apenas PDFs de texto, onde cada documento é lido, convertido para texto, normalizado e, se necessário, dividido em segmentos com sobreposição controlada, o que preserva contexto entre segmentos.

Os segmentos resultantes são transformados em vetores pelo modelo de *embeddings* e inseridos num índice vetorial local. Esta etapa possibilita pesquisa semântica eficiente na fase de consulta. Em paralelo, mantém-se meta informação mínima o que permite retomar à *pipeline* sem repetir trabalho já concluído.

Além do índice vetorial, o CAIS extrai entidades e relações do texto via LLM, e constrói um grafo de conhecimento. Esta camada estrutural é mais dispendiosa, mas oferece ganhos de contexto global e de raciocínio relacional durante a pesquisa. A construção do grafo só ocorre quando a base é alterada.

4.2.2 Ciclo pergunta-resposta

Perante uma pergunta do utilizador, o sistema transforma-a num vetor e procura evidências no índice vetorial construído na fase de ingestão. Dependendo do modo de resposta selecionado, a recuperação devolve apenas os segmentos mais semelhantes, expande a vizinhança imediata para preservar continuidade textual, percorre subestruturas do grafo para agregar contextos dispersos ou combina ambas as abordagens. Antes de seguir, o contexto é normalizado e limitado em tamanho para caber na janela de contexto do modelo, preservando as passagens mais informativas.

O material recuperado é depois encapsulado numa entrada controlada, que inclui instruções para responder apenas com base nos excertos fornecidos, abster-se com uma formulação do tipo "não sei com base nos documentos" quando a evidência é insuficiente, evitar referências externas e opiniões não suportadas. A entrada organiza o contexto por origem/segmento para manter a rastreabilidade, aplica um formato conciso e, quando aplicável, incorpora pistas de estilo para estabilizar a geração.

A entrada consolidada é enviada ao LLM a correr no Ollama, que produz a resposta final. A resposta pode ser produzida em *streaming* para menor latência percebida e enriquecida com proveniência. O resultado é, por fim, devolvido ao utilizador.

4.3 Modos de Resposta

No presente subcapítulo, são expostos e discutidos os quatro modos de recuperação disponibilizados pelo LightRAG, que determinam como o CAIS encontra e organiza evidências antes de gerar a resposta com o LLM. A exposição consiste na descrição do mecanismo de funcionamento, os benefícios e limitações e os casos de uso recomendados que os modos podem apresentar. Por fim, será realizada uma comparação entre os modos, exposta sobre a forma de tabela, seguido de uma discussão.

4.3.1 Modo *Naive*

Neste modo, é apenas executada a pesquisa semântica vetorial sobre os segmentos indexados. A pergunta é *embedded* num vetor e compara-se por similaridade, normalmente o cosseno, com os vetores dos segmentos. Selecionam-se os *top-k* mais próximos e compõe-se a entrada com este contexto, sem qualquer enriquecimento estrutural do grafo de conhecimento.

Os principais benefícios consistem em ser o modo mais rápido e leve, ter menos pontos de falha e exigir menos recursos. Funciona bem quando a resposta está contida em poucos trechos explícitos.

No entanto, pode perder contexto global que são as ligações entre partes distantes, e não captar relações entre as entidades. É sensível à segmentação, caso a informação fique dividida a resposta pode sair retalhada.

Os casos de uso deste modo são em perguntas factuais simples, bases pequenas ou muito homogéneas, situações em que a latência curta é prioritária (Data Intelligence Lab@HKU, n.d.).

4.3.2 Modo Local

Este modo parte da mesma pesquisa vetorial realizada no modo *naive*, mas expande o contexto na vizinhança dos segmentos selecionados, por exemplo, nos segmentos adjacentes ou nas subestruturas locais do grafo correspondentes às mesmas entidades. O objetivo é preservar a continuidade sem percorrer todo grafo de conhecimento.

Os principais benefícios consistem na melhoria da coesão contextual em comparação ao modo *naive*, com os custos relativamente igualados e na redução do risco de respostas fora de contexto quando a evidência está próxima no documento.

No entanto, está limitado a uma perspetiva de curto alcance, caso a resposta exija integrar partes distantes ou relações não triviais, o modo pode falhar.

Os casos de uso deste modo são em perguntas factuais que beneficiam de algum contexto circundante, como definições, procedimentos, passos sequenciais, secções contíguas de um PDF (Data Intelligence Lab@HKU, n.d.).

4.3.3 Modo Global

Neste modo, o grafo de conhecimento é explorado. São identificas as entidades relevantes na pergunta e nos segmentos candidatos e é realizada uma procura nas relações e caminhos no grafo. O contexto é composto a partir destas ligações, o que pode incluir evidências dispersas por vários documentos.

Os principais benefícios deste modo constituem na oferta da abrangência e da capacidade relacional.

No entanto, depende da qualidade do grafo devido à extração das entidades e das relações e pode introduzir ruído se o grafo estiver com falhas ou for impreciso.

Os casos de uso deste modo são em perguntas analíticas que podem estar relacionadas a mais do que um documento, pedidos de explicações ou relações entre conceitos, ou quando o *naive* ou o local não recuperarem evidências suficientes (Data Intelligence Lab@HKU, n.d.).

4.3.4 Modo Híbrido

Neste modo, é combinado as duas vias de evidência numa única etapa. Começa por recuperar, via pesquisa vetorial, os segmentos mais semelhantes à pergunta e, em paralelo, identifica as entidades presentes na pergunta. Percorre o grafo de conhecimento para recolher passagens adicionais ligadas, seja por vizinhança ou por relações relevantes e em seguida, funde os candidatos provenientes das duas fontes e procede a uma reordenação conjunta, onde cada passagem recebe uma pontuação que pondera, por exemplo, a similaridade vetorial, a proximidade no grafo e a coerência local do contexto.

O principal benefício é a tendência de produzir as melhores respostas em termos de equilíbrio entre relevância imediata e a cobertura global. Compensa casos em que uma via sozinha, apenas vetorial ou apenas grafo de conhecimento, seria insuficiente.

No entanto, é o modo mais computacionalmente exigente e, em máquinas limitadas, pode aumentar a latência. Exige um grafo com qualidade o suficiente para realizar a componente global com benefícios.

Os casos de uso deste modo são na definição como modo por defeito quando a prioridade é a qualidade e a latência aceitável. Útil em bases heterogéneas, onde há perguntas simples e também perguntas que exigem contexto mais rico (Data Intelligence Lab@HKU, n.d.).

4.3.5 Comparação entre Modos

Os modos descritos neste subcapítulo encontram-se expostas na seguinte Tabela 9 - Comparação dos Modos, onde são sintetizadas diversas características, o que permite, deste modo, uma comparação facilitada e sucinta. Após a tabela, encontra-se uma discussão dos pontos levantados pela mesma.

Tabela 9 - Comparação dos Modos

Características	***Naive***	**Local**	**Global**	**Híbrido**
Mecanismo	*Top-k* por similaridade vetorial	Vetorial e vizinhança ou adjacência de segmentos	Exploração do grafo nas entidades e nas relações	Combinação vetorial e do grafo
Benefícios	Modo mais rápido com menor custo	Contexto coeso com custo moderado	Abrangência e raciocínio relacional	Melhor equilíbrio entre relevância e cobertura
Limitações	Perde contexto global; Sensível à segmentação	Alcance curto Pode falhar em integrações complexas	Depende da qualidade do grafo	Modo mais computacionalmente exigente; Latência superior em *hardware* limitado
Casos de Uso	Factual simples; Bases pequenas	Definições, passos, secções contíguas	Perguntas analíticas ou em múltiplos documentos	Modo por defeito quando se privilegia qualidade

Os quatro modos analisados neste subcapítulo apresentam, cada um deles, as suas próprias características para atender a diferentes necessidades e casos de uso.

O modo *naive* é a forma mais simples de recuperação. A pergunta é convertida em vetor e recuperam-se os segmentos mais semelhante para compor o contexto enviado ao LLM. É o modo mais rápido, sendo adequado quando a informação relevante está concentrada num excerto curto e explícito. Contudo é sensível à segmentação e tende a falhar quando a resposta exige ligar partes distantes ou integrar múltiplos documentos.

O modo local parte da pesquisa vetorial realizada pelo modo *naive*, mas acrescenta a vizinhança o que preserva a continuidade textual. Mantém custos próximos do *naive* e melhora a coesão do contexto, o que reduz as respostas truncadas. Continua, no entanto, a operar num alcance curto, caso a pergunta requer ligação entre secções afastadas ou documentos distintos, pode não recuperar evidências suficientes.

O modo global recorre o grafo de conhecimento para complementar as evidências. Ele identifica as entidades e as relações relevantes e agrega o contexto potencialmente disperso por vários documentos, o que inclui caminhos multi-salto no grafo. É o modo indicado para perguntas analíticas, explicações causa efeito e sínteses de múltiplos documentos. Possui uma latência superior e dependência da qualidade do grafo, pois as extrações pobres de entidades e de relações diminuem a qualidade.

Por fim, o modo híbrido combina as duas perspetivas do modo local e do modo global. Usa a recuperação vetorial para obter os segmentos imediatamente pertinentes e, em paralelo, expande o grafo para captar ligações semânticas não triviais, depois funde e reordena os candidatos com um critério combinado e escolhe os *top-n* para a entrada do LLM. Normalmente, oferece o melhor equilíbrio entre relevância imediata e cobertura relacional, à custa de maior custo computacional. Em *hardware* limitado, a latência pode aumentar, pelo que é recomendado alternar para o modo local quando a melhoria do desempenho for necessária.

4.4 Parametrizações para Combater Alucinações

No presente subcapítulo, é ilustrado um caso de alucinação que o CAIS demostrou, as respetivas parametrizações escolhidas para a otimização do sistema e para a mitigação deste fenómeno, expostas em formato de tabela, e por fim, a aplicação de engenharia de *prompting* para reduzir o efeito da alucinação.

4.4.1 Alucinação Detetada no CAIS

Após uma pergunta simples baseada nos PDFs carregados, o CAIS devolveu uma resposta aparentemente rigorosa e, no final, como ilustrado na Figura 12 - Alucinação apresentada pelo CAIS, apresentou uma secção de referências com ligações externas inventadas que não constavam de nenhum dos documentos. O CAIS não tem acesso à *web*, pois ele é local, o que comprova que estas referências foram fabricadas pelo LLM para preencher a estrutura da resposta.

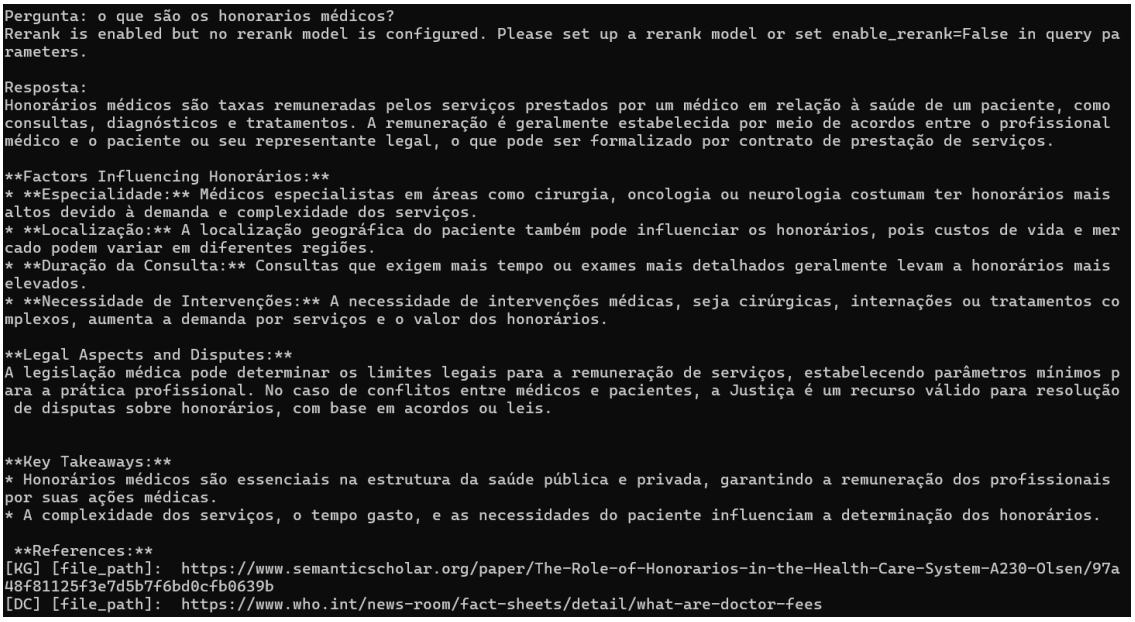


Figura 12 - Alucinação apresentada pelo CAIS

O LLM é um modelo generativo que aprende padrões de linguagem. Quando a entrada sugere, mesmo de forma implícita que deve haver bibliografia, o modelo tende a completar a resposta com elementos plausíveis, não necessariamente verdadeiros. Na *pipeline*, há ainda fatores que aumentam o risco como por exemplo entradas pouco restritivas, temperaturas elevadas que torna o LLM criativo, contextos longos com evidência difusa e falhas temporárias na recuperação devido a *timeouts* durante a extração de entidades e de relações.

4.4.2 Parametrização dos Modelos e Engenharia de *Prompting*

O Ollama necessita de algumas parametrizações, como exposto na Tabela 10 - Parametrização do CAIS, para garantir o funcionamento eficiente e para mitigar os casos de alucinação.

Tabela 10 - Parametrização do CAIS

Área	**Parâmetro**	**Valor escolhido**	**Observação**
LLM	llm_model_max_async	1	Menos concorrência implica a menos timeouts
	options.temperature	0.1	Menos criatividade
	options.top_p	0.1	Mais conservador
	options.repeat_penalty	1.2	Evita repetições
	timeout	600	Tempo máximo por chamada
	keep_alive	180	Tempo de o modelo estar quente
Embeddings	embedding_dim	768	Dimensão do vetor
	max_token_size	8192	Tamanho máximo de entrada
Segmentação	chunk_token_size	1200	Valor por defeito
	chunk_overlap_token_size	100	Valor por defeito

Foi configurado tempos para os *timeouts* de cinco (5) minutos nas chamadas ao LLM devido aos documentos extensos. Por fim, foi aumentado o *keep-alive* para três (3) minutos para minimizar os *cold starts* e assim reduzir os pré-aquecimentos necessários.

A concorrência nas requisições aos modelos foram ajustadas entre uma (1) a duas (2) requisições paralelas, para promover maior estabilidade no processador e evitar *timeouts* durante o processamento de entidades e relações.

Foi reduzida a temperatura e o top-p, ambos para 0.1 e aplicadas penalizações a repetições para 1.2. Estes ajustes tornam a geração menos criativa e mais conservadora, o que diminui a probabilidade de conteúdos inventados.

A segmentação e a sobreposição permanecem com os valores por defeito, mil e duzentos (1200) e cem (100) respetivamente. Estes dois parâmetros, quando demasiado curtos, podem perder contexto e, quando demasiados longos, podem diluir o sinal.

A dimensão dos *embeddings* foi definido como setecentos e sessenta e oito (768), adequados para a recuperação geral, e o tamanho máximo de entrada para oito mil cento e noventa e dois (8192).

No momento da consulta, a entrada inserida no LLM passou a ter a instrução de responder apenas com base nos documentos recuperados de forma explícita, e caso não haja evidências o suficiente, o LLM gerar que não sabe com base nos documentos.

5 Avaliação

No presente Capítulo 5, Avaliação, é apresentada a avaliação que a solução desenvolvida foi exposta.

A avaliação consiste em duas partes, sendo a primeira a avaliação do grafo de conhecimento gerado pela *framework* LightRAG e do grafo de conhecimento gerado pela ferramenta analítica InfraNodus. A segunda parte é a avaliação da geração das respostas do CAIS com a utilização do grafo de conhecimento criado pelo LightRAG, onde é verificado se as respostas geradas apresentam todos elementos do domínio que fazem sentido face à pergunta inserida.

5.1 Avaliação do Grafo de Conhecimento

No presente subcapítulo, é avaliado o grafo de conhecimento gerado.

Numa primeira parte, a avaliação consiste na análise do grafo gerado pelo LightRAG e do grafo gerado pela ferramenta analítica InfraNodus. Posteriormente, realiza-se a comparação entre ambos.

Em seguida, é explicada a manipulação manual que pode ser feita ao grafo através do LightRAG, para a melhoria ou personalização deste.

Por fim, é exposta a possibilidade de utilizar o grafo gerado pelo InfraNodus como fonte de conhecimento do CAIS para que este pare de ter a responsabilidade de o gerar, ou como grafo de autoridade.

5.1.1 Comparação do Grafo de Conhecimento LightRAG e InfraNodus

5.1.1.1 Método de Avaliação

A abordagem adotada para avaliar grafos de conhecimento consiste em dois planos complementares, o quantitativo e o qualitativo (Newman, 2003).

No plano quantitativo, são analisadas propriedades estruturais do grafo com base em indicadores formais de grafo, o que permite caracterizar a conectividade de forma objetiva e comparável entre diferentes grafos.

O grau médio, ou razão entre arestas e nós, calcula quantas ligações, em média, cada nó tem. Em grafos não dirigidos, calcula-se como $\overline{K} = 2E/N$ onde E é o número de arestas e N o número de nós. Quanto maior o $\overline{K}$ mais conectado o grafo é.

A densidade calcula a proporção de arestas observadas face ao máximo possível. Em grafos não dirigidos, calcula-se como $D = 2E/(N(N-1))$. Valores próximos a um (1) indicam grafos quase completos e valores próximos de zero (0) indicam grafos esparsos.

No plano qualitativo, é examinado a adequação semântica do grafo ao domínio, a granularidade e consistência dos rótulos, a plausibilidade das ligações, a cobertura temática e a estabilidade face a atualizações do *corpus*.

A granularidade das entidades mede o quão finas ou grossas são as unidades do grafo. Um valor considerado bom é a granularidade ser entre uma (1) a três (3) palavras, estar lematizado, estáveis e reutilizáveis entre documentos.

A plausibilidade das relações mede o quão coerentes e sustentadas são as arestas entre as entidades. O ideal é ter relações tipadas, ou seja, parte de, causa efeito, pertence a, com evidencia textual clara e consistência de tipo, pessoa organização, conceito sub-conceito.

5.1.1.2 Grafo de Conhecimento LightRAG

O grafo de conhecimento gerado pelo LightRAG, como ilustrado na Figura 13 - Grafo de Conhecimento LightRAG, resultou num total de cinquenta (50) nós e cinco (5) arestas. Significa que o sistema identificou poucas entidades e quase não registou ligações entre elas, o que limita a capacidade para responder a perguntas que exigem síntese entre documentos, cadeias causa efeito ou navegação por tópicos aparentados, mas não contíguos.

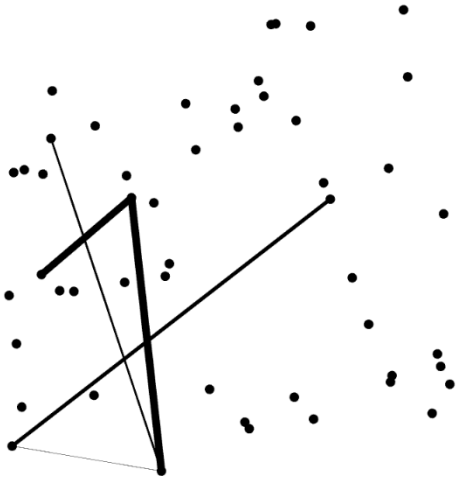


Figura 13 - Grafo de Conhecimento LightRAG

Existem vários motivos possíveis para esta pobre representação. Em primeiro lugar, a extração automática de entidades e relações tende a ser conservadora quando o modelo de linguagem é compacto, a janela de contexto é ampla, mas disputada, e os parâmetros de geração são restritivos. Esta combinação, embora benéfica para respostas factuais, leva frequentemente a sub-extração, o modelo arrisca menos o que faz propor menos entidades e relações.

Em segundo lugar, a granularidade das entidades extraídas foi inadequada. O extrator privilegiou frases completas e descrições longas, o que fez o grafo ficar povoado por rótulos pouco reutilizáveis e difíceis de ligar entre si. Uma entidade longa raramente coincide, palavra por palavra, com outras passagens relevantes, o que faz ilhas sem arestas. A ausência de um passo no fluxo do CAIS de seleção de palavras-chave ou de normalização por lematização/sinónimos agravou este efeito.

Por último, em execução local com concorrência baixa para evitar *timeouts*, a extração por vezes é interrompida com cortes de contexto, mesmo quando não falha, o sistema tende a adotar entradas mais curtas e objetivas, que favorecem a precisão, mas empobrecem o número de relações recolhidas. A inexistência de *re-ranking* dedicado na fase de extração também pode deixar oportunidades por explorar.

5.1.1.3 Grafo de Conhecimento InfraNodus

O grafo de conhecimento gerado pelo InfraNodus, como ilustrado na Figura 14 - Grafo de Conhecimento InfraNodus, resultou num total de trezentos e quarenta e três (343) nós e novecentos e trinta e sete (937) arestas. Ao transformar coocorrências de termos em conceitos e associações, a ferramenta gerou um mapa semântico denso que facilita a identificação de temas, subtemas e pontes entre tópicos que, no texto linear, podem surgir dispersos.

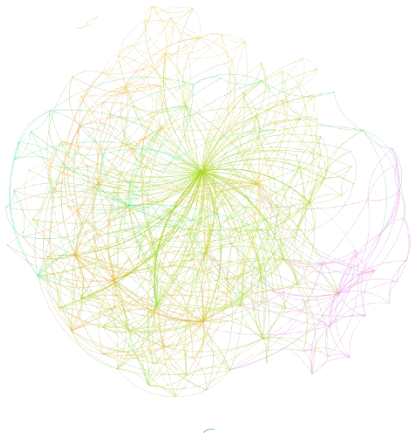


Figura 14 - Grafo de Conhecimento InfraNodus

Do ponto de vista metodológico, o InfraNodus aplica um *pipeline* de pré-processamento que consiste em normalização, remoção de *stopwords*, agregação de *n-grams* relevantes, e constrói o grafo com base na frequência e proximidade de coocorrências.

Este grafo oferece três benéficos. Primeiro, melhor cobertura pois mais nós significam maior probabilidade de o CAIS encontrar os conceitos que o utilizador invoca, o que reduz falhas por vocabulário. Segundo, contexto relacional mais robusto, as arestas captam padrões de coocorrência que orientam a seleção de evidências mesmo quando a resposta exige ligar secções afastadas ou documentos distintos. Terceiro, explicabilidade, comunidades e nós ponte tornam visível a estrutura do domínio, o que apoia justificações e navegação orientada por tópicos.

5.1.1.4 Comparação entre os Grafos de Conhecimento

Os grafos de conhecimento descritos nesta secção encontram-se expostas na seguinte Tabela 11 - Comparação dos Grafos de Conhecimento, onde são sintetizadas três características, o grau médio, a densidade e a granularidade, o que permite, deste modo uma comparação facilitada e sucinta. Após a tabela, encontra-se uma discussão dos pontos levantados pela mesma.

Tabela 11 - Comparação dos Grafos de Conhecimento

Características	**Grafo de Conhecimento LightRAG**	**Grafo de Conhecimento InfraNodus**
Grau médio	0,2	5,464
Densidade	0,00408	0,01598
Granularidade	*n-grams* longos (frases)	*n-grams* curtos (um a três)

No grafo de conhecimento do LightRAG, o grau médio é de 0,2 e a densidade é de 0,00408, o que indica um grafo esparso, poucos nós ligam-se entre si e grande parte permanece como ilha. Com entidades longas, os rótulos repetem-se pouco, o que dificulta a criação de arestas e de caminhos úteis. Isto favorece perguntas factuais e localizadas, em que a evidência está num único excerto, assim, os modos *naive* e local tendem a funcionar melhor. O modo global acrescenta pouco, e o híbrido só traz ganhos se a recuperação vetorial já encontrar boa evidência, portanto, para raciocínio multidocumento ou relações causa efeito, este grafo é inadequado.

No grafo de conhecimento do InfraNodus, o grau médio é de 5,464 e densidade é de 0,01598, o que indica um grafo mais conectado, cerca de vinte e sete vezes mais ligações por nó e quatro vezes mais denso. A granularidade promove reutilização e criação de pontes entre tópicos, o que melhora a cobertura e a navegabilidade. Este tipo de grafo beneficia os modos global e híbrido, adequados a perguntas que exigem síntese entre documentos, contexto relacional e explicabilidade. Os modos *naive* e local também ganham estabilidade por haver mais contexto coerente por perto, mas o diferencial surge quando o CAIS precisa de ligar conceitos dispersos e justificar o percurso no grafo.

5.1.2 Manipulação do Grafo de Conhecimento LightRAG de Forma Manual

O LightRAG permite realizar uma curadoria manual do grafo de conhecimento. Esta curadoria pode complementar, ou substituir em partes, a extração automática, o que permite acrescentar nós e arestas em falta, fundir duplicados e corrigir rótulos (Data Intelligence Lab@HKU, n.d.).

5.1.2.1 Antes da extração

É possível guiar o extrator para este privilegiar *n-grams* curtos e reutilizáveis em vez de frases longas, através da injeção de palavras-chaves ou de vocabulários controlados, que são constituídos por listas de termos, sinónimos ou abreviaturas. Também é possível realizar a normalização lexical e criar *stoplists* específicas do domínio para reduzir ruído.

Os benefícios são a não alteração da base do LightRAG e orientação da extração para uma granularidade mais útil.

5.1.2.2 Após a extração

É possível realizar a edição do grafo de conhecimento após a extração, isto é, adicionar nós e arestas, remover ligações espúrias, fundir rótulos equivalentes ou corrigir tipos de relação.

Estas edições devem respeitar a coerência referencial, como por exemplo, as arestas a apontar para identificadores válidos, e a coerência vetorial. Qualquer nó novo precisa de um vetor de

embedding gerado pelo mesmo modelo com a mesma dimensão usado no CAIS, no caso deste desenvolvimento 768 dimensões com nomic-embed-text via Ollama.

O benefício consiste na correção das lacunas reais do grafo sem reindexar todo o *corpus*.

5.1.3 Utilização do Grafo de Conhecimento InfraNodus

O grafo de conhecimento produzido pelo InfraNodus, ao oferecer uma melhor representação, pode servir tanto como fonte de enriquecimento para preencher lacunas, ou como referência de validação para detetar erros como nós redundantes ou ligações espúrias, ou como substituição total do grafo de conhecimento do LightRAG.

5.1.3.1 Fonte de Enriquecimento

O processo começa pela resolução de entidades e do mapeamento de relações do grafo exportado. Como o InfraNodus representa coocorrências, é necessário converter estas arestas para um tipo genérico e caso existam relações tipadas fiáveis no grafo do LightRAG, estas devem prevalecer, portanto, as coocorrências do InfraNodus entram como reforço ou sugestão a validar. Quando necessário, define-se um limiar de peso para filtrar coocorrências muito fracas, o que reduz ruído e mantem a legibilidade estrutural.

Para garantir compatibilidade na recuperação semântica, cada novo nó importado deve receber *embeddings* gerados com o mesmo modelo usado. Este cuidado, preserva a coerência do índice vetorial e evita degradações de desempenho.

5.1.3.2 Fonte de Referência

Ambos os grafos de conhecimento são comparados em termos de cobertura, centralidades e modularidade para gerar listas de curadoria, que podem representar pontes entre grafos, junção de duplicados ou a remoção de termos genéricos.

Após a geração da lista, é realizado a manipulação manual do grafo de conhecimento do LightRAG através dos métodos de curadoria que este oferece.

5.1.3.3 Substituição Total

O grafo de conhecimento externo pode ser totalmente adotado quando a qualidade deste é comprovadamente superior. Este recurso permite deslocar a responsabilidade de estruturação relacional do grafo do CAIS para uma aplicação externa, ficando somente como consumidor deste para a geração das respostas.

5.2 Avaliação do CAIS

No presente subcapítulo, é avaliado o desempenho do CAIS através da *framework* Ragas.

Numa primeira parte, é explicado o que é o Ragas e o seu funcionamento. Em seguida são explicadas algumas das principais métricas fornecidas pela *framework*, a gama dos valores possíveis de se obter e o significado deles.

Numa segunda parte, são expostas as perguntas construídas para o CAIS responder, a avaliação efetuada, a amostragem das métricas obtidas e uma discussão sobre os resultados obtidos, o que fez o CAIS atingir aqueles valores e o que eles significam na prática.

5.2.1 Ragas

O Ragas[8] é uma *framework* construída para avaliar CAISs construídos sobre o RAG. Em vez de comparar palavra a palavra, o Ragas atua com *embeddings* e, quando necessário, com um juiz LLM para julgar pertinência e fidelidade.

A finalidade do Ragas é oferecer métricas reprodutíveis que cubram três dimensões. A primeira dimensão é o conteúdo, que consiste se a resposta aborda os temas certos e alinha com uma referência. A segunda dimensão é o uso de contexto, que consiste em saber se o sistema recupera e utiliza os trechos certos. Por fim, a última dimensão é a fidelidade, que é se a resposta está ancorada no material fornecido (Ragas, n.d.-a).

5.2.2 Métricas de Avaliação do Ragas

O Ragas possui cinco métricas de avaliação, que determinam o quão corretas as respostas geradas pelo CAIS estão (Ragas, n.d.-b).

5.2.2.1 Answer_Similarity

Esta métrica avalia a proximidade semântica entre a resposta do CAIS e uma resposta de referência humana. Fá-lo através da projeção de ambos os textos num espaço vetorial de *embeddings* e calcula a similaridade. Os valores variam entre 0 e 1, quanto mais perto de 1, maior a sobreposição de conteúdo ou ideias, perto de 0 indica respostas que tratam temas diferentes. Esta métrica serve para avaliar a cobertura de conteúdo, mas não verifica, por si só, se a resposta está fiel ao contexto fornecido.

[8] https://docs.ragas.io/en/stable/

5.2.2.2 Context_Precision

Esta métrica mede a precisão da recuperação entre os trechos de contexto usados pelo CAIS, que proporção é realmente relevante para responder à pergunta. Marca-se cada segmento como relevante ou irrelevante e calcula-se a fração de relevantes. Vai de 0 a 1, onde 1 significa quase só sinal e pouco ruído, valores baixos revelam que o recuperador trouxe muito material inútil, o que pode confundir o gerador.

5.2.2.3 Context_Recall

Esta métrica mede a cobertura da recuperação, de todo o conteúdo relevante que existia na base, que parte foi efetivamente recuperada para o gerador. Na prática, estima-se um conjunto de trechos que deveria ter trazido e vê-se a fração que veio. Vai de 0 a 1, onde 1 implica que o CAIS trouxe tudo o que importava, valores baixos expõem perdas de informação chave que limitam o gerador.

5.2.2.4 Context_Entity_Recall

Esta métrica foca-se nas entidades presentes no contexto relevante e verifica se essas entidades aparecem ou são corretamente refletidas na resposta. Normalmente combina extração de entidades e correspondência semântica. Vai de 0 a 1, onde 1 sugere que os elementos críticos do contexto foram transportados para a resposta, valores baixos denunciam omissões de nomes e referências essenciais.

5.2.2.5 Context_Utilization

Esta métrica avalia em que medida a resposta utiliza o contexto recuperado, em vez de alucinar ou apoiar-se em conhecimento externo não fornecido. Implementa-se através da resposta com os segmentos utilizados e da agregação do grau de alinhamento. Também varia entre 0 a 1, onde 1 significa resposta fortemente ancorada no contexto, valores baixos sugerem extrapolação ou alucinações.

5.2.3 Perguntas e Respostas de Avaliação

Foram desenvolvidas dez (10) perguntas sobre o domínio dos laudos de honorários e deontologia médica para estas serem respondidas tanto pelo CAIS desenvolvido como por um ser humano com conhecimento no domínio.

As respostas geradas pelo CAIS serão avaliadas através das respostas do ser humano que servem como base de verdade, pelo contexto recuperado pelo CAIS para perceber o quão aproveitado ou alucinado a geração foi, e pela própria pergunta que serve como referência.

As perguntas desenvolvidas estão representadas na Tabela 12 - Perguntas de Avaliação e as respetivas respostas geradas pelo CAIS encontram-se nas secções abaixo.

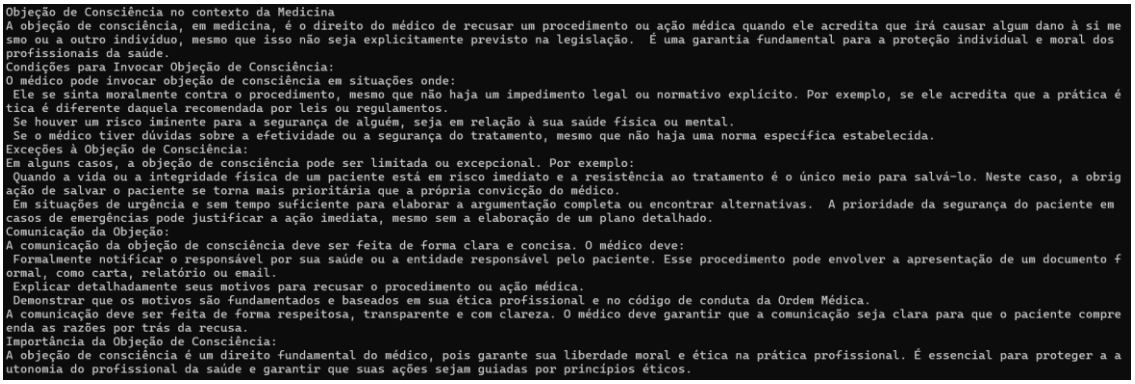
Tabela 12 - Perguntas de Avaliação

ID	Domínio	Pergunta
P1	Ética; Saúde; Pacientes	Em que condições um médico pode invocar objeção de consciência e quando é que não o pode fazer? Indica também como deve ser comunicada a objeção.
P2	Ética; Pacientes	O que é o segredo médico e qual o seu âmbito? Dá três exemplos de informação coberta e diz se o dever se mantém após a morte do doente?
P3	Ética; Saúde; Pacientes	Em que situações o médico pode ou deve revelar informação clínica sem violar o segredo?
P4	Ética; Pacientes	Quais são os requisitos de validade do consentimento do doente, deve existir intervalo de reflexão, e em que formas pode ser dado?
P5	Ética; Pacientes	Como deve o médico atuar quando o doente é incapaz de consentir? Qual o papel do representante legal, das diretivas antecipadas e do procurador de cuidados de saúde, e o que significa decidir de acordo com os melhores interesses do doente?
P6	Ética; Saúde; Pacientes	Numa teleconsulta de urgência com informação incompleta, quem é responsável pela decisão clínica? Que registos devem ser feitos na ficha clínica e que medidas de segurança/confidencialidade são exigidas?
P7	Ética; Saúde; Pacientes	Em fase de fim de vida, quando devem ser interrompidos meios extraordinários de manutenção da vida? Hidratação e alimentação artificiais são considerados extraordinários?
P8	Ética	Quais as regras para a publicidade da atividade médica e o uso de títulos profissionais? É lícito anunciar "resultados garantidos"?
P9	Ética; Honorários	Como devem ser fixados e cobrados os honorários em prática clínica e o que é vedado ao médico?
P10	Honorários	No Regulamento dos Laudos a Honorários, quem pode pedir um laudo, quais os pressupostos, e como deve ser instruído o pedido? É lícita a cobrança por falta à consulta programada sem aviso prévio?

5.2.3.1 P1

Esta pergunta insere-se sobretudo na área de Ética por envolver o equilíbrio entre autonomia moral do profissional de saúde e os princípios de beneficência e não-maleficência. Toca também na área de Saúde, pois delimita a obrigação de atuação em emergências e a continuidade de cuidados, garantindo a proteção da saúde pública quando a recusa é inadmissível. Por fim, relaciona-se com a área de Pacientes, já que a decisão do médico impacta diretamente o acesso do doente ao cuidado e a necessidade de comunicação clara e atempada para salvaguardar a dignidade e a autonomia do paciente.

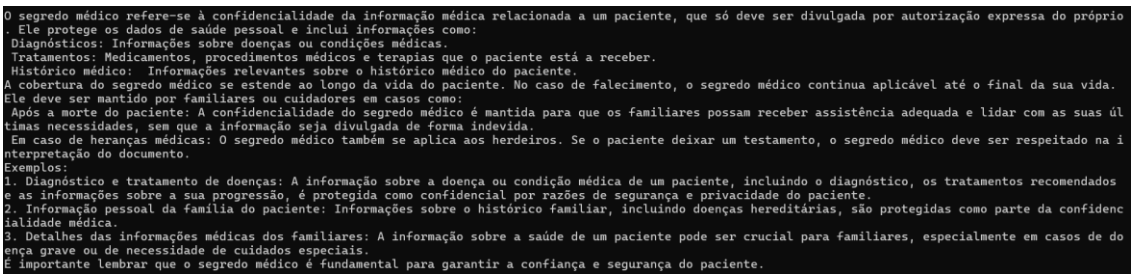
Resposta gerada pelo CAIS



5.2.3.2 P2

Esta pergunta insere-se sobretudo na área de Ética por se tratar do sigilo profissional e do respeito pela vida e dignidade. Enquadra-se igualmente na área de Pacientes, pois a confidencialidade é condição para a confiança na relação clínica e para o exercício da autonomia informada. Ao assegurar que os dados sensíveis permanecem protegidos, a pergunta reforça o respeito devido ao paciente nas suas dimensões pessoal e jurídica.

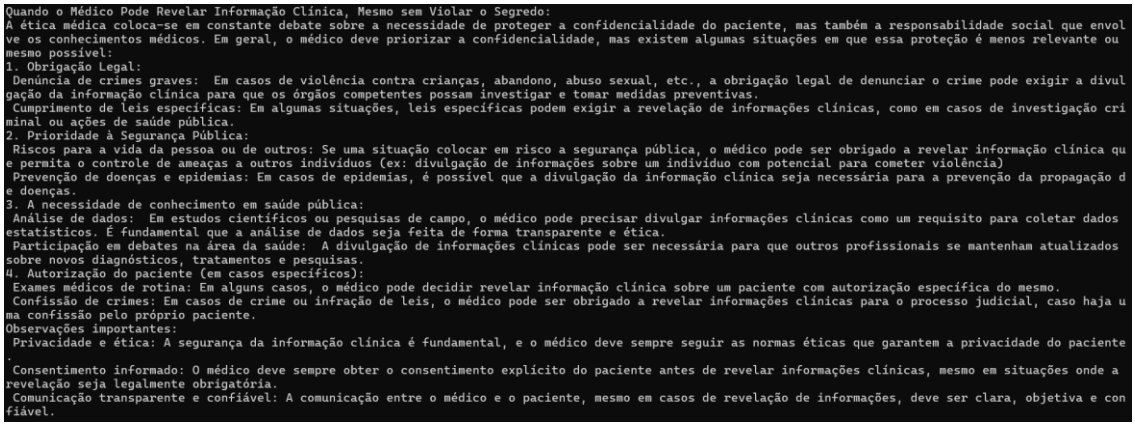
Resposta gerada pelo CAIS



5.2.3.3 P3

Esta pergunta insere-se sobretudo na área de Ética por ponderar justiça, beneficência e não-maleficência frente a terceiros potencialmente em risco. Cruza-se na área de Saúde ao refletir deveres de proteção da saúde pública e protocolos de melhores práticas. Permanece ligada à área de Pacientes, pois o consentimento é expressão direta da autonomia e da confiança na relação médico-doente.

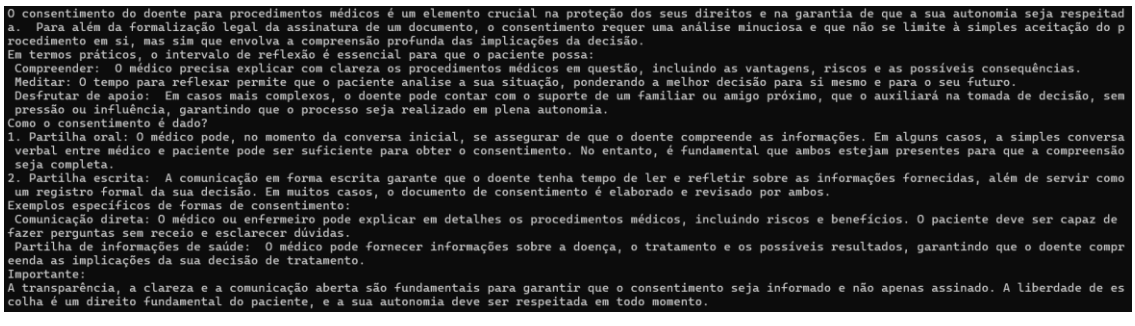
Resposta gerada pelo CAIS



5.2.3.4 P4

Esta pergunta insere-se sobretudo na área de Pacientes ao enfatizar autonomia, dignidade e livre escolha do médico. Ao mesmo tempo, cruza-se com a área de Ética, dado que o consentimento é a materialização dos princípios de autonomia e não-maleficência, o que orienta a conduta do profissional para decisões partilhadas e responsáveis.

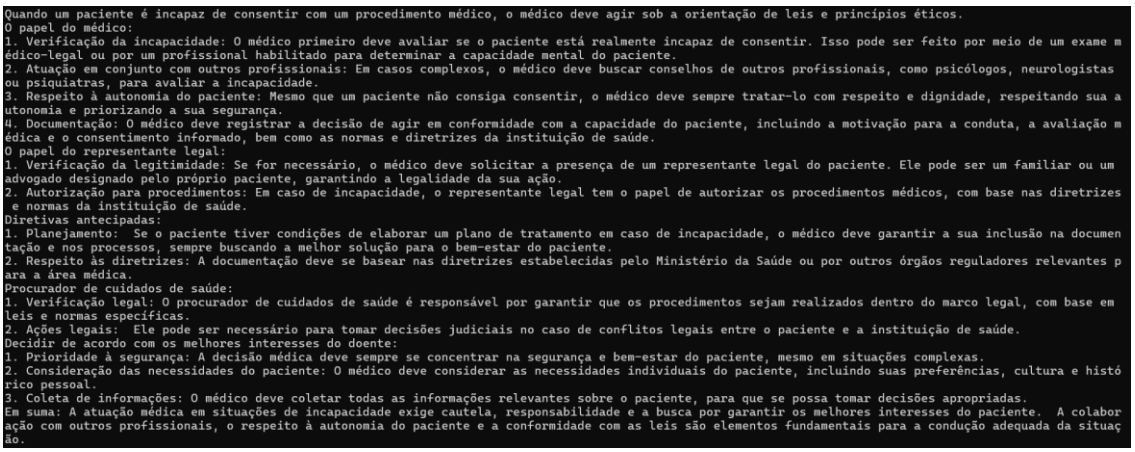
Resposta gerada pelo CAIS



5.2.3.5 P5

Esta pergunta insere-se sobretudo na área de Pacientes ao visar como decidir quando o doente não pode exprimir vontade pela tutela da autonomia previamente manifestada e pela proteção da dignidade. Cruza-se na área de Ética, pois requer ponderar beneficência, não-maleficência e justiça na determinação dos melhores interesses. O papel do representante legal e do procurador de cuidados assegura que a decisão respeita, tanto quanto possível, os valores do doente.

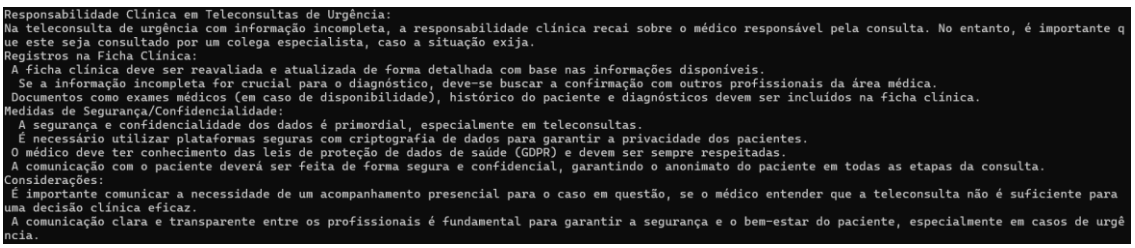
Resposta gerada pelo CAIS



5.2.3.6 P6

Esta pergunta insere-se tanto na área de Saúde, Ética e de Pacientes ao atribuir responsabilidade clínica e definir registos e salvaguardas de confidencialidade em, por tratar de melhores práticas, qualidade assistencial e obrigação de atuação em emergências mesmo em contextos digitais, envolve a proteção da confidencialidade e a clareza informativa numa interação direta médico e doente.

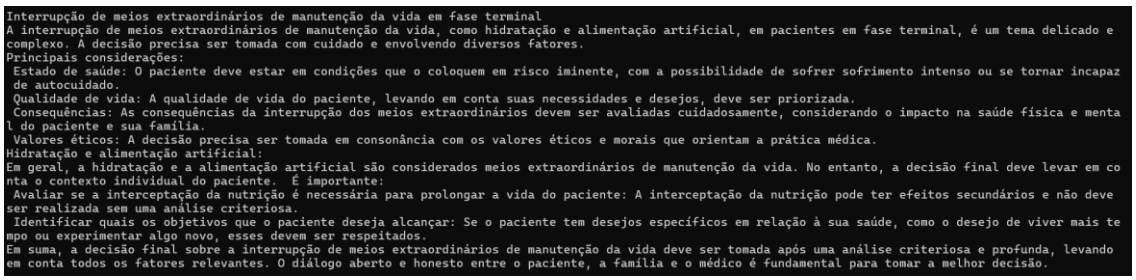
Resposta gerada pelo CAIS



5.2.3.7 P7

Esta pergunta insere-se sobretudo na área de Ética ao decidir quando interromper meios extraordinários de manutenção da vida, e como qualificar hidratação e alimentação artificiais, por lidar com dilemas de beneficência, não-maleficência e respeito pela vida e pela autonomia. Cruza-se com a área de Saúde, pois exige julgamento técnico sobre proporcionalidade de tratamentos e boas práticas clínicas, e cruza-se com a área de Pacientes dado o impacto direto sobre dignidade, sofrimento e vontade do doente e sua família.

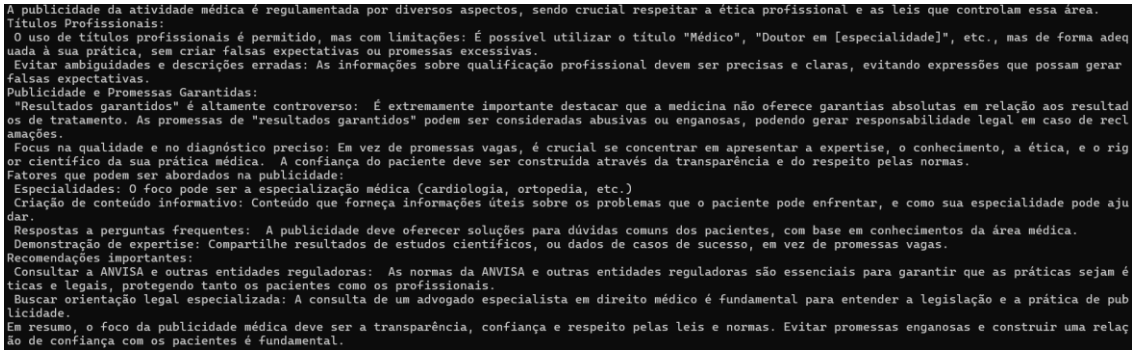
Resposta gerada pelo CAIS



5.2.3.8 P8

Esta pergunta insere-se somente na área de Ética devido às regras sobre comunicação pública, títulos e proibições a promessas de resultados garantidos, pois protegem a justiça e a honestidade profissional, o que previne práticas enganosas e preserva a confiança social no médico. Ao coibir informação imprópria, salvaguardam a tomada de decisão informada dos doentes.

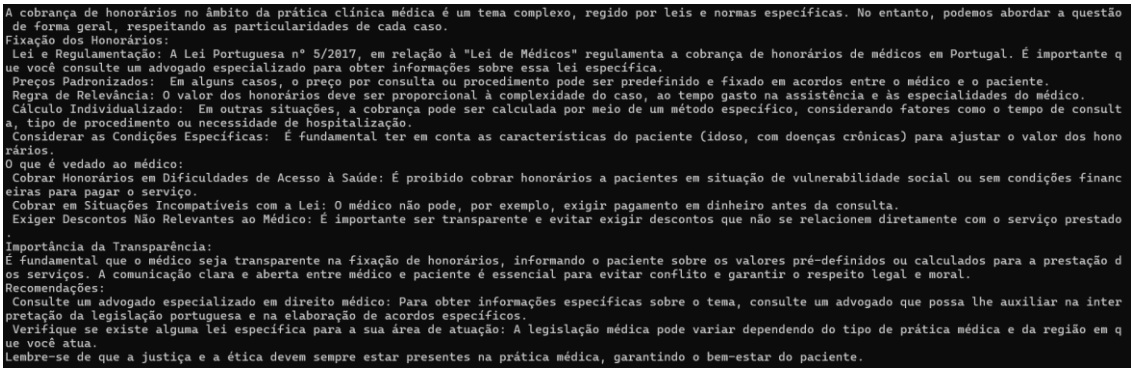
Resposta gerada pelo CAIS



5.2.3.9 P9

Esta pergunta insere-se sobretudo na área de Honorários devido à forma de definir preços, discriminar contas e vedar partilhas indevidas por referenciação, ao oferecer diretrizes de transparência e justa compensação pelo trabalho médico. Cruza-se com a área de Ética, pois práticas financeiras desleais violam justiça e integridade, diminuindo a confiança na profissão e interferindo com a independência clínica.

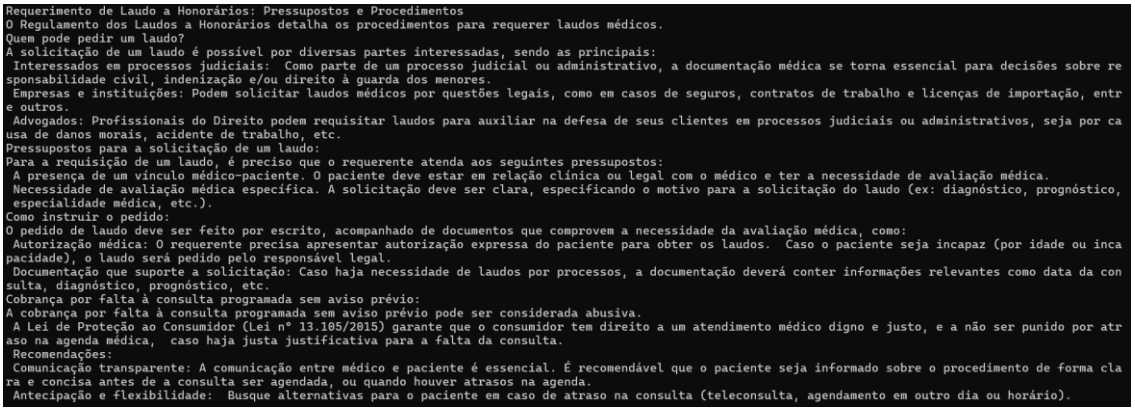
Resposta gerada pelo CAIS



5.2.3.10 P10

Esta pergunta insere-se somente na área de Honorários, pois compreende os requisitos para solicitar um laudo, os pressupostos temporais e a instrução do pedido, por regularem mecanismos de resolução de divergências e assegurar transparência e proporcionalidade nos valores cobrados.

Resposta gerada pelo CAIS



5.2.4 Discussão da Avaliação Obtida

Os resultados das métricas de avaliação do Ragas encontram-se exposta na seguinte Tabela 13 - Métricas obtidas Ragas

Tabela 13 - Métricas obtidas Ragas

Métrica	**Resultado**
answer_similarity	0,84
context_precision	0,24
context_recall	0,93
context_entity_recall	0,92
context_utilization	0,58

Na métrica answer_similarity o valor é de 0,84. Indica uma forte sobreposição semântica com a resposta de referência, o que significa que o CAIS cobre, em geral, os mesmos tópicos e conclusões. Existe uma boa cobertura do que interessa, mas não necessariamente máxima precisão factual ao nível de pormenores (Ragas, n.d.-a).

Na métrica context_precision o valor é de 0,24. O recuperador devolveu quase todo o *corpus* para as diferentes perguntas, o que torna a proporção de trechos realmente necessários para responder baixa. Existe muito ruído a acompanhar o sinal, o que leva a precisão para valores baixos. O efeito é a diluição da entrada no gerador, o que aumenta o custo, o risco de dispersão e a probabilidade de o modelo pegar em passagens menos relevantes (Ragas, n.d.-a).

Na métrica context_recall o valor é de 0,93. Como o recuperador devolveu quase todo o *corpus* em cada pergunta, a cobertura do que era relevante é praticamente total. Por isso, o *recall* aproxima-se de 1. Raramente faltam peças de informação importantes na entrada do gerador (Ragas, n.d.-a).

Na métrica context_entity_recall o valor é de 0,92. Ao incluir a maioria dos trechos, quase todas as entidades relevantes acabam por estar presentes e, em boa parte, a resposta também as reflete. Ficam lacunas apenas quando o modelo ignora algumas entidades por *overload* ou quando há muitas variantes de nome (Ragas, n.d.-a).

Na métrica context_utilization o valor é de 0,58. Apesar de ter o contexto certo na entrada para o gerador, o modelo não o usa totalmente, pois existe muito texto irrelevante, o que faz tender a apoiar-se em fragmentos mais superficiais ou em conhecimento geral. O efeito é uma resposta bem alinhada em termos de tema, mas com ancoragem só parcial no contexto, o suficiente para acertar o essencial, mas com margem para pequenas generalizações ou omissões (Ragas, n.d.-a).

6 Conclusão

O projeto desenvolvido demonstrou a viabilidade de um CAIS especializado em laudos de honorários e deontologia médica assente na *framework* LightRAG, que combina a técnica de RAG com a utilização de grafos de conhecimento.

Foram apresentadas abordagens com *grounding* para reduzir o risco de respostas alucinadas, apesar disto, a eficiência depende da qualidade do grafo, da capacidade do recuperador e do gerador. Com um grafo pobre composto por cinquenta (50) nós e cinco (5) relações, o que corresponde a um grau médio de 0.2 e uma densidade de 0.00408, e com modelos locais de baixa dimensão, foi obtido uma cobertura temática, de 84%, e uma recuperação do contexto e de entidades, de 93% e 92% correspondente, mas limitações na utilização efetiva do contexto, de 58%, e na precisão, de 24%.

6.1 Contextualização

O domínio de laudos de honorários e deontologia médica é complexo e normativamente denso, que envolve diversas diretrizes deontológicas, regulamentos internos, protocolos clínicos e regras de honorários que podem variar entre instituições. Esta variabilidade e complexidade gera dúvidas recorrentes entre os profissionais de saúde, cuja resolução recai tipicamente no departamento de HR, o faz gastar tempo e recursos humanos (Nádia Ventura, 2022). Em paralelo, os CAIS estão a ganhar destaque para este tipo de problemas de esclarecimento (Android Standard Blog, n.d.; Mordor Intelligence, n.d.).

A dissertação enquadra-se neste cenário de testar uma arquitetura que sirva respostas rápidas, alinhadas com o domínio e sustentadas em documentação institucional para garantir a factualidade.

6.2 Combate à Alucinação

Um dos problemas dos CAIS é o fenómeno de alucinação, onde cerca de 30% das respostas geradas por estes sistemas podem conter o fenómeno (Z. Ji et al., 2022). Este fenómeno consiste na geração de respostas plausíveis, mas factualmente erradas quando a evidência é ambígua, escassa ou mal recuperada.

Em domínios sensíveis, como o dos laudos de honorários e deontologia médica, é necessário diminuir a probabilidade deste fenómeno, medida imposta por exemplo, pelo AI Act (Parlamento Europeu e do Conselho, 2024), por isso, é preciso garantir o *grounding* documental e mecanismos de contenção.

Nesta dissertação adotou-se uma dupla estratégia, a utilização de grafos de conhecimento (S. Ji et al., 2022) para estruturar entidades, relações e restrições do domínio, e a utilização do RAG (Gao et al., 2023) para que cada resposta seja condicionada por trechos de contexto recuperados a partir do grafo de conhecimento.

6.3 Implementação

O CAIS foi implementado com base na *framework* LightRAG, que integra recuperação semântica e razão sobre um grafo de conhecimento (Guo et al., 2025).

Na fase de indexação, é construído o grafo de conhecimento e é gerado o índice de vetores. Os documentos são segmentados e submetidos à extração de entidades e de relações, seguidas de normalização e desambiguação para reduzir duplicados e garantir coerência lexical. As entidades validadas originam nós e as relações originam arestas, ambas enriquecidas com proveniência. O índice de vetores resultante, que armazena as representações semânticas de cada segmento, mantém versionamento e suporta atualizações incrementais para que o grafo evolua com o *corpus*.

A consulta do utilizador aciona o recuperador para que este obtenha as passagens relevantes, anote as entidades, ligue os conceitos e, quando possível, restringa ou priorize evidências. Durante a recuperação, o sistema compara e identifica os segmentos mais semânticos, o que maximiza a relevância. Após esta fase, o gerador gera a resposta ancorada nos excertos recuperados, utilizando os vetores como base para garantir a coerência e a precisão da resposta.

Esta arquitetura materializa o princípio de *grounding*, pois o modelo é incentivado a citar e usar o que foi recuperado, o que reduz a liberdade criativa que leva a alucinações, e permite rastrear a origem de cada afirmação (Data Intelligence Lab@HKU, n.d.; GeeksforGeeks, 2025b; Guo et al., 2025).

6.4 Limitações Encontradas

No presente subcapítulo, são expostos os principais desafios e limitações observados no desenvolvimento e na utilização do CAIS, bem como medidas de mitigação e soluções propostas.

6.4.1 Necessidade de PDFs Textuais

A *pipeline* assume PDFs com camada de texto, portanto, os documentos digitalizados não são legíveis pelo extrator usado e ficam, por omissão, fora do índice. Este constrangimento limita a cobertura.

As soluções passam em acrescentar um passo de OCR local anterior à indexação. Mas, mesmo com OCR, elementos como tabelas complexas, fórmulas e figuras exigem tratamento específico.

6.4.2 Modelos de Pequena Dimensão

O fraco desempenho do CAIS deveu-se à escolha de modelos locais e de pequena dimensão, tanto nos *embeddings* como no LLM. O grafo de conhecimento gerado pelo LightRAG resultou em cinquenta (50) nós e cinco (5) relações, o que corresponde a um grau médio de 0.2 e uma densidade de 0.00408, o que limita a capacidade de extração e de ligação de entidades. Um grafo pobre reduz a capacidade de desambiguar conceitos e de propagar restrições (Newman, 2003).

Na recuperação, os mesmos modelos comprometeram a seletividade, o que originou uma recuperação excessiva e irrelevante e, por consequência, a utilização parcial do que é relevante na geração. Estes efeitos observam-se nas métricas discutidas, foi obtida uma cobertura temática de 84%, mas uma precisão e utilização do contexto de 24% e 58% respetivamente, apesar de ter sido obtido uma recuperação do contexto e de entidades necessárias de 93% e 92% respetivamente.

O CAIS responde com coerência temática, mas ainda carece de maior precisão factual e melhor foco nas passagens nucleares, limitações devido à pequena capacidade dos modelos adotados.

6.4.3 Elevado Tempo de Indexação

A fase de preparação e ingestão é a mais exigente em recursos computacionais. Em *hardware* sem GPU, e com modelos locais e leves, esta etapa prolongou-se até cerca de uma hora e meia para coleções de um total de trinta e cinco (35) páginas de texto, o que condiciona a cadência de atualizações.

As mitigações passam pela evolução de modelos de maior dimensão e um *hardware* que os consiga suportar.

6.5 Trabalho Futuro

No presente subcapítulo, é exposto e discutido a evolução que o protótipo do CAIS desenvolvido e avaliado pode ter no futuro. Existem um conjunto de evoluções técnicas que podem elevar a robustez e o desempenho do CAIS ou do grafo de conhecimento.

6.5.1 Integração de OCR para PDFs

Neste momento, a ingestão pressupõe PDFs com camada de texto. A integração de um processo de OCR na *pipeline*, anterior à segmentação, permitiria indexar estes documentos, o que aumentaria a cobertura.

A implementação deve contemplar a seleção do motor de OCR, normalização do texto, e marcação dos segmentos com metadados de confiança, para que o utilizador saiba quando a resposta assenta em transcrição automática potencialmente ruidosa.

6.5.2 Re-ranking

A recuperação inicial por *embeddings* pode ser melhorada com um processo de *re-ranking* que reordena os candidatos através de um modelo de *cross-encoder*.

Esta etapa tende a melhorar a precisão do contexto passado ao LLM quando os segmentos são semanticamente próximos entre si (Data Intelligence Lab@HKU, n.d.).

6.5.3 Evolução dos Modelos de *Embeddings* e de LLM

Para reduzir alucinações e elevar a precisão, propõe-se migrar para modelos de *embeddings* e de LLM de maior dimensão. *Embeddings* maiores potenciarão um grafo de conhecimento mais denso, correto e desambiguado. Na recuperação com recuperação híbrida e com *reranking* deverá aumentar o context_precision sem sacrificar o *recall*. Um LLM mais capaz, com entradas orientadas a citações e penalização de conteúdo não suportado, melhorará a context_utilization e irá ancorar as respostas no *corpus*, o que reduzirá as alucinações.

Para comprovar os ganhos, propõe-se comparar sistematicamente as configurações atuais com modelos de maior dimensão através de um *benchmark* controlado, com mesmo conjunto de perguntas, mesmas fontes e a utilização das mesmas métricas, complementadas por avaliação humana.

7 Referências

Abhinav, K., Dubey, A., Jain, S., Bhatia, G. K., McCartin, B., & Bhardwaj, N. (2018). Crowdassistant: A virtual buddy for crowd worker. *Proceedings - International Conference on Software Engineering*, 17–20. https://doi.org/10.1145/3195863.3195865

Android Standard Blog. (n.d.). *Estatísticas de chatbots para 2023: uso, dados demográficos, tendências, mercados*. Retrieved January 18, 2025, from https://androidstandard.com/estatisticas-de-chatbots-para-2023-uso-dados-demograficos-tendencias-mercados/

Bai, X., He, S., Li, Y., Xie, Y., Zhang, X., Du, W., & Li, J. R. (2025). Construction of a knowledge graph for framework material enabled by large language models and its application. *Npj Computational Materials*, *11*(1). https://doi.org/10.1038/s41524-025-01540-6

Data Intelligence Lab@HKU. (n.d.). *LightRAG GitHub*. Retrieved August 24, 2025, from https://github.com/HKUDS/LightRAG/blob/main/README.md

Ganea, O., Hofmann, T. (2017). *Deep Joint Entity Disambiguation with Local Neural Attention*. https://arxiv.org/pdf/1704.04920

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). *Retrieval-Augmented Generation for Large Language Models: A Survey*. http://arxiv.org/abs/2312.10997

GeeksforGeeks. (2025a, February 20). *What is GraphRAG?* https://www.geeksforgeeks.org/blogs/what-is-graphrag/

GeeksforGeeks. (2025b, June 26). *LightRAG*. https://www.geeksforgeeks.org/artificial-intelligence/lightrag/

Goku Mohandas, & Philipp Moritz. (2023, October 25). *Building RAG-based LLM Applications for Production*. https://www.anyscale.com/blog/a-comprehensive-guide-for-building-rag-based-llm-applications-part-1

Guo, Z., Xia, L., Yu, Y., Ao, T., & Huang, C. (2025). *LightRAG: Simple and Fast Retrieval-Augmented Generation*. http://arxiv.org/abs/2410.05779

Han, H., Wang, Y., Shomer, H., Guo, K., Ding, J., Lei, Y., Halappanavar, M., Rossi, R. A., Mukherjee, S., Tang, X., He, Q., Hua, Z., Long, B., Zhao, T., Shah, N., Javari, A., Xia, Y., & Tang, J. (2025). *Retrieval-Augmented Generation with Graphs (GraphRAG)*. http://arxiv.org/abs/2501.00309

Hoang Tran. (2023, September 20). *Which is better, retrieval augmentation (RAG) or fine-tuning? Both.* https://snorkel.ai/blog/which-is-better-retrieval-augmentation-rag-or-fine-tuning-both/

Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2022). A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(2), 494–514. https://doi.org/10.1109/TNNLS.2021.3070843

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Chan, H. S., Madotto, A., & Fung, P. (2022). *Survey of Hallucination in Natural Language Generation*. https://doi.org/10.1145/3571730

Kolitsas, N,. Ganea, O. (2018). *End-to-End Neural Entity Linking*. https://arxiv.org/pdf/1808.07699

Lee, H., Joo, S., Kim, C., Jang, J., Kim, D., On, K.-W., & Seo, M. (2024). *How Well Do Large Language Models Truly Ground?* http://arxiv.org/abs/2311.09069

Liang, K., Meng, L., Liu, M., Liu, Y., Tu, W., Wang, S., Zhou, S., Liu, X., Sun, F., & He, K. (2024). A Survey of Knowledge Graph Reasoning on Graph Types: Static, Dynamic, and Multi-Modal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2024.3417451

Liu, J. C., Goetz, J., Sen, S., & Tewari, A. (2021). Learning from others without sacrificing privacy: Simulation comparing centralized and federated machine learning on mobile health data. *JMIR MHealth and UHealth*, *9*(3). https://doi.org/10.2196/23728

Meshram, S., Naik, N., Vr, M., More, T., & Kharche, S. (2021, June 25). Conversational AI: Chatbots. *2021 International Conference on Intelligent Technologies, CONIT 2021*. https://doi.org/10.1109/CONIT51480.2021.9498508

Mordor Intelligence. (n.d.). *Tamanho do mercado Chatbot & Análise de participação - Tendências de crescimento e previsões (2024 - 2029)*. Retrieved January 18, 2025, from https://www.mordorintelligence.com/pt/industry-reports/global-chatbot-market

Mousavi, S. M., Alghisi, S., & Riccardi, G. (2025). *LLMs as Repositories of Factual Knowledge: Limitations and Solutions*. http://arxiv.org/abs/2501.12774

Nádia Ventura. (2022, November 11). *Gestão de custos: o que é e qual o papel dos RH?* Factorial. https://factorialhr.pt/blog/gestao-de-custos/

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2024). *A Comprehensive Overview of Large Language Models*. http://arxiv.org/abs/2307.06435

Newman, M. E. J. (2003). *The structure and function of complex networks*.

Ordem dos Médicos. (n.d.). *Regulamento dos Laudos a Honorários*.

Ordem dos Médicos. (2016). *Regulamento de Deontologia Médica*.

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering*, *36*(7), 3580–3599. https://doi.org/10.1109/TKDE.2024.3352100

Parlamento Europeu e do Conselho. (2016). *REGULAMENTO (UE) 2016/67*.

Parlamento Europeu e do Conselho. (2024). *REGULAMENTO (UE) 2024/168*. http://data.europa.eu/eli/reg/2024/1689/oj

Rachana, T. V., Vishwas, H. N., & Nair, P. C. (2022). HR based Chatbot using Deep Neural Network. *5th International Conference on Inventive Computation Technologies, ICICT 2022 - Proceedings*, 130–139. https://doi.org/10.1109/ICICT54344.2022.9850474

Ragas. (n.d.-a). *Ragas Core Concepts*. Retrieved September 22, 2025, from https://docs.ragas.io/en/v0.1.21/concepts/index.html

Ragas. (n.d.-b). *Ragas Metrics*. Retrieved September 22, 2025, from https://docs.ragas.io/en/v0.1.21/concepts/metrics/

Sen, J. (2013). *Security and Privacy Issues in Cloud Computing*.

Shen, W., Wang, J., & Han, J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, *27*(2), 443–460. https://doi.org/10.1109/TKDE.2014.2327028

Stanford University. (2025). *AI Index 2025*. https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

Suakanto, S., Siswanto, J., Febrianti Kusumasari, T., Reza Prasetyo, I., & Hardiyanti, M. (2021, August 2). Interview Bot for Improving Human Resource Management. *8th International Conference on ICT for Smart Society: Digital Twin for Smart Society, ICISS 2021 - Proceeding*. https://doi.org/10.1109/ICISS53185.2021.9533248

Vakayil, S., Sujitha Juliet, D., Anitha, J., & Vakayil, S. (2024). RAG-Based LLM Chatbot Using Llama-2. *ICDCS 2024 - 2024 7th International Conference on Devices, Circuits and Systems*, 195–199. https://doi.org/10.1109/ICDCS59278.2024.10561020

Weller, O., Marone, M., Weir, N., Lawrie, D., Khashabi, D., & Van Durme, B. (2024). *"According to ...": Prompting Language Models Improves Quoting from Pre-Training Data*. http://arxiv.org/abs/2305.13252

White House. (2022). *Blueprint for an AI Bill of Rights*. https://www.whitehouse.gov/ostp/ai-bill-of-rights

World Economic Forum. (2023). *Future of Jobs Report 2023*.

Xu, S., Pang, L., Shen, H., & Cheng, X. (2024). *A Theory for Token-Level Harmonization in Retrieval-Augmented Generation*. http://arxiv.org/abs/2406.00944

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2025). *A Survey of Large Language Models*. http://arxiv.org/abs/2303.18223