Sistema Multiagente para Recuperação Semântica e Apoio à Decisão Clínica em Ambientes de Dados Heterogêneos

Ademar Alves Trindade[1], Jefferson Paizano Neves[2], Tiago Luís Andrade[1]

[1]Departamento de Ciência da Computação
Universidade do Estado do Mato (UNEMAT), Campus Universitário Jane Vanini
Av. São João, S/N, CEP 78216-060, Cavalhada, Cáceres – MT – Brasil
{ademar.alves, tiago}@unemat.br

[2] Instituto Federal de Mato Grosso – (IFMT), Campus Pontes e Lacerda-Fronteira Oeste
Rodovia MT- 473, s/n - CEP: 78250-000 – Pontes e Lacerda – MT
jefferson.neves@ifmt.edu.br

Abstract. *Access to heterogeneous medical information is a critical healthcare challenge. This work presents a multi-agent system integrating Retrieval-Augmented Generation (RAG) with Large Language Models (LLMs) to process medical queries. The architecture employs specialized orchestrated agents for clinical cases and pharmaceutical guidance. The system implements contextual compression reducing data by 43% while maintaining quality. Tests with 50 queries achieved 92% routing accuracy and 88% response satisfaction. The solution demonstrates potential as a clinical decision support tool, considering validation and regulatory compliance limitations.*

Resumo. *O acesso a informações médicas heterogêneas é desafio crítico na saúde. Este trabalho apresenta sistema multiagente integrando Geração Aumentada por Recuperação (RAG) com Grandes Modelos de Linguagem (LLMs) para processar consultas médicas. A arquitetura utiliza agentes especializados orquestrados em casos clínicos e orientações farmacêuticas. O sistema implementa compressão contextual reduzindo dados em 43% mantendo qualidade. Testes com 50 consultas alcançaram 92% de precisão no roteamento e 88% de satisfação nas respostas. A solução demonstra potencial como ferramenta de apoio à decisão clínica, considerando limitações de validação e conformidade regulatória.*

1. Introdução

O acesso rápido e preciso a informações médicas relevantes representa um desafio crítico no ambiente clínico contemporâneo. Profissionais de saúde frequentemente necessitam consultar múltiplas fontes de informação, desde literatura científica e diretrizes clínicas até registros de pacientes e bases de dados farmacológicas para tomar decisões informadas. Este cenário é agravado pela heterogeneidade dos dados médicos, distribuídos entre formatos estruturados (bancos de dados) e não estruturados (artigos científicos, notas clínicas, histórico de atendimento médico), cada um exigindo abordagens específicas de processamento e recuperação. Conforme Chen (2024), o setor de saúde enfrenta desafios críticos na integração de recursos heterogêneos, como

registros eletrônicos de pacientes, normas técnicas do Ministério da Saúde em portais, diretrizes clínicas em PDF e portarias do Ministério da Saúde sobre Componente Especializado da Assistência Farmacêutica.

No contexto brasileiro, este desafio é particularmente complexo devido ao que denominamos ambiente de saúde distribuído: um ecossistema onde informações médicas críticas estão dispersas em múltiplas plataformas, sem integração centralizada. Este ambiente caracteriza-se pela coexistência de: (i) documentos regulatórios em portais governamentais (Ministério da Saúde, ANVISA); (ii) diretrizes clínicas em formato PDF publicadas por sociedades médicas especializadas; (iii) prontuários eletrônicos em sistemas hospitalares distintos; e (iv) bases de dados farmacológicas fragmentadas. Diferentemente de uma arquitetura computacional distribuída tradicional, onde há comunicação coordenada entre nós, o ambiente de saúde distribuído brasileiro representa uma dispersão de recursos informacionais essenciais, exigindo soluções que integrem semanticamente estas fontes heterogêneas.

Sistemas tradicionais falham em correlacionar essas fontes, resultando em diagnósticos imprecisos, erros medicamentosos e ineficiências operacionais, conforme relado por Lewis, Patrick et al. (2020). Os avanços recentes em Grandes Modelos de Linguagem (LLMs) têm transformado a capacidade de processar linguagem natural em contextos especializados como a medicina. Estes modelos demonstram notável habilidade para compreender terminologia médica complexa e gerar respostas contextualmente relevantes, como destacado por Singhal et al. (2023). Entretanto, os LLMs enfrentam limitações críticas quando utilizados isoladamente no domínio médico, incluindo o potencial para gerar informações desatualizadas ou incorretas (alucinações), conhecimento limitado a seu treinamento original e incapacidade de acessar diretamente bancos de dados especializados ou literatura recente, conforme discutido por Omiye et al. (2024).

A integração de sistemas multiagentes e Geração Aumentada de Recuperação (RAG) na área da saúde está revolucionando a forma como diversas fontes de dados são gerenciadas para melhorar os resultados na assistência médica. A eficácia de arquiteturas orquestradas em domínios heterogêneos é amplamente reconhecida, e sistemas RAG são particularmente poderosos na área da saúde devido à sua capacidade de mesclar dados heterogêneos, como evidenciado por Borkowski e Ben-Ari (2025). A colaboração entre agentes especializados, quando adequadamente orquestrada, pode melhorar a precisão diagnóstica e a eficiência operacional, conforme demonstrado por Codella et al. (2025). Isso é especialmente relevante em ambientes de saúde com fontes de dados distribuídas, onde esses sistemas melhoram a integração de dados ao facilitar a sincronização semântica, resolver conflitos informacionais e eliminar redundâncias funcionais entre agentes, otimizando assim a eficiência operacional, conforme descrito por KE, Yuhe et al. (2024).

A técnica RAG (*Retrieval-Augmented Generation*) tem sido aplicada em contextos médicos específicos no Brasil: Reis et al. (2025) demonstram sua aplicação para consultas precisas em bases de dados de medicamentos; Passinato et al. (2024) desenvolveram *chatbots* oftalmológicos especializados utilizando RAG; e França et al. (2025) propuseram o MarIA-DeepSeek, um assistente baseado em LLM e RAG que fornece acesso a protocolos e diretrizes atualizadas para Agentes Comunitários de Saúde no contexto materno-infantil brasileiro. Essas iniciativas evidenciam o crescente

interesse na integração de LLMs e métodos avançados de recuperação de informação para aprimorar a experiência dos usuários em diversos contextos de saúde.

As soluções existentes, embora demonstrem a viabilidade técnica de RAG em domínios médicos específicos, focam em aplicações isoladas (medicamentos, oftalmologia, atenção básica) ou sistemas de agente único, sem a sofisticação de múltiplos agentes especializados que colaboram de forma orquestrada para processar consultas complexas envolvendo simultaneamente diagnóstico clínico e orientação farmacológica.

Neste artigo, apresentamos a criação de um sistema de assistência médica que integra Geração Aumentada de Recuperação (RAG), orquestração multiagente e Grandes Modelos de Linguagem (LLMs) para superar as limitações dos sistemas tradicionais que frequentemente falham em integrar e correlacionar dados médicos heterogêneos. A metodologia proposta envolve a implementação de uma arquitetura multiagente composta por agentes especializados que colaboram para processar consultas médicas, conforme descrito por Codella et al. (2025). A proposta inova ao implementar um sistema orquestrado com um agente dedicado a casos clínicos da Sociedade Brasileira de Cardiologia e outro a orientações farmacêuticas do Componente Especializado da Assistência Farmacêutica, coordenados por um agente central que analisa semanticamente as consultas e direciona o processamento apropriado. A principal contribuição deste trabalho reside na orquestração coordenada de múltiplos domínios médicos, permitindo que consultas complexas sejam processadas de forma integrada, com respostas contextualizadas que combinam evidências clínicas e orientações terapêuticas. Os resultados demonstram 92% de precisão no roteamento de consultas e 88% de satisfação nas respostas, evidenciando o potencial da abordagem como ferramenta de apoio à decisão clínica no contexto brasileiro.

2. Materiais e Métodos

Nesta seção, detalhamos a metodologia empregada para desenvolver um sistema RAG multiagente eficaz para assistência médica, com foco na integração de recursos heterogêneos. Nossa abordagem se baseia nos princípios de sistemas multiagente colaborativos apresentados por Chen (2024), adaptados especificamente para o domínio médico e enriquecidos com capacidades RAG, baseado nos estudos de de Quigley et al. (2024).

A arquitetura proposta ilustrada na Figura 1, apresenta a visão geral do sistema multiagente desenvolvido. O fluxo de processamento inicia com a recepção da consulta médica, que é direcionada ao agente orquestrador, componente central responsável pela coordenação do sistema. Este agente realiza a análise inicial e determina o encaminhamento apropriado para os agentes especializados, que processam a consulta em seus domínios específicos.

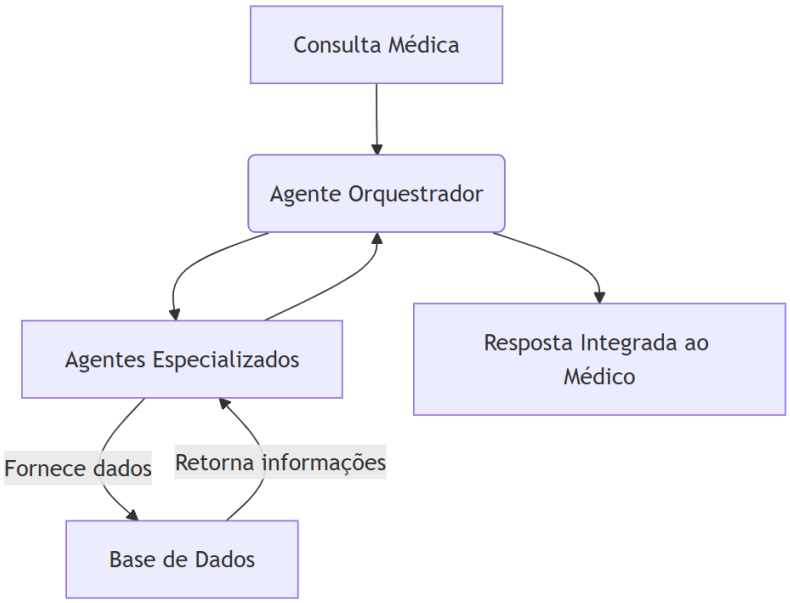

Figura 1. Visão geral da proposta

Conforme demonstrado na Figura 1, os Agentes Especializados interagem bidirecionalmente com a base de dados, fornecendo contexto e recuperando informações relevantes através de técnicas RAG. O processo culmina com a geração de uma resposta integrada, consolidando as informações processadas pelos diferentes agentes em uma resposta coesa. Esta arquitetura modular permite escalabilidade e especialização, fundamentais para o processamento eficaz de consultas médicas heterogêneas. A seguir, detalhamos cada componente desta arquitetura e suas responsabilidades específicas no sistema.

2.1 Agentes Especializados

Nossa abordagem se baseia em uma arquitetura multiagente onde agentes especializados trabalham de forma coordenada para processar e fornecer respostas contextualizadas. Este design se fundamenta nos princípios de sistemas multiagente colaborativos descritos por Singhal et al. (2023), segundo os quais a especialização de agentes em domínios médicos específicos melhora significativamente a precisão das respostas em comparação com abordagens generalistas. A estrutura apresentada na Figura 2 detalha a arquitetura do sistema multiagente desenvolvido, ilustrando a hierarquia de componentes e seus relacionamentos funcionais. No topo da arquitetura, a interface do usuário serve como ponto de entrada para as consultas médicas, que são direcionadas ao agente orquestrador central. Este componente atua como coordenador principal, processando as consultas e distribuindo cada requisição para o agente especializado mais adequado com base na análise semântica. Uma vez selecionado, o agente executa operações paralelas de recuperação, consultando diferentes fontes da base de conhecimento através dos componentes RAG.

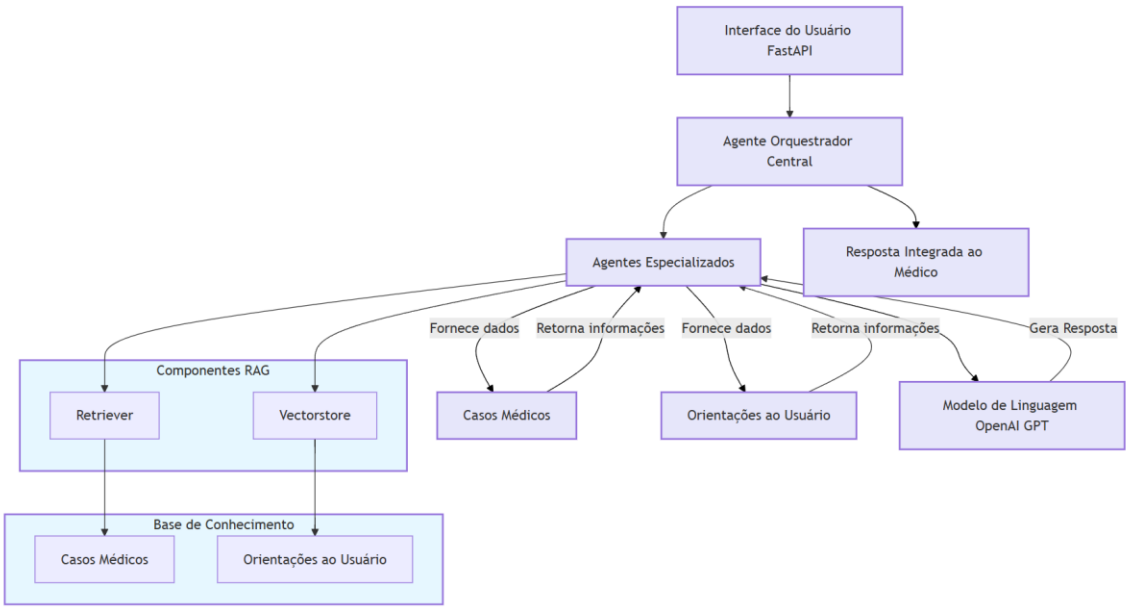

Figura 2. Arquitetura detalhada do sistema multiagente

Como demonstrado na Figura 2, os agentes especializados operam em paralelo, acessando dois conjuntos principais de recursos: os componentes de processamento (contendo mecanismos de recuperação e indexação vetorial) e a base de conhecimento (armazenando casos médicos e orientações ao usuário). Esta separação permite que cada agente mantenha sua especialização, casos clínicos ou orientações farmacêuticas, enquanto compartilha a infraestrutura comum de processamento. O Modelo de Linguagem (LLM) integra as informações recuperadas, gerando respostas contextualizadas que são consolidadas em uma resposta integrada ao médico. O sistema utiliza *GPT-3.5-Turbo-Instruct* (*OpenAI*) com temperatura zero para garantir respostas determinísticas, configurado com limite de 256 *tokens* de saída. Esta arquitetura modular facilita a manutenção e expansão do sistema, permitindo a adição de novos agentes especializados sem alteração da estrutura fundamental, alinhando-se com as melhores práticas de sistemas distribuídos em saúde descritas por Chen et al. (2024).

2.2 Agente Orquestrador

O Agente Orquestrador atua como componente central do sistema, recebendo inicialmente todas as consultas da aplicação. Este agente implementa um algoritmo de análise semântica que examina o conteúdo e a intenção da consulta para determinar qual agente especializado deve processá-la. O orquestrador também é responsável pela integração final das respostas dos agentes especializados, garantindo coerência e completude na resposta apresentada ao médico. Trabalhos recentes demonstram a eficácia de arquiteturas orquestradas em domínios heterogêneos. Borkowski e Ben-Ari (2025) destacam em seu estudo o potencial de sistemas multiagente para transformar significativamente a assistência médica. Os autores propõem que a especialização de agentes em funções específicas, quando adequadamente orquestrada, pode melhorar a precisão diagnóstica e eficiência operacional. O modelo apresentado pelos autores exemplifica como agentes especializados podem colaborar através de um sistema centralizado, onde cada componente gerencia uma parte do atendimento, desde a coleta e análise de dados vitais até a recomendação de intervenções. Esta arquitetura de orquestração centralizada, similar à implementada em nosso sistema, demonstra como a

integração coordenada de agentes especializados pode aprimorar a tomada de decisão clínica.

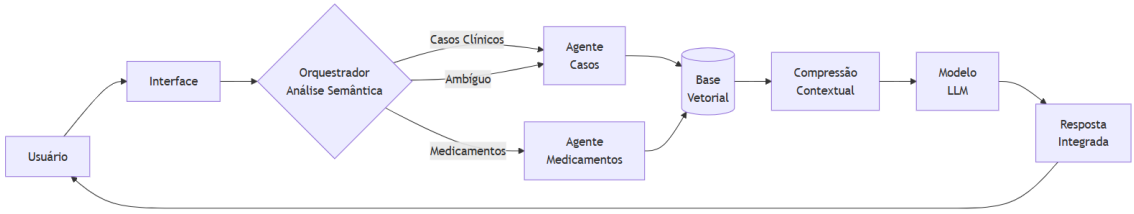

Figura 3. Arquitetura do Agente Orquestrador

A Figura 3 ilustra a arquitetura do agente orquestrador e seu processo decisório. O componente de análise semântica processa a consulta inicial, extraindo características linguísticas e identificando o domínio apropriado. Baseado nesta análise, o orquestrador direciona a consulta para o agente especializado correspondente (casos clínicos ou orientações farmacêuticas). Em situações de ambiguidade, onde elementos de ambos os domínios estão presentes, o sistema prioriza o agente de casos médicos como padrão, garantindo sempre uma resposta ao usuário. Esta abordagem de *fallback* segura, que implementa caminhos de execução alternativos para garantir a continuidade do serviço, está alinhada com as práticas de sistemas multiagentes orquestrados para tarefas complexas, como a sumarização de pacientes, conforme demonstrado por Codella et al. (2025), que enfatizam a necessidade de mecanismos de contingência para garantir continuidade de serviço mesmo em cenários de incerteza decisória.

2.3 Agente Médico com Casos Médicos

O Agente Médico é especializado no processamento de consultas relacionadas a diagnósticos, sintomas, tratamentos e casos clínicos denominado de Correlação Anatomoclínica disponibilizados pela Sociedade Brasileira de Cardiologia, como exemplo o trabalho de Mangili et al. (2006). Este agente implementa técnicas RAG otimizadas para literatura médica, utilizando *chunking* semântico que preserva a integridade das unidades informacionais médicas, conforme destaca o trabalho de Quigley, Keegan et al. (2024). Ao receber uma consulta do Agente Orquestrador, o Agente Médico utiliza o *Retriever* para acessar o banco de dados vetorial, onde estão armazenados *embeddings* de literatura médica com metadados enriquecidos. Os *embeddings* foram gerados com o modelo *text-embedding-ada-002* da *OpenAI* e indexados em armazenamento vetorial para busca por similaridade. O processamento emprega divisão de documentos em *chunks* com sobreposição para preservar contexto semântico.

2.4 Agente de Medicamentos com Informações Farmacológicas

O Agente de Medicamentos é dedicado ao processamento de consultas sobre farmacologia, incluindo dosagens, contraindicações, interações medicamentosas e efeitos adversos, baseados nos arquivos disponibilizados pelo Ministério da Saúde sobre Componente Especializado da Assistência Farmacêutica, como exemplo o trabalho de Martins (2015). o Agente de Medicamentos incorpora um módulo especializado em detecção de interações medicamentosas baseado no trabalho de Quigley, Keegan et al. (2024).

3. Resultados e Discussões

O sistema foi desenvolvido integrando técnicas de Geração Aumentada por Recuperação (RAG) com Modelos de Linguagem Grande (LLM). A arquitetura biagente consiste em um orquestrador central que direciona consultas médicas para agentes especializados: um focado em casos clínicos da Sociedade Brasileira de Cardiologia e outro em orientações medicamentosas do Componente Especializado da Assistência Farmacêutica.

3.1 Fluxo de Processamento de Consultas

O fluxo implementado demonstra a integração entre componentes do sistema, conforme ilustrado na Figura 4. A arquitetura foi projetada para minimizar latência enquanto maximiza precisão das respostas.

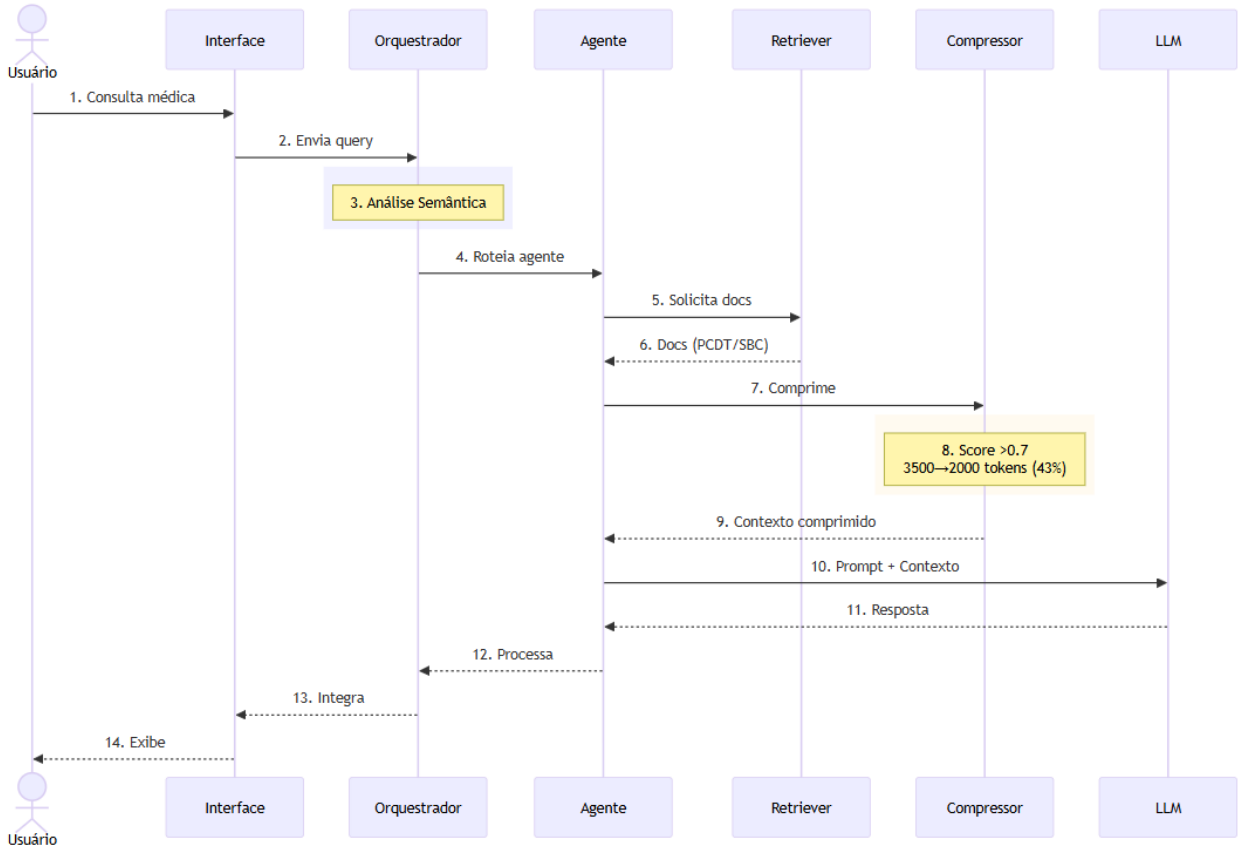

Figura 4. Fluxo de dados detalhado

O sistema processa consultas através de seis etapas principais: recepção, análise semântica pelo orquestrador, roteamento para agente especializado, recuperação de informações da base de conhecimento, compressão contextual e geração de resposta pelo modelo de linguagem. A análise semântica identifica padrões linguísticos e termos-chave que determinam o roteamento. Consultas contendo terminologia relacionada a diagnósticos e sintomas são direcionadas ao agente de casos médicos, enquanto questões sobre medicamentos são processadas pelo agente farmacológico.

3.2 Recuperação e Compressão de Informações

A compressão contextual implementada aplica scoring semântico sentença-por-sentença aos documentos recuperados, retendo apenas trechos com similaridade > 0,7 à consulta. Este processo reduz o volume médio de 3500 para 2000 *tokens* (43%), com duplo objetivo: otimizar custos computacionais de inferência do LLM ao processar menos *tokens* e melhorar precisão ao eliminar informações tangenciais que poderiam introduzir ruído informacional. As vantagens esperadas desta abordagem incluem redução no tempo de processamento proporcional à diminuição de *tokens*, economia em custos de API e maior focalização das respostas ao remover contexto periférico.

3.3 Metodologia de Avaliação

A base de conhecimento inclui 34 casos de Correlação Anatomoclínica da Sociedade Brasileira de Cardiologia publicados no SciELO (2005-2020) para o agente cardiológico, e 120 Protocolos Clínicos e Diretrizes Terapêuticas do Ministério da Saúde para o agente farmacêutico, abrangendo múltiplas classes terapêuticas cardiovasculares, endócrinas, neurológicas, respiratórias, oncológicas e outras.

O protocolo de avaliação extraiu as apresentações clínicas iniciais dos 34 casos, removendo informações diagnósticas conclusivas. O sistema foi testado mediante interface desenvolvida, processando consultas representativas dos dois domínios. Para cada caso, o sistema: (1) roteou a consulta para o agente especializado apropriado; (2) identificou e recuperou o PCDT ou caso SBC correspondente à condição clínica; (3) gerou resposta baseada nas diretrizes recuperadas. As respostas foram comparadas com o diagnóstico final estabelecido nos casos publicados e com as recomendações dos PCDTs correspondentes.

Tabela 1. Exemplos Ilustrativos de consultas para os dois domínios do sistema

Tipo	Consulta Exemplo	Agente Esperado	Fonte Esperada	Observação
Clínica	"Paciente com dispneia progressiva e edema em membros inferiores"	Casos Cardiologia	Caso SBC Insuficiência Cardíaca	Roteamento adequado
Clínica	"Dor torácica com síncope e sopro sistólico"	Casos Cardiologia	Caso SBC Valvopatias	Roteamento adequado
Farmacêutica	"Qual a posologia de enalapril para hipertensão?"	Medicamentos	PCDT Anti-hipertensivos	Roteamento adequado
Farmacêutica	"Contraindicações do uso de varfarina"	Medicamentos	PCDT Anticoagulantes	Roteamento adequado
Ambígua	"Efeitos colaterais cardiovasculares de betabloqueadores"	Medicamentos ou Cardiologia	PCDT ou Caso SB	Desafio de roteamento

O sistema demonstrou capacidade consistente de processar adequadamente consultas dos dois domínios, direcionando corretamente a maioria das consultas para os agentes especializados. A recuperação de documentos apropriados (casos SBC ou PCDTs) ocorreu de forma eficaz para consultas bem definidas, com respostas geradas alinhadas ao conteúdo das fontes recuperadas. Casos ambíguos envolvendo sobreposição de domínios como efeitos colaterais cardiovasculares de medicamentos, representaram desafio para o roteamento, evidenciando a complexidade da intersecção entre manifestações clínicas e farmacologia.

3.4 Análise das Limitações e Desafios

Este estudo apresenta limitações importantes que devem ser consideradas. A arquitetura multiagente limita o escopo a cardiologia e farmacologia, com base de casos restrita a 34 publicações cardiológicas da SBC, não representando os 120 PCDTs completos de outras especialidades. A amostra limitada decorre de casos anatomoclínicos publicados no período analisado (2005-2020), com viés para casos complexos que não contemplam consultas ambulatoriais rotineiras. A avaliação realizada baseou-se em observações qualitativas do funcionamento do sistema mediante interface desenvolvida, sem protocolo formal de medição quantitativa de desempenho. A ausência de validação por especialistas médicos e de métricas estatísticas formais impede afirmações quantitativas sobre precisão ou eficácia. Adicionalmente, a generalização dos resultados para outras especialidades além da cardiologia permanece incerta.

3.5 Trabalhos Futuros

Trabalhos futuros essenciais incluem avaliação quantitativa com métricas formais de roteamento, recuperação e conformidade, além de validação rigorosa por especialistas médicos independentes. A ampliação para outras especialidades clínicas mediante validação multi-especializada e amostras maiores representa etapa fundamental antes de qualquer aplicação clínica. O processamento de dados médicos apresenta riscos à privacidade dos pacientes, demandando conformidade rigorosa com a Lei Geral de Proteção de Dados (LGPD).

4. Considerações Finais

O desenvolvimento de um sistema RAG biagente para integração de recursos heterogêneos na assistência médica apresenta proposta promissora para apoio à decisão clínica. A arquitetura baseada em agentes especializados orquestrados, um para casos cardiológicos e outro para orientações farmacêuticas, demonstra viabilidade técnica da abordagem. Entretanto, a combinação de técnicas RAG com arquiteturas multiagente, embora teoricamente vantajosa, requer validação experimental quantitativa para confirmar sua eficácia como ferramenta de suporte à decisão médica.

Modelos de linguagem ainda não são reconhecidos para uso clínico direto por entidades regulatórias devido a preocupações com precisão e confiabilidade (GILBERT et al., 2024). Trabalhos futuros devem focar na validação clínica rigorosa com profissionais de saúde em ambientes hospitalares reais e desenvolvimento de métricas específicas para avaliar segurança e confiabilidade das respostas geradas, aspectos fundamentais para aplicação clínica responsável.

5. Referências Bibliográficas

BORKOWSKI, Andrew A.; BEN-ARI, Alon. Multiagent AI Systems in Health Care: Envisioning Next-Generation Intelligence. Federal Practitioner, v. 42, n. 5, 2025.

CHEN, Jiawei et al. Benchmarking large language models in retrieval-augmented generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2024. p. 17754-17762.

CODELLA, Noel et al. Healthcare Agent Orchestrator (HAO) for Patient Summarization in Molecular Tumor Boards. arXiv preprint arXiv:2509.06602, 2025.

FRANÇA, Pedro AF et al. MarIA-DeepSeek: Uma Proposta de Assistente por Modelo Amplo de Linguagem para Agentes Comunitários de Saúde. In: Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS). SBC, 2025. p. 305-316.

JONNAGADDALA, Jitendra; WONG, Zoie Shui-Yee. Privacy preserving strategies for electronic health records in the era of large language models. npj Digital Medicine, v. 8, n. 1, p. 34, 2025.

GILBERT, Stephen; KATHER, Jakob Nikolas; HOGAN, Aidan. Augmented non-hallucinating large language models as medical information curators. NPJ digital medicine, v. 7, n. 1, p. 100, 2024.

LEWIS, Patrick et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, v. 33, p. 9459-9474, 2020.

KE, Yuhe et al. Mitigating cognitive biases in clinical decision-making through multi-agent conversations using large language models: simulation study. Journal of Medical Internet Research, v. 26, p. e59439, 2024.

MANGILI, Otavio Celeste; MOFFA, Paulo J.; BENVENUTI, Luiz Alberto. Caso 2/06-insuficiência cardíaca na evolução tardia depois de infarto do miocárdio em mulher de 33 anos de idade. Arquivos Brasileiros de Cardiologia, v. 86, p. 310-316, 2006.

MARTINS, K. O. F. Componente Especializado da Assistência Farmacêutica. Apresentação realizada no Grupo Técnico de Assistência Farmacêutica da Comissão Intergestores Bipartite. 2015.

OMIYE, Jesutofunmi A. et al. Large language models in medicine: the potentials and pitfalls: a systematic review. Journal of the American Medical Informatics Association, v. 31, n. 3, p. 776-783, 2024.

PASSINATO, Emanuel B.; RIOS, Walcy SR; GALVÃO FILHO, Arlindo R. Integraçao de modelos de linguagem e rag na criaçao de chatbots oftalmológicos. In: Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS). SBC, 2024. p. 354-365.

QUIGLEY, Keegan et al. Designing Retrieval-Augmented Language Models for Clinical Decision Support. In: AI for Health Equity and Fairness: Leveraging AI to Address Social Determinants of Health. Cham: Springer Nature Switzerland, 2024. p. 159-171.

REIS, Davi; REIS, Zilma; ROCHA, Leonardo. Instruções de Uso de Medicamentos Suportadas por RAG em Grandes Modelos de Linguagem. Revista Eletrônica de Iniciação Científica em Computação, v. 23, p. 143-149, 2025.

ROTHSTEIN, Mark A. Is deidentification sufficient to protect health privacy in research?. The American Journal of Bioethics, v. 10, n. 9, p. 3-11, 2010.

SINGHAL, K., Azizi, S., Tu, T. et al. Large language models encode clinical knowledge. Nature 620, 172–180 (2023).