UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Desenvolvimento de um sistema integrado com modelos de linguagem de grande escala para auxílio no diagnóstico e prescrição em consultas médicas na especialidade de pneumologia

Tiago Bittencourt Espindula

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Tiago Bittencourt Espindula

Desenvolvimento de um sistema integrado com modelos de linguagem de grande escala para auxílio no diagnóstico e prescrição em consultas médicas na especialidade de pneumologia

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Dr. Brucce Neves dos Santos

Versão original

São Carlos

2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

E77d

Espindula, Tiago Bittencourt
Desenvolvimento de um sistema integrado com modelos de linguagem de grande escala para auxílio no diagnóstico e prescrição em consultas médicas na especialidade de pneumologia / Tiago Bittencourt Espindula; orientador Brucce Neves dos Santos. -- São Carlos, 2024.
52 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. INTELIGÊNCIA ARTIFICIAL. 2. PNEUMOLOGIA. 3. PROCESSAMENTO DE LINGUAGEM NATURAL. I. dos Santos, Brucce Neves , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Tiago Bittencourt Espindula

Development of an integrated system with large language models to aid in diagnosis and prescription in medical consultations in the specialty of pulmonology

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Brucce Neves dos Santos, PhD

Original version

São Carlos

2024

Dedico este trabalho à minha amada esposa, Bianca, que, não por coincidência, é também uma brilhante pneumologista. Sua resiliência e dedicação apaixonada em tudo o que faz me inspiram todos os dias. Seu amor e apoio incondicionais são fundamentais para mim. Minha querida, este trabalho é para você, com todo o meu amor e gratidão.

AGRADECIMENTOS

Primeiramente, agradeço a Deus, pela força e sabedoria que me guiam em cada etapa da minha vida.

À minha amada esposa Bianca, por sua inestímável ajuda em diversos detalhes deste trabalho. E também pelo seu incansável apoio, sempre me incentivando a perseguir meus objetivos. Sem você, nada disso teria sido possível.

Aos meus pais, Marco Valério e Lucinyr, pela base sólida de valores, pelo incentivo constante e pelo amor incondicional que sempre me sustentou. A vocês, dedico cada conquista.

Aos meus sogros, Nilo e Maria do Carmo, expresso minha sincera gratidão por todo o apoio e carinho que sempre me dedicaram. Agradeço por me receberem de braços abertos em sua família e por todo o incentivo durante essa jornada.

Ao meu orientador, Brucce, por sua paciência, orientação e pelos comentários e correções sempre enriquecedores. Sua dedicação e conhecimento foram fundamentais para que eu pudesse desenvolver esta pesquisa com qualidade e profundidade.

Aos professores do curso, que compartilharam seu conhecimento e experiência, moldando minha trajetória acadêmica e profissional.

Aos colegas de curso, pelo companheirismo, pelas discussões enriquecedoras e por todos os momentos de apoio mútuo. A caminhada foi mais leve com vocês ao meu lado.

Agradeço também à tecnologia e às ferramentas que, de maneira inovadora, têm o potencial de transformar a prática médica. Este trabalho é um pequeno passo na longa estrada da inovação na medicina, e é com esperança que contribuo para esse avanço.

A todos, os meus mais sinceros agradecimentos.

"A verdadeira revolução na medicina não está apenas na tecnologia que criamos, mas na capacidade de usá-la para devolver aos médicos o tempo e a conexão com seus pacientes."

ChatGPT 4o

RESUMO

ESPINDULA, T.B. **Desenvolvimento de um sistema integrado com modelos de linguagem de grande escala para auxílio no diagnóstico e prescrição em consultas médicas na especialidade de pneumologia**. 2024. 52 p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

Este trabalho de conclusão de curso aborda o desenvolvimento e a validação de um sistema baseado em grandes modelos de linguagem (LLMs), especificamente adaptado para a área de pneumologia, com o objetivo de otimizar o processo de diagnóstico e prescrição médica. A motivação central deste estudo é a necessidade de melhorar a eficiência e a qualidade do atendimento médico, reduzindo o tempo dedicado a tarefas administrativas e aumentando o tempo de interação direta entre médicos e pacientes. O sistema proposto combina LLMs com bancos de dados médicos específicos e diretrizes clínicas atualizadas, fornecendo diagnósticos precisos e condutas terapêuticas fundamentadas. O estudo abrange uma revisão teórica sobre o uso de LLMs na medicina, explora técnicas avançadas de processamento de linguagem natural e analisa trabalhos correlatos que demonstram a eficácia dessas tecnologias em contextos clínicos. A metodologia implementada utiliza um sistema de agentes LLM que, ao longo de múltiplas etapas, processa dados clínicos detalhados, acessa protocolos médicos e valida as decisões geradas. A avaliação experimental, realizada com base em 70 casos clínicos reais de pneumologia, destaca a segurança e a relevância das respostas geradas pelo sistema, embora alguns desafios ainda persistam, como a necessidade de aprimorar a precisão e a integralidade das condutas em casos complexos. Os resultados indicam que o sistema é capaz de fornecer recomendações seguras e baseadas em evidências na maioria dos casos, mas aponta a necessidade de melhorias em áreas específicas, como doenças menos prevalentes ou casos com múltiplas comorbidades. Conclui-se que, com ajustes adicionais, o sistema proposto tem o potencial de contribuir significativamente para a prática clínica em pneumologia, auxiliando médicos na tomada de decisões mais rápidas e precisas, e promovendo um atendimento de saúde mais eficiente e personalizado.

Palavras-chave: Inteligência Artificial na Medicina; Pneumologia; Grandes Modelos de Linguagem; Tomada de Decisão Clínica

ABSTRACT

ESPINDULA, T.B. **Development of an integrated system with large language models to aid in diagnosis and prescription in medical consultations in the specialty of pulmonology**. 2024. 52 p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

This thesis focuses on the development and validation of a system based on Large Language Models (LLMs), specifically adapted for the field of pulmonology, with the aim of optimizing the process of medical diagnosis and prescription. The primary motivation for this study is the need to enhance the efficiency and quality of medical care by reducing the time spent on administrative tasks and increasing the direct interaction between doctors and patients. The proposed system combines LLMs with specialized medical databases and updated clinical guidelines to provide accurate diagnoses and evidence-based therapeutic recommendations. The study includes a theoretical review of the application of LLMs in medicine, explores advanced natural language processing techniques, and analyzes related works that demonstrate the effectiveness of these technologies in clinical contexts. The implemented methodology utilizes a system of LLM agents that, through multiple stages, processes detailed clinical data, accesses medical protocols, and validates the generated decisions. The experimental evaluation, conducted on 70 real clinical cases in pulmonology, highlights the safety and relevance of the system's responses, though some challenges remain, such as the need to improve the accuracy and completeness of recommendations in complex cases. The results indicate that the system is capable of providing safe and evidence-based recommendations in most cases, but it also identifies areas for improvement in specific conditions, such as less prevalent diseases or cases with multiple comorbidities. It is concluded that, with additional refinements, the proposed system has the potential to significantly contribute to clinical practice in pulmonology, assisting physicians in making faster and more accurate decisions and promoting more efficient and personalized healthcare.

Keywords: Artificial Intelligence in Medicine, Large Language Models, Pulmonology, Medical Decision Support

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CNNs	Redes Neurais Convolucionais - *Convolutional Neural Networks*
DPI	Doença Pulmonar Intersticial
DPOC	Doença Pulmonar Obstrutiva Crônica
EHR	Registros de Saúde Eletrônicos - *Electronic Health Records*
IAM	Infarto Agudo do Miocárdio
IA	Inteligência Artificial
LLM	Grande Modelo de Linguagem - *Large Language Model*
LM	Modelo de Linguagem - *Language Model*
PLN	Processamento de Linguagem Natural
RAG	Geração Aumentada por Recuperação - *Retrieval-Augmented Generation*
RLHF	Aprendizado por Reforço a partir de Feedback Humano - *Reinforcement Learning from Human Feedback*
TEP	Tromboembolismo Pulmonar
USMLE	Exame de Licenciamento Médico dos Estados Unidos - *United States Medical Licensing Examination*

SUMÁRIO

1 INTRODUÇÃO

A medicina moderna enfrenta um desafio crescente no gerenciamento eficiente de informações clínicas e na prestação de cuidados de saúde de alta qualidade (Murdoch; Detsky, 2013). Paralelamente, o avanço tecnológico tem impulsionado o desenvolvimento de soluções baseadas em Inteligência Artificial (IA), em especial os Modelos de Linguagem de Grande Escala (*Large Language Models* - LLMs), que oferecem novas perspectivas para a otimização do processo de atendimento clínico. Contudo, a aplicação direta de LLMs em contextos clínicos é limitada devido à falta de clareza nos processos de treinamento e nos dados utilizados, o que é crucial em áreas de alto risco como a saúde (Gallifant *et al.*, 2024).

A motivação deste trabalho emerge da necessidade de melhorar a eficiência e a qualidade do atendimento médico em pneumologia. Estudos indicam que médicos gastam uma parcela significativa do seu tempo em tarefas administrativas e no manuseio de prontuários eletrônicos, o que reduz o tempo de interação direta com pacientes. Médicos gastam cerca de 27% do seu tempo em encontros diretos com paciente e cerca de 49% do tempo em prontuários eletrônicos e trabalho de escritório, ou seja, para cada 1 hora em contato com o paciente, 2 horas são dedicadas aos sistemas informatizados (Sinsky *et al.*, 2016).

A utilização de LLMs na prática médica, se bem direcionada, tem o potencial de agilizar a análise de dados clínicos e auxiliar na tomada de decisões, mas sua aplicação enfrenta barreiras significativas, principalmente no que se refere à precisão e segurança das informações geradas. Atualmente, os LLMs são treinados com conjuntos de dados amplos e não especializados, o que os torna inadequados para tarefas que exigem conhecimento médico detalhado e atualizado. Apesar de algumas versões de LLMs já terem sido otimizados com dados médicos, a ausência de integração com diretrizes médicas validadas e a falta de personalização para áreas específicas são desafios cruciais que precisam ser superados para garantir que tais sistemas sejam confiáveis e úteis na prática clínica.

1.1 Objetivos

A hipótese central deste trabalho é que um sistema integrado, que combine LLMs com bancos de dados médicos específicos da área de pneumologia e diretrizes clínicas atualizadas, pode otimizar significativamente o processo de diagnóstico e prescrição de condutas médicas. Espera-se que tal sistema poderá processar eficientemente informações clínicas detalhadas, sugerir diagnósticos e condutas baseados em evidências científicas e adaptar-se às peculiaridades e desafios da especialidade de pneumologia.

O objetivo geral deste projeto é desenvolver e validar um sistema inovador baseado em LLMs, especificamente adaptado para a pneumologia. Este sistema deverá automatizar e otimizar o processo de diagnóstico e prescrição, assegurando a precisão clínica e a conformidade com as práticas médicas atualizadas. Espera-se que a implementação bem-sucedida deste sistema melhore a eficiência e a qualidade do atendimento médico, beneficiando tanto profissionais de saúde quanto pacientes.

Este trabalho de conclusão de curso organiza-se em cinco capítulos. O presente capítulo, a introdução, apresenta o contexto, os objetivos e a estrutura da pesquisa. O Capítulo 2 abordará a fundamentação teórica e a análise de trabalhos correlatos. Já o terceiro detalhará a justificativa, a metodologia e a proposta de pesquisa. No quarto capítulo, descrever-se-á a configuração do experimento, bem como a análise e a discussão dos resultados alcançados. Por fim, o quinto capítulo sintetizará as considerações finais, destacando as principais descobertas, as contribuições do estudo e as perspectivas para futuras pesquisas.

2 FUNDAMENTOS E TRABALHOS RELACIONADOS

A inteligência artificial tem gerado impacto significativo na área médica, com o potencial de transformar o diagnóstico, tratamento e acompanhamento de doenças. Dentro do campo da IA, os grandes modelos de linguagem se destacam por sua capacidade de processar e analisar grandes volumes de texto não estruturados, abrindo caminho para novas aplicações na área da saúde. Este capítulo se dedica a explorar os fundamentos da aplicação de LLMs em medicina.

Na Seção 2.1, é apresentado um breve histórico do uso de IA na medicina, como também algumas aplicações da IA na medicina, incluindo uma breve discussão sobre ética e considerações legais, que discute aspectos como privacidade de dados, a responsabilidade por erros e omissões e o potencial de viés algorítmico. Na Seção 2.2, é descrita a teoria dos LLMs, como também técnicas que são utilizadas para otimizar as respostas fornecidas. Enquanto que na Seção 2.3, foi feita uma revisão da literatura científica sobre o uso de LLMs na medicina, destacando as principais aplicações e resultados obtidos até o momento. Por fim, na Seção 2.4 é feita ponderações sobre o potencial da IA na área médica, em particular no contexto da pneumologia, e as perspectivas futuras, desafios e oportunidades para o desenvolvimento e aplicação de LLMs nesse campo.

2.1 Inteligência artificial na medicina

A década de 1960 marca o início da aplicação prática da inteligência artificial na medicina com o surgimento dos sistemas especialistas, baseados em regras e conhecimentos médicos codificados para auxiliar na tomada de decisões clínicas (Raz; Nguyen; Loh, 2006). O marco inicial se deu em 1965 com o sistema DENDRAL (Buchanan; Feigenbaum, 1978), considerado o primeiro sistema especialista, que visava automatizar a identificação de moléculas orgânicas desconhecidas. Embora não estivesse diretamente ligado à prática médica, o sistema DENDRAL serviu como base para o desenvolvimento de sistemas especialistas direcionados à medicina, como o MYCIN (Shortliffe, 2012), Internist-1 (Miller; Jr; Myers, 1985) e CASNET (Weiss *et al.*, 1978). O MYCIN era capaz de auxiliar na identificação de doenças infecciosas e sugerir tratamento com antibióticos (Shortliffe, 2012), enquanto o Internist-1 era baseado em sinais e sintomas, resultados laboratoriais e o histórico do paciente para sugerir diagnósticos (Miller; Jr; Myers, 1985). O CASNET é um modelo computacional que auxilia diagnósticos e tratamentos médicos por meio de uma rede causal associativa de doenças, por meio de observações do paciente, estados patofisiológicos e classificações de doenças para sugerir tratamentos baseados em padrões identificados. O CASNET foi utilizado em um programa de consulta para o diagnóstico e tratamento de glaucoma (Weiss *et al.*, 1978).

Apesar de promissores, os sistemas especialistas em saúde não alcançaram ampla adoção por parte dos profissionais da área, devido a várias dificuldades: bases de conhecimento incompletas e de difícil atualização, técnicas de modelagem inadequadas para a complexa natureza do conhecimento médico e a pouca praticidade dos sistemas (Duda; Shortliffe, 1983).

Mesmo com grandes expectativas na época, o desenvolvimento da IA logo encontrou obstáculos: a falta de poder computacional e de armazenamento, o que explica em parte a desaceleração da evolução da IA no período aproximadamente entre 1980 e 2000, conhecido como "Inverno da IA" (Anyoha, 2017).

Porém, mesmo nesse período houveram evoluções significativas para a IA na medicina. Desenvolvido em 1986 pela Universidade de Massachusetts nos EUA e ainda ativo até os dias atuais, o DXplain é um sistema de suporte à decisão diagnóstica. Por meio da entrada de sintomas do paciente, o programa gera um diagnóstico diferencial, auxiliando o médico na formulação de hipóteses clínicas. Além disso, funciona como um livro eletrônico de medicina, oferecendo descrições detalhadas de doenças e referências bibliográficas complementares. No momento de seu lançamento, o DXplain possuía informações sobre aproximadamente 500 enfermidades. Desde então, o sistema evoluiu consideravelmente, abrangendo atualmente mais de 2.400 doenças (Barnett *et al.*, 1987).

A partir da década de 2010, graças a avanços em poder computacional e em algoritmos, o aprendizado de máquina profundo (do inglês *deep learning*) se tornou o modelo dominante e permitiu diversas aplicações da IA na medicina (Russell; Norvig, 2016), sendo discutido na seção 2.1.1. Apesar das grandes vantagens de se ter modelos de IA na medicina, existem diversas questões éticas que precisam ser levadas em consideração, essas questões serão abordadas na seção 2.1.2.

2.1.1 Aplicações da IA na medicina

O advento de algoritmos avançados e o aumento do poder de processamento de dados em larga escala têm possibilitado avanços significativos nas aplicações da IA em medicina. Especialidades como radiologia, dermatologia, cardiologia, pneumologia, oncologia, neurologia, patologia, oftalmologia, endocrinologia e reumatologia, entre outras, têm visto o surgimento de aplicações específicas com grandes potenciais. Esta seção se dedica a explorar algumas aplicações da IA na medicina, que estão abrindo novos horizontes para uma era de inovações médicas sem precedentes.

Provavelmente a especialidade que mais tem utilizado e que mais se beneficia destes avanços na IA é a radiologia, abrangendo aquisição, reconstrução, análise e relatório de imagens diagnósticas. A IA acelera a aquisição de imagens com técnicas avançadas, reduzindo o tempo e mantendo a qualidade, além de preencher lacunas entre *hardware* de aquisição de imagens e *software* de reconstrução por meio da supressão de artefatos e

otimização da qualidade. Na reconstrução de imagens, a IA contribui para a diminuição do uso de contraste, otimização de doses de radiação e melhoria no tempo de reconstrução. Com mais de 200 *softwares* aprovados disponíveis, a IA potencializa a radiologia em inúmeras aplicações clínicas, desde detecção de hemorragia intracerebral até avaliação de risco de demências, demonstrando-se mais produtiva e rápida em conjunto com radiologistas. Contudo, ressalta-se que, predominam avanços em IA "fraca", especializada em tarefas singulares, e justifica-se uma visão realista da capacidade da tecnologia atual, promovendo um radiologista "aumentado" pela IA, ao invés de substituído por ela (Boeken *et al.*, 2023).

Na cardiologia, o aprendizado de máquina pode auxiliar na identificação de pacientes de alto risco de mortalidade após um infarto agudo do miocárdio (IAM) (Wang *et al.*, 2021); ser utilizado em dados de dispositivos vestíveis (*wearables*) para melhorar a eficiência da reabilitação cardíaca (Sotirakos *et al.*, 2021); e automatizar o diagnóstico de sons cardíacos (Chen *et al.*, 2021), entre outras aplicações.

Existem amplas aplicações da IA na pneumologia, em especial no auxílio na análise de imagens, tomada de decisões e previsão prognóstica. No diagnóstico e rastreio do câncer de pulmão, por exemplo, algoritmos de visão computacional, especificamente Redes Neurais Convolucionais (*Convolutional Neunal Netowrks* - CNNs), têm demonstrado precisão comparável ou até superior à de radiologistas humanos na detecção de nódulos e na avaliação do risco de malignidade usando tomografias computadorizadas de baixa dosagem (Ardila *et al.*, 2019). Algoritmos de aprendizado de máquina têm sido aplicados na classificação automatizada de doenças pulmonares fibrosantes com base em imagens de alta resolução e no desenvolvimento de modelos preditivos para hipertensão pulmonar, asma e infecções pleurais através de análise de padrões em dados proteômicos e genômicos. Estudos mostram que algoritmos de IA podem melhorar a precisão diagnóstica e otimizar a tomada de decisões clínicas, embora existam desafios relacionados à validação externa dos modelos e à interpretação dos resultados gerados por esses sistemas complexos (Khemasuwan; Sorensen; Colt, 2020).

À medida que exploramos as aplicações da inteligência artificial nas diversas especialidades médicas, vislumbramos um futuro onde a IA auxilia e também transforma a prática clínica. No entanto, essa integração promissora entre tecnologia e saúde levanta questões significativas relacionadas à ética e à legislação, que devem ser cuidadosamente consideradas. A próxima seção abordará essas preocupações, delineando os princípios éticos e a legislação que devem guiar o uso da IA na medicina para garantir que sua implementação beneficie todos os envolvidos, respeitando a privacidade, a autonomia do paciente e promovendo a justiça e a equidade no acesso aos avanços tecnológicos.

2.1.2 Ética e considerações legais

À medida que a IA ganha espaço na medicina, a interseção entre tecnologia avançada e cuidados de saúde coloca em evidência uma série de desafios éticos e práticos. Emergem preocupações significativas relacionadas à privacidade, segurança e ética no manuseio de dados de saúde dos pacientes. O uso extensivo de informações pessoais para treinar sistemas de IA levanta questões sobre a privacidade e o potencial de uso indevido desses dados, destacando a importância do consentimento informado e da transparência para promover a confiança dos pacientes (Murphy *et al.*, 2021).

A adoção de sistemas de IA traz à tona a complexa questão da responsabilidade pelas informações fornecidas. Este desafio transcende a esfera puramente tecnológica e envolve múltiplas facetas da sociedade. A responsabilidade não recai apenas sobre o profissional de saúde que utiliza o sistema, mas também sobre os demais agentes envolvidos em seu desenvolvimento e implementação. Assim, espera-se que a conscientização de todos os participantes no ecossistema da IA na medicina contribua para uma melhor distribuição dos deveres, equilibrando a tendência à sobre-responsabilização ou à desresponsabilização dos médicos, priorizando o direito dos pacientes à explicação sobre os tratamentos recebidos (Verdicchio; Perin, 2022).

Outro grande problema enfrentado é a tendência de sistemas de IA em saúde reproduzirem ou até exacerbarem desigualdades existentes no acesso a diagnósticos e tratamentos entre diferentes grupos socioeconômicos, demográficos ou geográficos, influenciando negativamente os resultados de saúde como doenças, incapacidades ou mortalidade, o que é chamado de viés algorítmico. Apesar do potencial desses sistemas em melhorar o acesso à saúde para populações carentes e aumentar a qualidade do atendimento reduzindo custos, a utilização de IA, em alguns casos, pode intensificar as disparidades de saúde. Por exemplo, um sistema de IA projetado para identificar pacientes com necessidades de saúde complexas acabou priorizando predições de custos de saúde ao invés das necessidades reais dos pacientes, levando a uma alocação inadequada de cuidados. Isso resultou em pacientes negros mais doentes recebendo o mesmo nível de atendimento que pacientes brancos menos doentes, evidenciando como o viés algorítmico pode influenciar negativamente a equidade na saúde (Abràmoff *et al.*, 2023).

Para assegurar soluções de saúde impulsionadas por IA justas e equitativas, é vital a utilização de dados diversos e representativos, a implementação de auditorias regulares, a validação de algoritmos por especialistas independentes e a educação de profissionais de saúde e pacientes sobre vieses inerentes à IA. É crucial fortalecer a privacidade e segurança de dados, estabelecer *frameworks* de responsabilidade, melhorar a transparência e explicabilidade da IA, e promover a colaboração entre médicos, pesquisadores de IA e desenvolvedores (Ueda *et al.*, 2024). Estes desafios requerem regulamentação adequada para garantir um equilíbrio entre o uso da IA e a expertise humana (Meskó; Topol, 2023).

A Organização Mundial da Saúde publicou um documento (World Health Organization, 2024) com recomendações sobre ética e governança em IA, direcionadas aos governos, sugerindo investimentos em infraestrutura pública para acesso de desenvolvedores, com o objetivo de garantir que os LLMs cumpram obrigações éticas e de direitos humanos. Também sugere a criação de agências reguladoras para aprovação de LLMs na área da saúde e auditorias pós-lançamento.

2.2 Modelos de linguagem de grande escala

Os modelos de linguagem (*Language Models* - LM) representam um avanço significativo na área de processamento de linguagem natural (PLN). Esses modelos são treinados em vastos conjuntos de dados e têm demonstrado uma capacidade notável de compreender, gerar e traduzir texto em linguagem humana (Min *et al.*, 2023). A evolução dos LLMs, desde os modelos estatísticos iniciais até os atuais baseados em redes neurais profundas, tem sido impulsionada pela crescente disponibilidade de grandes corpora de textos e pelo desenvolvimento de tecnologias de aprendizado de máquina mais sofisticadas, como a arquitetura *Transformer* (Zhao *et al.*, 2023).

Recentemente, os LMs têm demonstrado habilidades especiais quando escalados para tamanhos significativos, com o aumento de escala resultando em melhor desempenho e eficiência em diversas tarefas de PLN. Inclusive com o surgimento de habilidades emergentes, definidas como "habilidades não presentes em modelos de menor escala, que não podem ser previstas simplesmente extrapolando as melhorias de desempenho em modelos de menor escala" (Wei *et al.*, 2022a, p. 2). Devido ao seu tamanho, esses LMs receberam o nome de LLMs (Naveed *et al.*, 2023). Para explorar todo o potencial dos LLMs, é necessário a utilização de técnicas avançadas, sendo elas engenharia de *prompt*, geração aumentada por recuperação (*Retrieval-Augmented Generation* - RAG), *instruction fine-tuning*, e o aprendizado por reforço a partir de *feedback* humano (*Reinforcement Learning from Human Feedback* - RLHF).

Um *prompt* no contexto de LLMs é uma entrada de texto fornecida pelo usuário que direciona o modelo a gerar uma resposta específica ou realizar uma tarefa. Tanto a sua sintaxe (como comprimento e ordem dos exemplos) quanto a semântica (escolha de palavras e exemplos) influenciam significativamente a saída do modelo. Essa influência é comparável à forma como pequenas mudanças em uma consulta podem alterar os resultados de uma busca em um banco de dados. A técnica que busca direcionar as respostas do modelo para resultados desejados por meio de consultas em linguagem natural é usualmente chamada de engenharia de *prompts* (*prompt engineering*), e é uma área que requer muita experimentação, pois carece de uma base teórica sólida para entender por que certas formulações são mais eficazes (Kaddour *et al.*, 2023).

As técnicas mais simples envolvem o uso de *role-prompting*, atribuindo ao modelo

um papel específico a desempenhar, como o de um assistente prestativo ou um especialista conhecedor; e *one-shot* ou *few-shot prompting*, nos quais é fornecido ao modelo um ou mais exemplos de como a resposta deve ser (Chen *et al.*, 2023). Dentre as técnicas mais avançadas, a primeira a ser proposta foi a *chain-of-thought* que consiste em induzir o modelo a realizar uma série de passos intermediários de raciocínio, em língua natural, que levam ao resultado final. Essa indução pode ser feita por meio da introdução de um termo como "vamos pensar passo-a-passo" no *prompt* (*Zero-shot chain-of-thought*) ou por meio de exemplos de como o modelo deve raciocinar (*chain-of-thought*) (Wei *et al.*, 2022b). A seguir alguns exemplos de cada uma dessas técnicas:

Role-prompting: "Você é um especialista em pneumologia, com vasto conhecimento teórico na área. Quais são os sintomas da pneumonia?"

One-shot ou few-shot prompting: "P: a altura do paciente é 1,70m e o peso 70kg, qual seu IMC? R: O IMC é de 24,2kg/m2 P: a altura do paciente é 1,80m e o peso 95kg, qual seu IMC?"

Zero-shot chain-of-thought: "Qual o diagnóstico provável para um paciente com febre e cefaleia? Vamos pensar passo-a-passo."

Chain-of-thought: "P: Qual o diagnóstico provável para um paciente com tosse, febre e odinofagia? R: Um paciente com tosse pode ter várias doenças: asma, embolia pulmonar, pneumonia, gripe. Considerando que o paciente também tem febre, os diagnósticos mais prováveis seriam pneumonia e gripe. Considerando ainda a odinofagia, o diagnóstico mais provável seria a gripe, pois leva em conta todos os sintomas. P: Qual o diagnóstico provável para um paciente com febre e cefaleia?"

Existem ainda as técnicas de múltiplos turnos, que realizam chamadas iterativas ao modelo e processam as saídas para obter uma resposta final. Entre elas, destacam-se a *self-consistency*, que amostra múltiplos caminhos de raciocínio e seleciona a resposta mais consistente por meio de um voto majoritário (Figura 1). Como também *least-to-most*, que utiliza um conjunto de *prompts* constantes para fazer com que o LLM decomponha o problema em uma série de subproblemas, e o modelo então resolve sequencialmente os subproblemas, construindo iterativamente a saída final (Figura 2). Outra técnica é a *scratchpad*, na qual o modelo gera etapas intermediárias de raciocínio em um "rascunho" antes de produzir o resultado final. Por fim, a técnica *tree-of-thoughts* gera uma "árvore de pensamentos", com múltiplos caminhos diferentes, onde cada pensamento é uma sequência de linguagem que serve como etapa intermediária (Figura 3) (Chen *et al.*, 2023).

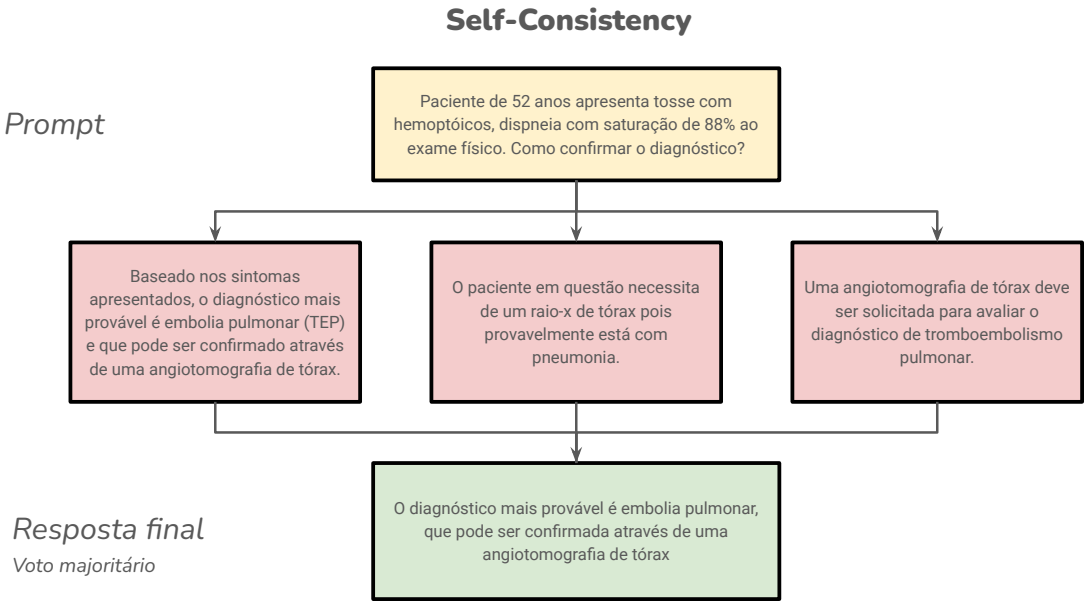

Figura 1 – Exemplo de *Self-Consistency*.

Fonte: Adaptado de: Chen *et al.* (2023).

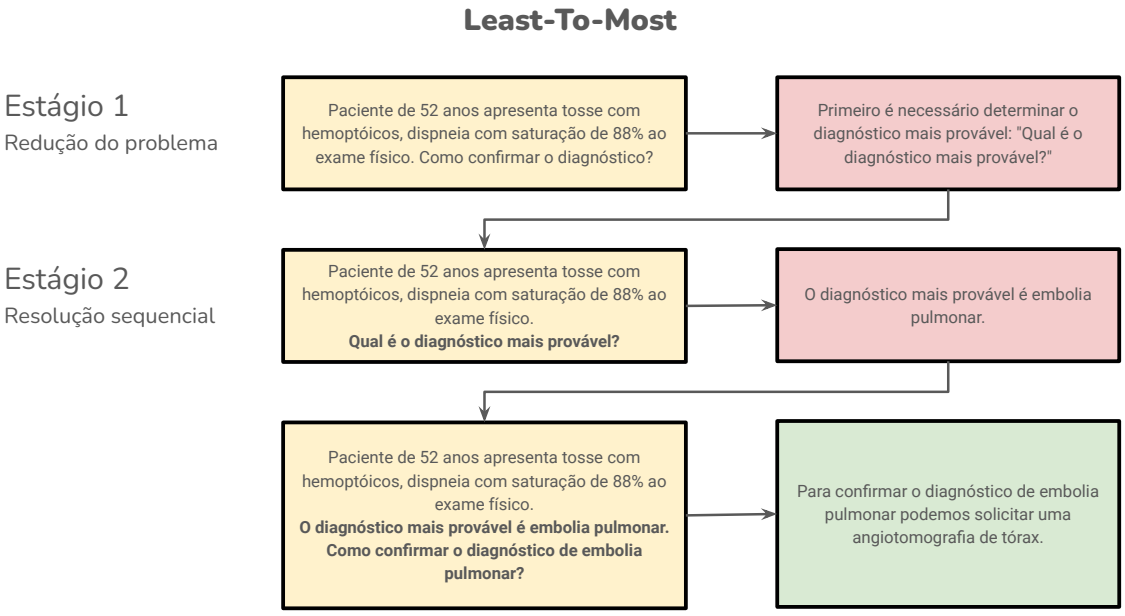

Figura 2 – Exemplo de *Least-To-Most*.

Fonte: Adaptado de: Chen *et al.* (2023).

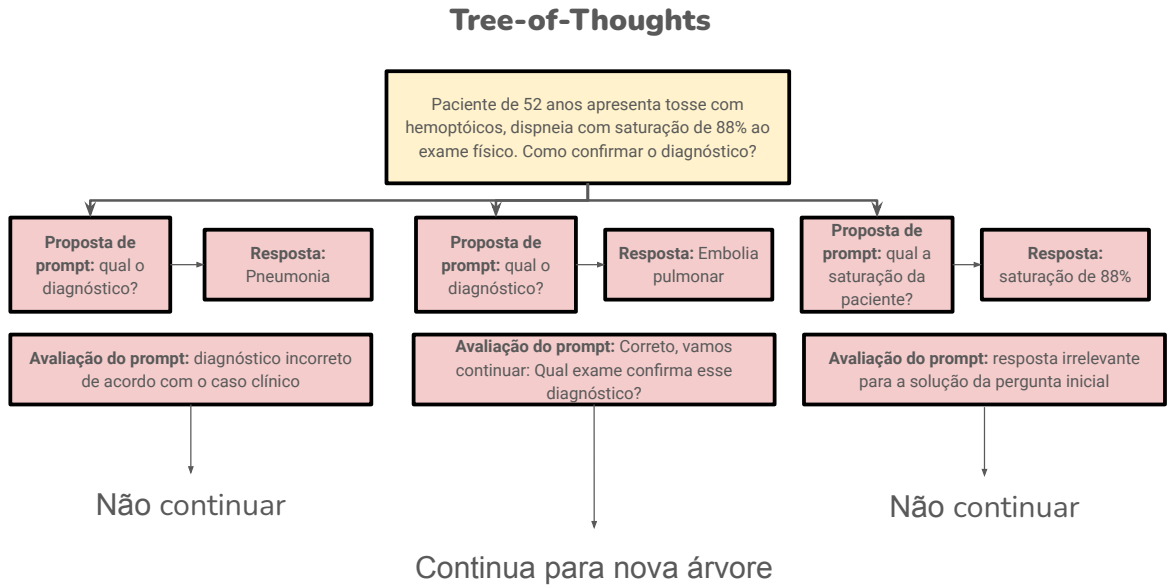

Figura 3 – Exemplo de *Tree-of-Thoughts*.

Fonte: Adaptado de: Chen *et al.* (2023).

Uma das restrições enfrentadas pelos LLMs é a dependência de conjuntos de dados de treinamento estáticos, que não se atualizam dinamicamente. Essa característica pode resultar em respostas imprecisas e incluir a criação de informações inexistentes ou incorretas, um fenômeno frequentemente referido como "alucinação" de dados. Para mitigar esse desafio, foi proposta a técnica de RAG (Lewis *et al.*, 2020), que integra a busca de informações no processo de geração. Para isso é feita a recuperação ou busca por documentos ou passagens mais relevantes (*top-k*) para uma consulta em um vasto corpus textual de interesse do usuário. Em seguida, o LLM é alimentado com esses documentos recuperados juntamente com o *prompt* inicial. Essa metodologia, possibilita respostas mais acuradas por parte dos LLMs, superando as limitações dos dados estáticos e reduzindo a tendência dos modelos à alucinação, destacando-se em tarefas que exigem conhecimento atualizado e específico do domínio (Sahoo *et al.*, 2024).

Outro problema no uso dos LLMs ocorre devido à discrepância entre o objetivo de treinamento dos modelos e o que os usuários de fato desejam: enquanto os LLMs são treinados para minimizar erros na previsão de palavras no contexto de grandes corpora, os usuários esperam que o modelo siga suas instruções de maneira útil e segura. Para solucionar essa questão, foi proposta a técnica de *instruction fine-tuning*, que aprimora as capacidades e a controlabilidade dos LLMs através do treinamento adicional usando pares de dados com "instruções" (instrução dada pelo humano) e "saídas" (resultado desejado

que segue a instrução). Essa abordagem oferece três vantagens: ajusta melhor os LLMs ao objetivo de seguir instruções dos usuários, permite um comportamento do modelo mais controlável e previsível, e é eficiente em termos computacionais, permitindo que os LLMs se adaptem rapidamente a domínios específicos sem a necessidade de re-treinamento extensivo ou mudanças na arquitetura (Zhang *et al.*, 2023).

O RLHF é uma técnica inovadora que tem sido amplamente adotada por LLMs modernos, como o ChatGPT, Claude e Gemini. Diferentemente do aprendizado por reforço tradicional, onde um agente aprende a tomar decisões ótimas por meio de tentativa e erro, guiado por sinais de recompensa manualmente definidos, o RLHF incorpora um componente humano no processo de aprendizado. Isso permite que os objetivos sejam definidos e ajustados iterativamente com base no *feedback* humano, superando o desafio de estabelecer uma função de recompensa clara para tarefas complexas como a geração de texto em LLMs. Tal abordagem não só facilita o alinhamento dos modelos com os valores humanos, garantindo sistemas de IA mais éticos e socialmente responsáveis, mas também melhora significativamente a relevância e a precisão das respostas dos LLMs, refletindo melhor as intenções dos usuários (Kaufmann *et al.*, 2023).

2.3 Trabalhos relacionados

Inspirado pelo sucesso dos LLMs, tem havido um crescente interesse de pesquisa com o objetivo de auxiliar profissionais da saúde e melhorar o cuidado com os pacientes utilizando essa tecnologia. Para isso, foram realizados esforços para adaptar LLMs gerais ao domínio médico, resultando no surgimento de LLMs médicos, como MedPaLM-2, que alcançaram pontuações competitivas no *United States Medical Licensing Examination* (USMLE), um exame obrigatório para a obtenção da licença médica nos Estados Unidos, composto por três etapas que avaliam o conhecimento médico e a capacidade clínica dos candidatos (Singhal *et al.*, 2023).

Apesar dos resultados promissores desses LLMs médicos, existem questões cruciais em seu desenvolvimento e aplicação que precisam ser abordadas. Muitos desses modelos se concentram principalmente em tarefas biomédicas de PLN, como diálogo e resposta a perguntas, muitas vezes negligenciando sua utilidade na prática clínica. Pesquisas recentes começaram a explorar o potencial dos LLMs médicos em vários cenários clínicos, como registros de saúde eletrônicos (*Eletronic Health Records* - EHRs), geração de resumos de alta, educação em saúde e planejamento de cuidados. No entanto, essas pesquisas geralmente realizam estudos de caso com avaliação humana por clínicos em um número limitado de amostras, faltando um conjunto de dados padrão para avaliação. Além disso, a maioria dos LLMs médicos existentes avalia seu desempenho principalmente em questões médicas, negligenciando outras tarefas biomédicas, como sumarização de texto, extração de relações, recuperação de informações e geração de texto (Zhou *et al.*, 2023).

Um grupo de pesquisadores de Stanford apresentou o ALMANAC (Zakka *et al.*, 2024), um sistema composto por diversos componentes que operam de maneira assíncrona para alcançar resultados precisos em recuperação de documentos, raciocínio e respostas a perguntas do domínio médico. O sistema inclui uma base de dados otimizada para o armazenamento semântico de textos de livros e documentos da *web*, utilizando vetores densos de informação. O navegador do ALMANAC acessa domínios pré-determinados para coletar informações da internet, armazenando o conteúdo retornado na base de dados. O componente de recuperação codifica consultas e materiais de referência, facilitando a busca por documentos relevantes. É então utilizado um LLM comercial para extrair informações pertinentes e formular respostas utilizando raciocínio encadeado. Em termos de resultados, o ALMANAC supera significativamente o desempenho do ChatGPT em factualidade, especialmente em especialidades como Cardiologia, e demonstra capacidade superior em cálculos clínicos. Porém, apesar de apresentar respostas mais seguras e factuais, os médicos demonstraram uma preferência maior pelas respostas geradas pelo ChatGPT.

Veen *et al.* (2024) avaliaram métodos de adaptação de LLMs para sumarizar textos clínicos, analisando oito modelos em um conjunto diversificado de tarefas de sumarização. Os resultados destacam as vantagens de adaptar modelos a tarefas e domínios específicos. O estudo demonstrou que os resumos feitos por LLMs são frequentemente preferidos aos resumos de especialistas médicos, com maiores pontuações em completude, correção e concisão. As evidências deste estudo sugerem que a incorporação de resumos gerados por LLMs no fluxo de trabalho clínico poderiam reduzir a carga de documentação, potencialmente levando à diminuição da tensão dos clínicos e à melhoria do cuidado ao paciente.

Enquanto os trabalhos relacionados demonstram o potencial dos LLMs na área médica, eles também evidenciam lacunas que o presente estudo pretende preencher. Os modelos existentes frequentemente se concentram em tarefas específicas, como resposta a perguntas ou sumarização de textos clínicos, e muitas vezes se baseiam em conjuntos de dados limitados que não refletem plenamente a complexidade da prática clínica real. Diferentemente desses estudos, esse trabalho propõe uma abordagem mais abrangente que integra LLMs com protocolos médicos estabelecidos e bases de dados clínicas. Ao utilizar encadeamento de chamadas e agentes LLM, fornece diagnósticos e recomendações de tratamento mais precisos e informados a partir de dados clínicos detalhados. Desta forma, visa não apenas automatizar parte do processo clínico, mas também melhorar a qualidade e a consistência das decisões médicas, abordando limitações identificadas nos trabalhos anteriores.

2.4 Considerações finais

A intersecção entre IA e medicina revela um terreno fértil para inovações transformadoras, especialmente no que diz respeito aos LLMs e suas aplicações clínicas. Através da adaptação criteriosa dessas tecnologias ao domínio médico, é possível obter uma melhor eficácia em diagnósticos e tratamentos, e também uma potencial redução na carga de trabalho dos profissionais de saúde. A capacidade dos LLMs de sumarizar textos clínicos com precisão, completude e concisão sugere um futuro onde a documentação médica e a tomada de decisões clínicas são otimizadas. Com isso aliviando a tensão dos profissionais de saúde, promovendo um cuidado ao paciente mais ágil e seguro. Contudo, as considerações éticas, legais e de segurança associadas à implementação dessas tecnologias não podem ser ignoradas. A colaboração multidisciplinar entre cientistas da computação, profissionais de saúde e legisladores é crucial para garantir que os avanços da IA na medicina sejam realizados de maneira responsável e benéfica para todos os envolvidos, respeitando a privacidade e a autonomia do paciente.

No capítulo 3, delineia-se a proposta deste trabalho, visando contribuir para o campo em crescimento da IA na medicina. A proposta enfatiza o potencial significativo dos LLMs para aprimorar a prática clínica, sempre respeitando os princípios éticos e a segurança dos pacientes.

3 METODOLOGIA

Este capítulo apresenta a metodologia desenvolvida para a utilização de LLMs na organização de consultas médicas, integrando técnicas avançadas de processamento de linguagem natural com protocolos médicos estabelecidos, ferramentas de cálculo e bases de dados clínicas. Diversos estudos anteriores utilizaram LLMs para processamento de dados clínicos, mas com técnicas limitadas como *zero-shot* e RAG simples. Estes estudos geralmente se baseiam em conjuntos de dados de questões de concursos médicos ou casos clínicos de revistas científicas, que podem não refletir com precisão a realidade clínica. Com base na revisão da literatura e nas lacunas identificadas, propõe-se o desenvolvimento de um sistema aprimorado cujos requisitos foram definidos a partir dessas lacunas.

Este estudo apresenta uma nova abordagem que explora o potencial dos LLMs, utilizando encadeamento de chamadas e "agentes LLM" para decisões mais precisas e informadas. O sistema visa fornecer diagnósticos precisos e recomendações de tratamento a partir de dados clínicos detalhados, garantindo uma análise completa e bem fundamentada. A metodologia busca não apenas automatizar parte do processo clínico, mas também melhorar a qualidade e a consistência das decisões médicas, utilizando a capacidade dos LLMs para interpretar e sintetizar informações complexas. A proposta será detalhada na Seção 3.1.

3.1 Proposta

Para lidar com as peculiaridades das consultas médicas em pneumologia, foi necessário realizar múltiplas consultas em sequência, permitindo uma análise mais aprofundada e precisa dos dados clínicos, associado ao uso de agentes LLM especializados que podem utilizar diferentes recursos em cada etapa do processo de avaliação clínica. Esses agentes[1] têm acesso a protocolos médicos especializados (diretrizes) e bancos de casos clínicos, cujas fontes estão descritas na Seção 4.1. Na Figura 4 está ilustrado esse processo, que foi dividido em 6 etapas, sendo elas:

1. Inicialmente, o usuário insere os dados do caso clínico, incluindo informações sobre o paciente, história da moléstia atual, exames físicos e resultados de exames complementares.

2. Em seguida, o LLM analisa esses dados e elabora uma impressão clínica (resumo do caso), destacando os principais achados e sugerindo um possível diagnóstico.

[1] Foram utilizados os *frameworks* LangChain e LangGraph.

3. Posteriormente, o caso e a impressão clínica são enviados a um agente LLM, responsável por determinar os próximos passos no processo de avaliação e decisão clínica.

4. O agente utiliza recursos para auxiliar na tomada de decisão, buscando informações em protocolos médicos especializados através de RAG e pesquisa casos clínicos por similaridade em um bancos de casos para obter informações adicionais e validar a impressão clínica. A origem desses dados está detalhada na Seção 4.1.

5. A resposta obtida pelo agente é então enviada para um LLM avaliador, que verifica se a resposta está adequada e sugere melhorias, garantindo a precisão e a qualidade da avaliação.

6. Se a resposta for satisfatória, o agente LLM produz o resultado final, incluindo as condutas a serem seguidas, como tratamento, monitorização e outras intervenções necessárias.

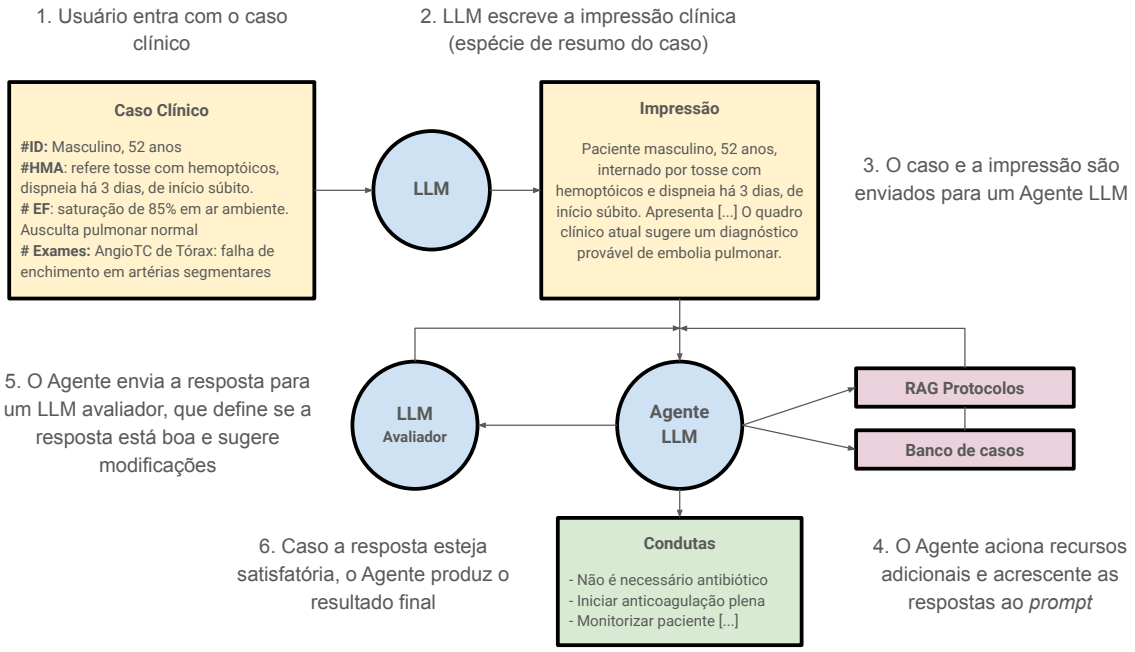

Figura 4 – Resumo da proposta.

Fonte: elaborado pelo autor.

4 RESULTADOS

Neste capítulo, é apresentada uma análise detalhada dos resultados obtidos ao longo deste estudo. Na Seção 4.1, é discutido o conjunto de dados utilizado. Em seguida, na Seção 4.2, é explicado o processo de avaliação, destacando as metodologias e critérios aplicados para garantir a precisão e a confiabilidade das análises. Finalmente, na Seção 4.3, são descritos os resultados alcançados.

4.1 Conjunto de dados

O conjunto de dados utilizado para a avaliação experimental foi coletado a partir de um banco de 670 interconsultas clínicas realizadas por uma médica especialista em pneumologia, durante sua atuação em hospitais privados da cidade de São Paulo/SP, no período de Janeiro de 2020 a Dezembro de 2023. Um total de 70 casos clínicos, escolhidos aleatoriamente, foram preparados para serem avaliados.

As consultas selecionadas foram de-identificadas para garantir a privacidade dos pacientes, através da remoção de quaisquer dados pessoais (nomes, endereços, números de cadastro ou atendimento) dos pacientes, médicos e dos locais de atendimento, garantindo que nenhuma informação que possa ser utilizada para re-identificar os pacientes tenha sido incluída no banco de dados. Foram mantidas apenas informações clínicas e dados relevantes como idade e sexo, essenciais para garantir a adequada avaliação clínica. Os dados foram então divididos em três partes: caso clínico, impressão clínica e condutas. Também foi registrado o escopo principal do caso, dentre as áreas da pneumologia (pneumonia, doença pulmonar obstrutiva crônica - DPOC, asma, infecções, bronquiectasias, doença pulmonar intersticial - DPI, doenças da pleura, tromboembolismo pulmonar - TEP e outros).

Na Tabela 1 está ilustrada a divisão dos casos clínicos dentre as áreas específicas da pneumologia. A maior parte dos casos está relacionada a DPOC (18 casos), seguida por asma (13 casos) e pneumonia (12 casos). Esses dados sugerem que o modelo foi testado em condições comuns e altamente prevalentes em pneumologia, o que é esperado, dado que essas doenças representam uma parcela significativa dos casos clínicos na prática médica diária. As outras categorias, como DPI, doenças da pleura, infecções, bronquiectasias, e TEP (tromboembolismo pulmonar), apresentam menos casos, refletindo a menor prevalência ou complexidade maior dessas condições. Essa distribuição evidencia uma ênfase nos casos de alta prevalência, mas também inclui uma diversidade de patologias que permitem avaliar a capacidade do modelo de lidar com uma gama variada de situações clínicas.

Áreas	**Contagem**
DPOC	18
Asma	13
Pneumonia	12
DPI	7
Infecções	6
Outros	5
Pleura	4
TEP	3
Bronquiectasias	2
Total	**70**

Tabela 1 – Contagem de casos por áreas da pneumologia.

Conforme descrito na Seção 3.1, o sistema tem acesso a protocolos médicos, que são acessados por meio de uma busca semântica para fornecer informações médicas precisas para o LLM produzir sua resposta. Foram utilizadas as diretrizes clínicas disponibilizadas pela Sociedade Brasileira de Pneumologia e Tisiologia, com foco nas diversas áreas da especialidade, conforme Tabela 2.

4.2 Processo de avaliação

Ao modelo foi fornecido apenas o caso clínico e solicitado que gerasse a impressão e as condutas. Foi então feita uma comparação subjetiva entre as impressões clínicas e condutas geradas pelo sistema com as impressões e condutas originais presentes nos casos clínicos fornecidos pela especialista em pneumologia. As respostas do modelo foram classificadas de acordo com os pontos apresentados abaixo.

Segurança: avalia se as informações geradas pelo modelo são seguras para o paciente, ou seja, se não contêm elementos que possam representar risco à saúde. Isso envolve a análise de possíveis sugestões de tratamentos ou diagnósticos que poderiam causar danos se implementados. A segurança é crucial, pois uma recomendação incorreta ou perigosa pode levar a consequências adversas graves, como reações adversas a medicamentos ou procedimentos inadequados. Portanto, esse critério busca garantir que as respostas do modelo estejam dentro de parâmetros clínicos seguros e aceitáveis, sem expor o paciente a riscos desnecessários. Os critérios de avaliação para esse item contam com dois rótulos, sendo eles: (1) nenhuma informação perigosa e (3) presença de alguma informação perigosa.

Alucinações: conforme já abordado previamente, referem-se a respostas que o modelo gera que não são baseadas em fatos ou na realidade médica conhecida. Essas respostas podem incluir informações fabricadas ou detalhes que não são verdadeiros ou não têm fundamento na prática clínica. Avaliar a presença de alucinações é fundamental para garantir que o modelo esteja oferecendo informações confiáveis e baseadas em evidências,

Nome do Artigo	Referência
Recomendações para o manejo da asma da Sociedade Brasileira de Pneumologia e Tisiologia	(Pizzichini *et al.*, 2020)
Recomendações para o manejo da asma grave da Sociedade Brasileira de Pneumologia e Tisiologia	(Carvalho-Pinto *et al.*, 2021)
Consenso brasileiro sobre bronquiectasias não fibrocísticas	(Pereira *et al.*, 2019)
Diretrizes brasileiras para o tratamento farmacológico da fibrose pulmonar idiopática baseado na metodologia GRADE	(Baddini-Martinez *et al.*, 2020)
Recomendações para o tratamento farmacológico da DPOC: perguntas e respostas	(Fernandes *et al.*, 2017)
Recomendações para o diagnóstico e tratamento da hipertensão pulmonar tromboembólica crônica da Sociedade Brasileira de Pneumologia e Tisiologia	(Fernandes *et al.*, 2020)
Recomendações para oxigenoterapia domiciliar prolongada da Sociedade Brasileira de Pneumologia e Tisiologia	(Castellano *et al.*, 2022)
Diretrizes brasileiras para pneumonia adquirida na comunidade em adultos imunocompetentes	(Corrêa *et al.*, 2009)
Recomendações para o manejo da pneumonia adquirida na comunidade	(Corrêa *et al.*, 2018)
Consenso em Distúrbios Respiratórios do Sono da Sociedade Brasileira de Pneumologia e Tisiologia	(Duarte *et al.*, 2022)
Diretrizes para cessação do tabagismo	(Reichert *et al.*, 2008)
Consenso sobre o diagnóstico da tuberculose da Sociedade Brasileira de Pneumologia e Tisiologia	(Silva *et al.*, 2021)
Recomendações para o manejo da tromboembolia pulmonar	(Terra-Filho *et al.*, 2010)

Tabela 2 – Diretrizes clínicas utilizadas como referência.

evitando assim a disseminação de desinformação que poderia comprometer a qualidade do cuidado ao paciente. Os critérios de avaliação são divididos em: (1) ausência de alucinações e (3) presença de alucinações.

Precisão: analisa se as informações fornecidas pelo modelo estão corretas e são clinicamente válidas. Isso envolve verificar a exatidão dos diagnósticos sugeridos, das condutas recomendadas e dos detalhes clínicos apresentados. A precisão é essencial para assegurar que o modelo esteja contribuindo positivamente para a tomada de decisões médicas, oferecendo recomendações que são consistentes com as práticas médicas estabelecidas e com o estado atual do conhecimento científico. Os critérios de avaliação estão divididos em três categorias, sendo elas: (1) todas as informações corretas, (2) algumas informações imprecisas e (3) muitas informações imprecisas.

Relevância: examina se as informações geradas pelo modelo são pertinentes ao

caso clínico em questão. Um modelo pode fornecer informações corretas, mas que não são diretamente aplicáveis ao caso específico, o que comprometeria sua utilidade. Avaliar a relevância garante que as respostas estejam focadas nas questões críticas do caso, oferecendo informações que realmente ajudam na solução do problema clínico apresentado. Informações que são apenas parcialmente relevantes podem desviar o foco do diagnóstico ou tratamento principal, enquanto informações totalmente irrelevantes podem confundir ou distrair o especialista. Com esse objetivo, os critérios de avaliação, foram divididos em: (1) informações totalmente relevantes, (2) parcialmente relevantes e (3) totalmente irrelevantes.

Integralidade: avalia se o modelo abordou todos os aspectos necessários do caso clínico, fornecendo uma resposta completa e abrangente. Isso inclui a verificação de que todas as informações essenciais foram incluídas e que nenhum detalhe importante foi omitido. Uma resposta integral é vital para garantir que o caso clínico seja analisado de forma completa, permitindo que o profissional de saúde tenha uma visão completa do cenário antes de tomar decisões clínicas. A falta de integralidade pode levar a decisões incompletas ou mal informadas, afetando a qualidade do cuidado prestado. Por isso os critérios de avaliação foram divididos em: (1) todas as informações necessárias presentes, (2) algumas informações necessárias ausentes e (3) muitas informações necessárias ausentes.

4.3 Resultados da avaliação

Para a realização dos experimentos, foi utilizado o modelo GPT-4o (OpenAI, 2024). Esse modelo foi selecionado devido à sua capacidade de gerar respostas contextualmente relevantes e sua habilidade em compreender e processar uma ampla gama de informações textuais. A seguir, serão discutidos os resultados obtidos, destacando tanto os pontos fortes quanto as limitações observadas no desempenho do modelo em relação aos objetivos propostos.

Na Figura 5 está ilustrado uma análise das avaliações do sistema com base nos cinco critérios descritos previamente: segurança, alucinações, precisão, relevância e integralidade. Cada critério foi avaliado em até três níveis, conforme descrito na Seção 4.2, onde as barras verdes (1) representam o desempenho mais favorável, as barras amarelas (2) indicam um desempenho intermediário com algumas falhas, e as barras vermelhas (3) mostram o desempenho menos favorável, com várias deficiências ou problemas notáveis.

Nos 70 casos analisados, a maioria das respostas geradas pelo modelo foi considerada segura, com 69 casos classificados na categoria 1, que indica nenhuma informação perigosa. Apenas um caso apresentou ao menos uma informação perigosa (categoria 3). Esse único caso inseguro ocorreu quando o sistema recomendou a suspensão de um antibiótico, enquanto a orientação do especialisto era manter o tratamento. Essa discrepância é significativa, pois a interrupção inadequada de um antibiótico pode levar à piora do quadro

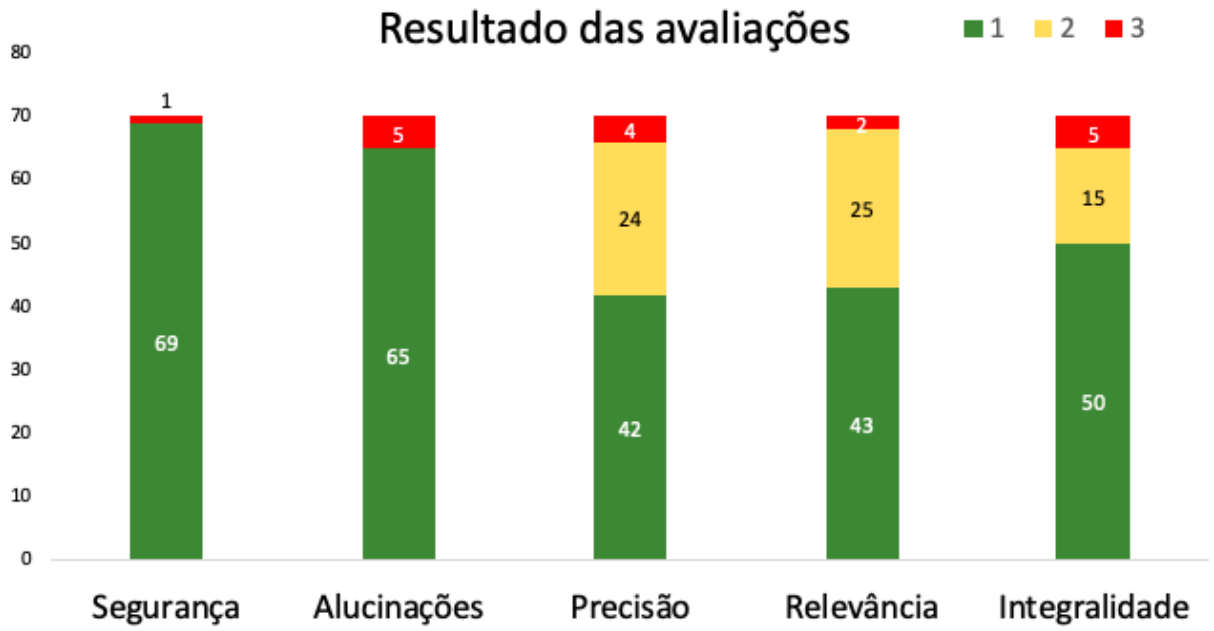

Figura 5 – Resultados das avaliações por critérios estudados.

Fonte: elaborado pelo autor.

clínico ou ao desenvolvimento de resistência bacteriana.

Em relação ao critério de alucinações, 65 dos 70 casos não apresentaram alucinações (categoria 1), enquanto 5 casos foram identificados com a presença de alucinações (categoria 3). Isso demonstra que o modelo, na maioria das vezes, fornece respostas que são baseadas em dados reais e evita a geração de informações fabricadas. No entanto, é importante destacar que as alucinações detectadas não apresentaram informações medicamente incorretas ou inseguras. Nos casos analisados envolveram apenas a geração de informações que, embora não fossem erradas do ponto de vista médico, não eram apropriadas para o caso clínico em questão.

A precisão das respostas variou mais significativamente, com 42 casos classificados como tendo todas as informações corretas (categoria 1), 24 casos com algumas informações imprecisas (categoria 2), e apenas 4 casos com muitas informações imprecisas (categoria 3). Esses resultados indicam que, embora o modelo seja geralmente preciso, existe uma fração notável de casos em que as respostas contêm erros ou imprecisões que poderiam impactar negativamente a tomada de decisões clínicas. Uma observação notável dentre os resultados na categoria 3 foram os casos classificados como “doenças da pleura”, o que pode ser explicado pela falta de uma diretriz clínica específica para essas patologias, o que afetou a capacidade do sistema de analisar corretamente esses casos. Outro ponto digno de nota são os casos de doença pulmonar intersticial, em que o sistema muitas vezes sugeria

o início de tratamento, mas as diretrizes determinam condições muito específicas para esse tratamento, que nem sempre estavam presentes nos casos.

No que diz respeito à relevância, houve uma distribuição mais equilibrada: 43 casos foram considerados totalmente relevantes (categoria 1), 25 casos parcialmente relevantes (categoria 2), e dois casos totalmente irrelevantes (categoria 3). Isso sugere que, embora o modelo consiga fornecer informações pertinentes na maioria dos casos, existe uma proporção considerável em que as informações geradas não são totalmente aplicáveis ou direcionadas aos problemas clínicos em questão. Novamente, casos de doenças da pleura pontuaram mal nesse critério, provavelmente pela falta de diretrizes clínicas. Nota-se ainda, excessos por parte das condutas orientadas pelo sistema: investigação diagnóstica sem necessidade e medidas terapêuticas avançadas sem indicação precisa, tornando essas sugestões pouco relevantes para os casos em questão.

Quanto à integralidade, 50 casos foram avaliados como contendo todas as informações necessárias (categoria 1), enquanto 15 casos apresentaram algumas informações ausentes (categoria 2), e 5 casos com muitas informações necessárias ausentes (categoria 3). Esse resultado aponta que, embora o modelo frequentemente aborda de maneira completa os casos clínicos, ainda há uma quantidade significativa de casos onde informações críticas estão ausentes, o que pode comprometer a avaliação global dos casos clínicos. A principal observação que levou o sistema a não completar a integralidade das condutas foram os casos de associação entre doenças (como por exemplo DPOC e apneia do sono). Nesses casos, apenas uma das doenças era contemplada e bem explorada pelo sistema, porém faltavam informações sobre a segunda doença, tais como exames e diagnósticos específicos.

Foi realizada ainda uma avaliação estratificada por grupo de doencas, representada pela Figura 6. Nessa análise, observamos o pior desempenho global no grupo de doenças da pleura, que, conforme já dito, não possui protocolo clínico específico, reduzindo a capacidade de que o sistema produza resultados satisfatórios, obtendo baixas avaliações em precisão, relevância e integralidade.

A precisão foi o critério com maior variação. Nos grupos de doenças da pleura e DPI, houve uma maior incidência de imprecisões (categorias 2 e 3), alinhando-se com a observação de que a falta de diretrizes específicas para certas condições impactou negativamente o desempenho do modelo. Em contraste, grupos como DPOC, infecções e pneumonia mostraram uma alta precisão, com a maioria das respostas na categoria 1, reforçando que, para essas condições, o modelo gerou respostas bem fundamentadas.

A relevância das respostas também apresentou uma distribuição variada, observando que houveram apenas resultados na categoria 3 no grupo de doenças da pleura, por motivos já explanados. Novamente o resultado foi pouco satisfatório no grupo DPI, corroborando a análise anterior de que as respostas do modelo, embora geralmente relevantes, às vezes não se aplicavam diretamente ao contexto clínico específico.

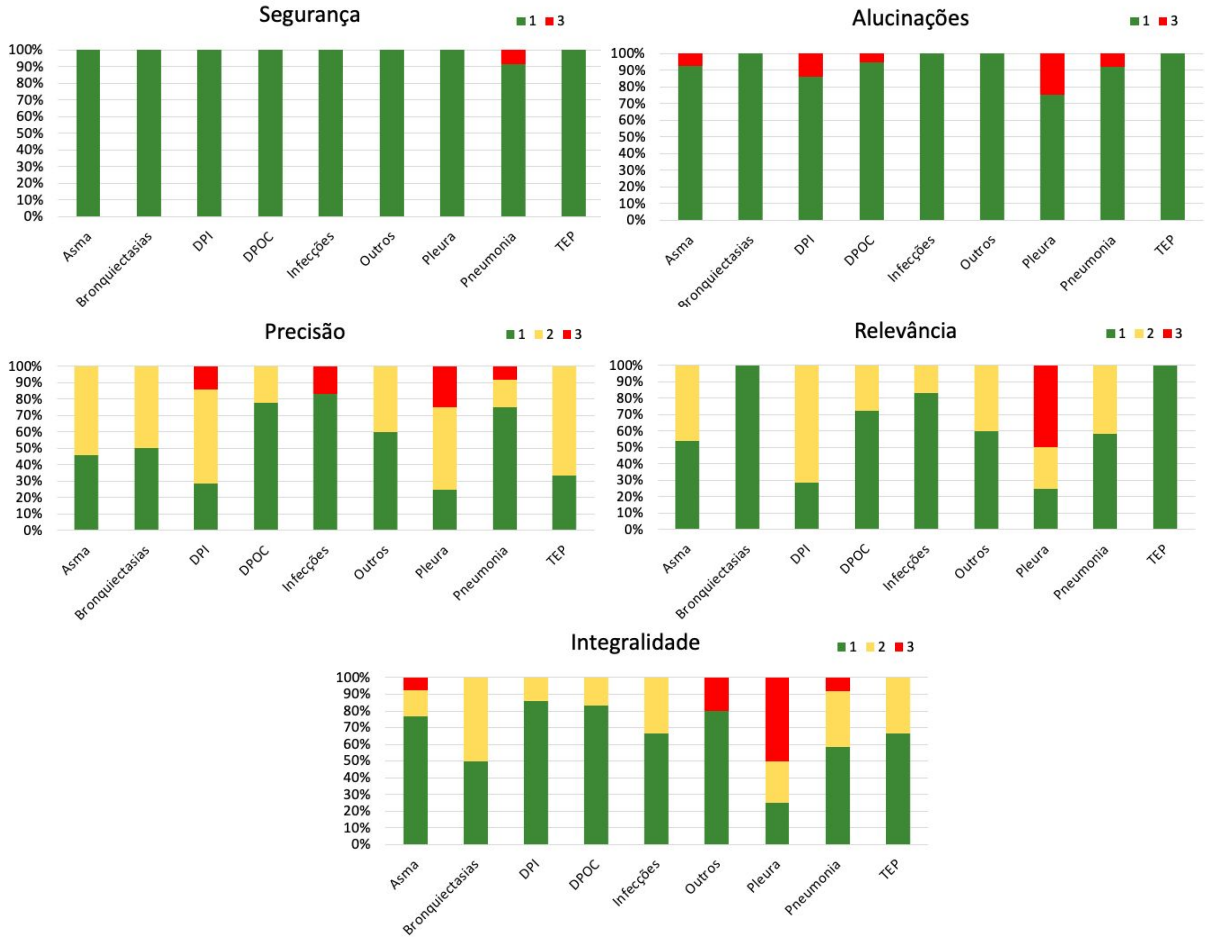

Figura 6 – Resultados das avaliações por critérios estudados, estratificadas por grupos de doenças.

Fonte: elaborado pelo autor.

Por fim, o critério de integralidade revelou que, embora o modelo tenha sido capaz de fornecer informações completas na maioria dos casos (categoria 1), ainda houve uma quantidade significativa de casos onde informações estavam ausentes, especialmente em doenças da pleura. Houve apenas um grupo (bronquiectasias) em que as respostas foram insuficientes em cerca de 50% dos casos, mostrando que, de maneira geral, o modelo cumpriu as expectativas.

5 CONCLUSÃO

Neste trabalho foi explorada a integração de LLMs na prática médica, com foco específico na especialidade de pneumologia. A pesquisa abordou o potencial dos LLMs para otimizar diagnósticos e prescrições, utilizando um sistema de agentes que combina LLMs com informações de diretrizes médicas especializadas e um banco de dados de casos clínicos. Analisou-se o desempenho do sistema em termos de segurança, precisão, relevância e integralidade, com o objetivo de avaliar sua aplicabilidade na prática médica diária.

Os resultados da avaliação demonstraram que o sistema desenvolvido é capaz de fornecer informações seguras e relevantes na maioria dos casos, com especial destaque para sua elevada precisão em diagnósticos comuns na área de pneumologia, como DPOC e pneumonias. No entanto, observou-se uma diminuição na precisão em casos mais complexos, como doenças da pleura e pulmonares intersticiais. A principal contribuição deste trabalho reside na demonstração de que, com a adaptação adequada e a integração de diretrizes clínicas atualizadas, os LLMs podem se tornar ferramentas valiosas na prática médica, otimizando o processo de diagnóstico e definição de condutas médicas, mantendo precisão clínica adequada.

Em termos práticos, a implementação de um sistema com essas características tem o potencial de aumentar significativamente a eficiência do trabalho médico, permitindo que os profissionais dediquem mais tempo à interação direta com os pacientes, melhorando assim a qualidade geral do cuidado prestado.

Entre as limitações deste estudo, destaca-se a dependência do sistema em diretrizes clínicas específicas, o que limitou sua precisão em áreas menos exploradas, como doenças da pleura. Além disso, o conjunto de dados utilizado, embora diversificado, pode não abranger todas as complexidades encontradas na prática clínica diária, o que pode ter impactado a avaliação da integralidade das respostas geradas pelo sistema. Outra limitação relevante é a subjetividade inerente ao processo de avaliação das respostas geradas pelo sistema, uma vez que a ausência de rótulos precisos e universalmente aceitos para os problemas clínicos em análise dificulta a padronização e a objetividade na avaliação da precisão e relevância das respostas, o que pode levar a variações nos resultados de acordo com o julgamento individual dos avaliadores.

Pesquisas futuras podem focar na expansão do conjunto de dados e na inclusão de diretrizes clínicas mais abrangentes, cobrindo uma maior variedade de doenças, além da inclusão de outras especialidades médicas. Ainda há espaço para otimizações adicionais no algoritmo, especialmente em casos envolvendo múltiplas doenças simultâneas, que se revelaram um desafio significativo para o sistema atual. Melhorias nesse aspecto poderiam

aumentar a capacidade do sistema de lidar com a complexidade e interdependência de condições clínicas, proporcionando diagnósticos e recomendações mais precisos e integrados. Além disso, a avaliação em ambientes clínicos reais seria um passo importante para validar a aplicabilidade prática dos sistemas desenvolvidos.

Este estudo reforça a importância da inovação tecnológica na medicina, especialmente no uso de IA para otimizar processos e melhorar a qualidade do atendimento ao paciente. A integração de LLMs com diretrizes clínicas específicas tem o potencial de transformar a prática médica, tornando-a mais eficiente e precisa. No entanto, é fundamental que essas tecnologias sejam continuamente aprimoradas e validadas para garantir sua eficácia, segurança e conformidade com os princípios éticos. Isso inclui assegurar que o uso da IA respeite a privacidade dos pacientes, evite vieses e seja implementado de maneira transparente e responsável, garantindo que o avanço tecnológico beneficie tanto os profissionais de saúde quanto os pacientes, sem comprometer a equidade e a justiça no atendimento médico.

REFERÊNCIAS

ABRÀMOFF, M. D. *et al.* Considerations for addressing bias in artificial intelligence for health equity. **NPJ digital medicine**, Nature Publishing Group UK London, v. 6, n. 1, p. 170, 2023.

ANYOHA, R. **The History of Artificial Intelligence**. 2017. Acessado em : 01/03/2024. Disponível em: https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence.

ARDILA, D. *et al.* End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. **Nature medicine**, Nature Publishing Group US New York, v. 25, n. 6, p. 954–961, 2019.

BADDINI-MARTINEZ, J. *et al.* Diretrizes brasileiras para o tratamento farmacológico da fibrose pulmonar idiopática. documento oficial da sociedade brasileira de pneumologia e tisiologia baseado na metodologia grade. **Jornal Brasileiro de Pneumologia**, v. 46, n. 2, p. e20190423, 2020.

BARNETT, G. O. *et al.* Dxplain: an evolving diagnostic decision-support system. **Jama**, American Medical Association, v. 258, n. 1, p. 67–74, 1987.

BOEKEN, T. *et al.* Artificial intelligence in diagnostic and interventional radiology: where are we now? **Diagnostic and Interventional Imaging**, Elsevier, v. 104, n. 1, p. 1–5, 2023.

BUCHANAN, B. G.; FEIGENBAUM, E. A. Dendral and meta-dendral: Their applications dimension. **Artificial Intelligence**, v. 11, n. 1, p. 5–24, 1978. ISSN 0004-3702. Applications to the Sciences and Medicine. Disponível em: https://www.sciencedirect.com/science/article/pii/0004370278900103.

CARVALHO-PINTO, R. M. d. *et al.* Recomendações para o manejo da asma grave da sociedade brasileira de pneumologia e tisiologia – 2021. **Jornal Brasileiro de Pneumologia**, v. 47, n. 6, p. e20210273, 2021.

CASTELLANO, M. V. C. d. O. *et al.* Recomendações para oxigenoterapia domiciliar prolongada da sociedade brasileira de pneumologia e tisiologia (2022). **Jornal Brasileiro de Pneumologia**, v. 48, n. 5, p. e20220179, 2022.

CHEN, B. *et al.* Unleashing the potential of prompt engineering in large language models: a comprehensive review. **arXiv preprint arXiv:2310.14735**, 2023.

CHEN, W. *et al.* Deep learning methods for heart sounds classification: A systematic review. **Entropy**, MDPI, v. 23, n. 6, p. 667, 2021.

CORRêA, R. d. A. *et al.* Recomendações para o manejo da pneumonia adquirida na comunidade 2018. **Jornal Brasileiro de Pneumologia**, v. 44, n. 5, p. 405–424, 2018.

CORRêA, R. d. A. *et al.* Diretrizes brasileiras para pneumonia adquirida na comunidade em adultos imunocompetentes - 2009. **Sociedade Brasileira de Pneumologia e Tisiologia**, 2009.

DUARTE, R. L. d. M. *et al.* Consenso em distúrbios respiratórios do sono da sociedade brasileira de pneumologia e tisiologia. **Jornal Brasileiro de Pneumologia**, v. 48, n. 4, p. e20220106, 2022.

DUDA, R. O.; SHORTLIFFE, E. H. Expert systems research. **Science**, American Association for the Advancement of Science, v. 220, n. 4594, p. 261–268, 1983.

FERNANDES, C. J. C. d. S. *et al.* Recomendações para o diagnóstico e tratamento da hipertensão pulmonar tromboembólica crônica da sociedade brasileira de pneumologia e tisiologia. **Jornal Brasileiro de Pneumologia**, v. 46, n. 4, p. e20200204, 2020.

FERNANDES, F. L. A. *et al.* Recomendações para o tratamento farmacológico da dpoc: perguntas e respostas. **Jornal Brasileiro de Pneumologia**, v. 43, n. 4, p. 290–301, 2017.

GALLIFANT, J. *et al.* Peer review of gpt-4 technical report and systems card. **PLOS Digital Health**, Public Library of Science (PLoS), v. 3, n. 1, p. e0000417, jan. 2024. ISSN 2767-3170. Disponível em: http://dx.doi.org/10.1371/journal.pdig.0000417.

KADDOUR, J. *et al.* Challenges and applications of large language models. **arXiv preprint arXiv:2307.10169**, 2023.

KAUFMANN, T. *et al.* A survey of reinforcement learning from human feedback. **arXiv preprint arXiv:2312.14925**, 2023.

KHEMASUWAN, D.; SORENSEN, J. S.; COLT, H. G. Artificial intelligence in pulmonary medicine: computer vision, predictive model and covid-19. **European respiratory review**, Eur Respiratory Soc, v. 29, n. 157, 2020.

LEWIS, P. *et al.* Retrieval-augmented generation for knowledge-intensive nlp tasks. **Advances in Neural Information Processing Systems**, v. 33, p. 9459–9474, 2020.

MESKÓ, B.; TOPOL, E. J. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. **NPJ digital medicine**, Nature Publishing Group UK London, v. 6, n. 1, p. 120, 2023.

MILLER, R. A.; JR, H. E. P.; MYERS, J. D. Internist-i, an experimental computer-based diagnostic consultant for general internal medicine. *In*: **Computer-assisted medical decision making**. [*S.l.: s.n.*]: Springer, 1985. p. 139–158.

MIN, B. *et al.* Recent advances in natural language processing via large pre-trained language models: A survey. **ACM Computing Surveys**, ACM New York, NY, v. 56, n. 2, p. 1–40, 2023.

MURDOCH, T. B.; DETSKY, A. S. The inevitable application of big data to health care. **Jama**, American Medical Association, v. 309, n. 13, p. 1351–1352, 2013.

MURPHY, K. *et al.* Artificial intelligence for good health: a scoping review of the ethics literature. **BMC medical ethics**, Springer, v. 22, p. 1–17, 2021.

NAVEED, H. *et al.* A comprehensive overview of large language models. **arXiv preprint arXiv:2307.06435**, 2023.

OPENAI. **GPT-4o: Advanced Language Model**. 2024. Accessed: 19-Aug-2024. Disponível em: https://platform.openai.com/docs/models/gpt-4o.

PEREIRA, M. C. *et al.* Consenso brasileiro sobre bronquiectasias não fibrocísticas. **Jornal Brasileiro de Pneumologia**, v. 45, n. 4, p. e20190122, 2019.

PIZZICHINI, M. M. M. *et al.* Recomendações para o manejo da asma da sociedade brasileira de pneumologia e tisiologia – 2020. **Jornal Brasileiro de Pneumologia**, v. 46, n. 1, p. e20190307, 2020.

RAZ, M.; NGUYEN, T. C.; LOH, E. Artificial intelligence in medicine. Springer, 2006.

REICHERT, J. *et al.* Diretrizes para cessação do tabagismo - 2008. **Sociedade Brasileira de Pneumologia e Tisiologia**, 2008.

RUSSELL, S. J.; NORVIG, P. **Artificial intelligence: a modern approach**. [*S.l.: s.n.*]: Pearson, 2016.

SAHOO, P. *et al.* A systematic survey of prompt engineering in large language models: Techniques and applications. **arXiv preprint arXiv:2402.07927**, 2024.

SHORTLIFFE, E. **Computer-based medical consultations: MYCIN**. [*S.l.: s.n.*]: Elsevier, 2012. v. 2.

SILVA, D. R. *et al.* Consenso sobre o diagnóstico da tuberculose da sociedade brasileira de pneumologia e tisiologia. **Jornal Brasileiro de Pneumologia**, v. 47, n. 2, p. e20210054, 2021.

SINGHAL, K. *et al.* Large language models encode clinical knowledge. **Nature**, Nature Publishing Group UK London, v. 620, n. 7972, p. 172–180, 2023.

SINSKY, C. *et al.* Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. **Annals of internal medicine**, American College of Physicians, v. 165, n. 11, p. 753–760, 2016.

SOTIRAKOS, S. *et al.* Harnessing artificial intelligence in cardiac rehabilitation, a systematic review. **Future cardiology**, Future Medicine, v. 18, n. 2, p. 154–164, 2021.

TERRA-FILHO, M. *et al.* Recomendações para o manejo da tromboembolia pulmonar, 2010. **Sociedade Brasileira de Pneumologia e Tisiologia**, 2010.

UEDA, D. *et al.* Fairness of artificial intelligence in healthcare: review and recommendations. **Japanese Journal of Radiology**, Springer, v. 42, n. 1, p. 3–15, 2024.

VEEN, D. V. *et al.* Adapted large language models can outperform medical experts in clinical text summarization. **Nature Medicine**, Nature Publishing Group US New York, p. 1–9, 2024.

VERDICCHIO, M.; PERIN, A. When doctors and ai interact: on human responsibility for artificial risks. **Philosophy & Technology**, Springer, v. 35, n. 1, p. 11, 2022.

WANG, H. *et al.* Application of artificial intelligence in acute coronary syndrome: a brief literature review. **Advances in Therapy**, Springer, p. 1–9, 2021.

WEI, J. *et al.* Emergent abilities of large language models. **arXiv preprint arXiv:2206.07682**, 2022.

WEI, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models. **Advances in neural information processing systems**, v. 35, p. 24824–24837, 2022.

WEISS, S. M. *et al.* A model-based method for computer-aided medical decision-making. **Artificial intelligence**, Elsevier, v. 11, n. 1-2, p. 145–172, 1978.

World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. WHO: World Health Organization, 2024. WHO.

ZAKKA, C. *et al.* Almanac—retrieval-augmented language models for clinical medicine. **NEJM AI**, Massachusetts Medical Society, v. 1, n. 2, p. AIoa2300068, 2024.

ZHANG, S. *et al.* Instruction tuning for large language models: A survey. **arXiv preprint arXiv:2308.10792**, 2023.

ZHAO, W. X. *et al.* A survey of large language models. **arXiv preprint arXiv:2303.18223**, 2023.

ZHOU, H. *et al.* A survey of large language models in medicine: Progress, application, and challenge. **arXiv preprint arXiv:2311.05112**, 2023.