25°enancib
Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação
Informação, Decolonialidade e Direitos Difusos para o Desenvolvimento Sustentável
3 a 7 de novembro de 2025 - Rio de Janeiro

XXV ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO - XXV ENANCIB

GT 11 – Informação & Saúde

MODELOS DE LINGUAGEM (LLMs) COM GERAÇÃO AUMENTADA POR RECUPERAÇÃO (RAG) NA SAÚDE: A CIÊNCIA DA INFORMAÇÃO NO CENTRO DOS DESAFIOS E OPORTUNIDADES

LANGUAGE MODELS (LLMs) WITH RETRIEVAL-AUGMENTED GENERATION (RAG) IN HEALTHCARE: INFORMATION SCIENCE AT THE HEART OF CHALLENGES AND OPPORTUNITIES

José Eduardo Santarem Segundo – Universidade de São Paulo (USP); Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP)

Modalidade: Trabalho Completo

Resumo: A rápida evolução dos Modelos de Linguagem de Grande Escala (LLMs) e sua integração com técnicas de Geração Aumentada por Recuperação (RAG) tem despertado crescente interesse no domínio da saúde, especialmente devido ao potencial de transformação da gestão do conhecimento médico e o apoio a tomada de decisão clínica. Este artigo tem como objetivo examinar essa convergência tecnológica sob uma perspectiva multidisciplinar, destacando seus impactos, limitações e possibilidades, de forma a apresentar alguns elementos de abordagem, o estado da arte, oportunidades e desafios, do uso e aplicação de LLMs na área da Saúde, e de como a Ciência da Informação pode desfrutar na tríade, Informação, Saúde e Inteligência Artificial. Destaca-se, ainda, o papel central da Ciência da Informação na estruturação, recuperação e disseminação de dados médicos complexos, essencial para garantir a eficácia e a confiabilidade desses sistemas. Como metodologia, foi conduzida uma análise exploratória baseada em uma revisão sistemática de 197 artigos científicos, resultando na identificação de oito eixos temáticos principais que demonstram as aplicações de RAG em saúde, incluindo diagnóstico assistido, gestão de registros eletrônicos e suporte à pesquisa médica. Além disso, propõe-se uma reflexão aprofundada sobre como a Ciência da Informação pode contribuir para a organização, representação e mediação do conhecimento em meio à crescente digitalização da saúde, assegurando que os avanços tecnológicos sejam aliados da equidade, transparência e segurança na prática médica.

Palavras-chave: inteligência artificial; geração aumentada por recuperação (RAG); modelos de linguagem de grande escala (LLM); inteligência artificial na saúde; aplicações clínicas de LLM.

Abstract: The fast evolution of Large Language Models (LLMs) and their integration with Retrieval-Augmented Generation (RAG) techniques has garnered increasing interest in the healthcare domain, particularly due to their potential to transform medical knowledge management and support clinical decision-making. This article aims to examine this technological convergence from a multidisciplinary perspective, highlighting its impacts, limitations, and possibilities. It presents key analytical approaches, the state of the art, opportunities, and challenges regarding the use and application of LLMs in healthcare. Furthermore, it explores how Information Science can engage with the triad of Information, Healthcare, and Artificial Intelligence. It further highlights the central role of Information Science in structuring, retrieving, and disseminating complex medical data, which is essential to ensure the efficacy and reliability of these systems. Methodologically, an exploratory analysis was conducted based on a systematic review of 197 scientific articles, resulting in the identification of eight key

thematic areas demonstrating RAG applications in healthcare, including assisted diagnosis, electronic health record management, and medical research support. Additionally, the study proposes an in-depth reflection on how Information Science can contribute to organizing, representing, and mediating knowledge amid the ongoing digital transformation of healthcare, ensuring that technological advancements promote equity, transparency, and safety in medical practice.

Keywords: artificial intelligence; retrieval-augmented generation (RAG); large language models (LLMs); artificial intelligence in healthcare; clinical applications of LLMs.

1 INTRODUÇÃO

O Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação, que se apresenta em sua vigésima quinta edição, consolidou-se como a maior e mais importante reunião de pesquisadores da Ciência da Informação do Brasil, discutindo temas importantes em toda sua transversalidade, em torno dos seus 12 Grupos de Trabalho.

É notável como a Ciência da Informação vem ampliando seu leque de discussões, abordando temáticas diferentes e é ainda mais interessante observar como muitas das discussões e temas de pesquisa da CI Brasileira nascem neste evento, para depois ganhar destaque e aderência dentro dos grupos de pesquisa espalhados pelas universidades brasileiras do sul ao norte do país.

O Grupo de Trabalho Informação & Saúde, GT11, que apresenta em sua ementa a questão da produção, organização, gestão, disseminação e compartilhamento de dados, informação e conhecimento relacionados a saúde humana, ambiental e animal, permitindo avanços em pesquisas da Ciência da Informação no Ecossistema de Informação em Saúde, tem sido um ambiente profícuo de discussões na CI com abordagem sobre saúde.

Ao observar o que tem sido discutido dentro do GT11, percebe-se que a Ciência da Informação, no Brasil e no Mundo, ainda tem muitos desafios e oportunidades de pesquisa e de avanços científicos relacionados a informação em saúde.

Diante disso, esta pesquisa apresenta um contexto que pretende ampliar a discussão no âmbito da informação em saúde, principalmente devido a crescente digitalização das práticas de saúde, impulsionada por avanços em tecnologias de inteligência artificial, que tem provocado profundas mudanças nos paradigmas clínicos e científicos.

Dessa forma, entende-se que a inteligência artificial, mais especificamente os Modelos de Linguagem de Grande Escala (LLMs, do inglês *Large Language Models*), base das chamadas inteligências artificiais generativas, podem ser mais um importante elemento de discussão,

para formar uma tríade, com informação e com saúde, de oportunidades e desafios de pesquisa, e um terreno fértil para discussão no âmbito da Informação em Saúde. As LLMs são a grande novidade e avanço das IAs, que vem se transformando desde 1943, chegando no modo de atuação que conhecemos atualmente.

De acordo com Russel e Norvig (2004), o primeiro trabalho atualmente reconhecido como IA foi desenvolvido por Warren McCulloch e Walter Pitts em 1943. Nesse estudo, os autores propuseram um modelo matemático de neurônios artificiais baseado na lógica booleana, e foi fundamental para a criação das redes neurais artificiais, servindo como base para o desenvolvimento da IA, especialmente no campo do aprendizado de máquina. O verdadeiro marco da IA como disciplina científica formal ocorreu em 1956, na Conferência de Dartmouth, organizada por John McCarthy, Marvin Minsky, Nathaniel Rochester e Claude Shannon. Nesse evento, o termo "Inteligência Artificial" foi cunhado (Russel; Norvig, 2004).

As LLMs, tornaram-se elementos centrais em aplicações que requerem compreensão, extração e geração de informação textual em larga escala. Esses modelos, treinados com bilhões de parâmetros e trilhões de *tokens*, são capazes de simular respostas com elevado grau de coerência e contextualização. Eles podem atuar de forma genérica, como são conhecidos na sociedade ferramentas como *ChatGPT, DeepSeek, CoPilot, Gemin*i entre outras, ou ainda podem ser utilizados dentro dos chamados ambientes privados, com delimitação de temas e fontes de dados.

Há claramente um entusiasmo na área da saúde em relação ao avanço das LLMs, isso é perceptível principalmente para apoio a tomada de decisão clínica e gerenciamento de pacientes, mas há uma diversidade imensa de possibilidades de aplicações. No entanto, limitações inerentes como alucinações, contexto clínico desatualizado e referências não confiáveis, problemas éticos e de condução na aplicação representam preocupações sérias para sua utilidade clínica, e por isso os modelos de Geração Aumentada por Recuperação (RAG, do inglês *Retrieval-Augmented Generation*) se posicionam como uma solução viável para resolver algumas dessas limitações, melhorando de forma significativa a precisão, relevância e transparência do conteúdo gerado como resposta.

De acordo com Giuffrè (2024, p 2114, tradução nossa), "apesar do potencial, a implementação segura de LLMs na saúde ainda precisa superar obstáculos relacionados à precisão, indicando a necessidade de estratégias que integrem *feedback* humano com treinamento avançado de modelos".

A RAG é uma abordagem hibrida, que combina mecanismos de busca e geração de linguagem com possibilidade de consultas a bases de dados externas em tempo real, enriquecendo as respostas com conteúdos recentes, confiáveis e em escopo fechado.

Na área da saúde de uma forma geral, mas também com abordagens muito específicas, a combinação entre LLMs e RAG emerge como solução promissora. É importante ressaltar que os fazeres da Ciência da Informação, tanto no contexto da Representação como da Recuperação da Informação, tem se apresentado como elementos importantes no contexto dos projetos de dados e inteligência artificial.

Discussões acerca da inserção da Inteligência Artificial na Recuperação da Informação foram realizadas ao longo das últimas décadas, ainda que com abordagens diferentes da proposta neste estudo. No entanto, as novas abordagens de IA, como os *Large Language Models* e o *Retrieval-Augmented Generatio*n, trazem a oportunidade de adequar esses processos às demandas atuais, como o tratamento de grandes volumes de dados e a interação mais natural entre usuários e sistemas (Tavares *et al*., 2024).

Assim, apresenta-se o problema de pesquisa que motivou o desenvolvimento desta pesquisa: "Quais são os desafios, oportunidades e impactos da aplicação de Modelos de Linguagem de Grande Escala (LLMs) e Geração Aumentada por Recuperação (RAG) no ecossistema de Informação em Saúde, sob a perspectiva da Ciência da Informação?"

Desta forma é urgente que possamos discutir as oportunidades e desafios do uso de LLMs dentro do contexto da Informação e Saúde, e por isso este artigo tem como objetivo examinar essa convergência tecnológica sob uma perspectiva multidisciplinar, destacando seus impactos, limitações e possibilidades, de forma a apresentar alguns elementos de abordagem, o estado da arte, oportunidades e desafios, do uso e aplicação de LLMs na área da Saúde, e de como a Ciência da Informação pode desfrutar na tríade, Informação, Saúde e Inteligência Artificial.

O trabalho desenvolvido aqui, apresenta-se com uma metodologia exploratória e uma revisão bibliográfica, tem a pretensão de destacar tópicos de pesquisa que possam ser avenidas de desenvolvimento de pesquisa em aplicações de Inteligência Artificial Generativa na área de Informação em Saúde.

2 MODELOS DE LINGUAGEM E RAG: CONCEITOS E FUNCIONALIDADES

Os LLMs são sistemas baseados em redes neurais profundas que aprendem padrões complexos da linguagem a partir de grandes volumes de dados textuais.

Esses modelos tornaram-se onipresentes nos últimos anos, permitindo acesso gratuito a praticamente qualquer usuário da internet. Eles são revolucionários na geração de texto semelhante ao humano em escala e velocidade sem precedentes. A aplicação de LLMs em praticamente todas as áreas do conhecimento e domínios da computação expôs multidões às capacidades e limitações da IA emergente atual. O acesso fácil e econômico a LLMs, seja por meio de APIs acessíveis via *web, sandboxes* ou *kits* de ferramentas de código aberto, permitiu que uma geração de desenvolvedores e pesquisadores integrasse a IA moderna baseada em LLMs em aplicações cotidianas (Erickson, 2025).

Arquiteturas como *GPT, Bert, DeepSeek* e T5 utilizam um mecanismo/arquitetura Transformer para processar texto com alta eficiência, empregando atenção distribuída e codificação contextual. Esse processamento permite que os modelos gerem, resumam, traduzam e classifiquem informações com alto grau de fluência e coerência. Com isso, os modelos são capazes de compreender e gerar textos altamente relevantes para tarefas como resumo automático, classificação e tradução.

A arquitetura *Transformer*, que foi publicada em 2017 por um conjunto de oito pesquisadores da Google, no artigo seminal “Attention is All”, revolucionou o processamento de linguagem natural ao substituir estruturas recorrentes (RNNS e LSTMs) por mecanismos de atenção, que permitem processamento paralelo e captura de dependências contextuais longas com alta dependência.

> Os modelos dominantes de transdução de sequência são baseados em redes neurais recorrentes ou convolucionais complexas, em uma configuração de codificador-decodificador. Os modelos com melhor desempenho também conectam o codificador e o decodificador por meio de um mecanismo de atenção. Nós propomos uma nova arquitetura de rede simples, o Transformer, baseada exclusivamente em mecanismos de atenção, eliminando completamente a recursividade e as convoluções. Experimentos em duas tarefas de tradução automática mostram que esses modelos são superiores em qualidade, além de serem mais paralelizáveis e exigirem significativamente menos tempo para treinamento (Vaswani *et al.*, 2017, p. 1, tradução nossa).

Os grandes modelos de linguagem generativos (LLMs) revolucionaram a IA ao permitir a geração rápida de textos semelhantes aos humanos, mas enfrentam desafios, incluindo o gerenciamento de informações inacuradas (Fu *et al.*, 2024).

No entanto, mesmo os LLMs mais sofisticados enfrentam limitações quanto à atualidade e completude da informação. Como são treinados com *corpus* estáticos, eles não podem acessar dados posteriores ao seu corte temporal de treinamento. Para contornar essa limitação, surgiu a abordagem de RAG, que combina mecanismos de recuperação semântica com modelos gerativos.

A RAG introduz um módulo de busca embutido ao modelo, permitindo a consulta em repositórios como bases científicas, prontuários eletrônicos, diretrizes clínicas e documentos técnicos.

Diferentemente de sistemas tradicionais de busca, que retornam documentos completos, na prática a RAG utiliza um *pipeline* em duas etapas: primeiro, recupera trechos relevantes de bases de dados externas por meio de *embeddings* vetoriais; em seguida, incorpora os trechos mais relevantes diretamente na geração textual, que o utiliza para gerar respostas baseadas em evidências. A aplicação de RAG permite uma ancoragem informacional mais precisa, reduzindo alucinações (ou seja, respostas factualmente incorretas) e promovendo transparência nas fontes. Tal capacidade é de vital importância para ambientes clínicos, onde precisão e confiabilidade são essenciais.

Na área da saúde, a RAG representa um avanço particularmente significativo. O domínio biomédico é caracterizado por um volume massivo de literatura científica, diretrizes clínicas, registros eletrônicos de pacientes, exames laboratoriais e dados de sensores fisiológicos — uma base de conhecimento complexa, heterogênea e dinâmica.

Os modelos RAG integram mecanismos de recuperação que consultam um banco de dados curado. Essa metodologia representa uma direção promissora para a utilidade clínica de IAs baseadas em LLM, ao integrar fontes externas de conhecimento confiável diretamente em seu *pipeline* de geração. Um mecanismo de recuperação consulta um banco de dados curado contendo literatura médica validada, diretrizes clínicas, protocolos institucionais e orientações específicas para contextos clínicos. Trechos ou documentos relevantes recuperados desse corpus curado servem como contexto autoritativo para um LLM, permitindo a produção de respostas firmemente embasadas em evidências clínicas validadas (Ozmen; Mathur, 2025).

Apresenta-se a seguir uma abordagem sobre a pilha tecnológica básica da estrutura de um *framework* de implementação de LLM com RAG.

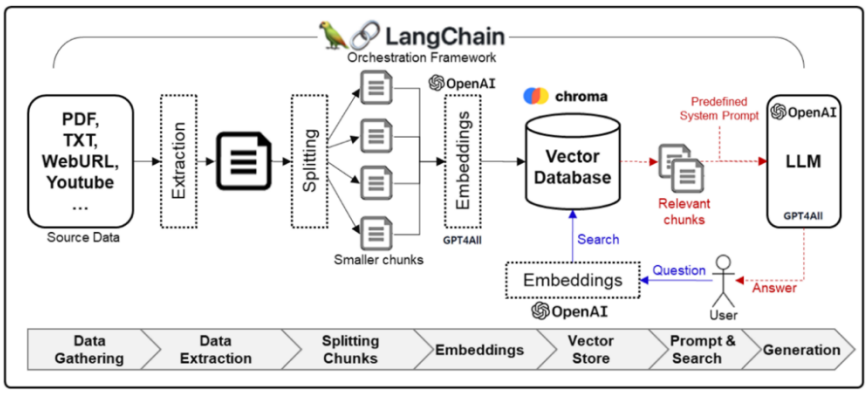
2.1 Framework LLM com RAG

A infraestrutura de uma LLM com RAG deve ser construída por meio de um *framework/pipeline* que possa implementar o serviço de IA generativa usando o modelo de RAG, de forma a orquestrar a integração das tecnologias.

Toda a infraestrutura é baseada em fases ou processos, e cada processo pode ser implementado com um uma grande variedade de possibilidades de soluções (*softwares*), tanto abertas (*software* livre) como proprietárias (pagas).

Para a base da infraestrutura, ou seja, a ferramenta que vai realizar a orquestração dos processos, também há uma diversidade de possibilidades: *Langchain, LangFlow, LlamaIndex, Ragas, Haystack*, N8N, entre outras. A figura 1 é o exemplo de um *framework* orquestrado com a ferramenta *LangChain*, e apresenta um conjunto de fases/processos apresentados a seguir.

Figura 1 – Estrutura de um framework de implementação de LLM com RAG



Fonte: Extraído de Jeong (2023).

No processo de *Data Gathering*, ou coleta de dados, são escolhidas as fontes, baseado nos interesses que se tem no projeto, sendo importante a clareza sobre os objetivos do projeto, se o conjunto de dados será mais abrangente ou mais específico. Dados estruturados e não estruturados podem ser selecionados, isso inclui bancos de dados, vocabulários médicos (SNOMED, CID, MeSH), bases de medicamentos, registros clínicos de prontuários, diretrizes clínicas, documentos técnicos, prescrições, laudos de exames, bases de dados com artigos, áudios, vídeos, entre outros documentos.

A seleção, assim como a estruturação e organização de documentos, é uma gestão de coleções, tarefa estritamente correlacionada a Ciência da Informação.

O processo de *Data Extraction*, é onde são organizados os *scripts* e ou procedimentos manuais para efetivar as entradas dos dados para o *framework*. Nesta fase há preparação dos dados, procedimentos como a conversão de documentos PDFs, transcrição de áudios e vídeos, geração de texto por consultas SQL. Para cada tarefa será necessário elencar ferramentas ou aplicações, já disponíveis como *software* livre ou proprietário. Essa é uma tarefa de Organização da Informação.

Splitting chunks é o processo de dividir documentos grandes (como artigos, laudos, prescrições, documentos técnicos, entre outros) em partes menores, chamadas de "*chunks*", que são mais fáceis de indexar e buscar. Esta etapa visa melhorar a recuperação e a geração de respostas por uma LLM. A ferramenta que hospeda o *framework* geralmente dispõe de algoritmos capazes de realizar a tarefa. Importante ressaltar que as LLMs têm um número máximo de *tokens* por *prompt*, por isso é tão importante quebrar os documentos longos. Isso favorece uma resposta baseada apenas no conteúdo realmente selecionado, evitando ruídos.

Embeddings é a o processo no qual os dados em nível de *chunk* são transformados em representações vetoriais que capturam semântica e relações contextuais. São a "ponte" entre dados não estruturados e modelos de IA, permitindo que RAG acesse e utilize informações de forma eficiente. As bibliotecas da *OpenAI*, que são proprietárias, são as mais utilizadas nesse processo, mas há outras como *SBERT, FastText e Glove*. Durante as consultas, o sistema compara *embeddings* da pergunta com os vetores armazenados para recuperar os trechos mais relevantes, melhorando a relevância da reposta com base em dados externos.

Vector Store (banco de dados vetorial), é um sistema projetado para armazenar, gerenciar e indexar grandes volumes de dados vetoriais de alta dimensionalidade. Os bancos de dados vetoriais são otimizados especificamente para o gerenciamento de informações espaciais representadas como formas geométricas, como linhas e polígonos, é a estrutura onde os *chunks* vetorizados serão armazenados.

Bancos de dados vetoriais são caracterizados por sua velocidade, capazes de encontrar semelhança semântica em bilhões de vetores em milissegundos. Pode-se dizer que são o "cérebro" do RAG para recuperação de informações. Há várias ferramentas que implementam bancos de dados vetoriais, como: *FAISS, Pinecone, PGVector, ChromaDB, Milvus*, entre outros.

Prompt & Search, ou engenharia de *prompt*, refere-se ao processo metódico de elaborar e aperfeiçoar instruções ou entradas para modelos de linguagem, com o objetivo de obter respostas ou comportamentos desejados. Esta técnica ganhou relevância significativa no campo do Processamento de Linguagem Natural (PLN), especialmente com o advento dos grandes modelos de linguagem (LLMs), como a série GPT (*Generative Pre-trained Transformer*) da *OpenAI* e o BERT (*Bidirectional Encoder Representations from Transformers*) do Google.

Este processo demanda uma compreensão clara sobre como uma LLM pode atuar, além do entendimento sobre a coleção de documentos que formaram o corpus da LLM. Entende-se que com conhecimento sobre recuperação da informação, área importante da Ciência da Informação, pode melhorar a formulação do prompt e consequentemente as respostas da LLM. O processo de integração do *prompt* e resultados da busca é uma etapa crítica em sistemas RAG, onde a qualidade da resposta final depende diretamente da harmonização entre a consulta do usuário e as informações recuperadas.

As ferramentas para engenharia de *prompt* podem ser encontradas em APIs de *Advanced Prompting* como do GPT da *OpenAI, Anthropic* (Claude 3) e *Meta* (Llama 3), mas também podem estar disponíveis diretamente em frameworks como *LangChain*. Há uma diversidade de opções para implementar engenharia de *prompt*.

Generation: Em RAG, a geração de respostas pela LLM é o estágio final e mais crítico do *framework*. Antes disso o *framework* recupera *chunks* relevantes, filtra e classifica os resultados por relevância e concatena um contexto procurando dar um formato estruturado ao prompt enviado pelo usuário. Nesta fase, é possível especificar o tipo, o comprimento e o estilo linguístico do texto gerado, é também onde se determinam os parâmetros de controle como temperatura (criatividade x precisão) e limite do tamanho da resposta.

Há uma infinidade de possibilidade de uso de algoritmos para implementar o processo de geração. Apesar das ferramentas mais conhecidas da *OpenAI* (GPT), é possível acessar repositórios como *Hugging Face*[1] e encontrar/escolher entre uma infinidade de ferramentas disponíveis, de acordo com a necessidade e possibilidade do projeto.

O framework de LLM com RAG, orquestrado com a ferramenta *LangChain*, apresentado na Figura 1, não é um padrão único para esse tipo de implementação, mas é uma proposta de modelo a ser utilizada. No exemplo o autor escolhe o *LangChain*, escolhe também

[1] https://huggingface.co/

o *GPT4All* da *OpenAI* para *Embeddings, ChromaDB* para o banco de dados de Vetor e repete o *GPT4All da OpenAi* para a LLM, para recuperar documentos relevantes e gerar as respostas.

Essa abordagem abrangente cobre coleta de dados, processamento, vetorização, busca e geração de respostas, garantindo que o *framework* forneça respostas precisas e contextualmente relevantes às consultas dos usuários. Quanto melhor for a gestão da coleção de documentos do corpus, melhores serão as repostas entregues.

3 PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa caracterizada como exploratória e descritiva, adota como método a Análise Exploratória. Foram aceitos estudos que abordassem, tanto de maneira secundária como principal, a aplicação de *Retrieval Augmented Generation* (RAG) na área da saúde ou medicina. Optou-se por essa abordagem mais abrangente pois a temática ainda é recente e pouco explorada. O protocolo adotado pode ser observado no quadro 1.

Quadro 1 – Protocolo de pesquisa

Objetivos	Identificar o estado da arte e potenciais oportunidades e desafios do uso de RAG e LLM em aplicações para Informação em Saúde.
Questões de Pesquisa	Qual é o estado da arte e os principais impactos do uso de RAGs em LLMs no contexto de Informação em Saúde? Quais são as principais questões relacionadas a oportunidades e desafios para uso de RAGs em LLMs na área da saúde?
Palavras-Chave	("*Retrieval Augmented Generation*" OR "*Retrieval-Augmented Generation*") AND (*healthcare* OR *health* OR *medicine*)
Idioma	Inglês
Seleção	Leitura do título, palavras-chave e resumo dos documentos;
Bases de Dados	*Web of Science e PubMed*
Critérios de seleção	(I) Discussões aplicadas do uso de RAG na área da Saúde (I). Discussões teóricas que abordem o uso de RAG na área de Saúde (E) Não está nos idiomas estabelecidos para a pesquisa; (E) Não aplica ou não discute o tema da pesquisa (E) Não é possível ter acesso ao documento completo; (E) Documentos anteriores a 2023.
Tipos Documentais	Artigos em periódicos
Campos de Extração	1) o objetivo principal do estudo e forma como a temática foi abordada 2) Definições de RAG/LLM e conceitos correlatos 3) Potenciais oportunidades de uso de RAG na área da Saúde 4) Potenciais desafios de uso de RAG na área da Saúde. 5) Referências Relevantes
Sumarização dos Resultados	Os dados serão agrupados em eixos temáticos que permitam uma abordagem mais específica sobre RAG em Saúde.

Fonte: Autor (2025).

Apresentados os procedimentos adotados, os resultados são apresentados na próxima seção.

4 RESULTADOS E DISCUSSÃO DA ANÁLISE EXPLORATÓRIA

Com base no protocolo apresentado, foram recuperados 307 documentos, dos quais 44 foram identificados como duplicados, 66 foram recusados e 197 foram aceitos para compor o corpus da pesquisa, conforme descrito no Quadro 1. Com base nos estudos levantados, foram identificados potenciais oportunidades e desafios da aplicação de RAG na área da Saúde, desta forma realizou-se uma análise temática por eixos, com subtemas de aplicação. Essa organização visa facilitar o entendimento sobre como os LLM com RAG estão sendo aplicados na área da saúde. Na sequência destaca-se desafios e oportunidades que foram identificadas, com destaques a questões que envolvem os conceitos, métodos e práticas da Ciência da Informação. A análise temática baseada na análise exploratória nos levou aos eixos temáticos discutidos a seguir:

O eixo **diagnóstico e apoio a tomada de decisão clínica**, é baseado no uso de LLM para melhorar a precisão diagnóstica, interpretação de exames e suporte a decisão médica. Neste eixo temos avaliações de LLM em radiologia, oftalmologia e patologia, usando imagens médicas, e também apoio a decisões na área de oncologia e determinação de padrões para sintomas psiquiátricos e doenças raras.

O eixo **personalização do tratamento e medicina de precisão** trata da adaptação de LLMs para terapias individualizadas, farmacogenômica e acompanhamento de doenças crônicas, com destaque para integração com *knowledge graphs* (como *carcerkg.org*) e dados genômicos. As LLMs podem cruzar dados de farmacogenômica do paciente com milhões de estudos científicos, sugerindo medicamentos e dosagens ideais enquanto alertam para riscos de reações adversas baseadas no perfil genético individual.

O eixo **educação médica e treinamento,** segmenta as LLMs como ferramentas de ensino, simulação de casos e preparação de exames, incluindo geração de materiais didáticos e tutores virtuais para especialidades.

O eixo ***chatbots* e assistentes virtuais em saúde,** aborda o desenvolvimento de agentes conversacionais para pacientes e profissionais e uso de LLM para adaptação cultural/linguística, como agentes com tradutores. Esses agentes conversacionais, baseados em LLM, podem agendar consultas, esclarecer dúvidas sobre sintomas, lembrar sobre medicamentos e até auxiliar no diagnóstico preliminar. Para médicos e enfermeiros, essas ferramentas ajudam na triagem de casos e acesso rápido a protocolos clínicos.

O eixo de **aplicação em saúde global e populações específicas** aborda a adaptação a contextos locais e grupos vulneráveis, principalmente para regiões de idioma não inglês. Em

regiões com poucos recursos, LLMs com RAG podem auxiliar profissionais de saúde a diagnosticar doenças com base em sintomas locais, considerando variações epidemiológicas. LLMs com RAG podem ser treinados em línguas minoritárias ou dialetos locais, democratizando o acesso à informação em saúde para comunidades indígenas ou refugiadas.

Processamento de dados clínicos e registros eletrônicos, trata de extração, normalização e sumarização. A quantidade de artigos sobre o tema é robusta para mantê-lo como eixo de discussão, podendo ser agregado em qualquer dos eixos anteriores.

O eixo de **aprimoramento de LLMs com técnicas avançadas**, aborda estratégias para melhorar o desempenho das respostas. Uso *de fine-tuning* e *feedback* com humanos são algumas das opções. Esse eixo agrega uma grande quantidade de estudos, porque esse refinamento torna-se cada vez mais importante, para dar mais precisão as respostas.

Por fim, a discussão sobre **limitação e riscos éticos** é um tema que tem gerado muitas pesquisas, porque envolve críticas e desafios na implementação clínica. O uso de IA em contextos clínicos e biomédicos, intensifica discussões sobre responsabilidade ética, arcabouço regulatório e fundamentos epistemológicos da prática médica. Algumas questões: consentimento explícito do titular; anonimização e pseudonimização: que envolve o uso de técnicas para reduzir o risco de reidentificação dos pacientes; e a transparência e acesso, que trata do direito do paciente em saber como e por que seus dados estão sendo processados.

Baseado nos textos analisados é perceptível que a maioria dos eixos privilegia uma visão biomédica e tecnocêntrica, não abordando fatores como renda, educação, racismo estrutural e acesso desigual a tecnologias. Uma exceção parcial é o eixo de saúde global, que toca no tema, mas sem aprofundamento crítico sobre os fatores.

Ainda sobre os resultados, predomina uma visão técnica, com governança implícita baseada em eficiência e escalabilidade, não em equidade, que também abre espaço de oportunidades e desafios de pesquisa.

4.1 As oportunidades e desafios e como a Ciência da Informação se posiciona contexto das LLMs e RAGs

Ao realizar a segmentação em eixos, fica claro que há muitas oportunidades para desenvolvimento de pesquisas na aplicação de LLMs com RAG na área de informação em saúde. Os 197 artigos selecionados permitem ampliar os horizontes e pensar em muitas

perguntas de pesquisa que podem gerar inúmeras oportunidades de atuação, principalmente no contexto da Ciência da Informação.

Em todos os eixos é possível pensar na curadoria e seleção de documentos que poderiam alimentar um *corpus* de uma LLM e dessa forma podem surgir inúmeras questões de pesquisa como: Como determinar a qualidade e confiabilidade dos documentos que servirão como corpus de aprendizado de uma LLM? Modelos de metadados podem favorecer ou acelerar a geração das relações semânticas entre os documentos que compõe o corpus de uma LLM? Qual o impacto do uso de ferramentas de organização do conhecimento como tesauros e ontologias em projetos de LLM? Como construir e aplicar uma política de garantia de ética tanto no uso de dados de pacientes como no uso das respostas por agentes de saúde? Como definir grupos de estudo e análise que possam formalizar e validar o uso de LLM com RAG na área da saúde, e dessa forma determinar a qualidade das respostas oferecidas por LLMs? Como determinar estratégias para construir coleções de documentos que sejam realmente relevantes para o tema específico em que a LLM com RAG deve atuar? Quais são os critérios que determinam as licenças e cessões de direito de autor para que documentos possam integrar o corpus de LLM? Como construir boas práticas e critérios para que os documentos selecionados mitiguem os problemas de análises enviesadas ou então repostas discriminatórias?

Seria possível escrever um texto completo apenas com perguntas que nos inquietam sobre a atuação do profissional da informação no contexto do uso de LLMs com RAG na saúde. Esse é um contexto que mistura oportunidade e desafios que a área precisa avançar, determinar atenção e atuar.

As perguntas apresentadas não envolvem nenhum tipo de desafio tecnológico, *software*, *hardware* ou qualquer coisa do gênero, aborda-se apenas o contexto da informação, um conjunto de oportunidades e desafios, muitos deles ligados a estruturação, organização e representação adequada da informação.

A Ciência da Informação fornece os fundamentos teóricos e metodológicos para estruturar, classificar, indexar e recuperar conhecimento de forma eficiente. Há disciplinas de estudos de usuários, recursos informacionais e gestão de coleções. É onde se discute governança de documentos.

Camillo e Castro Filho (2023, p. 5) indicam que:

> o processo de gestão de coleções é uma atividade de planejamento em bibliotecas que requer a constante tomada de decisão do bibliotecário sobre o acervo, visando aos itens, aos motivos, às finalidades e aos públicos-alvo que justifiquem esforços em favor de uma determinada coleção.

No âmbito da saúde, a diversidade e complexidade dos dados exigem a adoção de estruturas robustas como as ontologias (SNOMED CT, ICD, LOINC) e modelos semânticos interoperáveis, ferramentas que garantem consistência terminológica, e permitem que RAG compreendam e operem sobre dados heterogêneos.

Além disso, a Ciência da Informação adquire novas formas na era digital: curadoria automatizada, validação de fontes, determinação de proveniência para documentos que serão utilizados em corpus que alimentam LLMs, rastreabilidade de evidências e gestão de ciclo de vida dos dados passam a ser funções cruciais.

A curadoria digital envolve a manutenção, preservação e agregação de valor aos documentos digitais ao longo de todo o seu ciclo de vida. O gerenciamento ativo de documentos digitais reduz as ameaças ao seu valor científico em longo prazo e mitiga os riscos de obsolescência digital (DCC, 2025).

É fato que o principal desafio é estar em ambientes que nos permitam desenvolver pesquisa dessa natureza. Pesquisas dessa natureza, com LLM e RAG demandas colaboração com grupos que tenham profissionais de TI além de ambientes tecnológicos capazes de processar pesquisas dessa natureza. É fundamental que os profissionais da informação integrem grupos heterogêneos que envolvam também profissionais de TI e saúde, de forma que seja possível construir modelos que realmente sejam eficientes para os profissionais de saúde e para a sociedade. Além do que está diretamente relacionado a Ciência da Informação, há muitos outros desafios a serem ultrapassados quando pensamos no uso de LLMs com RAG na área de saúde.

É possível pensar na equidade e acesso, visto que há risco de que populações marginalizadas sejam sub-representadas nos dados; a questão da indeterminação de responsabilidades legais entre desenvolvedores, provedores de serviços e profissionais de saúde. A introdução de sistemas automatizados de recomendação impacta a relação entre conhecimento tácito (experiência clínica) e conhecimento explícito (evidências baseadas em dados), exigindo uma revisão crítica dos fundamentos da decisão médica.

Podemos pensar em problemas técnicos como garantir que dados clínicos sensíveis consultados em tempo real não sejam indevidamente memorizados pelos modelos, além

disso novas soluções de IA requer padronização, interoperabilidade e adaptação dos fluxos de trabalho. Modelos com RAG operam com lógica probabilística, dificultando a explicação causal de suas recomendações. Isso pode gerar resistência por parte dos profissionais de saúde e desconfiança por parte dos pacientes.

O uso de dados como base para recomendações automatizadas desloca o eixo do saber clínico da experiência tácita e observação empírica para a inferência estatística. Esse processo desafia os modos tradicionais de validação do conhecimento médico e requer uma crítica epistemológica contínua.

5 CONSIDERAÇÕES FINAIS

Os Modelos de Linguagem com Geração Aumentada têm o potencial de redesenhar profundamente a forma como a informação é utilizada na prática clínica, na investigação científica e na gestão da saúde. No entanto, esse potencial não se concretiza sem enfrentamento dos desafios técnicos, éticos, institucionais e epistemológicos que essa transformação implica.

Mais do que uma mudança tecnológica, a introdução de RAG na medicina exige uma reconfiguração cultural e estrutural, que inclui desde o desenho de sistemas informacionais interoperáveis até a formação de profissionais capazes de atuar nesse novo ecossistema.

A Ciência da Informação pode ocupar papel central na construção de soluções eficazes, equitativas e seguras. Cabe, portanto, fomentar pesquisas interdisciplinares que fortaleçam a integração entre tecnologias de IA e práticas de saúde baseadas em evidências, sensibilidade ética e compromisso social.

É fato a compreensão de que do ponto de vista tecnológico (*software* e *hardware*) temos uma estrutura que está pronta para implementar sistemas de geração aumentada na área da saúde, mas os principais desafios e oportunidades que se apresentam estão ligados a contextos diretamente relacionados a organização da informação. Vive-se uma das revoluções tecnológicas da história da humanidade, e dentro deste contexto há muito trabalho a ser realizado pelos profissionais da Ciência da Informação, tanto neste grupo de trabalho da Associação Nacional de Pesquisa em Ciência da Informação (ANCIB), como também em cada ambiente clínico, hospitalar ou educacional que esteja usando em vias de usar inteligência artificial na área de saúde.

REFERÊNCIAS

CAMILLO, Everton da Silva; CASTRO FILHO, Claudio Marcondes de. Potenciais da coleção infantil sobre meio ambiente na escola: a gestão de coleções na biblioteca escolar. **Revista Brasileira de Biblioteconomia e Documentação**, *[s. l.]*, v. 19, p. 1–26, 2023. Disponível em: https://rbbd.febab.org.br/rbbd/article/view/1860. Acesso em: 15 ago. 2025.

DIGITAL CURATION CENTRE. **What is Digital Curation?.** c2025. Disponível em: https://www.dcc.ac.uk/about/digital-curation. Acesso em: 15 ago. 2025.

ERICKSON, John S. *et al*. LLM experimentation through knowledge graphs: towards improved management, repeatability, and verification. **Journal of Web Semantics**, [*s.L.*], v. 100853, n. 85, p. 100853, maio 2025. Disponível em: http://dx.doi.org/10.1016/j.websem.2024.100853. Acesso em: 13 maio 2025.

FU, Zhaoli *et al*. Application of large language model combined with retrieval enhanced generation technology in digestive endoscopic nursing. **Frontiers In Medicine**, [*s.L.*], v. 1, n. 11, p. 1-14, 6 nov. 2024. Disponível em: http://dx.doi.org/10.3389/fmed.2024.1500258. Acesso em: 10 maio 2025.

GIUFFRÈ, Mauro *et al*. Optimizing large language models in digestive disease: strategies and challenges to improve clinical outcomes. **Liver International**, [*s.L.*], v. 44, n. 9, p. 2114-2124, 31 maio 2024. Disponível em: http://dx.doi.org/10.1111/liv.15974. Acesso em: 12 maio 2025.

JEONG, Cheonsu. A Study on the Implementation of Generative AI Services Using an Enterprise Data-Based LLM Application Architecture. **Advances In Artificial Intelligence And Machine Learning**, [*s.L.*], v. 3, n. 4, p. 1588-1618, 2023. Disponível em: https://doi.org/10.54364/AAIML.2023.1191. Acesso em: 13 maio 2025.

OZMEN, Berk B.; MATHUR, Piyush. Evidence-based artificial intelligence: implementing retrieval-augmented generation models to enhance clinical decision support in plastic surgery. **Journal of Plastic, Reconstructive & Aesthetic Surgery**, [s.L.], v. 104, p. 414-416, maio 2025. Disponível em: http://dx.doi.org/10.1016/j.bjps.2025.03.053. Acesso em: 23 maio 2025.

RUSSELL, Stuart; NORVIG, Peter. **Inteligência Artificial**. Rio de Janeiro: Elsevier, 2004. 1324 p.

TAVARES, Henrique Leal *et al*. Recuperação da informação e inteligência artificial generativa com large language model e retrieval-augmented generation. *In*: ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, 24., 2024, Vitória, ES. **Anais [...].** [*s.L.*]: ANCIB, 2024. p. 1-16. Disponível em: https://enancib.ancib.org/index.php/enancib/xxivenancib/paper/view/2690. Acesso em: 10 maio 2025.

VASWANI, Ashish *et al*. Attention Is All You Need. *In*: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 31., 2017, Long Beach. **Anais [...].** Long Beach, 2017. p. 1-11.