Arquitetura Multiagente de Apoio ao Diagnóstico Diferencial

1st Etelvina Costa Santos Sá Oliveira
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
etelvina.oliveira2003@gmail.com

2nd Wagner Meira Junior
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
meira@dcc.ufmg.br

***Abstract*—O diagnóstico diferencial constitui uma das etapas mais importantes da prática clínica, exigindo a integração de grandes volumes de informações e o raciocínio sobre múltiplas hipóteses. Embora os Grandes Modelos de Linguagem demonstrem potencial na síntese de dados médicos, desafios como alucinações e a falta de rastreabilidade limitam sua aplicação segura. Nesse contexto, o presente trabalho apresenta uma arquitetura multiagente de apoio ao diagnóstico diferencial baseada no paradigma de IA Agêntica e em Geração Aumentada por Recuperação. O sistema decompõe o raciocínio clínico em um fluxo de trabalho colaborativo composto por três agentes especializados: um Analista de Sintomas, responsável pela estruturação dos dados do paciente, um Minerador de Evidências, que realiza buscas iterativas em bases de conhecimento médico curadas (Textbooks e StatPearls) utilizando o recuperador MedCPT, e um Sintetizador de Conduta, que consolida as informações e gera hipóteses ordenadas com justificativas clínicas e referências. A arquitetura foi avaliada utilizando o dataset DDXPlus, testando quatro diferentes LLMs como motores de inferência. Os resultados indicam que o modelo Qwen-9B obteve o melhor desempenho, atingindo 41% de acurácia *top-1* e 60% em *top-5*. A pesquisa demonstra que a estrutura multiagente provê uma camada de explicabilidade essencial para o contexto médico, transformando a saída dos modelos em um fluxo de raciocínio justificável, pautado em evidências verificáveis e adequado para o auxílio à tomada de decisão clínica.**

***Index Terms*—Agentes de IA, Diagnóstico Diferencial, LLMs, RAG, IA Agêntica, Raciocínio Clínico.**

I. Introduction

A evolução da Inteligência Artificial (IA) tem promovido uma transformação significativa na forma como sistemas computacionais são projetados e utilizados, marcando a transição de modelos estáticos de processamento de dados para sistemas autônomos capazes de perceber o ambiente, tomar decisões e executar ações orientadas a objetivos. Nesse contexto, agentes de IA são definidos como entidades computacionais capazes de perceber informações do ambiente, raciocinar sobre essas informações e agir de forma autônoma para alcançar metas previamente estabelecidas [11]. Esse paradigma, fundamentado na teoria de agentes racionais, tornou-se um dos modelos conceituais centrais no desenvolvimento de sistemas inteligentes e adaptativos.

Nos últimos anos, especialmente a partir de 2022, o conceito de agentes de IA ganhou nova relevância com a popularização dos LLMs (*Large Language Models* - LLMs). Esses modelos ampliaram significativamente as capacidades de sistemas baseados em linguagem natural, permitindo não apenas a geração de texto, mas também a realização de tarefas complexas que envolvem raciocínio sequencial, planejamento e tomada de decisão assistida por ferramentas externas. A integração de LLMs com arquiteturas baseadas em agentes possibilitou o desenvolvimento de sistemas capazes de utilizar APIs, executar chamadas de funções, consultar bases de conhecimento externas e realizar recuperação de informações em tempo real, ampliando significativamente o escopo de aplicações da IA.

Mais recentemente, a partir do final de 2023, esses avanços levaram ao surgimento do paradigma conhecido como IA Agêntica(*Agentic AI*). Esse paradigma caracteriza-se pela utilização de múltiplos agentes especializados que colaboram entre si para resolver problemas complexos. Em arquiteturas multiagentes, diferentes agentes podem assumir papéis específicos dentro de um fluxo de trabalho, decompondo tarefas, compartilhando informações e coordenando ações de maneira cooperativa [12]. Essa abordagem permite a construção de sistemas distribuídos capazes de lidar com objetivos complexos e dinâmicos, representando um avanço significativo rumo a sistemas computacionais mais autônomos, adaptativos e colaborativos.

A prática clínica envolve um processo iterativo e interativo de raciocínio sobre um diagnóstico diferencial (*Differential Diagnosis* - DDx), ponderando múltiplas possibilidades diagnósticas à luz de quantidades crescentes de informações clínicas ao longo do tempo [16]. A criticidade do DDx reside na ambiguidade semântica e na sobreposição de sinais e sintomas. Nesse cenário, LLMs destacam-se por sua capacidade de compreender, sintetizar e integrar informações clínicas complexas, contribuindo para a geração estruturada de diagnósticos diferenciais. [1].

No contexto da área da saúde, essas capacidades tornam-se particularmente relevantes. Embora sistemas de IA tenham demonstrado grande eficiência no reconhecimento de padrões em grandes volumes de dados médicos, ainda existem desafios importantes relacionados ao raciocínio clínico, à integração de múltiplas fontes de evidência e ao suporte à tomada de decisões médicas [13]. Nesse cenário, agentes de IA surgem como uma abordagem promissora para aux-

iliar profissionais de saúde na interpretação de informações clínicas complexas e na identificação de possíveis diagnósticos e tratamentos. Esses sistemas podem ser projetados para analisar dados clínicos, consultar bases de conhecimento médico e gerar recomendações alinhadas a diretrizes clínicas e evidências científicas, respeitando restrições de segurança, regulamentação e governança [14].

Nesse cenário, técnicas de Geração Aumentada por Recuperação (*Retrieval-Augmented Generation* - RAG) surgem como uma forma de viabilizar essa autonomia, permitindo que o sistema consulte repositórios de literatura médica e prontuários eletrônicos em tempo real para fundamentar suas recomendações. Ao integrar a recuperação de dados externos especializados, o RAG auxilia na limitação das bases de treinamento estáticas dos modelos de linguagem, garantindo que as respostas sejam pautadas em evidências científicas atualizadas e específicas para o histórico de cada paciente [17]. Dessa forma, a convergência entre a autonomia dos sistemas agênticos e a robustez informacional do RAG torna-se a base estrutural deste trabalho.

O objetivo deste trabalho é investigar o uso de arquiteturas multiagentes no suporte ao raciocínio clínico, com ênfase na geração de diagnósticos diferenciais. Será desenvolvida uma arquitetura baseada em RAG com agentes especializados responsáveis pela análise de sintomas, recuperação de evidências e raciocínio clínico.

A importância deste projeto reside em seu potencial de:

- **Otimizar a eficiência clínica**, minimizando o tempo necessário na busca, triagem e análise de informações médicas complexas;
- **Elevar a acurácia diagnóstica**, auxiliando na formulação de hipóteses diferenciais fundamentadas na síntese e convergência de múltiplas fontes de evidência;

II. Objetivos

A. Objetivo Geral

Desenvolver e avaliar uma arquitetura multiagente de apoio ao diagnóstico diferencial, com base em evidências científicas e diretrizes clínicas.

B. Objetivos Específicos

- Implementar a extração de informações clínicas, por meio de um agente responsável por identificar sintomas e histórico do paciente.
- Desenvolver um sistema de recuperação de evidências científicas, integrando um agente capaz de consultar bases de dados médicas, manuais clínicos e diretrizes médicas, a fim de sustentar ou refutar hipóteses diagnósticas.
- Implementar um mecanismo de síntese clínica, no qual um agente consolida as informações obtidas e organiza as hipóteses diagnósticas por probabilidade
- Avaliar o desempenho do sistema, analisando a acurácia das hipóteses diagnósticas

III. Referencial Teórico

A. Modelos de Linguagem no Domínio Biomédico

O aumento no volume de dados biomédicos disponíveis na literatura, junto à evolução de técnicas de aprendizado profundo aplicadas ao processamento de linguagem natural, possibilitou o desenvolvimento e a popularização de modelos capazes de atuar em tarefas como recuperação de conhecimento, apoio à decisão clínica e sumarização de achados relevantes [3], [4]. Nesse contexto, observa-se a emergência de modelos de linguagem especializados no domínio clínico, desenvolvidos a partir da adaptação e do treinamento de LLMs em grandes conjuntos de dados biomédicos e registros de saúde. As vantagens associadas ao uso de LLMs decorrem, sobretudo, de sua capacidade de processar grandes volumes de dados heterogêneos, fornecer informações contextualizadas e oferecer suporte à tomada de decisão. Entretanto, a adoção desses modelos no domínio da saúde exige uma análise crítica de suas limitações e implicações éticas. A tendência dos LLMs de produzirem conteúdo convincente mas factualmente impreciso, alucinações, representando um risco considerável em contextos de alta responsabilidade, como o diagnóstico clínico. A natureza estatística e preditiva desses modelos, aliada à sua limitada interpretabilidade, impõe desafios adicionais à validação de sua confiabilidade e à rastreabilidade de suas decisões [7].

B. LLMs como Suporte ao Diagnóstico Diferencial

O diagnóstico diferencial constitui uma das etapas mais importantes do raciocínio clínico, fazendo-se necessário a integração de múltiplas fontes de informação, a consideração simultânea de múltiplas hipóteses e a atualização iterativa das suspeitas diagnósticas à medida que novos dados são obtidos. Nesse contexto, Natarajan et al. [16] introduzem o *Articulate Medical Intelligence Explorer* (AMIE), um modelo de linguagem otimizado especificamente para raciocínio diagnóstico. Em estudo clínico controlado com 20 médicos avaliando 302 casos reais extraídos de relatos de caso publicados, o AMIE demonstrou desempenho superior ao dos clínicos sem assistência, atingindo 59,1% de acurácia no *top-10* frente a 33,6% dos médicos não assistidos. Clínicos assistidos pelo AMIE alcançaram 51,7% de acurácia, em contraste com 36,1% dos clínicos sem assistência e 44,4% daqueles apoiados apenas por ferramentas de busca tradicionais. Esses resultados demonstram que LLMs podem servir como ferramentas de apoio ao diagnóstico de forma significativa e motivam a investigação de arquiteturas que estendam esse raciocínio para além da geração textual centralizada.

C. Arquiteturas Multiagente para Diagnóstico

Apesar do potencial demonstrado pelo AMIE, sua arquitetura centralizada e a ausência de mecanismos explícitos de recuperação de conhecimento externo impõem limitações à transparência e à adaptabilidade do processo diagnóstico. Abordagens mais recentes têm explorado o uso de sistemas multiagente baseados em LLMs como forma de superar essas restrições e aproximar o diagnóstico automatizado do

raciocínio clínico real. Rose et al. [19] propõem o *MEDDxAgent*, um framework para diagnóstico diferencial interativo. A arquitetura integra três componentes principais: um orquestrador central denominado DDxDriver, um simulador de anamnese e dois agentes especializados, para recuperação de conhecimento e de estratégia diagnóstica. Uma contribuição central do trabalho é o rompimento com a premissa de perfis clínicos completos desde o início da consulta: o sistema coleta informações progressivamente e refina as hipóteses diagnósticas a cada iteração, de forma análoga à prática clínica real. Avaliado em um *benchmark* abrangendo doenças respiratórias, dermatológicas e raras, o MEDDxAgent alcança melhorias superiores a 10 pontos percentuais de acurácia frente a abordagens não iterativas, tanto em modelos de grande e pequeno porte. A ênfase em explicabilidade e na rastreabilidade do processo decisório aborda diretamente os requisitos de confiabilidade exigidos em aplicações clínicas.

D. Sistemas Agênticos para Condições de Alta Complexidade

Em cenários de maior complexidade diagnóstica, como o das doenças raras, as limitações dos sistemas baseados em LLMs isolados ou em arquiteturas multiagente tornam-se ainda mais evidentes. A baixa prevalência dessas condições, combinada à heterogeneidade dos dados clínicos e genéticos envolvidos, exige sistemas capazes de integrar fontes de conhecimento especializadas e garantir rastreabilidade das inferências. Zhao et al. [9] apresentam o DeepRare, um sistema multiagente que utiliza o *Model Context Protocol* (MCP) para coordenar mais de 40 ferramentas especializadas e fontes de conhecimento atualizadas. O sistema processa entradas clínicas heterogêneas, incluindo descrições em linguagem livre, termos estruturados da *Human Phenotype Ontology* (HPO) e resultados de testes genéticos, para gerar hipóteses diagnósticas ranqueadas com raciocínio transparente vinculado a evidências médicas verificáveis. Os resultados obtidos são expressivos: em tarefas baseadas em HPO, o DeepRare atinge média de Recall@1 de 57,18%, superando o segundo melhor método em 23,79 pontos percentuais; em testes multimodais, alcança 69,1% frente a 55,9% do Exomiser. A revisão especializada das cadeias de raciocínio geradas pelo sistema obteve 95,4% de concordância, confirmando sua validade e rastreabilidade.

E. Posicionamento da Proposta

A análise de trabalhos recentes evidencia uma evolução das abordagens baseadas em IA aplicada ao diagnóstico clínico. Inicialmente, destacam-se modelos fundamentados em LLMs utilizados de forma isolada para o raciocínio diagnóstico, como o AMIE. Em seguida, surgem arquiteturas que empregam múltiplos agentes especializados, a exemplo do MEDDxAgent, capazes de dividir tarefas e aprimorar o processo de tomada de decisão. Mais recentemente, abordagens como o DeepRare incorporam conhecimento estruturado para lidar com casos de maior complexidade, especialmente no contexto de doenças raras.

Nesse contexto, a presente proposta investiga uma arquitetura multiagente em que a recuperação de evidências clínicas por meio de mecanismos de RAG desempenha papel central na construção, validação e refinamento das hipóteses diagnósticas. O objetivo é permitir que cada etapa do raciocínio seja fundamentada em conhecimento biomédico recuperado dinamicamente, promovendo maior transparência, justificabilidade e alinhamento com a literatura científica.

IV. Metodologia

O sistema proposto é implementado como um *pipeline* multiagente orquestrado pelo *framework* LangGraph [20], no qual três agentes especializados atuam como nós em um grafo de estados direcionado: o Analista de Sintomas, o Minerador de Evidências e o Sintetizador de Conduta. O fluxo de execução é iniciado no nó de entrada (`START`), onde o Analista de Sintomas processa a descrição textual do caso clínico. Em seguida, o Minerador de Evidências opera em um ciclo iterativo de busca e autoavaliação, refinando suas consultas dinamicamente até atingir um limiar de suficiência ou o número máximo de tentativas permitido. Por fim, o fluxo converge para o Sintetizador de Conduta, que consolida as informações e gera o relatório clínico final, encerrando o processo no nó de término (`END`). A arquitetura completa do sistema é apresentada na Figura 1.

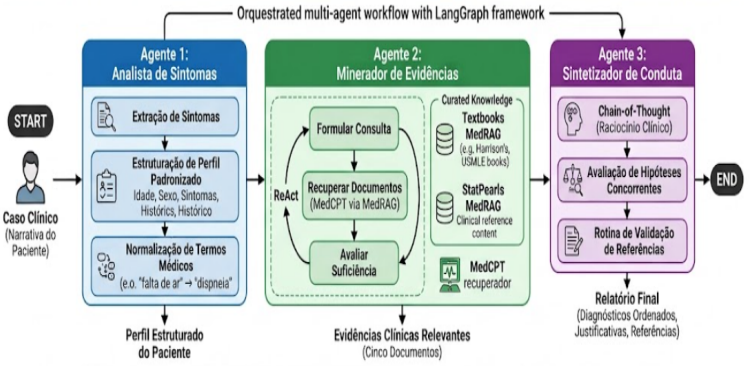

Fig. 1. Arquitetura do sistema multiagente proposto para suporte ao diagnóstico diferencial, ilustrando a decomposição do pipeline em módulos especializados e o fluxo sequencial de informações entre os agentes de análise clínica, recuperação de evidências e síntese diagnóstica.

A. Agente 1: Analista de Sintomas

O agente inicial do *pipeline* é responsável por extrair e estruturar os dados da narrativa clínica em linguagem natural. O resultado desse processamento é um perfil padronizado do paciente que compreende: idade, sexo, lista de sintomas, tempo de evolução e histórico clínico relevante. O *prompt* de instrução orienta o modelo a realizar a normalização de termos coloquiais para a terminologia médica formal (por exemplo, convertendo ”falta de ar” para ”dispneia”) e a decompor sintomas compostos em entidades atômicas, mitigando ambiguidades e assegurando a consistência para as etapas subsequentes.

B. Agente 2: Minerador de Evidências

O segundo agente realiza a recuperação de conhecimento clínico contextualizado, constituindo o componente de RAG do sistema. A busca ocorre por meio do *toolkit* MedRAG [21], um *framework* desenvolvido especificamente para padronizar e avaliar pipelines de recuperação de informação no domínio médico, integrando diferentes bases de conhecimento, recuperadores e modelos de geração sob uma interface unificada. A escolha do MedRAG se justifica por sua adoção consolidada como referência de avaliação em RAG médico, o que confere maior comparabilidade aos resultados obtidos neste trabalho frente a estudos correlatos da literatura.

O recuperador empregado é o MedCPT [22], um modelo de recuperação densa contrastiva treinado especificamente sobre dados biomédicos, a partir de centenas de milhões de pares de consulta-artigo extraídos de logs de busca do PubMed. Diferentemente de recuperadores densos de propósito geral, o MedCPT foi otimizado para capturar similaridade semântica em terminologia clínica e biomédica, sendo composto por dois sub-modelos especializados: um codificador de consultas e um codificador de artigos, ambos baseados em arquitetura *transformer*, que projetam consultas e documentos em um espaço vetorial denso comum, onde a relevância é estimada por similaridade do produto interno.

A recuperação é realizada sobre duas bases de dados médicos curados disponibilizados pelo MedRAG: *Textbooks*, compreendendo aproximadamente 125 mil fragmentos textuais extraídos de 18 livros-texto de referência amplamente utilizados na preparação para o exame USMLE, e *StatPearls*, um repositório de conteúdo clínico orientado à prática e a condutas no ponto de atendimento, frequentemente atualizado por profissionais da área. A combinação desses dois corpora busca equilibrar conhecimento médico fundamental e estruturado (Textbooks) com informação clínica mais aplicada e orientada à tomada de decisão (StatPearls).

A estratégia de busca fundamenta-se nos sintomas e no histórico previamente estruturados pelo Agente 1. Para cada *corpus*, os k documentos mais semanticamente semelhantes, segundo o escore de similaridade do MedCPT, são recuperados de forma independente. Dado que as escalas de similaridade podem oscilar em função do tamanho e do domínio de cada base de dados, os escores brutos são normalizados para o intervalo $[0, 1]$ localmente, dentro de cada *corpus*, antes da fusão. Posteriormente, as listas de candidatos são unidas em uma única lista ordenada pelos escores normalizados, das quais são selecionadas as cinco evidências de maior relevância global, consolidando em um único ranqueamento informação proveniente de fontes de naturezas distintas.

Para aprimorar a recuperação, integrou-se um mecanismo de autoavaliação baseado na arquitetura ReAct (*Reasoning and Acting*) [23]. Após cada iteração de busca, o próprio modelo de linguagem avalia se as evidências obtidas fornecem cobertura, especificidade e qualidade suficientes para embasar o diagnóstico diferencial. Caso a informação seja julgada insuficiente, o modelo formula uma nova consulta de busca aprimorada, em vez de repetir a consulta original. Esse ciclo iterativo repete-se consecutivamente até que o critério de suficiência seja satisfeito ou que o limite de tentativas seja atingido, prevenindo a ocorrência de laços infinitos.

C. Agente 3: Sintetizador de Conduta

O terceiro agente atua na consolidação do perfil estruturado do paciente, obtido pelo Agente 1, e das evidências clínicas recuperadas pelo Agente 2, produzindo um relatório estruturado de hipóteses diagnósticas. Antes de definir o ranking final das hipóteses diagnósticas, o agente é instruído a elaborar uma cadeia de raciocínio (*chain-of-thought*) estruturada, na qual avalia cada diagnóstico diferencial com base no quadro apresentado. Nesse processo, o agente analisa como as evidências recuperadas contribuem para sustentar ou refutar cada hipótese considerada, justificando a posição atribuída a cada uma na ordenação final. Ao exigir essa etapa de raciocínio explícito antes da geração da lista de diagnósticos, busca-se reduzir a tendência do modelo de convergir prematuramente para uma única hipótese, incentivando uma avaliação mais abrangente, crítica e sistemática das alternativas diagnósticas.

O relatório final é composto por hipóteses diagnósticas ordenadas da mais para a menos provável. Cada hipótese é acompanhada de uma justificativa clínica concisa, fundamentada nos sintomas e no histórico do paciente, e de referências explícitas às evidências que a sustentam, conferindo rastreabilidade ao processo de decisão. Uma rotina de validação verifica se os índices de evidência referenciados correspondem a posições válidas na lista recuperada, disparando uma nova tentativa de geração com instrução reforçada nos casos em que referências inválidas são detectadas.

V. Procedimento de Avaliação Experimental

A validação quantitativa da arquitetura proposta foi conduzida utilizando-se uma amostra estratificada de 500 casos selecionados aleatoriamente a partir do conjunto de teste do *dataset* DDXPlus [24]. Para cada cenário, o *pipeline* foi instanciado a partir de uma narrativa clínica em primeira pessoa, gerada a partir dos parâmetros clínicos do *dataset*, simulando o relato real de um paciente.

O DDXPlus é um conjuntos de dados públicos que permite avaliar sistemas de diagnóstico diferencial baseados em IA. Ele foi criado para refletir a complexidade de uma consulta médica real, cobrindo dezenas de patologias, comuns e raras, e associando cada caso a evidências possíveis, como sintomas e histórico clínico.

Com o intuito de avaliar a robustez da arquitetura proposta frente a diferentes modelos de linguagem clínicos, o experimento foi replicado utilizando-se quatro LLMs distintos para inferência dos três agentes do *pipeline*: Llama3-8B, Qwen-9B, MedGemma-4b e Med42-8b. A escolha desses modelos justifica-se pela diversidade de abordagens de especialização representadas: enquanto Llama3-8B e Qwen-9B servem como referência de um modelos de propósito geral sem ajuste fino específico para o domínio clínico, MedGemma e Med42

foram desenvolvidos ou adaptados especificamente para tarefas biomédicas, permitindo investigar em que medida a especialização do modelo de base impacta o desempenho do sistema multiagente como um todo.

As hipóteses diagnósticas produzidas pelo Sintetizador de Conduta foram extraídas e comparadas ao diagnóstico de referência. O desempenho do sistema foi avaliado por meio de duas métricas amplamente utilizadas em tarefas de ranqueamento: a taxa de acerto na primeira posição (`top-1 hit`) e a taxa de acerto entre as cinco primeiras posições (`top-5 hit`). Essas métricas permitem verificar a capacidade do sistema de identificar corretamente o diagnóstico verdadeiro e posicioná-lo entre as hipóteses diagnósticas mais relevantes apresentadas.

VI. Resultados

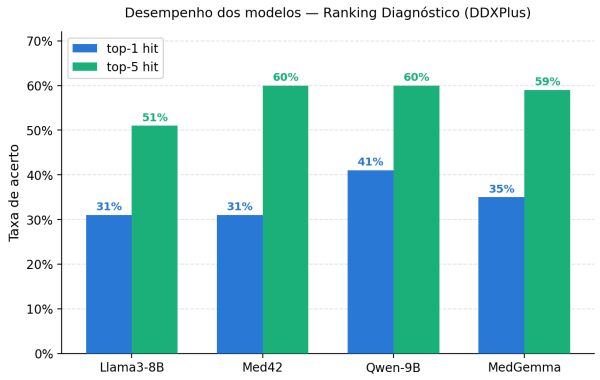

Fig. 2. Comparação entre modelos de linguagem clínicos quanto às métricas de acerto *top-1* e *top-5*.

A comparação entre as quatro configurações do pipeline agêntico, diferenciadas pelo modelo de linguagem base utilizado como motor de inferência, evidencia diferenças relevantes na orquestração do fluxo de trabalho diagnóstico. O pipeline baseado no Qwen-9B apresentou o melhor desempenho em acurácia *top-1* (41%), superando as demais implementações. Em seguida, o fluxo orquestrado pelo MedGemma atingiu 35%, enquanto as arquiteturas baseadas no Llama3-8B e no Med42 obtiveram 31% nessa métrica.

Na avaliação da métrica *top-5*, os pipelines baseados no Med42 e no Qwen-9B empataram no melhor desempenho (60%), seguidos pelo MedGemma (59%), enquanto o Llama3-8B apresentou o menor valor (51%).

Um resultado central para a validação da arquitetura agêntica é que o Qwen-9B, um modelo de propósito geral sem ajuste fino específico para o domínio biomédico, obteve a melhor acurácia *top-1*, além de igualar o melhor desempenho em *top-5*. Sob a perspectiva de sistemas agênticos, esse achado sugere que, para o perfil de casos do *dataset* DDXPlus, predominantemente composto por patologias comuns e bem documentadas, a eficácia do pipeline em coordenar etapas de raciocínio, extrair entidades clínicas e manter consistência no histórico de sintomas depende da capacidade geral de raciocínio estruturado do modelo base. Nesse contexto, o fluxo agêntico demonstra robustez ao compensar a ausência de especialização médica por meio de uma orquestração eficaz do processo de decisão.

Por fim, o desempenho consistentemente inferior do pipeline baseado no Llama3-8B estabelece uma linha de base (*baseline*) relevante para comparação. Sendo o único a permanecer abaixo de 31% em *top-1* e de 51% em *top-5*, esse resultado evidencia limitações na capacidade do modelo subjacente de executar adequadamente tarefas agênticas fundamentais, como o seguimento de instruções estruturadas e a transição consistente entre estados do grafo de decisão. O contraste entre o Qwen-9B e o Llama3-8B reforça que, embora a arquitetura agêntica defina a lógica do sistema, seu desempenho final permanece fortemente dependente da capacidade intrínseca do modelo em compreender e operar sobre o fluxo de raciocínio estruturado.

VII. Discussão

Os resultados experimentais evidenciam a interação entre arquiteturas agênticas e bases de conhecimento dinâmicas no contexto do raciocínio clínico assistido por LLMs. Embora o Qwen-9B tenha apresentado forte desempenho na orquestração estruturada das etapas do pipeline, os ganhos observados vão além da acurácia isolada e estão diretamente relacionados à forma como os modelos de linguagem são integrados ao sistema.

Nesse sentido, a incorporação de um pipeline de RAG baseado em MedRAG e MedCPT desempenha um papel central ao reduzir o risco de alucinações e ao ancorar as saídas do modelo em evidências verificáveis. Em aplicações clínicas, a utilidade de um sistema não se limita à previsão correta, mas depende também da rastreabilidade das informações utilizadas. Ao vincular hipóteses diagnósticas a fontes como textos médicos consolidados (Textbooks) e bases de referência como StatPearls, o sistema adiciona uma camada importante de confiabilidade e explicabilidade, essencial para contextos de decisão em saúde.

Além disso, o componente de ranqueamento produzido pelo Sintetizador de Conduta se alinha de forma mais natural ao processo de raciocínio clínico, que raramente é binário ou determinístico. Na prática médica, trabalha-se com hipóteses concorrentes ordenadas por plausibilidade e gravidade. A apresentação estruturada em top-k preserva essa característica, permitindo que o usuário considere diagnósticos menos prováveis, porém clinicamente relevantes, especialmente em cenários de maior risco.

Por fim, a avaliação também evidencia um desafio recorrente na aplicação de IA em medicina, a diferença entre a formalização rígida dos benchmarks e a flexibilidade da linguagem clínica real. O dataset DDXPlus utiliza rótulos padronizados, enquanto na prática existem múltiplas formas de nomear a mesma condição ou espectros clínicos relacionados. Assim, casos semanticamente equivalentes podem ser penalizados em métricas como exact match por diferenças terminológicas. Isso sugere que métricas baseadas apenas em correspondência textual podem subestimar o desempenho

clínico real dos sistemas, apontando para a necessidade de avaliações apoiadas em ontologias médicas, capazes de capturar equivalência semântica de forma mais adequada.

VIII. Conclusões

Este estudo investigou o impacto de uma arquitetura multiagente no aprimoramento da qualidade e da segurança do raciocínio clínico aplicado ao suporte ao diagnóstico diferencial. A decomposição estruturada do pipeline em três agentes com funções especializadas permitiu simular o processo iterativo característico da prática médica, transformando descrições clínicas não estruturadas em um percurso diagnóstico mais interpretável, transparente e alinhado às diretrizes da prática baseada em evidências.

Os experimentos de validação quantitativa realizados no dataset DDXPlus demonstraram um ganho significativo na acurácia diagnóstica do sistema. O modelo Qwen-9B destacou-se com 41% de acerto na primeira hipótese diagnóstica (top-1 hit), superando modelos intrinsecamente otimizados para o domínio biomédico, ao mesmo tempo em que manteve um desempenho robusto na inclusão do diagnóstico correto entre as cinco principais hipóteses (60% em top-5 hit), com desempenho competitivo em relação ao Med42.

Do ponto de vista clínico, esses resultados sugerem que o avanço de sistemas de suporte à decisão médica baseados em IA não depende apenas do uso de modelos especializados, mas também, de forma relevante, do desenho de arquiteturas de orquestração de agentes. Essa abordagem favorece a estruturação do raciocínio diagnóstico em etapas explícitas e rastreáveis, reduzindo ambiguidades e contribuindo para a mitigação de riscos clínicos ao ancorar hipóteses diferenciais em cadeias de justificativa fundamentadas em literatura médica.

IX. Trabalhos Futuros

Apesar dos resultados promissores, a arquitetura proposta abre caminhos para investigações adicionais que visam aumentar sua robustez e aplicabilidade clínica. Entre as direções futuras, destacam-se:

- Ampliação e Diversificação das Bases de Evidência: A implementação atual utiliza bases consolidadas, como *StatPearls* e livros-texto médicos. Como evolução natural do sistema, propõe-se a integração de fontes dinâmicas e de atualização contínua, incluindo o *PubMed* e repositórios de diretrizes clínicas recentes.
- Escalabilidade para Modelos de Larga Escala: Os experimentos realizados até o momento foram conduzidos com modelos de até 9 bilhões de parâmetros, priorizando eficiência computacional e viabilidade experimental. Como extensão, propõe-se a avaliação de modelos de maior escala, como Llama-3-70B ou GPT-4o, com o objetivo de investigar possíveis ganhos decorrentes da ampliação da capacidade paramétrica no contexto agêntico. Em particular, busca-se analisar se modelos mais robustos em raciocínio abstrato podem aprimorar a etapa do *Sintetizador de Conduta*, produzindo cadeias de raciocínio mais estruturadas (*Chain-of-Thought*) e identificando relações clínicas sutis entre sintomas e evidências que podem não ser capturadas por modelos de menor porte.

References

[1] B. Bhasuran, Q. Jin, Y. Xie, C. Yang, K. Hanna, J. Costa, C. Shavor, W. Han, Z. Lu, and Z. He, "Preliminary analysis of the impact of lab results on large language model generated differential diagnoses," *npj Digital Medicine*, vol. 8, no. 1, p. 166, 2025.

[2] H. Wan, J. Zhang, A. A. Suria, B. Yao, D. Wang, Y. Coady, and M. Prpa, "Building llm-based ai agents in social virtual reality," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–7.

[3] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[4] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[5] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," *arXiv preprint arXiv:1904.05342*, 2019.

[6] Y. Luo, J. Zhang, S. Fan, K. Yang, Y. Wu, M. Qiao, and Z. Nie, "Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine," *arXiv preprint arXiv:2308.09442*, 2023.

[7] J. Haltaufderheide and R. Ranisch, "The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs)," *NPJ Digital Medicine*, vol. 7, no. 1, p. 183, 2024.

[8] Y. H. Ke, L. Jin, K. Elangovan, H. R. Abdullah, N. Liu, A. T. H. Sia, C. R. Soh, J. Y. M. Tung, J. C. L. Ong, C.-F. Kuo, *et al.*, "Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness," *npj Digital Medicine*, vol. 8, no. 1, p. 187, 2025.

[9] W. Zhao, C. Wu, Y. Fan, P. Qiu, X. Zhang, Y. Sun, X. Zhou, S. Zhang, Y. Peng, Y. Wang, *et al.*, "An agentic system for rare disease diagnosis with traceable reasoning," *Nature*, pp. 1–10, 2026.

[10] H. Wang, S. Zhao, Z. Qiang, N. Xi, B. Qin, and T. Liu, "Beyond direct diagnosis: Llm-based multi-specialist agent consultation for automatic diagnosis," *arXiv preprint arXiv:2401.16107*, 2024.

[11] M. Wooldridge, *An Introduction to MultiAgent Systems*, 2nd ed. Chichester, UK: John Wiley & Sons, 2009.

[12] R. Sapkota, K. I. Roumeliotis, and M. Karkee, "Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges," *Information Fusion*, p. 103599, 2025.

[13] N. Karunanayake, "Next-generation agentic AI for transforming healthcare," *Informatics and Health*, vol. 2, no. 2, pp. 73–83, 2025.

[14] M. Finio and A. Downie, "AI Agents Healthcare," *Ibm.com*, Feb. 2026. [Online]. Available: https://www.ibm.com/think/topics/ai-agents-healthcare

[15] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," *NPJ Digital Medicine*, vol. 3, no. 1, p. 17, 2020.

[16] D. McDuff, M. Schaekermann, T. Tu, A. Palepu, A. Wang, J. Garrison, K. Singhal, Y. Sharma, S. Azizi, K. Kulkarni, *et al.*, "Towards accurate differential diagnosis with large language models," *Nature*, vol. 642, no. 8067, pp. 451–457, 2025.

[17] L. M. Amugongo, P. Mascheroni, S. Brooks, S. Doering, and J. Seidel, "Retrieval augmented generation for large language models in healthcare: A systematic review," *PLOS Digital Health*, vol. 4, no. 6, p. e0000877, 2025.

[18] Y. Shavit, S. Agarwal, M. Brundage, S. Adler, C. O'Keefe, R. Campbell, T. Lee, P. Mishkin, T. Eloundou, A. Hickey, *et al.*, "Practices for governing agentic AI systems," Research Paper, OpenAI, 2023.

[19] D. Rose, C.-C. Hung, M. Lepri, I. Alqassem, K. Gashteovski, and C. Lawrence, "MEDDxAgent: A unified modular agent framework for explainable automatic differential diagnosis," *arXiv preprint arXiv:2502.19175*, 2025.

[20] LangChain, "LangGraph: agent orchestration framework for reliable AI agents," *Langchain.com*, 2026. Disponível em: https://www.langchain.com/langgraph.

[21] G. Xiong, Q. Jin, Z. Lu, and A. Zhang, "Benchmarking retrieval-augmented generation for medicine," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 6233–6251.

[22] Q. Jin, W. Kim, Q. Chen, D. C. Comeau, L. Yeganova, W. J. Wilbur, and Z. Lu, "MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval," *Bioinformatics*, vol. 39, no. 11, p. btad651, 2023.

[23] YAO, Shunyu et al. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

[24] TCHANGO, Arsene Fansi et al. DDXPlus: A new dataset for automatic medical diagnosis. Advances in Neural Information Processing Systems (NeurIPS), v. 35, p. 31306–31318, 2022.