Hybrid Retrieval RAG System for Clinical Decision Support in Portuguese

Lucas Henrique Gonzaga de Sales[1], Pablo Azevedo Sampaio[1], Maria Taciana Holanda Cavalcanti[2]

[1]Departamento de Computação – Universidade Federal Rural de Pernambuco (UFRPE) – Recife – PE – Brazil

[2]Departamento de Medicina Veterinária – Universidade Federal Rural de Pernambuco (UFRPE) – Recife – PE – Brazil

lucas.gonzaga@ufrpe.br, pablo.sampaio@ufrpe.br, maria.vsoares@ufrpe.br

Resumo. *Sistemas de apoio à decisão clínica (CDSS) baseados em modelos de linguagem em sua maioria dependem de bases estáticas e de domínio único, herdando do RAG convencional a fragmentação do conhecimento em chunks. Este trabalho propõe, em caráter de prova de conceito, um CDSS baseado em Retrieval-Augmented Generation (RAG) para recomendação de medicamentos em português, com conhecimento inteiramente fornecido pelo usuário. A arquitetura combina recuperação híbrida (BM25 e embeddings densos com fusão RRF), alimentada por fichas em JSON extraídas de artigos clínicos por um LLM. Em 50 casos sintéticos no domínio fitoterápico, a configuração híbrida recuperou o tratamento-alvo em 95,3% ± 4,6% dos casos (n=3; IC95% de Wilson [86,5–98,9%]), com HR@3 de 0,88 a 0,94, triplicando o baseline lexical. O sistema demonstrou abstenção diante de condições sem evidência na base. Por se apoiarem em casos sintéticos e em base compacta, os resultados caracterizam um estudo preliminar e não uma validação clínica, sendo o sistema destinado exclusivamente a profissionais habilitados.*

Abstract. *Clinical decision support systems (CDSS) based on language models typically rely on static, single-domain knowledge bases, inheriting from conventional RAG the fragmentation of knowledge into chunks. This work proposes, as a proof of concept, a CDSS based on Retrieval-Augmented Generation (RAG) for medication recommendation in Portuguese, with knowledge entirely provided by the user. The architecture combines hybrid retrieval (BM25 and dense embeddings with RRF fusion), fed by JSON records extracted from clinical articles by an LLM. Across 50 synthetic cases in the herbal medicine domain, the hybrid configuration retrieved the target treatment in 95.3% ± 4.6% of cases (n=3; Wilson 95% CI [86.5–98.9%]), with HR@3 of 0.88 to 0.94, tripling the lexical baseline. The system also demonstrated abstention when facing conditions without evidence in the base. Since they rest on synthetic cases and a compact base, the results characterize a preliminary study rather than clinical validation, and the system is intended exclusively for qualified professionals.*

1. Introdução

A inteligência artificial vem sendo utilizada em diversos contextos da saúde como radiologia, desenvolvimento de medicamentos e sistemas de apoio à decisão clínica (CDSS - Clinical Decision Support System). Esses sistemas auxiliam médicos, enfermeiros e demais clínicos na tomada de decisões informadas, aumentando sua eficácia e eficiência por meio de recomendações durante o processo decisório [Gupta et al. 2020].

Parcela significativa dos CDSS contemporâneos fundamenta-se, do ponto de vista tecnológico, em modelos de linguagem de grande escala (LLMs - *Large Language Models*). Uma limitação estrutural recorrente está na tendência à geração de informações factualmente incorretas, fenômeno denominado alucinação, bem como na ausência de rastreabilidade das respostas produzidas. O *Retrieval-Augmented Generation* (RAG) apresenta uma abordagem que mitiga as limitações dos LLMs ao fornecer um contexto documental selecionado no momento da inferência, tornando as respostas rastreáveis às fontes e reduzindo alucinações [Gupta et al. 2024]. Sistemas RAG aplicados à saúde demonstraram resultados superiores à abordagem *zero-shot* em tarefas de consulta clínica [Passinato et al. 2024, Farber et al. 2025, Reis et al. 2025].

Apesar desses avanços, os CDSS descritos na literatura são, em geral, construídos para um domínio fixo (Ex.: oftalmologia) e operam sobre bases de conhecimento definidas previamente, o que limita sua reutilização quando o profissional precisa atuar em outra área, filosofia de tratamento ou quando surgem novas formas de tratamento.

Soma-se a isso uma limitação do RAG convencional: a base de conhecimento é particionada em trechos (*chunks*) de tamanho fixo, o que pode cortar uma informação clínica ao meio e dispersá-la entre unidades distintas. Como esses trechos não são preparados especificamente para o domínio de saúde, mesmo quando unidades sequenciais e complementares são recuperadas a relação entre elas tende a se perder, e uma unidade pode chegar ao gerador com parte da informação relevante faltando, comprometendo a recomendação final. A arquitetura proposta evita esse problema ao garantir que cada unidade da base seja uma ficha completa, reunindo todas as informações necessárias para sustentar uma *recomendação de tratamento*, termo adotado ao longo deste trabalho para designar a saída do sistema.

Diante desse cenário, esta pesquisa tem dois objetivos centrais: propor uma arquitetura de CDSS baseada em LLMs e RAG, composta por um módulo de ingestão de documentos, voltado à construção da base de conhecimento, e por um módulo de recuperação e recomendação de tratamento; e validar sua relevância em comparação a uma abordagem RAG simples no contexto de sistemas de apoio à decisão clínica. A base de conhecimento é constituída pelo próprio profissional a partir de documentos de seu domínio de atuação, e cada recomendação produzida é rastreável às fontes que a fundamentam, garantindo transparência ao processo decisório.

O artigo está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados, a Seção 3 detalha a arquitetura proposta, com os módulos de extração e de recomendação, a Seção 4 descreve a base documental, o *dataset* de avaliação e o design experimental, a Seção 5 discute os resultados, a Seção 6 expõe as limitações e a Seção 7 conclui o trabalho.

2. Trabalhos Relacionados

A revisão da literatura é conduzida a partir de dois eixos: sistemas de apoio à decisão clínica baseados em LLMs e abordagens de resposta a perguntas clínicas com emprego de RAG.

Os sistemas de apoio à decisão clínica têm uma trajetória crescente na inteligência artificial aplicada à saúde, principalmente com o advento dos LLMs. Trabalhos como Med-PaLM 2 [Singhal et al. 2025] e Almanac [Zakka et al. 2024] demonstraram que modelos de grande porte podem atingir desempenho próximo ao humano em questões clínicas estruturadas. O Med-PaLM 2 alcançou 86,5% no *benchmark* MedQA e, em avaliação humana, os médicos preferiram suas respostas às de outros médicos em oito dos nove eixos clínicos avaliados. O Almanac, por sua vez, estrutura o conhecimento clínico em bases estáticas de consulta, o que, assim como o Med-PaLM 2, pode não se adaptar tão bem a formulários regionais ou bases proprietárias não representadas no pré-treinamento.

No que diz respeito ao emprego de RAG em perguntas clínicas, o paradigma tem se mostrado eficaz para suprir as limitações de bases estáticas, permitindo que o modelo recupere informações contextualizadas no momento da geração. Reis et al. (2025) avaliaram RAG sob demanda para geração de instruções de medicamentos a partir de bulas da ANVISA, extraindo apenas a seção de administração e sumarizando-a antes da geração. Com cinco médicos avaliando 119 casos, o modelo RAG (Llama 3.1 8b) atingiu adequação plena (100) contra 93,0 do modelo base, o que evidencia a viabilidade de conteúdo clínico confiável a partir de documentação técnica especializada.

Analisando aplicações de saúde em língua portuguesa, Passinato, Rios e Galvão Filho (2024) desenvolveram um *chatbot* oftalmológico em português baseado em modelos de código aberto com RAG. Com 40 perguntas simulando pacientes sobre documentos de portais hospitalares, compararam três variantes, Naive RAG, HYDE e Rewrite-Retrieve-Read (RRR), avaliadas pelo *framework* Ragas: a RRR destacou-se na relevância do contexto e todas superaram o GPT-3.5 *zero-shot*, o que evidencia a viabilidade de soluções locais e de baixo custo em domínios clínicos sensíveis. Mais recentemente, Farber et al. (2025) propuseram o IaraMed, *chatbot* de saúde feminina em português que integra um classificador de especialização com um *pipeline* RAG composto por recuperação semântica e geração por Gemini 2.0 Flash. O classificador atua como filtro de escopo, redirecionando consultas fora do domínio feminino sem acionar o gerador, o que reduz o risco de alucinações em tópicos não cobertos. Avaliado via *LLM-as-judge*, o sistema atingiu 95,35% de acerto, evidenciando a viabilidade dessas arquiteturas em domínios médicos sensíveis para o português.

Em relação a esses trabalhos, a plataforma proposta neste artigo se diferencia em três aspectos principais. Primeiro, é independente da filosofia de tratamento: enquanto trabalhos anteriores são construídos para domínios específicos, a arquitetura proposta funciona igualmente para fitoterapia, medicina convencional ou qualquer outro sistema que possua documentação especializada, o que torna a plataforma reutilizável em qualquer especialidade. Segundo, é alimentada pelo próprio profissional: não há uma base de conhecimento pré-definida no sistema, o especialista fornece os documentos de seu domínio, e o terceiro, talvez o mais distintivo, é a ausência de pressuposição quanto ao formato dos documentos de entrada. A plataforma não requer que os dados sigam um

padrão tipográfico ou estrutural uniforme, como o das bulas de medicamentos, o que amplia consideravelmente o escopo de aplicação do sistema.

3. Proposta

A plataforma é composta por dois módulos principais, ilustrados na Figura 1. O módulo de extração é executado em etapa inicial pelo profissional e destina-se à construção da base de conhecimento a partir do corpus documental de referência, conforme detalhado na Seção 3.1. O módulo de recomendação é acionado em tempo de uso, recebendo a descrição clínica em português e produz, como saída, uma recomendação de tratamento fundamentada nos registros recuperados, assegurando rastreabilidade e aderência ao contexto documental previamente indexado.

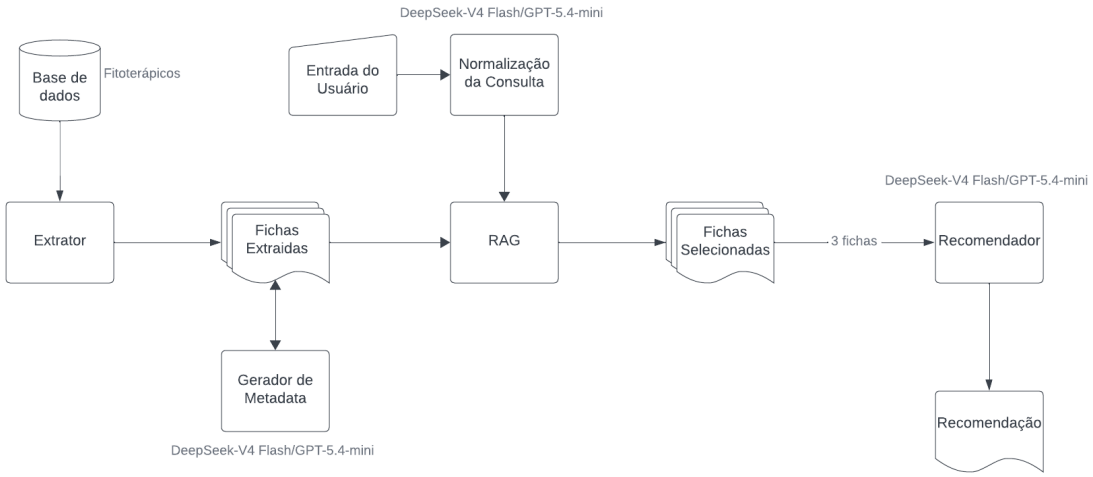


Figura 1. Arquitetura geral da plataforma.

3.1. Extração

A alimentação da base é atribuição do próprio profissional e antecede qualquer consulta. O usuário reúne, no seu domínio de atuação, os documentos que considera confiáveis, como artigos científicos, protocolos clínicos, diretrizes terapêuticas e compêndios especializados, e os submete ao sistema na forma de um arquivo individual ou de um diretório com múltiplos arquivos, sem exigência de padronização tipográfica ou estrutural prévia. A base resultante é incremental: como o processamento registra o progresso por bloco e ignora os blocos já convertidos, novos documentos podem ser acrescentados a qualquer momento sem reprocessar o material anterior. A seleção e a curadoria do corpus permanecem, portanto, sob responsabilidade do especialista, e a qualidade das recomendações fica condicionada à qualidade das fontes escolhidas.

Foi demonstrado por Sundararajan et al. (2025) que a estrutura do *input* responde por mais de 80% da variância nas taxas de erro de LLMs, com o JSON produzindo consistentemente os melhores resultados entre os formatos testados. Assim, o módulo de extração converte o corpus fornecido em registros estruturados e indexáveis, empregando um LLM para interpretar seu conteúdo e produzir fichas no formato JSON.

3.1.1. Processo de Extração

Os documentos são lidos página a página e segmentados em blocos de 4 páginas cada. Cada bloco é enviado ao modelo Gemini (gemini-3.1-flash-lite) via API, acompanhado de um *prompt* de extração que instrui o modelo a: identificar todos os remédios presentes no bloco, preencher para cada remédio os campos da ficha (princípio ativo, dosagem, indicações clínicas, sintomas, contraindicações e posologia), usando o termo "Não encontrado" para campos ausentes, nunca inferindo ou usando conhecimento externo ao trecho enviado. Blocos sem conteúdo clínico reconhecível (capas, índices, glossários) são majoritariamente detectados por pré-filtro de palavras-chave e ignorados sem chamar o modelo, reduzindo o consumo de cota de API.

O *prompt* de extração estrutura-se em duas regras obrigatórias. A primeira determina que todas as indicações de um mesmo princípio ativo sejam consolidadas em uma única ficha, representando variações de indicação e via de administração como listas internas. A segunda instrui o modelo a realizar avaliação crítica do documento antes de classificar cada tratamento, atribuindo ao campo *status_evidencia* um dos seis valores possíveis: *eficaz*, *superior*, *inferior*, *descartado*, *inconclusivo* ou *não avaliado*. Em situações de ambiguidade, o modelo é orientado a preferir *inconclusivo* a *eficaz*.

A estrutura de saída é garantida por Saída Estruturada com modelos Pydantic que definem contratualmente os campos esperados em cada ficha e geram automaticamente o esquema JSON correspondente, que é enviado ao Gemini via parâmetro response_format. A API do Gemini retorna apenas JSON aderente a esse esquema. O processo é idempotente: ao concluir cada bloco, um marcador de progresso é gravado no diretório de saída. Em execuções subsequentes ou após interrupções, blocos já processados são ignorados automaticamente. Ao final, fichas do mesmo remédio que aparecem em múltiplos blocos são unificadas por uma chave de normalização baseada no nome do princípio ativo. Um segundo passo de enriquecimento adiciona metadados estruturados a cada ficha: condições tratadas, categorias terapêuticas e sintomas normalizados sem acentos, melhorando a qualidade lexical para o módulo BM25 do *retrieval*.

Cada ficha produzida pelo módulo de extração é um registro estruturado em JSON que descreve um único tratamento. Seus campos contemplam a identificação do tratamento, indicações clínicas, sinais e sintomas associados, esquemas posológicos, contraindicações e avaliação crítica da evidência, composta por conclusão e justificativa extraída do documento de origem. Atributos de procedência identificam o documento-fonte, o número de fichas consolidadas e a data de geração, enquanto o bloco de metadados reúne os dados de enriquecimento já descritos.

Após a unificação dos princípios ativos, aplica-se um filtro para fichas poluentes e os registros que não possuem *posologia* preenchida são descartados. Esse critério elimina fichas que são geradas a partir de menções acidentais, em especial medicamentos que aparecem nas seções de contraindicação e são capturados pelo modelo como se fossem remédios completos da base extraída. Para o *status_evidencia*, aplica-se regra de precedência conservadora: o status mais restritivo prevalece, de modo que evidências conflitantes resultem sempre na classificação mais cautelosa. O filtro de evidência exclui do índice de *retrieval* as fichas classificadas como *não avaliado* ou *inconclusivo*, que são preservadas em disco para auditoria com o *flag descartado_pelo_artigo: true.* As fichas classificadas como *descartado* ou *inferior* são descartadas, apenas fichas com status *eficaz* e *superior* alimentam as recomendações.

3.2. Recomendação

O módulo de recomendação é acionado a cada consulta. Recebe a descrição em linguagem natural, recupera as fichas mais relevantes e gera uma recomendação de tratamento estruturada fundamentada nessas fichas.

3.2.1 RAG

Em sua configuração básica, o RAG disponibiliza os documentos recuperados como contexto ao gerador no momento da inferência [Gupta et al. 2024].

Rathod (2024) destaca que o RAG endereça o desafio crítico de manter as respostas geradas por IA atualizadas e factualmente precisas.

Song et al. (2025) avaliaram sete modelos em questões clínicas de oftalmologia, com e sem RAG: a acurácia subiu em todos eles, com paridade entre DeepSeek-V3 (80,7%) e GPT-4o (80,9%), o que evidencia a competitividade de modelos *open-source* frente a sistemas proprietários.

A arquitetura proposta parte desse princípio, operando sobre fichas estruturadas, e o expande com duas etapas adicionais ao fluxo convencional: a normalização clínica da consulta antes da recuperação e a recuperação híbrida com reordenação das fichas recuperadas. Essas modificações, descritas em detalhe nas seções seguintes, visam aumentar a precisão e a relevância das respostas geradas pelo sistema.

3.2.2 Recuperação Híbrida com RRF

A recuperação híbrida combina dois métodos complementares via Reciprocal Rank Fusion (RRF), estratégia validada por Mala et al. (2025), que demonstram redução de alucinações em LLMs ao fundir resultados de BM25 com a capacidade semântica dos modelos de *embedding*.

BM25 (lexical): recupera fichas por correspondência de termos. Forte para nomes exatos de remédios e termos técnicos. Os textos são normalizados com remoção de acentos via Unicode, de modo que "cefaleia" e "cefaléia" sejam tratados como equivalentes. Já a recuperação por *dense embeddings* (semântico): recupera fichas por similaridade de cosseno no espaço vetorial. Forte para sinônimos e paráfrases, por exemplo, "dor de cabeça" e "cefaleia" ficam próximos no espaço vetorial mesmo sem compartilhar tokens.

Para a representação vetorial das fichas e das consultas, adotou-se o modelo BERTimbau [Souza et al. 2020], por sua cobertura do português e por sua compatibilidade com GPUs de consumo. Por se tratar de um estudo de caráter inicial, conduzido sob restrição de tempo, outros modelos de *embeddings* não foram comparados experimentalmente. Os vetores das fichas são pré-computados uma única vez e mantidos em memória para consulta em tempo real.

A fusão por RRF combina os dois rankings pela fórmula abaixo, com k = 60:

$$score(d) = \sum \frac{1}{k\ + rank_i(d)}$$

A recuperação híbrida adota o método RetHyb-RRF proposto por Mala et al. (2025), que estende o RRF clássico com ponderação dinâmica baseada em um *query specificity score*: consultas específicas favorecem BM25, enquanto consultas abertas em linguagem natural favorecem o retriever denso.

Após o cálculo dos *scores*, um estágio de *reranking* por *cross-encoder* refina a ordenação das 6 fichas recuperadas. O modelo utilizado, BAAI/bge-reranker-v2-m3 [CHEN et al., 2024], é multilíngue e recomendado para *pipelines* híbridos. O modelo é um *cross-encoder* que analisa cada par em conjunto, produzindo *scores* de relevância mais precisos, ainda que com maior custo de inferência. As 3 fichas mais bem classificadas são encaminhadas ao modelo gerador.

3.2.3 LLM

Os Grandes Modelos de Linguagem (LLMs, do inglês *Large Language Models*) são sistemas de aprendizado profundo treinados sobre corpora textuais extensos [BROWN et al., 2020] que, fundamentados na arquitetura Transformer, exibem capacidades como raciocínio encadeado e generalização a tarefas não vistas [ZHAO et al., 2023], o que amplia sua aplicabilidade a domínios técnicos especializados. Neste trabalho, os modelos DeepSeek-V4 Flash e GPT-5.4-mini foram adotados como geradores primários de recomendações, operando sobre o contexto recuperado pela abordagem RAG, por representarem versões recentes com melhor relação custo-tempo para um estudo de caráter inicial.

3.2.4 Geração da Recomendação

O *pipeline* de recomendação encadeia a normalização da consulta, a recuperação híbrida descrita na Seção 3.2.2 e a geração da recomendação. Na etapa de normalização da consulta (*Step 0*, pré-retrieval), o LLM converte a descrição informal do paciente em terminologia clínica estruturada, produzindo sintomas normalizados, modalidades de agravamento e melhora, hipóteses diagnósticas e características diferenciais. Essa saída enriquece tanto a *query* de recuperação quanto os *prompts* subsequentes. A temperatura foi ajustada por etapa: 0,1 na normalização clínica e na extração estruturada, para respostas determinísticas e reprodutíveis, e 0,3 na geração da recomendação, para maior naturalidade do texto narrativo; em produção, recomenda-se reduzir esses valores.

A etapa de geração realiza uma única chamada ao LLM, que produz a recomendação a partir da leitura direta das fichas recuperadas na etapa anterior. A resposta é estruturada em até seis seções: (1) remédio(s) recomendado(s), (2) posologia, (3) justificativa clínica fundamentada nas fichas recuperadas, (4) tomada de decisão, preenchida quando há empate técnico, diferença de *score* do *reranker* entre o primeiro e o segundo colocado inferior a 0,08, (5) limitações e (6) alternativas, apresentadas condicionalmente quando há outras opções clinicamente válidas. Caso nenhuma ficha cubra a condição apresentada, o sistema abstém-se explicitamente, sinalizando ausência de evidência suficiente em vez de recomendar.

4. Experimentos

A validação da plataforma exige uma base documental com cobertura temática suficiente para diferenciar casos clínicos próximos, condição crítica para métricas de *ranking* terem poder discriminativo. Nessa experimentação, a base foi construída com Memento Fitoterápico da Farmacopeia Brasileira [BRASIL. ANVISA] como catálogo de referência: documento oficial que relaciona as espécies vegetais reconhecidas na fitoterapia brasileira, com indicações, sintomas-alvo e posologia. A partir desse catálogo, foram identificadas as espécies vegetais a serem cobertas e coletados artigos científicos indexados para cada uma delas. Esses artigos, utilizados exclusivamente nos

experimentos, foram submetidos ao módulo de extração descrito na Seção 3.1, que produziu as fichas estruturadas que alimentam o *pipeline*.

O processo de extração produziu aproximadamente 167 fichas brutas. Após a aplicação do filtro de fichas poluentes e do filtro de evidência conforme descrito na Seção 3.1, que manteve apenas os registros com status eficaz ou superior, a base final foi reduzida a 47 fichas úteis.

4.1 Dataset de Avaliação

Foram utilizados 50 casos clínicos em linguagem natural, construídos a partir das indicações registradas no Memento Fitoterápico. Os casos reproduzem a apresentação típica de um profissional ao sistema: descrição de sintomas, modalidades e perfil do paciente, sem menção direta ao remédio esperado. Para cada caso, o remédio-alvo foi definido com base nas indicações clínicas do Memento, servindo como gabarito (*rank esperado*) para cálculo automático das métricas de posicionamento a cada execução. Além dos 50 casos, foram construídos 5 casos-armadilha adicionais, referentes a remédios esperados classificados como *inferior*, *descartado*, *inconclusivo* ou *não avaliado.* Nesses casos, espera-se que o sistema se abstenha de recomendar, por não dispor de evidência comprovada; essa métrica é designada *abstenção.*

4.2 Design dos Experimentos

O design experimental segue uma estratégia de ablação progressiva controlada: a partir do *baseline* mínimo, cada configuração sucessiva introduz exatamente um novo componente arquitetural, isolando sua contribuição marginal nas métricas. A Tabela 1 detalha a progressão:

Tabela 1. Descrição dos experimentos realizados.

ID	Teste	Base de contexto	RAG	Modelo
A1	PDF Bruto BM25	PDFs sem extração	BM25	DeepSeek
B1	Ficha Extraída BM25	Fichas extraídas do PDF	BM25	DeepSeek
B2	Ficha Extraída BM25 com *Metadata*	Fichas extraídas do PDF + *Metadata*	BM25	DeepSeek
B3	Híbrido sem *rerank*	Fichas extraídas do PDF + *Metadata*	BM25 + Denso (RRF dinâmico)	DeepSeek
B4	Híbrido sem *rerank*	Fichas extraídas do PDF + *Metadata*	BM25 + Denso (RRF dinâmico)	GPT-5.4-mini
C1	Híbrido com *rerank*	Fichas extraídas do PDF + *Metadata*	BM25 + Denso (RRF dinâmico) + *cross-encoder*	DeepSeek

O modelo gerador é o DeepSeek-V4 Flash em todas as configurações, o GPT-5.4-mini executa somente o teste Híbrido sem Rerank, como segunda referência de geração sem demandar a execução completa da ablação.

4.3 Métricas

As métricas de *ranking* são calculadas sobre o *rank* esperado retornado pelo *pipeline* e mensuram a qualidade do posicionamento do medicamento-alvo na lista gerada pelo sistema. Duas métricas compõem esse conjunto avaliativo: Acerto e *Hit Rate*@K (HR@K), para $K \in \{1, 3\}$.

A taxa de Acerto é uma métrica binária que verifica se o medicamento-alvo aparece em qualquer posição do conjunto recomendado; a métrica agregada é a proporção de acertos sobre o total de consultas. Como o sistema pode sugerir mais de um medicamento por consulta, a métrica é considerada válida sempre que o alvo figure em qualquer posição do conjunto retornado.

O HR@K, por sua vez, restringe a avaliação às K primeiras posições da lista ranqueada, medindo a fração de consultas em que o alvo aparece dentro desse intervalo, com cortes mais restritivos (HR@1, HR@3) capturando a precisão do topo.

Para avaliar a resposta gerada e não apenas o *retrieval* foi utilizado *LLM-as-Judge* com DeepSeek-V4 PRO como avaliador externo. O avaliador pontua três dimensões em uma escala de 0 a 100: *Context Relevance* (30%), *Faithfulness* (40%) e *Answer Relevance* (30%).

Context Relevance mede a pertinência dos documentos recuperados em relação à consulta. *Faithfulness* verifica se as afirmações da resposta estão ancoradas nesses documentos, recebendo o maior peso por ser o principal vetor de alucinação em recomendações medicamentosas. *Answer Relevance* avalia se a resposta endereça de forma direta e completa a consulta.

Cabe registrar que essas três dimensões constituem operacionalizações aproximadas, apuradas por *LLM-as-Judge* e, portanto, sujeitas à variabilidade do avaliador automático, e não métricas formalmente padronizadas. Sua definição detalhada e o respectivo protocolo de cálculo encontram-se em Passinato, Rios e Galvão Filho (2024).

5. Resultados

A Tabela 2 reúne os resultados das seis configurações testadas nos 50 casos clínicos, com destaque para as métricas de acerto e segurança avaliadas nos 5 casos-armadilha adicionais. Os resultados referentes à qualidade das respostas são detalhados na Tabela 3.

Tabela 2. Resultados comparativos: métricas de *ranking*.

ID	Teste	Modelo	HR@1	HR@3	Acerto	*Abstenção*
A1	PDF Bruto BM25	DeepSeek	-	-	82%	3/5
B1	Ficha Extraída BM25	DeepSeek	0,10	0,30	54%	5/5
B2	Ficha Extraída BM25 com *Metadata*	DeepSeek	0,18	0,32	72%	5/5
B3	Híbrido sem *rerank*	DeepSeek	0,80	0,88	**95,3%±4,6**	4/5
B4	Híbrido sem *rerank*	GPT-5.4-mini	**0,86**	**0,94**	94%	5/5
C1	Híbrido com *rerank*	DeepSeek	0,78	0,92	96%	3/5

No *baseline* A1 (PDF Bruto), HR@1 e HR@3 não se aplicam: sem extração, recuperam-se trechos, e não fichas discretas, de modo que não há posição do alvo a ranquear; o acerto

(82%) indica apenas menção, sem ordenação. A ausência dessas métricas evidencia o ganho da extração, que torna a recuperação mensurável e rastreável.

O ganho decisivo ocorre na transição da recuperação lexical para a híbrida com *embeddings* densos e fusão RRF (B3), em que o *Hit Rate*@3 sobe de aproximadamente 0,30 para 0,90, cerca do triplo. O *reranking* por *cross-encoder* refina a ordenação intermediária (HR@3 de 0,88 a 0,92), mas não eleva o acerto final, indicando que, em base compacta e temática, a fusão híbrida já recupera quase todo o conteúdo útil. Observou-se paridade entre DeepSeek-V4 Flash (95,3% ± 4,6%) e GPT-5.4-mini (94%), com HR praticamente idêntico sob o mesmo *pipeline*.

O acerto reduzido dos *baselines* BM25 (54% e 72%) decorre de abstenções, não de recomendações incorretas: sob a mesma instrução de abstenção, a recuperação lexical fraca leva o sistema a não recomendar quando inexiste correspondência confiável, comportamento prudente que produz abstenção correta em 5/5 armadilhas. A plataforma demonstrou abstenção em condições sem evidência na base (3 a 5 dos 5 casos-armadilha, conforme o gerador), nunca recomendando tratamento sem comprovação, e as falhas residuais (imunomodulador para quadros virais) concentram-se em condições ausentes da base.

Para quantificar a incerteza, limitada pelo número reduzido de casos (50 de acerto e 5 armadilhas), o experimento vencedor (B3, DeepSeek) foi repetido, por questões de custo apenas nessa configuração, em três execuções independentes, com intervalos de confiança de Wilson. O acerto foi de 95,3% ± 4,6% (média ± DP entre 3 execuções independentes), sendo o valor da execução única de 98% o extremo otimista da distribuição. Para a incerteza amostral, cada execução individual carrega um IC95% de Wilson de aproximadamente [86,5–98,9%] (48/50). A variância concentra-se na seleção final do LLM, pois os componentes de recuperação são determinísticos, o que mantém o HR@K estável. Pelo teste de McNemar (com correção de continuidade), apenas a transição para a recuperação híbrida é significativa (B2 vs B3, $p = 0,003$), pois nem o *reranking* (B3 vs C1, $p = 0,62$) nem a troca de gerador (DeepSeek vs GPT-5.4-mini, $p = 1,0$) diferem do acaso, razão pela qual não se afirma superioridade entre essas configurações. Na qualidade das respostas, descrita na Tabela 3, a configuração B3 teve desvio-padrão de ±1,2 ponto entre execuções: a diferença DeepSeek vs GPT (0,1 ponto) situa-se nesse ruído, enquanto a queda de B3 para C1 (cerca de 5 pontos) o excede, indicando que o *reranking* degrada a relevância do contexto.

Tabela 3. Resultados comparativos: métricas de qualidade de contexto e resposta gerada.

ID	Teste	Modelo	*Context Relevance*	*Faithfulness*	*Answer Relevance*	*Total*
B3	Híbrido sem *rerank*	DeepSeek	69,6±3,4	39,8±2,2	**91,3±1,2**	**64,2±0,9**
B4	Híbrido sem *rerank*	GPT-5.4-mini	**70,9**	**43,9**	86,5	64,8
C1	Híbrido com *rerank*	DeepSeek	51	42,3	90,4	59,3

As variantes híbridas apresentaram alta *Answer Relevance* (87–93) mas *Faithfulness* moderado (42–44): parte das fichas eficaz/superior carecem de posologia explícita,

levando o gerador a complementar a dose com conhecimento externo não ancorado na fonte principal, que é um vetor de risco de alucinação.

6. Limitações

As principais limitações do presente trabalho são:

- **Dataset de avaliação sintético:** os cenários clínicos foram construídos pelos próprios pesquisadores, o que pode introduzir viés de confirmação. A validação com casos reais ou com curadoria de especialistas é necessária para resultados mais robustos.
- **Construção da base de artigos:** realizada manualmente por um profissional capacitado, a seleção dos materiais exige senso crítico para avaliar a relevância e a adequação de cada obra ao escopo da pesquisa, além de demandar tempo considerável, o que pode limitar a quantidade de trabalhos incorporados à base e a abrangência das análises.
- **Possível Circularidade dos testes:** como os 55 casos de teste e seus gabaritos derivam do Memento Fitoterápico, mesma fonte da base de fichas, há riscos de circularidade, de modo que os resultados refletem a recuperação do alvo em casos derivados do catálogo de referência, e não o acerto clínico em condições reais.
- **Escopo inicial da avaliação de *embeddings*:** por restrição de tempo, a representação densa foi avaliada com um único modelo (BERTimbau), sem comparação com alternativas multilíngues, de modo que não se pode afirmar que a configuração adotada seja a melhor disponível.

7. Conclusão

Este trabalho implementou e avaliou, como prova de conceito, uma plataforma CDSS baseada em RAG para recomendação de remédios em português. O principal achado é que a qualidade da recuperação e não o modelo gerador determina o desempenho, a recuperação híbrida (BM25 + densa via RRF) triplicou o *Hit Rate@*3 frente ao *baseline* lexical, e modelos *open-source* e proprietário foram equivalentes quando ancorados pelo mesmo *pipeline*. A abordagem estruturada agregou, além da acurácia, rastreabilidade e filtro de evidência ausentes no acesso direto ao PDF, incluindo a capacidade de abster-se diante de condições não cobertas. As limitações remanescentes, *Faithfulness* moderado por fichas sem posologia e abstenção imperfeita em condições fora da base, orientam os trabalhos futuros. Cabe ressaltar que os achados derivam de casos sintéticos e de uma base compacta e, portanto, sustentam a viabilidade da arquitetura, mas não a validação clínica da plataforma, que depende de avaliação humana com profissionais habilitados, além de completar a posologia das fichas e restringir a geração ao conteúdo recuperado.

8. Referências

BRASIL. Agência Nacional de Vigilância Sanitária. Memento Fitoterápico da Farmacopeia Brasileira. 1. ed. Brasília: Anvisa, 2016. Disponível em: https://www.gov.br/anvisa/en/pharmacopeia/arquivos/memento-fitoterapico.pdf. Acesso em: jan. 2026.

Brown, T. B., Mann, B., Ryder, N. et al. (2020) "Language Models are Few-Shot Learners", arXiv:2005.14165. DOI: 10.48550/arXiv.2005.14165.

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D. e Liu, Z. (2024) "BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation", In: Findings of the ACL 2024, ACL. DOI: 10.18653/v1/2024.findings-acl.137.

Farber, M. et al. (2025) "IaraMed: A Women's Healthcare Chatbot for Portuguese Speakers", In: Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS). DOI: 10.5753/sbcas.2025.7897.

Gupta, P., Ramachandran, A., Keerthi, A., Dave, P. S., Giridhar, S., Kallapur, S. S. e Saikia, A. (2020) "An Overview of Clinical Decision Support System (CDSS) as a Computational Tool and Its Applications in Public Health", In: EAI/Springer Innovations in Communication and Computing, Springer.

Gupta, S., Ranjan, R. e Singh, S. N. (2024) "A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions", arXiv:2410.12837. https://arxiv.org/abs/2410.12837.

Mala, C. S., Gezici, G. e Giannotti, F. (2025) "Hybrid Retrieval for Hallucination Mitigation in Large Language Models: A Comparative Analysis", arXiv:2504.05324. https://arxiv.org/abs/2504.05324.

Passinato, B., Rios, R. e Galvão Filho, M. V. (2024) "Integração de Modelos de Linguagem e RAG na Criação de Chatbots Oftalmológicos", In: Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS). DOI: 10.5753/sbcas.2024.2228.

Rathod, P. e Anurag (2024) "Efficient Usage of RAG Systems in the World of LLMs", TechRxiv. DOI: 10.36227/techrxiv.171625877.73379410/v1.

Reis, D., Reis, Z. e Rocha, L. (2025) "Instruções de Uso de Medicamentos Suportadas por RAG em Grandes Modelos de Linguagem", Revista Eletrônica de Iniciação Científica em Computação, v. 23, n. 1. DOI: 10.5753/reic.2025.6039.

Singhal, K., Tu, T., Gottweis, J. et al. (2025) "Toward Expert-Level Medical Question Answering with Large Language Models", Nature Medicine, v. 31, p. 943–950. DOI: 10.1038/s41591-024-03423-7.

Song, S., Peng, K. C., Wang, E. T. e Liu, T. Y. A. (2025) "Open-Source DeepSeek-R1 Outperforms Proprietary Non-Reasoning Large Language Models With and Without Retrieval-Augmented Generation", medRxiv. DOI: 10.1101/2025.09.12.25334809.

Souza, F., Nogueira, R. e Lotufo, R. (2020) "BERTimbau: Pretrained BERT Models for Brazilian Portuguese", In: Intelligent Systems – BRACIS 2020, LNCS, v. 12319, Cham, Springer, p. 403–417. DOI: 10.1007/978-3-030-61377-8_28.

Sundararajan, B., Sripada, S. e Reiter, E. (2025) "Input Matters: Evaluating Input Structure's Impact on LLM Summaries of Sports Play-by-Play", arXiv:2510.21034. https://arxiv.org/abs/2510.21034.

Zakka, C., Shad, R., Chaurasia, A. et al. (2024) "Almanac – Retrieval-Augmented Language Models for Clinical Medicine", NEJM AI, v. 1, n. 2, p. AIoa2300068. DOI: 10.1056/AIoa2300068.

Zhao, W. X., Zhou, K., Li, J. et al. (2023) "A Survey of Large Language Models", arXiv:2303.18223. DOI: 10.48550/arXiv.2303.18223.