

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO

SAMAI SOARES PEREIRA

USO DE SISTEMAS MULTIAGENTES PARA APOIO À DECISÃO CLÍNICA

TRABALHO DE CONCLUSÃO DE CURSO

Belo Horizonte

2026

SAMAI SOARES PEREIRA

USO DE SISTEMAS MULTIAGENTES PARA APOIO À DECISÃO CLÍNICA

TRABALHO DE CONCLUSÃO DE CURSO

Trabalho de Conclusão de Curso apresentado ao curso de Graduação em Sistemas de Informação da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Orientador:
Prof. Frederico Gadelha Guimarães

Belo Horizonte

2026

Dedico este trabalho, primeiramente, à minha esposa e à minha filha, que são meus pilares. Dedico também aos meus pais, fonte dos meus valores, conhecimentos e saber.

Graças a eles, tornei-me alguém capaz de lutar pelos meus sonhos e objetivos, sem jamais desanimar. Considero-me forte porque eles me ensinaram a ser forte.

AGRADECIMENTOS

Agradeço, primeiramente, à minha esposa, que, apesar de todas as dificuldades enfrentadas ao longo desta jornada, permaneceu ao meu lado, oferecendo apoio, compreensão e incentivo constantes. À minha filha, que diariamente me inspira a ser uma pessoa melhor e mais responsável, dedico também este esforço. Expresso minha gratidão aos meus pais, que sempre fizeram o possível para me orientar, apoiar e motivar nos estudos desde a infância. Sem o amor, a paciência e a confiança de cada um deles, este trabalho não seria possível.

RESUMO

Este trabalho apresenta o desenvolvimento de uma plataforma inteligente de apoio à decisão clínica baseada em modelos de linguagem de grande porte (Large Language Models – LLMs). A solução tem como objetivo realizar uma pré-triagem por meio da condução automatizada da anamnese inicial, auxiliando profissionais de saúde na obtenção de informações relevantes e contribuindo para a otimização do fluxo de atendimento.

A arquitetura proposta integra uma abordagem multiagente com técnicas de *Retrieval-Augmented Generation* (RAG), permitindo que as respostas sejam fundamentadas em bases de conhecimento médicas confiáveis. Além disso, emprega o paradigma *Agent-as-a-Judge*, no qual um agente especializado realiza a avaliação da qualidade, consistência e fundamentação das respostas produzidas pelos demais agentes, buscando reduzir alucinações e aumentar a confiabilidade das informações fornecidas.

Espera-se que a plataforma contribua para o aumento da eficiência do processo de triagem, forneça suporte à tomada de decisão clínica e demonstre o potencial da combinação entre arquiteturas multiagentes e modelos de linguagem na área da saúde.

Palavras-chave: Inteligência artificial. Modelos de linguagem. Arquitetura multiagente. RAG. Agent-as-a-Judge. Apoio à decisão clínica. Saúde digital.

ABSTRACT

This work presents the development of an intelligent clinical decision support platform based on Large Language Models (LLMs). The proposed solution aims to perform an automated pre-triage by conducting the initial medical interview, assisting healthcare professionals in collecting relevant patient information and improving clinical workflow.

The proposed architecture combines a multi-agent system with Retrieval-Augmented Generation (RAG), enabling responses to be grounded in reliable medical knowledge sources. Furthermore, it incorporates the Agent-as-a-Judge paradigm, in which a dedicated evaluation agent assesses the quality, consistency, and factual grounding of the responses generated by the other agents, reducing hallucinations and improving reliability.

The proposed platform is expected to enhance the efficiency of the triage process, support clinical decision-making, and demonstrate the potential of combining multi-agent architectures with large language models in healthcare applications.

This work presents the development of an intelligent clinical decision support platform based on Large Language Models (LLMs). The proposed solution aims to perform an automated pre-triage by conducting the initial medical interview, assisting healthcare professionals in collecting relevant patient information and improving clinical workflow.

The proposed architecture combines a multi-agent system with Retrieval-Augmented Generation (RAG), enabling responses to be grounded in reliable medical knowledge sources. Furthermore, it incorporates the Agent-as-a-Judge paradigm, in which a dedicated evaluation agent assesses the quality, consistency, and factual grounding of the responses generated by the other agents, reducing hallucinations and improving reliability.

The proposed platform is expected to enhance the efficiency of the triage process, support clinical decision-making, and demonstrate the potential of combining multi-agent architectures with large language models in healthcare applications.

Key-words: Artificial intelligence. Large Language Models. Multi-agent systems. Retrieval-Augmented Generation. Agent-as-a-Judge. Clinical decision support. Digital health.

SUMÁRIO

1	INTRODUÇÃO	8
1.1	APRESENTAÇÃO	8
1.2	DESCRIÇÃO DO PROBLEMA	8
1.3	JUSTIFICATIVA	8
1.4	OBJETIVOS	9
1.4.1	Objetivo geral	9
1.4.2	Objetivos específicos	9
1.5	METODOLOGIA	9
2	FUNDAMENTAÇÃO TEÓRICA	11
2.1	Inteligência Artificial aplicada à Saúde	11
2.2	Sistemas de Apoio à Decisão Clínica	12
2.3	Modelos de Linguagem de Grande Escala	12
2.4	Retrieval-Augmented Generation	13
2.5	Arquiteturas Multiagentes	13
2.6	Agent-as-a-Judge	13
2.7	RAG de Medicamentos e Re-ranqueamento Demográfico	14
2.7.1	Construção da Base Vetorial de Bulas	14
2.7.2	Representação Vetorial Utilizando Embeddings Multilíngues	15
2.7.3	Organização da Base Vetorial no ChromaDB	16
3	Visão Geral da Arquitetura	18
3.1	Módulo de Triagem Conversacional e Extração de Anamnese	21
3.2	Motor <i>Medical RAG</i>	22
3.2.1	Construção da Base Vetorial	25
3.2.2	Representação Vetorial dos Documentos	26
3.2.3	Armazenamento no ChromaDB	26
3.2.4	Primeiro Estágio: Recuperação das Hipóteses Diagnósticas	27
3.2.5	Segundo Estágio: Recuperação de Diretrizes Terapêuticas	28
3.2.6	Construção da Resposta Clínica	29
3.2.7	Vantagens da Recuperação em Dois Estágios	30
3.3	RAG de Medicamentos e Re-ranqueamento Demográfico	30
3.3.1	Construção da Base Vetorial de Bulas	31

3.3.2	Representação Vetorial Utilizando Embeddings Multilíngues	32
3.3.3	Organização da Base Vetorial no ChromaDB	33
3.3.4	Recuperação Semântica das Informações Farmacológicas	34
3.3.5	Heurística de Re-ranqueamento Demográfico	34
3.3.6	Formulação Matemática do Re-ranqueamento	36
3.3.7	Exemplo de Aplicação da Heurística	36
3.3.8	Geração das Informações Farmacológicas	37
3.3.9	Estrutura da Resposta Produzida	38
3.3.10	Integração com o Agente Auditor	38
3.3.11	Limitações da Estratégia Proposta	39
3.4	Agente Auditor Clínico (<i>Agent-as-a-Judge</i>)	39
3.4.1	Arquitetura do Agente Auditor	40
3.4.2	Dados Fornecidos ao Agente Auditor	41
3.4.3	Processo de Auditoria Clínica	42
3.4.4	Geração dos Escores de Auditoria e Parecer Clínico	43
3.5	Fluxo Completo de Processamento e Geração dos Relatórios	45
4	RESULTADOS	49
4.1	Funcionamento geral da aplicação	49
4.2	Avaliação do Medical RAG	50
4.3	Avaliação do Drug RAG	50
4.4	Avaliação do Agent-as-a-Judge	50
4.5	Discussão dos resultados	51
5	CONCLUSÃO	52
	Referências	55

1 INTRODUÇÃO

1.1 APRESENTAÇÃO

Os avanços recentes dos Modelos de Linguagem de Grande Escala (*Large Language Models* – LLMs) impulsionaram uma transformação significativa na forma como sistemas inteligentes processam informações e interagem com usuários. Essas tecnologias têm sido incorporadas em diversos setores para automatizar tarefas complexas, apoiar a tomada de decisão e aumentar a eficiência operacional. Na área da saúde, entretanto, sua adoção ainda enfrenta desafios relacionados à confiabilidade das respostas, à ocorrência de alucinações, à rastreabilidade das informações geradas e à necessidade de fundamentação em evidências científicas.

Nesse contexto, surgem arquiteturas baseadas em sistemas multiagentes, capazes de distribuir responsabilidades entre agentes especializados, bem como técnicas como *Retrieval-Augmented Generation* (RAG), que permitem fundamentar as respostas em bases de conhecimento confiáveis. Complementarmente, abordagens como *Agent-as-a-Judge* possibilitam que um agente avaliador analise criticamente as respostas produzidas pelos demais agentes, aumentando sua consistência e confiabilidade.

1.2 DESCRIÇÃO DO PROBLEMA

A crescente demanda por atendimentos médicos, especialmente em serviços públicos de saúde, impõe elevada carga de trabalho aos profissionais, reduzindo o tempo disponível para coleta detalhada de informações clínicas e aumentando o risco de inconsistências durante a tomada de decisão. Embora os LLMs apresentem elevado potencial para auxiliar nesse processo, sua utilização direta em aplicações médicas ainda é limitada devido à possibilidade de geração de informações incorretas ou não fundamentadas.

Dessa forma, torna-se necessário investigar arquiteturas que aumentem a confiabilidade das respostas produzidas por modelos de linguagem, permitindo sua utilização como ferramenta de apoio ao profissional de saúde durante a realização da anamnese e da pré-triagem clínica.

1.3 JUSTIFICATIVA

A utilização de inteligência artificial como ferramenta de apoio à decisão clínica representa uma oportunidade para aumentar a eficiência dos atendimentos, reduzir o tempo gasto na coleta de informações e fornecer suporte baseado em evidências ao profissional de saúde.

Neste cenário, a combinação de arquiteturas multiagentes com técnicas de RAG e mecanismos de avaliação baseados em *Agent-as-a-Judge* pode contribuir para minimizar alucinações, aumentar a confiabilidade das respostas geradas e tornar os modelos de linguagem mais adequados para aplicações na área médica.

Além da relevância prática, o trabalho possui contribuição científica ao investigar uma arquitetura que integra múltiplas técnicas recentes de inteligência artificial generativa, avaliando seu potencial como sistema de apoio à decisão clínica.

1.4 OBJETIVOS

1.4.1 Objetivo geral

Desenvolver e avaliar uma arquitetura multiagente baseada em Modelos de Linguagem de Grande Escala (LLMs), integrada às técnicas de Retrieval-Augmented Generation (RAG) e Agent-as-a-Judge, para apoiar a realização da anamnese e da pré-triagem clínica, buscando aumentar a confiabilidade das respostas produzidas pelo sistema.

1.4.2 Objetivos específicos

- Estudar a aplicação de LLMs em sistemas de apoio à decisão clínica.
- Projetar uma arquitetura baseada em múltiplos agentes especializados.
- Integrar técnicas de Retrieval-Augmented Generation (RAG) para fundamentar as respostas em literatura médica.
- Implementar um agente avaliador utilizando o paradigma Agent-as-a-Judge para validar a qualidade das respostas produzidas.
- Desenvolver um protótipo funcional capaz de realizar a condução da anamnese inicial.
- Avaliar a arquitetura proposta quanto à qualidade, consistência e confiabilidade das respostas geradas.

1.5 METODOLOGIA

A pesquisa caracteriza-se como aplicada, de natureza experimental, envolvendo o desenvolvimento de um protótipo de sistema inteligente para apoio à decisão clínica.

Inicialmente será realizada uma revisão bibliográfica sobre Modelos de Linguagem de Grande Escala, sistemas multiagentes, Retrieval-Augmented Generation (RAG) e Agent-as-a-Judge

aplicados à saúde. Em seguida será projetada e implementada uma arquitetura composta por agentes especializados responsáveis pela condução da anamnese, recuperação de conhecimento em bases médicas e avaliação das respostas produzidas.

Posteriormente, o sistema será submetido a experimentos utilizando casos clínicos, permitindo analisar aspectos como consistência das respostas, fundamentação em evidências, redução de alucinações e qualidade geral do processo de apoio à decisão clínica.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os principais conceitos que fundamentam o desenvolvimento da arquitetura proposta neste trabalho. Inicialmente são discutidas as aplicações da Inteligência Artificial na área da saúde e os Sistemas de Apoio à Decisão Clínica. Em seguida, são abordados os Modelos de Linguagem de Grande Escala (LLMs), as arquiteturas baseadas em sistemas multiagentes, a técnica de *Retrieval-Augmented Generation* (RAG) e o paradigma *Agent-as-a-Judge*, destacando suas contribuições para o aumento da confiabilidade de sistemas inteligentes aplicados ao contexto clínico.

2.1 INTELIGÊNCIA ARTIFICIAL APLICADA À SAÚDE

A Inteligência Artificial (IA) consolidou-se como uma das principais áreas de pesquisa e inovação tecnológica da atualidade, promovendo avanços significativos em setores como indústria, finanças, educação e saúde. O desenvolvimento dos Modelos de Linguagem de Grande Escala (*Large Language Models* – LLMs) ampliou significativamente a capacidade dos sistemas computacionais de compreender linguagem natural, sintetizar informações e apoiar processos complexos de tomada de decisão.

Na área da saúde, essas tecnologias apresentam potencial para apoiar atividades como triagem clínica, interpretação de protocolos, documentação médica e sistemas de apoio à decisão clínica. Contudo, sua utilização em cenários de alta criticidade ainda enfrenta desafios importantes relacionados à ocorrência de alucinações, inconsistências factuais, falta de rastreabilidade das respostas e necessidade de fundamentação em evidências científicas. Essas limitações restringem a adoção de LLMs como ferramentas autônomas em ambientes clínicos, motivando o desenvolvimento de arquiteturas que integrem mecanismos de recuperação de conhecimento, validação das respostas e supervisão humana. Estudos recentes apontam que técnicas como *Retrieval-Augmented Generation* (RAG) e fluxos compostos por agentes especializados representam abordagens promissoras para aumentar a confiabilidade desses sistemas em aplicações médicas.

Entretanto, diversos estudos demonstram que, quando empregados de forma supervisionada, sistemas inteligentes podem contribuir para reduzir erros humanos, aumentar a eficiência operacional e oferecer suporte ao profissional durante atividades críticas do atendimento médico (1).

No contexto brasileiro, tais tecnologias apresentam potencial ainda maior devido aos desafios enfrentados pelo Sistema Único de Saúde (SUS). Segundo levantamento do Tribunal de Contas da União (TCU), entre 2019 e 2024 a eficiência média das unidades hospitalares

variou entre 32% e 50% (2). Esse cenário evidencia a necessidade de soluções tecnológicas capazes de otimizar o fluxo assistencial, reduzir o tempo de espera e fornecer apoio à tomada de decisão clínica.

2.2 SISTEMAS DE APOIO À DECISÃO CLÍNICA

Os Sistemas de Apoio à Decisão Clínica (SADC) constituem ferramentas computacionais desenvolvidas para auxiliar profissionais de saúde durante o processo de diagnóstico, tratamento e acompanhamento de pacientes. Segundo Shortliffe e Cimino (3), esses sistemas utilizam bases de conhecimento, protocolos clínicos e informações do paciente para fornecer recomendações que auxiliam a tomada de decisão sem substituir o julgamento médico.

Pesquisas recentes indicam que os SADC contribuem para a redução de erros médicos, aumento da segurança do paciente e melhoria da qualidade assistencial (4). Com o avanço da Inteligência Artificial Generativa, esses sistemas passaram a incorporar modelos de linguagem capazes de interpretar linguagem natural e sintetizar grandes volumes de informação clínica.

Entretanto, a utilização direta de LLMs em aplicações médicas ainda apresenta limitações relacionadas à ocorrência de alucinações, inconsistências e ausência de fundamentação em evidências científicas, tornando necessária a utilização de arquiteturas que aumentem sua confiabilidade.

2.3 MODELOS DE LINGUAGEM DE GRANDE ESCALA

Os Modelos de Linguagem de Grande Escala (*Large Language Models* – LLMs) representam uma evolução significativa na área de Processamento de Linguagem Natural. Treinados sobre grandes volumes de dados textuais, esses modelos são capazes de compreender contexto, responder perguntas, resumir documentos e produzir textos coerentes em linguagem natural.

Na área da saúde, os LLMs apresentam potencial para apoiar a realização da anamnese, auxiliar na interpretação de protocolos clínicos e fornecer suporte durante a tomada de decisão. Contudo, sua utilização exige mecanismos adicionais capazes de reduzir alucinações e garantir que as respostas estejam fundamentadas em literatura médica confiável.

Essas limitações motivaram o desenvolvimento de técnicas como o Retrieval-Augmented Generation (RAG), bem como arquiteturas compostas por múltiplos agentes especializados responsáveis por diferentes etapas do processo de raciocínio.

2.4 RETRIEVAL-AUGMENTED GENERATION

O Retrieval-Augmented Generation (RAG) combina modelos de linguagem com mecanismos de recuperação de informações em bases de conhecimento externas. Em vez de responder apenas com base no conhecimento adquirido durante o treinamento do modelo, o sistema recupera documentos relevantes e os utiliza como contexto para a geração da resposta.

Essa abordagem reduz significativamente a ocorrência de alucinações e aumenta a rastreabilidade das informações utilizadas pelo modelo, características fundamentais para aplicações médicas. A construção de sistemas RAG envolve etapas como processamento documental, geração de embeddings, armazenamento em bancos vetoriais, recuperação semântica e reclassificação das evidências recuperadas.

2.5 ARQUITETURAS MULTIAGENTES

Sistemas multiagentes distribuem diferentes responsabilidades entre agentes especializados, permitindo que tarefas complexas sejam executadas de maneira colaborativa. Em aplicações clínicas, diferentes agentes podem atuar na condução da anamnese, estruturação das informações clínicas, recuperação de literatura médica, consulta a bulas de medicamentos e avaliação da qualidade das respostas produzidas.

Essa divisão de responsabilidades reduz a complexidade de cada componente individual e permite que diferentes estratégias sejam empregadas em cada etapa do processo de apoio à decisão clínica.

2.6 AGENT-AS-A-JUDGE

O paradigma Agent-as-a-Judge consiste na utilização de um agente especializado responsável por avaliar criticamente as respostas produzidas pelos demais agentes da arquitetura. Em vez de aceitar automaticamente a saída do modelo gerador, o agente avaliador verifica aspectos como fidelidade ao contexto recuperado, coerência clínica, identificação de contraindicações, riscos terapêuticos e possíveis alucinações.

Essa estratégia representa uma camada adicional de segurança, particularmente importante em aplicações médicas, nas quais respostas incorretas podem impactar diretamente a qualidade da assistência prestada ao paciente.

2.7 RAG DE MEDICAMENTOS E RE-RANQUEAMENTO DEMOGRÁFICO

Embora o *Medical RAG* seja responsável por identificar hipóteses diagnósticas e recuperar diretrizes terapêuticas a partir da literatura médica, as informações obtidas nessa etapa ainda são insuficientes para apoiar a prescrição medicamentosa de maneira segura. Em grande parte dos casos, livros-texto e diretrizes clínicas concentram-se na indicação dos princípios ativos recomendados para determinada condição clínica, mas não apresentam detalhadamente aspectos relacionados à posologia, contraindicações, advertências, interações medicamentosas e precauções específicas de cada fármaco.

Com o objetivo de complementar essas informações, foi desenvolvido um segundo mecanismo de recuperação documental denominado *Drug RAG*. Esse módulo opera de forma independente do *Medical RAG* e possui como finalidade recuperar informações farmacológicas diretamente de bulas oficiais de medicamentos previamente indexadas em uma base vetorial própria.

A separação entre os dois mecanismos de recuperação constitui uma decisão arquitetural importante da solução proposta. Enquanto o *Medical RAG* trabalha sobre literatura médica destinada ao raciocínio clínico e ao estabelecimento de hipóteses diagnósticas, o *Drug RAG* atua sobre documentos regulatórios que descrevem as características farmacológicas dos medicamentos comercializados. Dessa forma, cada módulo utiliza uma base documental especializada para resolver um problema específico, reduzindo o ruído semântico durante a recuperação das informações.

Outro aspecto relevante dessa arquitetura é que o *Drug RAG* não realiza inferências clínicas sobre a condição do paciente. Seu papel consiste exclusivamente em recuperar informações oficiais relacionadas aos medicamentos sugeridos pelo módulo anterior, permitindo que recomendações terapêuticas sejam posteriormente avaliadas pelo agente auditor utilizando evidências provenientes das próprias bulas.

Ao utilizar duas bases documentais independentes, o sistema reduz a dependência de uma única fonte de conhecimento e aumenta a diversidade das evidências consideradas durante o processo de apoio à decisão clínica.

2.7.1 Construção da Base Vetorial de Bulas

A construção da base documental do *Drug RAG* iniciou-se com a obtenção de bulas oficiais de medicamentos disponibilizadas em formato digital. Esses documentos constituem a principal fonte de informações regulamentadas sobre princípios ativos, indicações terapêuticas,

contraindicações, posologia, advertências, precauções, reações adversas e demais informações farmacológicas exigidas pelos órgãos reguladores.

Assim como ocorreu na construção da base do *Medical RAG*, os documentos foram inicialmente convertidos para texto utilizando ferramentas de extração compatíveis com arquivos PDF. Após essa etapa, foi realizado um processo de limpeza textual destinado à remoção de elementos que não agregavam informação semântica relevante para o processo de recuperação, como cabeçalhos repetitivos, numeração de páginas, caracteres especiais e espaços em branco excedentes.

Entretanto, diferentemente da literatura médica utilizada anteriormente, as bulas apresentam uma organização bastante estruturada, sendo divididas em seções padronizadas. Aproveitando essa característica, optou-se por segmentar cada bula de acordo com suas seções principais, preservando a organização lógica do documento.

Entre as seções consideradas mais relevantes para este trabalho destacam-se:

- Posologia;
- Contraindicações;
- Advertências e Precauções;
- Reações Adversas;
- Interações Medicamentosas.

Cada uma dessas seções foi tratada como um documento independente durante o processo de indexação. Essa estratégia reduz a quantidade de informações irrelevantes recuperadas durante as consultas, permitindo que o mecanismo de busca retorne apenas o trecho da bula relacionado ao tipo de informação solicitado.

Além disso, a segmentação por seções facilita a interpretação posterior realizada pelo modelo de linguagem, uma vez que o contexto enviado ao LLM torna-se significativamente menor e mais específico.

Após a segmentação, cada trecho textual passa a representar uma unidade independente dentro da base vetorial, podendo ser recuperado individualmente durante o processo de busca.

2.7.2 Representação Vetorial Utilizando Embeddings Multilíngues

Para transformar os fragmentos textuais das bulas em representações numéricas foi utilizado o modelo de embeddings *BAAI/bge-m3*.

A escolha desse modelo ocorreu principalmente devido à natureza multilíngue da base documental utilizada. Embora parte da literatura médica empregada pelo *Medical RAG* esteja predominantemente em língua inglesa, as bulas utilizadas pelo *Drug RAG* encontram-se em português e apresentam grande quantidade de termos farmacológicos específicos, abreviações, unidades de medida e nomenclaturas técnicas.

O modelo *BAAI/bge-m3* apresenta elevado desempenho em tarefas de recuperação semântica multilíngue, sendo capaz de representar documentos escritos em diferentes idiomas dentro de um mesmo espaço vetorial. Essa característica reduz problemas relacionados à variação linguística e melhora a recuperação de documentos mesmo quando diferentes termos são utilizados para representar um mesmo conceito farmacológico.

Outro fator considerado durante a seleção do modelo foi sua capacidade de capturar relações semânticas entre expressões equivalentes. Dessa forma, consultas contendo termos como “dose”, “posologia”, “administração” ou “esquema terapêutico” tendem a recuperar documentos semanticamente relacionados, mesmo quando essas palavras não aparecem exatamente da mesma forma na bula indexada.

Após a geração dos embeddings, cada fragmento textual foi associado ao seu vetor correspondente, tornando possível a realização de consultas baseadas em similaridade semântica em vez de simples correspondência lexical.

2.7.3 Organização da Base Vetorial no ChromaDB

Todos os embeddings produzidos foram armazenados em uma coleção específica do banco vetorial ChromaDB denominada *drug_rag_v2*.

Cada registro armazenado nessa coleção contém o texto original do fragmento da bula, seu vetor de embedding e um conjunto de metadados utilizados durante o processo de recuperação. Entre esses metadados encontram-se o nome do medicamento, a seção da bula à qual o fragmento pertence e identificadores internos utilizados pelo sistema.

A utilização de uma coleção independente daquela empregada pelo *Medical RAG* evita que documentos de naturezas distintas sejam recuperados simultaneamente durante uma mesma consulta. Enquanto a coleção *medical_rag_v2* concentra informações provenientes da literatura médica, a coleção *drug_rag_v2* armazena exclusivamente documentos farmacológicos.

Essa separação simplifica o processo de recuperação, reduz o espaço de busca e aumenta a precisão dos resultados retornados pelo banco vetorial.

Ao final dessa etapa, a base vetorial encontra-se preparada para receber consultas contendo

os princípios ativos identificados pelo *Medical RAG*, possibilitando a recuperação das informações farmacológicas que serão utilizadas nas etapas subsequentes de re-ranqueamento e auditoria clínica.

3 VISÃO GERAL DA ARQUITETURA

O sistema desenvolvido neste trabalho foi concebido seguindo uma arquitetura modular baseada no paradigma de sistemas multiagentes, na qual cada componente é responsável por uma etapa específica do processo de triagem clínica e apoio à decisão médica. Essa divisão de responsabilidades reduz o acoplamento entre os módulos, facilita a manutenção do software e permite que cada etapa do processamento seja evoluída de forma independente, sem comprometer o funcionamento global da solução.

Diferentemente de sistemas tradicionais de perguntas e respostas baseados apenas em Modelos de Linguagem de Grande Escala (*Large Language Models* – LLMs), a arquitetura proposta busca reduzir a ocorrência de alucinações por meio da integração entre técnicas de *Retrieval-Augmented Generation* (RAG), bases documentais especializadas e um mecanismo adicional de auditoria automatizada baseado no paradigma *Agent-as-a-Judge*. Dessa forma, o modelo de linguagem deixa de responder exclusivamente com base em seu conhecimento paramétrico e passa a fundamentar suas respostas em evidências recuperadas dinamicamente de bases documentais previamente indexadas.

A arquitetura implementada organiza o fluxo de processamento em quatro grandes módulos, apresentados na Figura 1. Cada módulo possui responsabilidades bem definidas e comunica-se com os demais por meio de estruturas de dados padronizadas em formato JSON, permitindo que as informações produzidas em uma etapa sejam reutilizadas nas etapas subsequentes sem perda de contexto.

De forma resumida, o fluxo de execução inicia-se com a coleta das informações clínicas do paciente por meio de uma interface conversacional. Durante essa etapa, um agente especializado conduz a entrevista clínica de maneira dinâmica, adaptando as perguntas às respostas fornecidas pelo usuário e estruturando automaticamente as informações em um modelo padronizado de anamnese.

Após a conclusão da entrevista, a anamnese estruturada é encaminhada ao módulo denominado *Medical RAG*. Esse módulo realiza duas etapas distintas de recuperação semântica sobre uma base vetorial composta por livros-texto, protocolos clínicos e obras de referência em Medicina Interna. Na primeira etapa, o sistema identifica as hipóteses diagnósticas mais compatíveis com os sinais e sintomas apresentados pelo paciente. Na segunda etapa, utilizando a hipótese principal obtida anteriormente, uma nova consulta é realizada para recuperar diretrizes terapêuticas e opções farmacológicas descritas na literatura científica.

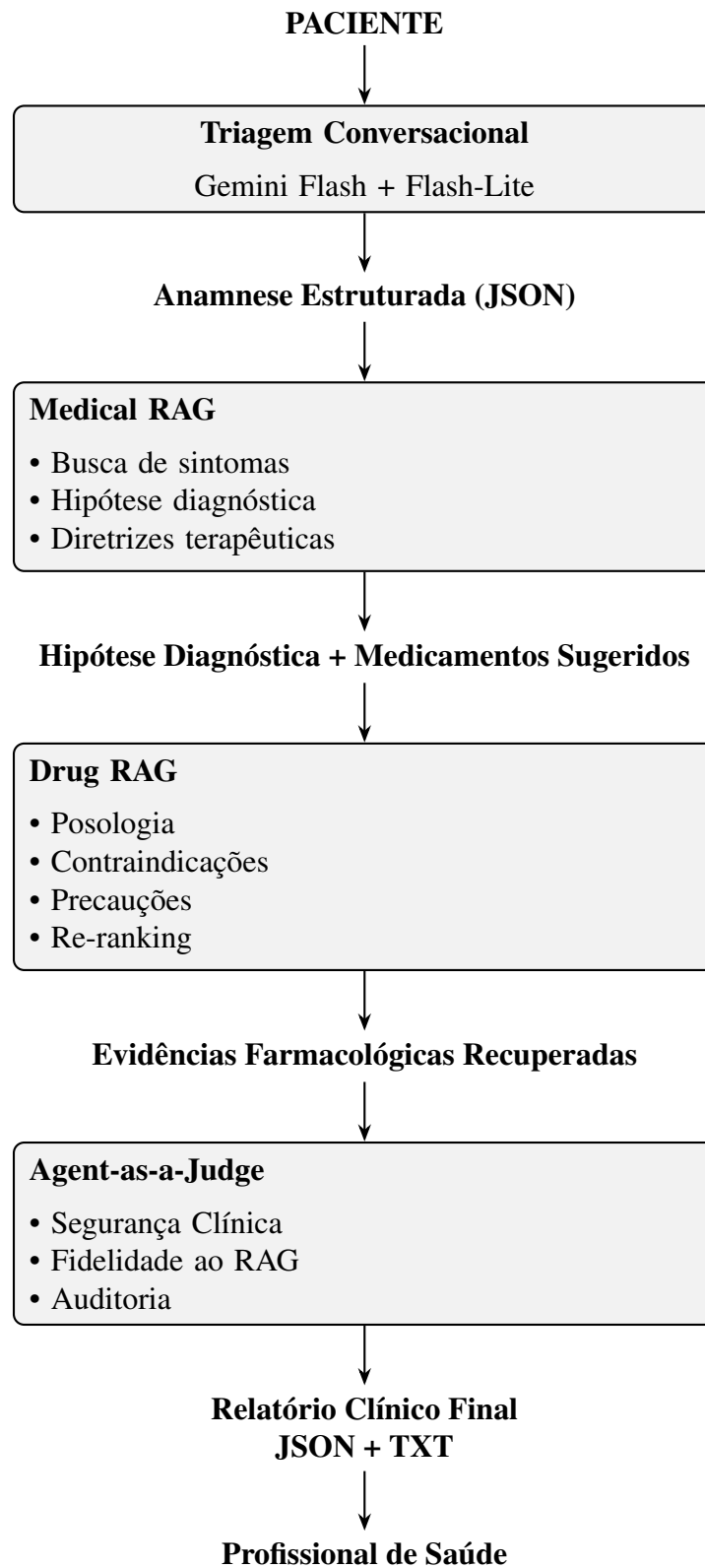


Figura 1: Arquitetura geral do sistema desenvolvido.

As opções terapêuticas identificadas pelo *Medical RAG* não são apresentadas diretamente ao profissional de saúde. Antes disso, elas são encaminhadas ao módulo denominado *Drug RAG*, cuja finalidade é recuperar informações detalhadas sobre cada princípio ativo recomendado. Diferentemente da primeira base documental, composta por literatura médica, essa segunda base é formada por bulas oficiais de medicamentos previamente processadas e indexadas em um banco vetorial independente.

O *Drug RAG* recupera informações referentes à posologia, contraindicações, advertências, precauções e demais informações farmacológicas relevantes para cada medicamento sugerido. Além da recuperação semântica tradicional, esse módulo incorpora um mecanismo de re-ranqueamento que considera características demográficas do paciente, como idade e peso corporal, priorizando automaticamente trechos de bula compatíveis com o perfil clínico informado durante a triagem.

Embora as duas etapas de recuperação forneçam informações fundamentadas em documentação especializada, ainda existe a possibilidade de interpretações incorretas pelo modelo de linguagem ou da utilização inadequada do contexto recuperado. Para minimizar esse risco, a arquitetura incorpora uma quarta camada denominada *Agent-as-a-Judge*.

Nesse estágio, um segundo agente baseado em LLM recebe simultaneamente a anamnese estruturada, as hipóteses diagnósticas produzidas pelo *Medical RAG*, os trechos recuperados da literatura médica, as evidências provenientes das bulas e a resposta clínica construída pelo sistema. O agente auditor realiza uma análise crítica dessas informações verificando se todas as recomendações estão efetivamente sustentadas pelos documentos recuperados, se existem contraindicações relevantes para o perfil do paciente e se alguma recomendação foi produzida sem respaldo nas evidências recuperadas.

Como resultado dessa auditoria são produzidos indicadores quantitativos de qualidade, incluindo um escore de segurança clínica e um escore de fidelidade ao contexto recuperado pelo RAG, além de uma análise textual justificando a decisão tomada pelo agente auditor. Essa etapa funciona como um mecanismo adicional de validação antes que qualquer informação seja disponibilizada ao profissional de saúde.

Concluídas todas as etapas do processamento, o sistema consolida os resultados em dois formatos complementares. O primeiro consiste em um arquivo JSON contendo todas as informações estruturadas produzidas durante a execução, incluindo a anamnese, hipóteses diagnósticas, medicamentos recuperados, evidências documentais e parecer do agente auditor. O segundo corresponde a um relatório textual organizado em linguagem natural, destinado à leitura pelo profissional médico, reunindo de forma objetiva o resumo da anamnese, as hipóteses

clínicas levantadas, os medicamentos encontrados na literatura, as principais evidências recuperadas das bulas e a avaliação produzida pelo módulo de auditoria.

Essa estratégia permite que o sistema funcione como um mecanismo de apoio à decisão clínica, fornecendo ao profissional de saúde informações organizadas, rastreáveis e fundamentadas em fontes documentais, sem substituir o julgamento médico. Todas as decisões terapêuticas permanecem sob responsabilidade do profissional responsável pelo atendimento, enquanto o sistema atua como uma ferramenta de suporte capaz de acelerar o acesso às evidências clínicas e farmacológicas relevantes para cada caso.

Sob o ponto de vista computacional, a separação da arquitetura em módulos independentes também facilita futuras expansões do sistema. Novas bases documentais podem ser incorporadas ao processo de recuperação sem necessidade de alterações significativas nos demais componentes, assim como novos agentes especializados podem ser adicionados ao fluxo para executar tarefas específicas, como interpretação de exames laboratoriais, análise de imagens médicas ou integração com prontuários eletrônicos. Essa característica torna a arquitetura suficientemente flexível para acomodar futuras evoluções da solução proposta.

3.1 MÓDULO DE TRIAGEM CONVERSACIONAL E EXTRAÇÃO DE ANAMNESE

A primeira etapa do sistema consiste na coleta das informações clínicas do paciente por meio de uma interface conversacional em linguagem natural. Diferentemente dos sistemas tradicionais baseados em formulários estáticos, a arquitetura proposta utiliza um agente conversacional capaz de conduzir dinamicamente a entrevista clínica, adaptando as perguntas às respostas fornecidas pelo paciente.

Esse módulo é responsável por gerenciar toda a interação inicial, interpretar as respostas do usuário e organizar progressivamente as informações coletadas em uma estrutura padronizada de anamnese. Durante toda a conversa, o sistema mantém o contexto da sessão, permitindo que novas perguntas sejam formuladas de acordo com as informações já obtidas, evitando repetições desnecessárias e proporcionando uma interação mais natural.

As respostas fornecidas pelo paciente são processadas por um modelo de linguagem, cuja função é identificar automaticamente as informações clínicas relevantes presentes em cada mensagem. Em vez de trabalhar apenas com texto livre, o sistema converte essas informações em uma estrutura organizada, permitindo que os módulos subsequentes utilizem dados padronizados durante o restante do processamento.

Após a extração das informações, é realizada uma etapa de validação para verificar se os dados obtidos são consistentes e suficientes para compor a anamnese. Caso alguma informação seja considerada incompleta ou ambígua, o sistema solicita esclarecimentos adicionais ao paciente antes de prosseguir, reduzindo a propagação de inconsistências para as etapas seguintes.

A condução do diálogo é realizada por um segundo modelo de linguagem otimizado para interações rápidas, responsável por formular perguntas de acompanhamento de forma contextualizada. Dessa maneira, a sequência das perguntas adapta-se continuamente ao estado da conversa e às características clínicas identificadas durante a entrevista, em vez de seguir uma ordem fixa previamente estabelecida.

Ao término da triagem, todas as informações coletadas são consolidadas em uma anamnese estruturada, organizada segundo um modelo previamente definido. Esse documento reúne dados demográficos, queixa principal, história da doença atual, sintomas associados, antecedentes clínicos e demais informações relevantes para a continuidade da avaliação.

A padronização da anamnese representa um dos principais elementos da arquitetura proposta, pois estabelece uma interface comum entre os diferentes módulos do sistema. Assim, o *Medical RAG*, o *Drug RAG* e o agente auditor passam a operar sobre uma mesma representação estruturada da condição clínica do paciente, garantindo consistência no fluxo de processamento e reduzindo a necessidade de adaptações entre os componentes da arquitetura.

3.2 MOTOR *MEDICAL RAG*

Após a consolidação da anamnese estruturada, inicia-se a etapa central do sistema proposta neste trabalho: o módulo denominado *Medical Retrieval-Augmented Generation (Medical RAG)*. Esse componente é responsável por recuperar evidências clínicas em literatura médica especializada que servirão de fundamento para a identificação de hipóteses diagnósticas e para a obtenção de recomendações terapêuticas. Diferentemente de abordagens convencionais baseadas exclusivamente em Modelos de Linguagem de Grande Escala (*Large Language Models – LLMs*), nas quais o conhecimento utilizado para responder às consultas está limitado aos parâmetros aprendidos durante o treinamento do modelo, a estratégia adotada neste trabalho fundamenta todas as respostas em documentos recuperados dinamicamente de uma base vetorial construída especificamente para esse propósito.

A utilização do paradigma *Retrieval-Augmented Generation* reduz significativamente a probabilidade de alucinações do modelo de linguagem, uma vez que a geração textual deixa de

depende exclusivamente do conhecimento paramétrico do LLM e passa a considerar informações provenientes de fontes documentais previamente selecionadas. Dessa forma, o modelo atua principalmente como um mecanismo de interpretação e síntese das evidências recuperadas, mantendo maior aderência à literatura médica utilizada como base de conhecimento.

Durante o desenvolvimento da solução verificou-se que realizar simultaneamente a identificação do diagnóstico e a recuperação do tratamento em uma única consulta resultava em perda de precisão. Isso ocorre porque consultas excessivamente abrangentes tendem a recuperar documentos relacionados tanto aos sintomas quanto às opções terapêuticas, reduzindo a relevância do contexto encaminhado ao modelo de linguagem. Para contornar essa limitação, optou-se pela implementação de um fluxo dividido em dois estágios independentes de recuperação semântica.

No primeiro estágio, o objetivo consiste exclusivamente na identificação da hipótese diagnóstica mais compatível com os sinais, sintomas e antecedentes presentes na anamnese estruturada. Em seguida, utilizando apenas o diagnóstico principal identificado, uma segunda recuperação é realizada sobre a mesma base documental, desta vez direcionada exclusivamente à obtenção de diretrizes terapêuticas e recomendações farmacológicas relacionadas à patologia encontrada. Essa estratégia reduz a ambiguidade da consulta e aumenta a qualidade dos documentos recuperados em cada etapa.

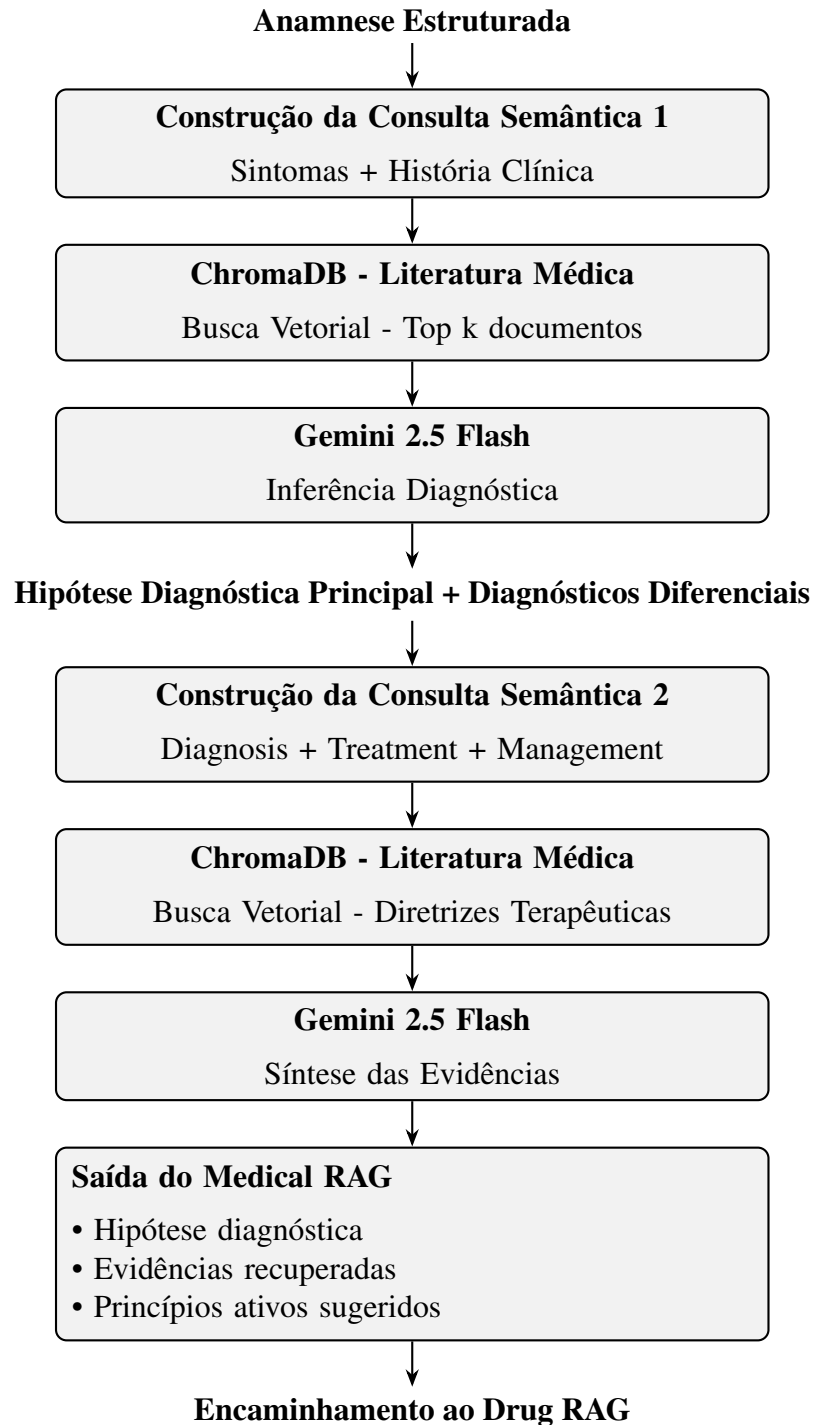


Figura 2: Fluxo interno do componente Medical RAG.

O funcionamento do módulo inicia-se imediatamente após a conclusão da etapa de triagem conversacional. A anamnese estruturada produzida pelo módulo anterior é utilizada como entrada para a construção da primeira consulta semântica. A partir dessa consulta, são recuperados os documentos mais relevantes da base vetorial, que posteriormente são utilizados como contexto pelo modelo de linguagem para inferência diagnóstica.

Uma vez determinada a hipótese clínica principal, o sistema constrói automaticamente

uma segunda consulta especializada contendo o nome da doença identificado, termos equivalentes em inglês e palavras-chave relacionadas ao manejo clínico da enfermidade. Essa nova consulta é utilizada para recuperar documentos contendo protocolos terapêuticos, esquemas farmacológicos e recomendações presentes na literatura médica indexada.

Como resultado, o módulo produz uma estrutura intermediária contendo a hipótese diagnóstica principal, diagnósticos diferenciais, evidências documentais recuperadas e uma lista preliminar de princípios ativos associados ao tratamento da condição clínica identificada. Essas informações são posteriormente encaminhadas ao módulo *Drug RAG*, responsável pela recuperação detalhada de informações farmacológicas presentes nas bulas oficiais dos medicamentos.

3.2.1 Construção da Base Vetorial

A qualidade de um sistema baseado em *Retrieval-Augmented Generation* depende diretamente da organização e da representação dos documentos utilizados durante o processo de recuperação. Dessa forma, antes da execução do fluxo de inferência clínica, foi necessária a construção de uma base vetorial contendo literatura médica especializada que pudesse servir como fonte de conhecimento para o modelo de linguagem.

A base documental utilizada neste trabalho foi composta por livros-texto de Medicina Interna, diretrizes clínicas e materiais de referência utilizados rotineiramente na prática médica. A escolha desse tipo de documento teve como objetivo fornecer informações abrangentes sobre fisiopatologia, manifestações clínicas, critérios diagnósticos, exames complementares e opções terapêuticas das principais doenças abordadas pelo sistema.

Como esses documentos encontram-se originalmente em formato PDF, foi necessário realizar um processo de extração textual antes da indexação. Essa etapa foi implementada utilizando a biblioteca *PyMuPDF*, responsável pela leitura das páginas e extração do conteúdo textual preservando, sempre que possível, a ordem lógica das informações presentes no documento.

Após a extração, os textos passaram por uma etapa de pré-processamento destinada à remoção de caracteres especiais, múltiplos espaços consecutivos, quebras de linha desnecessárias, cabeçalhos, rodapés e outros elementos que poderiam prejudicar o processo de indexação semântica.

Concluída essa etapa, os documentos foram segmentados em pequenos fragmentos (*chunks*) utilizando o algoritmo *RecursiveCharacterTextSplitter*, disponibilizado pela biblioteca *LangChain*.

O algoritmo foi configurado utilizando fragmentos de aproximadamente 1600 caracteres com sobreposição de 300 caracteres entre fragmentos consecutivos. A utilização de sobreposição reduz a perda de contexto causada pela divisão do texto, permitindo que informações localizadas

próximas às fronteiras entre dois fragmentos permaneçam acessíveis durante a recuperação.

Essa estratégia produz uma coleção significativamente maior de documentos menores, aumentando a granularidade da recuperação semântica e reduzindo a probabilidade de que informações relevantes permaneçam ocultas em grandes blocos textuais.

Após o particionamento, cada fragmento passa a representar uma unidade independente dentro da base vetorial, podendo ser recuperado individualmente durante as consultas realizadas pelo sistema.

3.2.2 Representação Vetorial dos Documentos

Após a etapa de segmentação, cada fragmento textual é convertido em uma representação vetorial utilizando um modelo de *embeddings*. O objetivo dessa etapa é representar semanticamente o conteúdo dos documentos em um espaço vetorial de alta dimensionalidade, permitindo que textos semanticamente semelhantes sejam posicionados próximos entre si e possam ser recuperados posteriormente por meio de buscas por similaridade.

Neste trabalho foi utilizado o modelo *BAAI/bge-large-en-v1.5*, executado localmente por meio do servidor de inferência Ollama. A escolha desse modelo deve-se ao seu elevado desempenho em tarefas de recuperação semântica (*Dense Retrieval*), sendo amplamente utilizado em arquiteturas baseadas em *Retrieval-Augmented Generation*. Além disso, como grande parte da literatura médica empregada neste trabalho encontra-se originalmente em língua inglesa, o modelo apresenta elevada capacidade de representação semântica para esse domínio.

Durante a geração dos embeddings foram seguidas as recomendações de utilização fornecidas pelos desenvolvedores do modelo, empregando a configuração indicada para tarefas de recuperação documental. Essa estratégia contribui para produzir representações vetoriais mais adequadas à busca por similaridade semântica, aumentando a qualidade dos documentos recuperados durante as consultas.

Após a vetorização, cada embedding é armazenado juntamente com seu respectivo fragmento textual e seus metadados no banco vetorial ChromaDB, tornando possível a realização de consultas por similaridade durante a execução do sistema.

3.2.3 Armazenamento no ChromaDB

Para armazenar os vetores gerados foi utilizado o banco de dados vetorial ChromaDB.

Diferentemente de bancos relacionais convencionais, o ChromaDB foi desenvolvido especificamente para operações envolvendo similaridade vetorial, permitindo consultas eficientes

sobre milhares de documentos representados por embeddings de alta dimensão.

Neste trabalho foi criada uma coleção denominada *medical_rag_v2*, responsável por armazenar todos os fragmentos da literatura médica utilizados durante o processo de recuperação.

Cada registro da coleção contém:

- o texto original do fragmento;
- o vetor de embedding correspondente;
- metadados associados ao documento de origem;
- identificadores utilizados durante a recuperação.

Durante a execução do sistema, nenhuma consulta é realizada diretamente sobre o texto dos documentos. Toda recuperação ocorre por meio da comparação entre embeddings, permitindo localizar trechos semanticamente semelhantes mesmo quando diferentes palavras são utilizadas para descrever um mesmo conceito clínico.

Essa característica representa uma das principais vantagens do paradigma RAG em relação às abordagens tradicionais baseadas apenas em busca por palavras-chave.

3.2.4 Primeiro Estágio: Recuperação das Hipóteses Diagnósticas

Após a conclusão da anamnese estruturada, inicia-se a primeira etapa do *Medical RAG*.

O objetivo desse estágio consiste em identificar quais doenças apresentam maior compatibilidade com os sinais e sintomas descritos pelo paciente.

Para isso, a anamnese estruturada é convertida em uma consulta textual contendo as principais informações clínicas coletadas durante a triagem, incluindo idade, sexo, sintomas principais, tempo de evolução, fatores agravantes, sintomas associados, antecedentes relevantes e demais informações disponíveis.

Essa descrição é então convertida em embedding utilizando o mesmo modelo empregado durante a indexação da base documental.

Em seguida, realiza-se uma busca por similaridade vetorial na coleção *medical_rag_v2*. Neste trabalho foram recuperados os trinta fragmentos mais próximos da consulta realizada.

A utilização de um número relativamente elevado de documentos recuperados reduz a probabilidade de perda de informações relevantes, permitindo que o modelo de linguagem tenha acesso simultaneamente a diferentes descrições clínicas provenientes da literatura médica.

Os fragmentos recuperados não representam necessariamente um diagnóstico único. Em muitos casos são encontrados documentos relacionados a diferentes doenças que compartilham manifestações clínicas semelhantes, refletindo a própria natureza do processo diagnóstico realizado por profissionais de saúde.

Todo o contexto recuperado é então organizado e encaminhado ao modelo de linguagem juntamente com instruções específicas que limitam sua atuação exclusivamente às informações presentes nos documentos recuperados.

Como saída, o modelo produz um objeto estruturado em formato JSON contendo a hipótese diagnóstica principal, diagnósticos diferenciais, justificativas clínicas e os respectivos termos médicos em português e inglês, que serão utilizados na segunda etapa do processo de recuperação.

3.2.5 Segundo Estágio: Recuperação de Diretrizes Terapêuticas

Embora a identificação da hipótese diagnóstica represente um passo importante para o apoio à decisão clínica, ela não fornece, isoladamente, informações suficientes para subsidiar recomendações terapêuticas. Em razão disso, a arquitetura proposta adota um segundo estágio independente de recuperação documental, cujo objetivo é localizar diretrizes clínicas e esquemas terapêuticos específicos para a doença identificada na etapa anterior.

Ao contrário do primeiro estágio, cuja consulta é construída a partir dos sintomas apresentados pelo paciente, a segunda etapa utiliza diretamente o nome da hipótese diagnóstica principal produzida pelo modelo de linguagem. Sempre que possível, utiliza-se a nomenclatura internacional da doença em inglês, uma vez que a maior parte da literatura médica indexada encontra-se nesse idioma.

A consulta enviada ao banco vetorial é construída dinamicamente a partir de modelos textuais semelhantes ao exemplo apresentado a seguir:

"How is Gastroesophageal Reflux Disease treated? Drug treatment of Gastroesophageal Reflux Disease."

Esse formato de consulta direciona a recuperação para capítulos e seções da literatura relacionados especificamente ao manejo clínico da doença, reduzindo a recuperação de documentos exclusivamente voltados ao diagnóstico ou à fisiopatologia.

Da mesma forma que na primeira etapa, a consulta textual é convertida em um vetor de embeddings utilizando o modelo *BAAI/bge-large-en-v1.5*. Em seguida, realiza-se uma nova

busca por similaridade vetorial na coleção *medical_rag_v2*, retornando os fragmentos mais relevantes relacionados às condutas terapêuticas.

Os documentos recuperados normalmente incluem informações referentes às opções farmacológicas indicadas, critérios de escolha do tratamento, doses usuais, terapias não medicamentosas, recomendações de acompanhamento e situações especiais previstas pelas diretrizes clínicas.

Todo esse contexto é então encaminhado ao modelo de linguagem, juntamente com a anamnese estruturada e a hipótese diagnóstica produzida anteriormente.

3.2.6 Construção da Resposta Clínica

Diferentemente de um sistema convencional baseado apenas em recuperação documental, o objetivo do modelo de linguagem nesta etapa não consiste em reproduzir integralmente os documentos recuperados, mas em sintetizar as evidências encontradas em uma resposta estruturada destinada às etapas posteriores do sistema.

Para isso, o prompt enviado ao modelo contém instruções explícitas restringindo a utilização exclusiva das informações presentes nos trechos recuperados pelo mecanismo de busca vetorial. Dessa forma, procura-se minimizar a utilização de conhecimento paramétrico armazenado durante o treinamento do modelo, reduzindo a probabilidade de alucinações.

Além disso, o modelo é instruído a produzir uma resposta estruturada contendo apenas princípios ativos farmacológicos, evitando nomes comerciais de medicamentos ou categorias terapêuticas excessivamente amplas.

Essa estratégia possui duas vantagens importantes.

A primeira consiste em facilitar a integração com o módulo *Drug RAG*, responsável por recuperar bulas específicas de cada princípio ativo identificado.

A segunda está relacionada à padronização dos resultados, uma vez que diferentes fabricantes podem comercializar o mesmo medicamento sob nomes distintos, enquanto o princípio ativo permanece inalterado.

O resultado produzido pelo modelo inclui informações como:

- hipótese diagnóstica principal;
- diagnósticos diferenciais, quando aplicáveis;
- justificativa baseada nas evidências recuperadas;
- princípios ativos sugeridos;

- recomendações não farmacológicas encontradas na literatura;
- observações clínicas relevantes.

Todas essas informações permanecem estruturadas em formato JSON, permitindo sua reutilização pelas etapas subsequentes da arquitetura.

3.2.7 Vantagens da Recuperação em Dois Estágios

A principal característica do *Medical RAG* desenvolvido neste trabalho consiste na separação entre o processo de identificação diagnóstica e a recuperação das diretrizes terapêuticas.

Embora fosse possível realizar ambas as tarefas em uma única consulta ao banco vetorial, observou-se durante o desenvolvimento do sistema que essa abordagem produzia resultados menos consistentes. Consultas excessivamente abrangentes tendiam a recuperar simultaneamente documentos relacionados a sintomas, fisiopatologia, exames complementares e tratamento, aumentando a quantidade de informações irrelevantes encaminhadas ao modelo de linguagem.

Ao dividir o processo em duas etapas independentes, cada consulta passa a possuir um objetivo específico. A primeira concentra-se exclusivamente na identificação da hipótese diagnóstica mais provável, enquanto a segunda busca apenas informações relacionadas ao manejo terapêutico da condição identificada.

Essa estratégia reduz significativamente o ruído documental presente no contexto recuperado, melhora a coerência das respostas produzidas pelo modelo e aumenta a rastreabilidade das recomendações terapêuticas apresentadas ao usuário.

Outro benefício importante está relacionado à reutilização dos resultados intermediários. Como a hipótese diagnóstica permanece armazenada de forma estruturada, diferentes módulos do sistema podem utilizá-la simultaneamente, incluindo o *Drug RAG* e o agente auditor, sem necessidade de repetir consultas ao banco vetorial.

Além disso, a separação entre diagnóstico e tratamento permite substituir futuramente qualquer uma das etapas sem impacto significativo sobre as demais. Novas bases documentais, novos modelos de embeddings ou até mesmo diferentes estratégias de recuperação podem ser incorporados individualmente, preservando a arquitetura geral do sistema.

3.3 RAG DE MEDICAMENTOS E RE-RANQUEAMENTO DEMOGRÁFICO

Embora o *Medical RAG* seja responsável por identificar hipóteses diagnósticas e recuperar diretrizes terapêuticas a partir da literatura médica, as informações obtidas nessa etapa ainda

são insuficientes para apoiar a prescrição medicamentosa de maneira segura. Em grande parte dos casos, livros-texto e diretrizes clínicas concentram-se na indicação dos princípios ativos recomendados para determinada condição clínica, mas não apresentam detalhadamente aspectos relacionados à posologia, contraindicações, advertências, interações medicamentosas e precauções específicas de cada fármaco.

Com o objetivo de complementar essas informações, foi desenvolvido um segundo mecanismo de recuperação documental denominado *Drug RAG*. Esse módulo opera de forma independente do *Medical RAG* e possui como finalidade recuperar informações farmacológicas diretamente de bulas oficiais de medicamentos previamente indexadas em uma base vetorial própria.

A separação entre os dois mecanismos de recuperação constitui uma decisão arquitetural importante da solução proposta. Enquanto o *Medical RAG* trabalha sobre literatura médica destinada ao raciocínio clínico e ao estabelecimento de hipóteses diagnósticas, o *Drug RAG* atua sobre documentos regulatórios que descrevem as características farmacológicas dos medicamentos comercializados. Dessa forma, cada módulo utiliza uma base documental especializada para resolver um problema específico, reduzindo o ruído semântico durante a recuperação das informações.

Outro aspecto relevante dessa arquitetura é que o *Drug RAG* não realiza inferências clínicas sobre a condição do paciente. Seu papel consiste exclusivamente em recuperar informações oficiais relacionadas aos medicamentos sugeridos pelo módulo anterior, permitindo que recomendações terapêuticas sejam posteriormente avaliadas pelo agente auditor utilizando evidências provenientes das próprias bulas.

Ao utilizar duas bases documentais independentes, o sistema reduz a dependência de uma única fonte de conhecimento e aumenta a diversidade das evidências consideradas durante o processo de apoio à decisão clínica.

3.3.1 Construção da Base Vetorial de Bulas

A construção da base documental do *Drug RAG* iniciou-se com a obtenção de bulas oficiais de medicamentos disponibilizadas em formato digital. Esses documentos constituem a principal fonte de informações regulamentadas sobre princípios ativos, indicações terapêuticas, contraindicações, posologia, advertências, precauções, reações adversas e demais informações farmacológicas exigidas pelos órgãos reguladores.

Assim como ocorreu na construção da base do *Medical RAG*, os documentos foram inicialmente convertidos para texto utilizando ferramentas de extração compatíveis com arquivos

PDF. Após essa etapa, foi realizado um processo de limpeza textual destinado à remoção de elementos que não agregavam informação semântica relevante para o processo de recuperação, como cabeçalhos repetitivos, numeração de páginas, caracteres especiais e espaços em branco excedentes.

Entretanto, diferentemente da literatura médica utilizada anteriormente, as bulas apresentam uma organização bastante estruturada, sendo divididas em seções padronizadas. Aproveitando essa característica, optou-se por segmentar cada bula de acordo com suas seções principais, preservando a organização lógica do documento.

Entre as seções consideradas mais relevantes para este trabalho destacam-se:

- Posologia;
- Contraindicações;
- Advertências e Precauções;
- Reações Adversas;
- Interações Medicamentosas.

Cada uma dessas seções foi tratada como um documento independente durante o processo de indexação. Essa estratégia reduz a quantidade de informações irrelevantes recuperadas durante as consultas, permitindo que o mecanismo de busca retorne apenas o trecho da bula relacionado ao tipo de informação solicitado.

Além disso, a segmentação por seções facilita a interpretação posterior realizada pelo modelo de linguagem, uma vez que o contexto enviado ao LLM torna-se significativamente menor e mais específico.

Após a segmentação, cada trecho textual passa a representar uma unidade independente dentro da base vetorial, podendo ser recuperado individualmente durante o processo de busca.

3.3.2 Representação Vetorial Utilizando Embeddings Multilíngues

Para transformar os fragmentos textuais das bulas em representações numéricas foi utilizado o modelo de embeddings *BAAI/bge-m3*.

A escolha desse modelo ocorreu principalmente devido à natureza multilíngue da base documental utilizada. Embora parte da literatura médica empregada pelo *Medical RAG* esteja predominantemente em língua inglesa, as bulas utilizadas pelo *Drug RAG* encontram-se em

português e apresentam grande quantidade de termos farmacológicos específicos, abreviações, unidades de medida e nomenclaturas técnicas.

O modelo *BAAI/bge-m3* apresenta elevado desempenho em tarefas de recuperação semântica multilíngue, sendo capaz de representar documentos escritos em diferentes idiomas dentro de um mesmo espaço vetorial. Essa característica reduz problemas relacionados à variação linguística e melhora a recuperação de documentos mesmo quando diferentes termos são utilizados para representar um mesmo conceito farmacológico.

Outro fator considerado durante a seleção do modelo foi sua capacidade de capturar relações semânticas entre expressões equivalentes. Dessa forma, consultas contendo termos como “dose”, “posologia”, “administração” ou “esquema terapêutico” tendem a recuperar documentos semanticamente relacionados, mesmo quando essas palavras não aparecem exatamente da mesma forma na bula indexada.

Após a geração dos embeddings, cada fragmento textual foi associado ao seu vetor correspondente, tornando possível a realização de consultas baseadas em similaridade semântica em vez de simples correspondência lexical.

3.3.3 Organização da Base Vetorial no ChromaDB

Todos os embeddings produzidos foram armazenados em uma coleção específica do banco vetorial ChromaDB denominada *drug_rag_v2*.

Cada registro armazenado nessa coleção contém o texto original do fragmento da bula, seu vetor de embedding e um conjunto de metadados utilizados durante o processo de recuperação. Entre esses metadados encontram-se o nome do medicamento, a seção da bula à qual o fragmento pertence e identificadores internos utilizados pelo sistema.

A utilização de uma coleção independente daquela empregada pelo *Medical RAG* evita que documentos de naturezas distintas sejam recuperados simultaneamente durante uma mesma consulta. Enquanto a coleção *medical_rag_v2* concentra informações provenientes da literatura médica, a coleção *drug_rag_v2* armazena exclusivamente documentos farmacológicos.

Essa separação simplifica o processo de recuperação, reduz o espaço de busca e aumenta a precisão dos resultados retornados pelo banco vetorial.

Ao final dessa etapa, a base vetorial encontra-se preparada para receber consultas contendo os princípios ativos identificados pelo *Medical RAG*, possibilitando a recuperação das informações farmacológicas que serão utilizadas nas etapas subsequentes de re-ranqueamento e auditoria clínica.

3.3.4 Recuperação Semântica das Informações Farmacológicas

Concluída a etapa de identificação dos princípios ativos recomendados pelo *Medical RAG*, inicia-se o processo de recuperação das informações farmacológicas correspondentes. Diferentemente da etapa anterior, em que a consulta era construída a partir dos sintomas apresentados pelo paciente, neste módulo a recuperação é realizada utilizando diretamente o nome do princípio ativo retornado pelo sistema.

Para cada medicamento identificado, o sistema realiza consultas independentes sobre a coleção *drug_rag_v2*, buscando recuperar fragmentos pertencentes às diferentes seções da bula. Em vez de recuperar o documento completo, são selecionados apenas os trechos semanticamente mais próximos da consulta realizada, reduzindo significativamente a quantidade de informação irrelevante enviada ao modelo de linguagem.

As consultas são executadas separadamente para as categorias de posologia, contraindicações e precauções. Essa separação permite que cada tipo de informação seja recuperado de maneira especializada, aumentando a precisão dos resultados obtidos e reduzindo a probabilidade de que informações pertencentes a outras seções da bula interfiram no processo de recuperação.

Inicialmente, o ChromaDB retorna um conjunto de fragmentos ordenados exclusivamente pela distância vetorial entre a consulta e os documentos indexados. Embora essa estratégia apresente excelente desempenho na recuperação semântica, observou-se durante o desenvolvimento do sistema que ela não considera características individuais do paciente, como idade e peso corporal, fatores fundamentais para a definição da posologia e para a identificação de contraindicações específicas.

Com o objetivo de incorporar essas características ao processo de busca, foi desenvolvido um mecanismo adicional de pós-processamento responsável por reorganizar os resultados retornados pelo banco vetorial antes de sua utilização pelo modelo de linguagem.

3.3.5 Heurística de Re-ranqueamento Demográfico

A principal contribuição deste módulo consiste na implementação de uma heurística de re-ranqueamento aplicada após a recuperação semântica realizada pelo banco vetorial.

Embora o ChromaDB seja capaz de recuperar documentos semanticamente semelhantes à consulta realizada, ele não possui conhecimento sobre características clínicas específicas do paciente. Consequentemente, dois fragmentos semanticamente semelhantes podem receber pontuações muito próximas, mesmo quando apenas um deles é adequado ao perfil demográfico

do indivíduo em atendimento.

Por exemplo, considere um paciente pediátrico de sete anos de idade. Durante a recuperação da posologia de determinado medicamento, podem ser encontrados simultaneamente trechos destinados a adultos e trechos destinados a crianças. Sob a perspectiva da similaridade vetorial, ambos podem apresentar elevada relevância. Entretanto, do ponto de vista clínico, apenas o trecho destinado à faixa etária pediátrica deve ser priorizado.

Para resolver esse problema foi desenvolvido um algoritmo de re-ranqueamento baseado em regras heurísticas. Após a recuperação inicial realizada pelo banco vetorial, cada fragmento recuperado é analisado individualmente em busca da ocorrência de palavras-chave associadas ao perfil demográfico do paciente.

Inicialmente, o sistema classifica automaticamente o paciente em um dos seguintes grupos:

- paciente pediátrico;
- paciente adulto.

Essa classificação utiliza como critério principal a idade informada durante a triagem conversacional. Atualmente considera-se como pediátrico todo paciente com idade inferior ou igual a 12 anos.

Uma vez identificado o perfil do paciente, o algoritmo percorre todos os fragmentos recuperados atribuindo bônus ou penalidades conforme os termos encontrados em cada trecho da bula.

Para pacientes pediátricos, expressões como “criança”, “crianças”, “uso pediátrico”, “mg/kg”, “peso corporal”, “gotas”, “lactentes” e “infantil” indicam maior aderência ao perfil clínico analisado, reduzindo a distância final atribuída ao documento.

Por outro lado, termos como “adultos”, “uso adulto”, “idosos” ou “dose fixa” recebem penalidades, deslocando esses fragmentos para posições inferiores do ranking.

Quando o paciente pertence ao grupo adulto, aplica-se a lógica inversa, priorizando documentos destinados à população adulta e reduzindo a relevância de trechos exclusivos para pacientes pediátricos.

Essa estratégia incorpora conhecimento clínico simples ao processo de recuperação sem alterar o funcionamento interno do banco vetorial, preservando a eficiência computacional da busca semântica.

3.3.6 Formulação Matemática do Re-ranqueamento

O algoritmo desenvolvido pode ser representado pela Equação 3.1.

$$S_{final} = S_{original} + \sum B_i - \sum P_j \quad (3.1)$$

em que:

- $S_{original}$ representa a distância vetorial retornada pelo ChromaDB;
- B_i representa os bônus atribuídos aos fragmentos compatíveis com o perfil do paciente;
- P_j representa as penalidades aplicadas aos fragmentos incompatíveis;
- S_{final} corresponde à pontuação utilizada para ordenar os resultados finais.

Como o ChromaDB utiliza distância vetorial como métrica de similaridade, valores menores indicam maior relevância. Dessa forma, a aplicação de bônus reduz a distância final do documento, enquanto a aplicação de penalidades aumenta essa distância.

É importante observar que esse procedimento não altera os embeddings nem interfere no mecanismo de recuperação vetorial implementado pelo banco de dados. Toda a modificação ocorre apenas após a recuperação inicial dos documentos, funcionando como uma etapa adicional de refinamento dos resultados.

Essa característica permite que novas regras clínicas sejam incorporadas futuramente sem necessidade de reconstrução da base vetorial ou novo treinamento dos modelos de embeddings.

3.3.7 Exemplo de Aplicação da Heurística

Considere um paciente de oito anos de idade com peso corporal de 27 kg.

Durante a recuperação da posologia de determinado medicamento, o banco vetorial retorna inicialmente dois fragmentos com elevada similaridade semântica.

O primeiro fragmento contém a seguinte orientação:

“Administrar 10 mg/kg/dia divididos em duas doses.”

O segundo apresenta a recomendação:

“Dose habitual para adultos: 500 mg a cada oito horas.”

Embora ambos apresentem elevada similaridade em relação à consulta realizada, apenas o primeiro é adequado ao perfil pediátrico do paciente.

Após identificar expressões como “mg/kg” e referências ao uso infantil, o algoritmo aplica bônus ao primeiro fragmento. Em contrapartida, a presença da expressão “adultos” faz com que o segundo documento receba uma penalidade.

Como consequência, o primeiro trecho passa a ocupar a posição mais alta do ranking, tornando-se o principal contexto encaminhado ao modelo de linguagem.

Esse procedimento reduz a probabilidade de que recomendações incompatíveis com o perfil demográfico do paciente sejam utilizadas durante a geração da resposta clínica, contribuindo para aumentar a consistência das informações recuperadas.

3.3.8 Geração das Informações Farmacológicas

Após a aplicação da heurística de re-ranqueamento, os fragmentos mais relevantes de cada medicamento são encaminhados ao modelo de linguagem para interpretação e consolidação das informações farmacológicas.

Embora os documentos recuperados contenham informações oficiais provenientes das bulas, sua estrutura textual nem sempre é adequada para utilização direta nas etapas posteriores do sistema. Em muitos casos, diferentes informações encontram-se distribuídas ao longo de vários parágrafos, exigindo um processo adicional de organização antes de serem disponibilizadas ao módulo de auditoria.

Com essa finalidade, o modelo de linguagem recebe como entrada o conjunto de fragmentos selecionados para cada princípio ativo juntamente com instruções que delimitam o formato esperado da resposta. Diferentemente do *Medical RAG*, cuja finalidade consiste em apoiar a construção das hipóteses diagnósticas, neste módulo o modelo atua exclusivamente como um mecanismo de organização das evidências recuperadas, não sendo autorizado a produzir informações que não estejam presentes nas bulas utilizadas como contexto.

Para cada medicamento pesquisado, o sistema solicita ao modelo que identifique e organize, sempre que disponíveis, informações referentes à posologia, contraindicações, advertências, precauções, principais reações adversas e observações relevantes para utilização clínica.

Essa estratégia permite transformar documentos extensos em uma representação estruturada, facilitando tanto o processamento automatizado quanto a posterior interpretação pelo profissional de saúde.

3.3.9 Estrutura da Resposta Produzida

As informações extraídas são armazenadas em uma estrutura padronizada em formato JSON, permitindo sua reutilização pelos módulos subsequentes da arquitetura.

Para cada princípio ativo recuperado são registrados, entre outros, os seguintes campos:

- nome do princípio ativo;
- trechos referentes à posologia;
- contraindicações identificadas;
- advertências e precauções;
- reações adversas relevantes;
- evidências documentais utilizadas na recuperação.

A utilização de uma estrutura padronizada elimina a necessidade de que os módulos posteriores interpretem novamente documentos em linguagem natural. Todas as informações farmacológicas passam a ser disponibilizadas de maneira organizada e consistente, reduzindo ambiguidades durante o processo de auditoria clínica.

Além disso, o armazenamento dos trechos recuperados preserva a rastreabilidade das recomendações produzidas pelo sistema, permitindo identificar posteriormente quais evidências documentais fundamentaram determinada orientação farmacológica.

3.3.10 Integração com o Agente Auditor

A saída produzida pelo *Drug RAG* constitui uma das principais entradas do módulo de auditoria clínica implementado nesta arquitetura.

Enquanto o *Medical RAG* fornece as hipóteses diagnósticas e os princípios ativos sugeridos pela literatura médica, o *Drug RAG* acrescenta informações detalhadas sobre a utilização segura desses medicamentos.

Essas informações permitem ao agente auditor verificar automaticamente se os medicamentos sugeridos apresentam contraindicações relacionadas à idade, ao peso corporal, às comorbidades registradas durante a anamnese ou a outros fatores clínicos identificados durante a entrevista.

Da mesma forma, os trechos referentes à posologia recuperados das bulas permitem verificar se eventuais doses mencionadas pelo modelo de linguagem permanecem compatíveis com aquelas descritas na documentação oficial do medicamento.

Essa integração entre os módulos aumenta significativamente a capacidade do sistema de identificar inconsistências antes da geração do relatório final, funcionando como uma camada adicional de segurança durante o processo de apoio à decisão clínica.

3.3.11 Limitações da Estratégia Proposta

Embora a utilização de uma base vetorial de bulas aumente a qualidade das informações farmacológicas recuperadas, algumas limitações permanecem presentes.

A primeira está relacionada à cobertura documental da base utilizada. Apenas medicamentos previamente indexados podem ser recuperados durante a execução do sistema. Caso determinado princípio ativo não esteja presente na coleção vetorial, nenhuma informação farmacológica poderá ser obtida automaticamente.

Outra limitação refere-se à própria natureza da recuperação semântica. Apesar do mecanismo de re-ranqueamento reduzir significativamente a recuperação de documentos incompatíveis com o perfil do paciente, ainda existe a possibilidade de fragmentos parcialmente relevantes serem selecionados, especialmente em medicamentos que apresentam diferentes esquemas terapêuticos para distintas indicações clínicas.

Além disso, o algoritmo de re-ranqueamento atualmente considera apenas informações demográficas básicas, como idade e peso corporal. Outros fatores clínicos importantes, incluindo função renal, função hepática, gravidez, lactação e doenças preexistentes, não participam diretamente do cálculo da pontuação dos documentos recuperados. Essas informações permanecem disponíveis para análise apenas na etapa de auditoria clínica.

Apesar dessas limitações, observou-se que a combinação entre recuperação vetorial, re-ranqueamento heurístico e organização das informações pelo modelo de linguagem produz resultados mais consistentes do que a utilização isolada da busca semântica.

3.4 AGENTE AUDITOR CLÍNICO (*AGENT-AS-A-JUDGE*)

Apesar da utilização de técnicas de *Retrieval-Augmented Generation* reduzir significativamente a ocorrência de alucinações em Modelos de Linguagem de Grande Escala, a simples recuperação de documentos relevantes não garante que todas as recomendações produzidas pelo modelo estejam integralmente fundamentadas nas evidências recuperadas. O modelo ainda pode interpretar incorretamente o contexto disponível, omitir informações importantes ou produzir recomendações parcialmente incompatíveis com o perfil clínico do paciente.

Com o objetivo de reduzir esse risco, a arquitetura proposta incorpora uma etapa adicional de validação baseada no paradigma *Agent-as-a-Judge*. Diferentemente da estratégia tradicional de utilizar um único modelo responsável por toda a tomada de decisão, essa abordagem introduz um segundo agente especializado cuja única função consiste em avaliar criticamente a resposta produzida pelo primeiro agente antes que ela seja apresentada ao profissional de saúde.

Esse paradigma foi recentemente formalizado por Zhuge et al. (5), que propõem substituir mecanismos tradicionais de avaliação por agentes capazes de analisar todo o processo de resolução de uma tarefa. Em vez de avaliar apenas o resultado final produzido por outro agente, o avaliador recebe acesso às etapas intermediárias do raciocínio, às evidências utilizadas durante o processamento e ao contexto completo da tarefa, permitindo uma análise muito mais consistente da qualidade da solução produzida.

Embora o trabalho de Zhuge et al. tenha sido desenvolvido para avaliação de agentes voltados ao desenvolvimento de software, o princípio arquitetural pode ser generalizado para outros domínios que exijam elevado grau de confiabilidade. Neste trabalho, esse conceito foi adaptado para o contexto de apoio à decisão clínica, no qual o agente auditor atua como uma camada adicional de verificação sobre todas as informações produzidas pelos módulos anteriores da arquitetura.

3.4.1 Arquitetura do Agente Auditor

O agente auditor constitui o último componente inteligente do pipeline de processamento. Sua execução ocorre somente após a conclusão das etapas de recuperação documental realizadas pelos módulos *Medical RAG* e *Drug RAG*.

Enquanto os módulos anteriores possuem como objetivo recuperar evidências clínicas e farmacológicas relevantes, o agente auditor não realiza novas buscas documentais. Sua responsabilidade consiste exclusivamente em verificar a consistência entre a anamnese estruturada, as evidências recuperadas e a resposta clínica produzida pelo sistema.

Essa separação de responsabilidades reduz o risco de que interpretações equivocadas realizadas durante a geração da resposta permaneçam despercebidas, funcionando como uma segunda camada independente de validação.

Para realizar essa auditoria, o sistema utiliza novamente o modelo Gemini 2.5 Flash, porém com um conjunto completamente distinto de instruções (*prompt*). Diferentemente do agente responsável pela geração da recomendação clínica, o auditor recebe instruções explícitas para assumir uma postura crítica, identificar inconsistências e rejeitar qualquer recomendação

que não esteja sustentada pelas evidências documentais disponibilizadas.

3.4.2 Dados Fornecidos ao Agente Auditor

Ao contrário de uma arquitetura convencional baseada em um único modelo de linguagem, na qual a mesma instância do modelo é responsável tanto pela recuperação das informações quanto pela elaboração da resposta final, a arquitetura proposta separa explicitamente as etapas de geração e validação. Essa separação permite que o agente auditor realize sua análise de forma independente, sem influência direta do processo utilizado para produzir a recomendação clínica inicial.

Para que essa auditoria seja realizada de maneira consistente, o agente recebe simultaneamente todas as informações produzidas durante as etapas anteriores do pipeline. Dessa forma, sua avaliação não depende exclusivamente da resposta textual gerada pelo sistema, mas considera também as evidências documentais que fundamentaram cada recomendação apresentada.

Os dados encaminhados ao agente auditor podem ser agrupados em cinco conjuntos principais:

1. A anamnese estruturada produzida durante a etapa de triagem conversacional;
2. As hipóteses diagnósticas identificadas pelo módulo *Medical RAG*;
3. Os trechos recuperados da literatura médica utilizados para fundamentar o diagnóstico e o tratamento sugeridos;
4. As informações farmacológicas recuperadas pelo módulo *Drug RAG*, incluindo posologia, contraindicações, advertências e precauções;
5. A resposta clínica completa produzida pelo agente responsável pela geração da recomendação.

A utilização simultânea desses diferentes conjuntos de dados permite que o agente auditor realize uma verificação cruzada entre a resposta produzida pelo sistema e as evidências efetivamente recuperadas durante o processo de busca documental. Dessa forma, torna-se possível identificar situações nas quais determinada recomendação clínica não encontra respaldo na literatura médica recuperada ou nas informações presentes nas bulas dos medicamentos.

Outro aspecto importante dessa arquitetura consiste no fato de que o auditor não recebe apenas informações resumidas. Sempre que possível, são encaminhados também os fragmentos textuais originais recuperados pelo banco vetorial, preservando o contexto utilizado durante a etapa de geração da resposta. Essa estratégia reduz a necessidade de inferências adicionais

por parte do modelo avaliador e melhora a rastreabilidade das decisões produzidas durante a auditoria.

Além disso, o envio da anamnese estruturada permite que o agente considere características específicas do paciente, como idade, sexo, peso corporal, medicamentos em uso, comorbidades previamente informadas e possíveis alergias. Essas informações são fundamentais para identificar contraindicações que poderiam passar despercebidas caso apenas a resposta textual produzida pelo sistema fosse analisada.

3.4.3 Processo de Auditoria Clínica

Após receber todas as informações produzidas pelas etapas anteriores do pipeline, o agente inicia um processo estruturado de avaliação da recomendação clínica. Diferentemente dos módulos de recuperação documental, cujo objetivo consiste em localizar informações relevantes na base de conhecimento, o agente auditor executa uma tarefa exclusivamente analítica.

Inicialmente, o modelo verifica se a hipótese diagnóstica apresentada é compatível com os sinais, sintomas e antecedentes registrados na anamnese estruturada. Embora não tenha como objetivo confirmar ou refutar o diagnóstico proposto, essa etapa permite identificar inconsistências evidentes entre os dados clínicos coletados e a hipótese levantada pelo sistema.

Na sequência, o agente compara cada recomendação terapêutica produzida pelo módulo *Medical RAG* com os trechos recuperados da literatura médica. O objetivo dessa etapa é verificar se os princípios ativos sugeridos realmente aparecem nas diretrizes clínicas recuperadas ou se foram introduzidos pelo modelo de linguagem sem respaldo documental.

Posteriormente, o agente confronta cada medicamento recomendado com as informações provenientes do módulo *Drug RAG*. São analisadas contraindicações, advertências, restrições por faixa etária, limitações relacionadas ao peso corporal, precauções específicas e demais informações farmacológicas relevantes recuperadas diretamente das bulas oficiais.

Durante essa etapa também são avaliadas possíveis incompatibilidades entre as características clínicas do paciente e os medicamentos sugeridos. Caso a anamnese indique, por exemplo, alergias previamente conhecidas, doenças pré-existentes ou utilização de outros medicamentos potencialmente conflitantes, essas informações são consideradas na elaboração do parecer final.

Além da análise clínica propriamente dita, o agente verifica se existem informações insuficientes para uma recomendação segura. Em situações nas quais dados essenciais estejam ausentes, como idade, peso corporal ou histórico medicamentoso, o sistema registra explicitamente essa limitação no relatório de auditoria, indicando que determinadas recomendações devem ser

interpretadas com cautela.

Ao final desse processo, todas as evidências identificadas são consolidadas em um relatório estruturado que será utilizado tanto para composição da resposta final ao profissional de saúde quanto para registro da execução da sessão clínica.

3.4.4 Geração dos Escores de Auditoria e Parecer Clínico

Concluída a etapa de verificação das evidências recuperadas, o agente auditor consolida sua avaliação em um relatório estruturado contendo indicadores quantitativos e qualitativos sobre a consistência da recomendação clínica produzida pelo sistema.

A utilização de indicadores numéricos facilita tanto a interpretação pelo profissional de saúde quanto futuras avaliações experimentais da arquitetura proposta. Em vez de produzir apenas uma justificativa textual, o agente sintetiza sua análise por meio de escores que representam diferentes dimensões da qualidade da resposta.

Nesta implementação são calculados dois indicadores principais: o *Score de Segurança Clínica* e o *Score de Fidelidade ao Contexto Recuperado*.

O primeiro indicador procura estimar o nível de segurança da recomendação produzida considerando exclusivamente as características clínicas do paciente e as informações farmacológicas recuperadas durante a etapa do *Drug RAG*. São analisados fatores como contraindicações, alergias relatadas, restrições por faixa etária, limitações relacionadas ao peso corporal, advertências específicas e demais informações presentes nas bulas dos medicamentos sugeridos.

Quanto maior o número de incompatibilidades identificadas entre o perfil clínico do paciente e as recomendações terapêuticas produzidas pelo sistema, menor tende a ser o escore de segurança atribuído pelo agente auditor.

O segundo indicador avalia a fidelidade da resposta produzida em relação às evidências efetivamente recuperadas durante o processo de busca documental. Seu objetivo é identificar possíveis alucinações do modelo de linguagem, verificando se os medicamentos, posologias e recomendações terapêuticas realmente encontram respaldo nos trechos recuperados pelo módulo *Medical RAG*.

Embora ambos os indicadores sejam produzidos pelo modelo de linguagem, eles seguem critérios previamente definidos nas instruções fornecidas ao agente auditor, garantindo que a avaliação seja realizada de maneira padronizada para diferentes execuções do sistema.

Do ponto de vista conceitual, o processo de avaliação pode ser representado pela Equação 3.2, na qual a qualidade global da resposta depende simultaneamente da consistência clínica observada

e da aderência às evidências recuperadas.

$$Q = f(S_{seg}, S_{fid}) \quad (3.2)$$

em que:

- Q representa a qualidade global da recomendação produzida;
- S_{seg} corresponde ao Score de Segurança Clínica;
- S_{fid} representa o Score de Fidelidade ao Contexto Recuperado.

Nesta implementação, ambos os escores assumem valores no intervalo de 0 a 10, sendo valores mais elevados indicativos de maior confiabilidade da recomendação produzida. Ressalta-se, entretanto, que esses indicadores não possuem finalidade diagnóstica nem substituem a avaliação médica, constituindo apenas mecanismos auxiliares para apoiar a interpretação dos resultados gerados pela arquitetura.

Além dos escores numéricos, o agente produz uma justificativa textual detalhando os principais fatores que influenciaram sua decisão. Essa justificativa descreve, por exemplo, contraindicações encontradas, ausência de informações clínicas relevantes, inconsistências entre os documentos recuperados e a resposta produzida ou evidências suficientes para aprovação da recomendação.

Como resultado final da auditoria, o sistema gera um objeto estruturado contendo todos os indicadores produzidos durante a avaliação. A Listagem 3.1 apresenta uma representação simplificada dessa estrutura.

Listing 3.1 – Estrutura simplificada do relatório produzido pelo agente auditor.

```
{
  "score_seguranca": 9.4,
  "score_fidelidade": 9.7,
  "riscos_identificados": [
    "...",
    "..."
  ],
  "analise_critica": "...",
  "aprovado_para_medico": true
}
```

O campo *score_seguranca* representa a avaliação da compatibilidade entre a recomendação produzida e as características clínicas do paciente. O campo *score_fidelidade* informa o grau de aderência da resposta às evidências recuperadas pelo sistema de recuperação documental. Já o vetor *riscos_identificados* reúne todas as inconsistências relevantes encontradas durante a auditoria, enquanto o campo *analise_critica* apresenta uma justificativa textual sintetizando a decisão do agente.

Por fim, o atributo booleano *aprovado_para_medico* indica se a recomendação foi considerada suficientemente consistente para ser apresentada ao profissional de saúde. Essa informação não representa uma aprovação automática de tratamento, mas apenas sinaliza que, segundo os critérios definidos para o agente auditor, não foram identificadas inconsistências relevantes que impeçam a utilização da resposta como ferramenta de apoio à decisão clínica.

Dessa forma, o paradigma *Agent-as-a-Judge* deixa de atuar apenas como um mecanismo de avaliação experimental de agentes inteligentes, conforme proposto por Zhuge et al. (5), passando a integrar efetivamente o pipeline de processamento da arquitetura desenvolvida neste trabalho. Em vez de avaliar agentes para fins de benchmarking, o auditor atua como uma camada adicional de segurança clínica, responsável por verificar a consistência das recomendações antes de sua disponibilização ao profissional médico.

3.5 FLUXO COMPLETO DE PROCESSAMENTO E GERAÇÃO DOS RELATÓRIOS

Após a implementação individual de cada módulo apresentado nas seções anteriores, a arquitetura proposta opera como um pipeline integrado de processamento, no qual as informações produzidas em cada etapa são utilizadas como entrada para o módulo subsequente. Essa organização permite que a solução execute todo o processo de apoio à decisão clínica de forma automática, desde a coleta inicial das informações do paciente até a geração de um relatório estruturado contendo as evidências utilizadas na recomendação clínica e a avaliação produzida pelo agente auditor.

A Figura 1 resume esse fluxo de execução, evidenciando a sequência de processamento adotada pela arquitetura desenvolvida.

O fluxo inicia-se quando o paciente estabelece uma nova sessão de atendimento por meio da interface conversacional. A primeira mensagem enviada é encaminhada ao módulo de triagem conversacional, responsável por conduzir a entrevista clínica utilizando linguagem natural. Durante toda a interação, o sistema mantém o histórico completo da conversa, permitindo

que cada nova resposta seja interpretada considerando o contexto acumulado da sessão.

A cada interação, o módulo de extração de informações clínicas identifica automaticamente os dados relevantes presentes na resposta do paciente, atualizando progressivamente a anamnese estruturada. Esse processo continua até que o agente conversacional determine que foram obtidas informações suficientes para subsidiar uma avaliação clínica inicial.

Concluída a entrevista, todas as informações coletadas são consolidadas em uma estrutura padronizada de anamnese. Esse documento representa o principal artefato produzido pela etapa de triagem e passa a constituir a entrada para os módulos responsáveis pela recuperação das evidências médicas.

Na sequência, a anamnese estruturada é encaminhada ao módulo *Medical RAG*. Inicialmente, são utilizados os sinais, sintomas, antecedentes e demais informações clínicas para construir uma consulta semântica destinada à identificação das hipóteses diagnósticas mais compatíveis com o quadro apresentado pelo paciente. A busca é realizada sobre a base vetorial composta por livros-texto e literatura médica previamente indexada.

Os fragmentos documentais recuperados são então utilizados pelo modelo de linguagem para identificar a hipótese diagnóstica considerada mais provável. Além da identificação do diagnóstico principal, o sistema produz uma lista de diagnósticos diferenciais e termos equivalentes em língua inglesa, permitindo a realização de uma segunda etapa de recuperação documental mais especializada.

Utilizando o diagnóstico identificado anteriormente, o módulo *Medical RAG* realiza uma nova consulta vetorial direcionada especificamente às diretrizes terapêuticas associadas à patologia encontrada. Como resultado, são recuperados novos fragmentos documentais contendo informações referentes ao tratamento farmacológico, às recomendações clínicas e às opções terapêuticas descritas na literatura científica.

A lista de princípios ativos obtida nessa etapa é posteriormente encaminhada ao módulo *Drug RAG*. Para cada medicamento identificado, o sistema realiza uma nova busca semântica sobre a base vetorial composta por bulas oficiais previamente processadas. Diferentemente da etapa anterior, cujo foco consiste na literatura médica, essa nova recuperação concentra-se exclusivamente nas informações farmacológicas referentes aos medicamentos sugeridos.

Durante essa etapa são recuperadas informações relacionadas à posologia, contraindicações, advertências, precauções, interações medicamentosas e demais orientações presentes nas bulas. Antes da seleção final dos documentos, os resultados passam pelo mecanismo de re-ranqueamento desenvolvido neste trabalho, responsável por priorizar automaticamente fragmentos compatíveis

com o perfil demográfico do paciente, considerando informações como idade e peso corporal.

Concluídas as duas etapas de recuperação documental, todas as evidências produzidas ao longo do processamento são reunidas para compor a resposta clínica preliminar do sistema. Essa resposta inclui a anamnese estruturada, a hipótese diagnóstica principal, os diagnósticos diferenciais, os medicamentos encontrados na literatura médica e as principais informações farmacológicas recuperadas das bulas.

Entretanto, antes que essas informações sejam disponibilizadas ao profissional de saúde, toda a documentação produzida é encaminhada ao agente auditor baseado no paradigma *Agent-as-a-Judge*. Nessa etapa, o sistema realiza uma segunda análise independente da recomendação produzida, verificando sua consistência com as evidências recuperadas durante as etapas anteriores.

O agente auditor recebe simultaneamente a anamnese estruturada, os documentos recuperados pelo *Medical RAG*, as informações farmacológicas provenientes do *Drug RAG* e a resposta clínica construída pelo sistema. Com base nesse conjunto de informações, são avaliadas possíveis inconsistências entre a recomendação produzida e os documentos utilizados como fundamentação.

Além da verificação da aderência às evidências documentais, o auditor identifica possíveis contraindicações relacionadas às características do paciente, incompatibilidades farmacológicas, ausência de informações clínicas relevantes e recomendações que não encontram respaldo na documentação recuperada. Como resultado dessa análise são produzidos o escore de segurança clínica, o escore de fidelidade ao contexto recuperado, a lista de riscos identificados e um parecer textual justificando a decisão tomada pelo agente.

Após a conclusão da auditoria, o sistema consolida todas as informações produzidas ao longo da execução em uma estrutura única de resultados. Essa estrutura reúne desde os dados obtidos durante a entrevista inicial até as avaliações realizadas pelo agente auditor, preservando todo o histórico necessário para rastrear a origem de cada recomendação produzida.

Por fim, os resultados são persistidos em dois formatos complementares. O primeiro corresponde a um arquivo estruturado em formato JSON, destinado ao armazenamento de todas as informações produzidas durante a execução da arquitetura. Esse arquivo contém a anamnese estruturada, as hipóteses diagnósticas, os documentos recuperados pelos módulos de RAG, as informações farmacológicas, os indicadores produzidos pelo agente auditor e demais metadados da sessão clínica.

O segundo formato consiste em um relatório textual destinado ao profissional de saúde. Diferentemente do arquivo JSON, cujo objetivo principal é facilitar o processamento computacional, esse relatório organiza as informações em linguagem natural, apresentando de maneira resumida

os principais dados da anamnese, as hipóteses diagnósticas levantadas, as recomendações terapêuticas encontradas na literatura, os cuidados farmacológicos identificados nas bulas e a avaliação produzida pelo agente auditor.

Essa estratégia permite que a arquitetura produza simultaneamente um registro estruturado para integração com outros sistemas computacionais e um documento de fácil interpretação para utilização durante o atendimento clínico. Além disso, a manutenção de todas as evidências utilizadas ao longo do processamento proporciona transparência ao funcionamento do sistema, permitindo que o profissional de saúde identifique quais documentos fundamentaram cada recomendação apresentada.

Sob o ponto de vista arquitetural, a organização do processamento em módulos independentes também favorece futuras extensões da solução proposta. Novas bases documentais podem ser incorporadas ao pipeline sem alterações significativas nos componentes existentes, assim como novos agentes especializados podem ser adicionados para executar tarefas complementares, como interpretação de exames laboratoriais, análise de exames de imagem ou integração com sistemas eletrônicos de prontuário. Essa característica reforça a flexibilidade da arquitetura desenvolvida, permitindo sua evolução para diferentes cenários de apoio à decisão clínica.

4 RESULTADOS

Diante das atividades desenvolvidas ao longo deste trabalho, os resultados obtidos demonstram que a arquitetura proposta foi implementada com sucesso e atende aos requisitos funcionais definidos inicialmente. A aplicação web apresenta interface acessível, fluida e intuitiva, tanto para usuários leigos quanto para profissionais de saúde, permitindo a realização completa do fluxo de triagem, processamento clínico e geração de relatórios.

A seguir, são apresentados os principais resultados obtidos, organizados por módulo e aspecto avaliado.

4.1 FUNCIONAMENTO GERAL DA APLICAÇÃO

O sistema opera por meio de um fluxo integrado que inicia com a triagem conversacional realizada pelo assistente virtual. Este módulo conduz uma entrevista em linguagem natural, coletando a queixa principal, história da doença atual, revisão de sintomas e antecedentes relevantes, consolidando automaticamente uma anamnese estruturada.

Após o processamento das informações pela arquitetura multiagente, o sistema gera um relatório clínico completo contendo hipóteses diagnósticas, recomendações terapêuticas, evidências recuperadas e avaliação do agente auditor.

A seguir, é apresentado um exemplo real de relatório gerado pelo sistema:

RELATÓRIO CLÍNICO DE TRIAGEM E AUDITORIA - SESSÃO:

ce86aebb-6c58-4f54-9f19-d7e684214e2d

Gerado em: 19/06/2026 às 18:18:39

1. DADOS DE IDENTIFICAÇÃO DO PACIENTE

Nome: Maria Conceição

Idade: 56 anos

Sexo: Feminino

2. HISTÓRICO DA QUEIXA PRINCIPAL

Sintomas: Dor no peito tipo queimação e azia.

Evolução: Piora após comer e ao se deitar, tosse seca/noturna.

3. HIPÓTESES PRELIMINARES IDENTIFICADAS

Hipótese 1: Doença do Refluxo Gastroesofágico (DRGE)

4. RELATÓRIO DO AGENTE AUDITOR (AGENT-AS-A-JUDGE)

Aprovado para encaminhamento médico: SIM

Score de Segurança Clínica: 9/10

Score de Fidelidade do RAG: 10/10

O agente auditor identificou corretamente a hipótese diagnóstica e considerou a recomendação terapêutica clinicamente adequada e fundamentada nas evidências recuperadas.

4.2 AVALIAÇÃO DO MEDICAL RAG

O módulo Medical RAG realizou a recuperação de informações em base documental composta por literatura médica e protocolos clínicos. Para o caso clínico apresentado, ao buscar o termo “Doença do Refluxo Gastroesofágico (DRGE)”, o sistema recuperou documentos relevantes, incluindo obras como *Decision Making in Medicine*, *Approach to Internal Medicine* e *Current Essentials of Medicine*.

A recuperação permitiu identificar a hipótese diagnóstica principal e obter estratégias terapêuticas consistentes com a prática clínica.

4.3 AVALIAÇÃO DO DRUG RAG

O Drug RAG efetuou buscas em base vetorial de bulas oficiais de medicamentos, recuperando informações de posologia, contraindicações e precauções para as classes terapêuticas sugeridas (Inibidores da Bomba de Prótons, Antagonistas dos Receptores H2, Antiácidos, Agentes Procinéticos e Alginato).

O mecanismo de re-ranqueamento baseado em características demográficas do paciente priorizou trechos mais relevantes das bulas, apresentando scores de similaridade e extratos específicos das seções consultadas.

4.4 AVALIAÇÃO DO AGENT-AS-A-JUDGE

O agente auditor atuou como camada de validação final, analisando de forma independente a consistência entre a anamnese, as hipóteses diagnósticas, as recomendações do Medical RAG e as evidências farmacológicas do Drug RAG.

No caso avaliado, o agente atribuiu os seguintes scores:

- Score de Segurança Clínica: 9/10

- Score de Fidelidade do RAG: 10/10

Além disso, produziu uma análise crítica detalhada, destacando pontos positivos da recomendação e alertando sobre a limitação decorrente de campos incompletos na anamnese.

4.5 DISCUSSÃO DOS RESULTADOS

A arquitetura multiagente com recuperação aumentada por geração (RAG) em dois estágios e auditoria automática demonstrou ser capaz de produzir recomendações fundamentadas em evidências documentais, aumentando significativamente a transparência e a rastreabilidade do processo.

A separação entre Medical RAG (focado em diagnóstico e diretrizes terapêuticas) e Drug RAG (focado em farmacologia com re-ranqueamento) mostrou-se alinhada ao raciocínio clínico. A inclusão do Agent-as-a-Judge adicionou uma importante camada de segurança ao identificar riscos e inconsistências.

Como principais limitações observadas, destacam-se a dependência da qualidade e abrangência das bases documentais utilizadas e a necessidade de uma anamnese completa para maior precisão das recomendações. Ressalta-se que o sistema foi desenvolvido como ferramenta de apoio à decisão clínica, não possuindo autonomia para realizar diagnósticos definitivos ou substituir o julgamento médico.

5 CONCLUSÃO

O avanço recente dos Modelos de Linguagem de Grande Escala (*Large Language Models* – LLMs) abriu novas possibilidades para o desenvolvimento de sistemas inteligentes capazes de auxiliar profissionais em tarefas complexas. Entretanto, quando aplicados diretamente ao domínio da saúde, esses modelos apresentam limitações importantes relacionadas à geração de informações não fundamentadas, conhecidas como alucinações, à ausência de mecanismos de rastreabilidade das respostas produzidas e à dificuldade de incorporar evidências clínicas atualizadas durante o processo de inferência. Essas limitações tornam inadequada a utilização isolada de LLMs em aplicações voltadas ao apoio à decisão clínica.

Diante desse cenário, este trabalho propôs e implementou uma arquitetura modular baseada em sistemas multiagentes para apoio à decisão clínica, integrando técnicas de *Retrieval-Augmented Generation* (RAG), recuperação documental especializada, auditoria automatizada por agentes inteligentes e geração estruturada de relatórios clínicos. A arquitetura foi concebida com o objetivo de reduzir a dependência do conhecimento paramétrico dos modelos de linguagem, fundamentando as respostas produzidas em evidências recuperadas dinamicamente de bases documentais compostas por literatura médica e bulas oficiais de medicamentos.

A primeira contribuição deste trabalho consiste na implementação de um módulo de triagem conversacional capaz de conduzir entrevistas clínicas utilizando linguagem natural. Diferentemente de formulários tradicionais, o sistema interpreta automaticamente as respostas do paciente, identifica informações relevantes, valida os dados coletados e consolida progressivamente uma anamnese estruturada. Essa abordagem proporciona maior flexibilidade durante a coleta das informações clínicas e reduz a necessidade de preenchimento manual de formulários extensos, tornando a interação mais natural para o usuário.

Como segunda contribuição, foi desenvolvido um mecanismo de recuperação documental em dois estágios denominado *Medical RAG*. A separação entre a identificação das hipóteses diagnósticas e a recuperação das diretrizes terapêuticas mostrou-se adequada para representar o raciocínio clínico empregado durante a prática médica. Enquanto a primeira etapa concentra-se na análise dos sinais e sintomas apresentados pelo paciente, a segunda realiza uma busca direcionada às estratégias terapêuticas descritas na literatura científica, permitindo que as recomendações produzidas sejam fundamentadas em documentos especializados previamente indexados.

Complementando essa arquitetura, foi implementado um segundo módulo de recuperação documental voltado especificamente para informações farmacológicas. O *Drug RAG* realiza

buscas em uma base vetorial composta por bulas oficiais de medicamentos, recuperando automaticamente informações relacionadas à posologia, contraindicações, advertências e precauções. Além da recuperação semântica convencional, esse módulo incorpora um mecanismo de re-ranqueamento baseado em características demográficas do paciente, como idade e peso corporal, permitindo priorizar automaticamente trechos de bula compatíveis com o perfil clínico informado durante a triagem.

Outra contribuição relevante deste trabalho consiste na incorporação do paradigma *Agent-as-a-Judge* ao processo de apoio à decisão clínica. Em vez de apresentar diretamente as recomendações produzidas pelo modelo de linguagem, a arquitetura introduz uma etapa adicional de auditoria automática responsável por verificar a consistência entre as hipóteses diagnósticas, as recomendações terapêuticas e as evidências documentais recuperadas pelos módulos de RAG. Essa estratégia aumenta a transparência do processo de inferência, permitindo identificar inconsistências, contraindicações relevantes e eventuais recomendações não fundamentadas na documentação consultada.

Como resultado do fluxo completo de processamento, o sistema produz relatórios estruturados que reúnem a anamnese do paciente, as hipóteses diagnósticas identificadas, as recomendações terapêuticas encontradas na literatura, os trechos relevantes das bulas consultadas e a avaliação realizada pelo agente auditor. A geração simultânea de relatórios em formato JSON e em formato textual facilita tanto a integração com outros sistemas quanto a interpretação das informações por profissionais de saúde, aumentando a rastreabilidade das decisões produzidas pela arquitetura.

Os resultados obtidos demonstram que a integração entre técnicas de recuperação documental, modelos de linguagem e agentes especializados constitui uma alternativa promissora para o desenvolvimento de sistemas de apoio à decisão clínica mais transparentes e fundamentados em evidências. Ao separar o processo em módulos independentes, a arquitetura reduz o acoplamento entre seus componentes, favorecendo a manutenção, a escalabilidade e a incorporação de novas funcionalidades sem alterações significativas na estrutura existente.

Entretanto, algumas limitações devem ser consideradas. A qualidade das recomendações produzidas depende diretamente da qualidade e da abrangência das bases documentais utilizadas durante o processo de recuperação. Da mesma forma, embora o mecanismo de auditoria reduza significativamente a probabilidade de inconsistências, ele não elimina completamente a possibilidade de interpretações incorretas produzidas pelos modelos de linguagem. Além disso, a arquitetura proposta foi desenvolvida como um sistema de apoio à decisão clínica, não possuindo o objetivo de substituir o julgamento médico nem realizar diagnósticos definitivos de forma autônoma.

Outra limitação refere-se ao fato de que a arquitetura ainda não realiza integração direta

com prontuários eletrônicos, sistemas hospitalares ou exames laboratoriais. Dessa forma, todas as informações clínicas utilizadas durante a inferência dependem dos dados fornecidos durante a entrevista conversacional, podendo existir limitações decorrentes da ausência de informações relevantes para determinados casos clínicos.

Como perspectivas para trabalhos futuros, destacam-se a integração da arquitetura com padrões de interoperabilidade em saúde, como o HL7 FHIR, permitindo sua comunicação com prontuários eletrônicos e sistemas hospitalares. Também se mostra promissora a incorporação de novos agentes especializados voltados à interpretação de exames laboratoriais, exames de imagem e monitoramento contínuo de pacientes, ampliando a capacidade de análise clínica da solução proposta.

Outra possibilidade consiste na atualização automática das bases vetoriais utilizadas pelos módulos de recuperação documental, incorporando continuamente novas diretrizes clínicas, protocolos terapêuticos e bulas de medicamentos publicados pelos órgãos reguladores. Essa estratégia contribuiria para manter o sistema alinhado às evidências científicas mais recentes sem necessidade de reprocessamentos completos da base documental.

Adicionalmente, futuras pesquisas poderão investigar a utilização de modelos multimodais capazes de integrar informações textuais, imagens médicas e dados estruturados provenientes de dispositivos de monitoramento clínico, ampliando significativamente o escopo da arquitetura apresentada neste trabalho.

Por fim, destaca-se que a principal contribuição desta pesquisa não reside apenas na utilização de modelos de linguagem para responder perguntas médicas, mas na proposição de uma arquitetura completa baseada em sistemas multiagentes, recuperação aumentada por geração e mecanismos automáticos de auditoria capazes de produzir recomendações clínicas fundamentadas em evidências documentais. Embora a decisão final permaneça sob responsabilidade do profissional de saúde, os resultados obtidos indicam que a abordagem proposta representa um passo relevante para o desenvolvimento de sistemas de apoio à decisão clínica mais confiáveis, transparentes, rastreáveis e alinhados aos princípios da medicina baseada em evidências.

REFERÊNCIAS

- 1 MCKINNEY, S. M. e. a. International evaluation of an ai system for breast cancer screening. *Nature*, v. 577, p. 89–94, 2020.
- 2 UNIÃO, T. de Contas da. *Eficiência de hospitais do SUS se manteve estável entre 2019 e 2024*. 2024. Notícia do Portal TCU. Disponível em: <<https://portal.tcu.gov.br/>>. Acesso em: 03 dez. 2025.
- 3 SHORTLIFFE, E.; CIMINO, J. *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*. 5. ed. [S.l.]: Springer, 2021.
- 4 SAÚDE, M. da. *Suporte à Decisão Clínica*. 2024. Disponível em: <<https://www.gov.br/saude>>. Acesso em: 03 dez. 2025.
- 5 ZHUGE, M. et al. Agent-as-a-judge: Evaluate agents with agents. *arXiv preprint arXiv:2410.10934*, 2024. Disponível em: <<https://doi.org/10.48550/arXiv.2410.10934>>.