Revista Eletrônica de Iniciação Científica em Computação, 2025, 23:1
ISSN: 3085-8461 • doi: 10.5753/reic.2025.6039

Artigo de Pesquisa/Research Paper

Instruções de Uso de Medicamentos Suportadas por RAG em Grandes Modelos de Linguagem

Drug Usage Instructions Supported by RAG in Large Language Models

Davi dos Reis [Universidade Federal de São João del-Rei | *davireisjesus@aluno.ufsj.edu.br*]
Zilma Reis [Universidade Federal de Minas Gerais |*zilma.medicina@gmail.com*]
Leonardo Rocha [Universidade Federal de São João del-Rei |*lcrocha@ufsj.edu.br*]

Departamento de Ciência da Computação, Universidade Federal de São João del-Rei, Praça Frei Orlando, 170, Centro, São João del-Rei, Minas Gerais, 36307-352, Brazil.

Resumo. Sistemas de prescrição de medicamentos padronizados visam melhorar a legibilidade das instruções, mas ainda há desafios na sua personalização. Este trabalho propõe uma nova abordagem baseada em Grandes Modelos de Linguagem com Geração Aumentada de Recuperação (RAG) utilizando bulas de medicamentos. Testamos três modelos em 119 casos ambulatoriais, com instruções avaliadas por médicos quanto à adequação, clareza e personalização. Nossa proposta melhorou significativamente a adequação (100 vs. 93,0) e a clareza (95,0 vs. 90,0), reduzindo erros e minimizando as alucinações. Nossa proposta aumenta a segurança das instruções integrando com informações confiáveis, mas a validação humana segue sendo essencial.

Abstract. Prescription systems improve medication treatment by standardizing content and enhancing legibility, but challenges remain in personalizing instructions. This study proposes and evaluates a new approach based on Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG) using patient information leaflets (PIL). Three models were tested on 119 outpatient cases, with instructions assessed by physicians for adequacy, clarity, and personalization. Our proposal significantly improved adequacy (100 vs. 93.0) and clarity (95.0 vs. 90.0), reduced errors while minimizing hallucinations. Our proposal can enhance medication safety by integrating authoritative information, but human validation remains essential for safe implementation.

Palavras-chave: Inteligência Artificial, Grandes Modelos de Linguagem, Retrieval-Augmented Generation, Prescrição Eletrônica, Segurança de Medicamentos.

Keywords: Artificial Intelligence, Large Language Models, Retrieval-Augmented Generation, Eletronic Prescribing, Medication Safety.

Recebido/Received: 16 May 2025 • **Aceito/Accepted:** 30 June 2025 • **Publicado/Published:** 11 July 2025

1 Introdução

Os sistemas de prescrição eletrônica são ferramentas importantes que mesclam tecnologia digital e assistência médica, fornecendo prescrições claras e padronizadas diretamente às farmácias. Essa transmissão direta reduz os riscos associados a pedidos manuscritos ilegíveis, uma fonte comum de erros de medicação Klepser *et al*. [2016], aderindo a padrões regulatórios e de segurança para garantir a conformidade. Todavia, sob a perspectiva dos pacientes, a falta de clareza na linguagem da prescrição dificulta a compreensão e pode confundir os pacientes sobre o uso adequado de seus medicamentos Rezende *et al*. [2020]. Pesquisas destacaram que as informações atuais escritas ao paciente nas prescrições frequentemente não comunicam os riscos essenciais associados ao uso indevido de medicamentos Gosselin *et al*. [2021] de forma eficaz. Além disso, as inconsistências comportamentais entre os prescritores e os dados fragmentados sobre o uso de medicamentos revelam a necessidade de mais aprimoramentos no sistema Shiffman *et al*. [2011]. Além da padronização, abordar a lacuna no fornecimento de instruções de medicação personalizadas, adaptadas a cada paciente é um fator essencial para melhorar os resultados da assistência médica.

Em um trabalho recente Reis *et al*. [2024], exploramos o potencial dos Grandes Modelos de Linguagem (*Large Language Models* - LLMs) para aprimorar os sistemas de prescrição eletrônica, gerando instruções personalizadas sobre como os pacientes devem usar os medicamentos. Adotando o padrão nacional brasileiro de prescrição eletrônica para a atenção primária, propomos e avaliamos três diferentes prompts - genéricos, refinados e com viés mitigado - considerando dois LLMs - um fechado (GPT4) e um aberto (Llama 3.1), em diversos cenários de pacientes. Os resultados demonstraram que as soluções baseadas nos LLMs podem fornecer instruções precisas. Além disso, observamos que modelos abertos, como o Llama 3.1, foram capazes de apresentar resultados bastante consistentes. Apesar dos bons resultados, as soluções avaliadas ainda apresentaram limitações em cenários complexos que exigem instruções altamente específicas ou condicionais, como medicamentos que exigem uso prolongado ou restrições de uso específicas.

Para mitigar essas limitações, nesse trabalho apresentamos uma avanço introduzindo o uso de Geração Aumentada de Recuperação (*Retrieval-Augmented Generation* - RAG) Xiong *et al*. [2024]; Zakka *et al*. [2024]. O RAG integra informações de fontes externas aos modelos, aprimorando significativamente o conhecimento contextual disponível para

os LLMs durante a preparação da prescrição. Especificamente, propomos o uso do RAG para introduzir informações obtidas por meio das bulas disponibilizadas pela Agência Nacional de Vigilância Sanitária (Anvisa) [1]. Utilizar todo o conteúdo disponível no banco de dados da Anvisa é um processo computacional caro, que pode aumentar o tempo de resposta e afetar a qualidade dos resultados ao incluir informações irrelevantes para a medicação. Dessa forma, propomos uma estratégia de RAG sob demanda, utilizando exclusivamente a seção específica relacionada à administração da bula do medicamento prescrito.

Para avaliar nossa abordagem, consideramos 119 cenários ambulatoriais para preparar prescrições como entrada para LLMs. Avaliamos as instruções do paciente geradas pelo Llama 3.1 de código aberto, usando um prompt padrão - modelo 1, prompts aprimorados com orientação estruturada - modelo 2 e prompts aprimorados pelo RAG incorporando informações das bulas - modelo RAG. Cinco médicos pontuaram independentemente as instruções geradas pelos modelos, avaliando adequação, clareza, personalização e qualidade em comparação com textos de referência. Nossos resultados indicaram que a introdução do RAG melhorou significativamente a adequação das instruções (93,0 vs. 94 vs. 100) (15,0), clareza (90,0 vs. 92 vs. 95,0), para o modelo 1, modelo 2 e modelo RAG, respectivamente. O modelo com RAG reduziu erros críticos, como instruções incorretas e incompletas e imprecisões factuais. Nossas descobertas ressaltam o papel da Inteligência Artificial na promoção do uso mais seguro de medicamentos, incorporando informações confiáveis sobre medicamentos, mantendo uma linguagem concisa e amigável ao paciente. A validação humana da saída garante uma implementação segura e sem erros.

2 Trabalhos Relacionados

Os Grandes Modelos de Linguagem (*Large Language Models* LLMs) têm sido amplamente adotados em várias áreas do conhecimento, inclusive na medicina. Eles já demonstraram potencial em tarefas complexas, como a previsão de possíveis doenças usando dados médicos e informações de sensores de dispositivos inteligentes Yang *et al*. [2024]. Notavelmente, os LLMs também têm desempenhado um papel significativo nas tarefas de redação de notas médicas, garantindo eficiência e precisão Goyal *et al*. [2024]. Entretanto, sua aplicação na prescrição médica e no cenário do presente trabalho ainda representa um campo emergente repleto de oportunidades de inovação e desenvolvimento.

Em um trabalho recente Reis *et al*. [2024] exploramos os LLMs usados para gerar instruções de medicação precisas e personalizadas, com o objetivo de melhorar os sistemas de prescrição eletrônica. O estudo avaliou uma série de cenários ambulatoriais, incluindo doenças crônicas, condições agudas e cuidados preventivos, dentro do sistema brasileiro de registro eletrônico de saúde (EHR) para a atenção primária à saúde. Esses cenários consideraram diferentes medicamentos, vias de administração e condições do paciente. As interações com os LLMs foram estruturadas em três prompts: um genérico inicial, uma versão aprimorada que incorporava sugestões de conteúdo e um prompt estruturado para mitigação de viés de preconceito Reis *et al*. [2024]. Nossos resultados indicaram que soluções baseadas em LLMs utilizando prompts cuidadosamente elaborados podem gerar instruções adequadas e personalizadas. No entanto, nosso estudo também destacou limitações, especialmente em cenários complexos que exigem instruções altamente específicas ou condicionais, como medicamentos para uso prolongado ou com restrições específicas. Essas informações geralmente estão ausentes dos conjuntos de dados de treinamento de modelos avançados, como o GPT-4, ou de alternativas de código aberto, como o LLaMA 3.1, que foram utilizados no estudo Reis *et al*. [2024].

Para atenuar a falta de informações específicas para um determinado contexto, a literatura vem utilizando técnicas de *RAG* Xiong *et al*. [2024]; Zakka *et al*. [2024], que adicionam informações contextuais ao prompt antes de submetê-lo ao LLM. O RAG ajuda a atenuar problemas como a geração de informações incorretas ou alucinações, fornecendo respostas mais precisas e confiáveis. Por exemplo, Xiong *et al*. [2024] apresentaram o MIRAGE, uma plataforma para comparar sistematicamente diferentes configurações de RAG, avaliando seu desempenho em tarefas médicas. Em Zakka *et al*. [2024], o RAG foi usado para explorar a segurança dos modelos LLM em ambientes clínicos. A proposta envolveu pesquisas on-line em vários domínios médicos, integrando informações de fontes externas confiáveis. A abordagem visa aumentar a precisão e a confiabilidade das tarefas de perguntas e respostas, permitindo que o modelo tenha acesso a conteúdo atualizado e relevante para fornecer respostas mais seguras e informadas.

Embora o RAG ofereça benefícios, sua implementação tradicional tem limitações. A incorporação de vários documentos como contexto geralmente leva a prompts excessivamente longos, o que pode aumentar a complexidade computacional, reduzir a eficiência e comprometer a relevância das informações fornecidas ao modelo Xiong *et al*. [2024]. Para superar essas limitações, nossa proposta apresenta uma abordagem eficiente de RAG sob demanda, na qual o geração aumentada é direcionada de acordo com cada medicamento prescrito, conforme detalhamos a seguir.

3 Proposta

Nossa proposta adota uma abordagem colaborativa, combinando o desenvolvimento de um prompt cuidadosamente planejado e guiado por critérios previamente definidos Reis *et al*. [2024] com o uso de RAG para incorporar conhecimento especializado.

3.1 Engenharia de Prompt

O principal método de interação com os LLMs é por meio de prompts, uma ferramenta essencial para a geração de resultados de alta qualidade Lu *et al*. [2024], especialmente quando são necessárias informações e instruções específicas para tratar de problemas complexos. O prompt desenvolvido neste estudo foi projetado para orientar as respostas do modelo de modo a abranger os principais aspectos e atenuar os erros relatados por Reis *et al*. [2024]. Com as contribuições de especialistas em saúde, definimos um conjunto de nove requisitos essenciais para o desenvolvimento do prompt:

[1] `https://consultas.anvisa.gov.br/#/bulario/`

1. Especificar o nome do medicamento e a forma de apresentação usando valores numéricos para doses e concentrações;
2. Use imperativo para comandos;
3. Fornecer instruções com base na via de administração;
4. Seguir a ordem cronológica das ações que o paciente deve executar;
5. Mencionar quaisquer etapas de preparação antes de explicar como administrar o medicamento;
6. Indicar requisitos especiais de armazenamento, se aplicável;
7. Especificar a frequência, a hora do dia e se o medicamento deve ser tomado nas refeições ou eventos diários;
8. Para tratamento de longo prazo, aconselhar o paciente a procurar atendimento médico antes de concluir a prescrição.
9. Instruir o paciente a armazenar o medicamento com segurança, fora do alcance das crianças e em sua embalagem original, e não compartilhá-lo com outras pessoas.

Esses requisitos foram usados para aprimorar a comunicação com o LLM por meio da engenharia de prompts, que consiste em desenvolver estrategicamente a entrada e estruturá-la para orientar o modelo na geração de respostas alinhadas com os resultados desejados. Por meio de um processo iterativo intensivo que envolveu sucessivas tentativas e ajustes, desenvolvemos um prompt otimizado, com o objetivo de produzir prescrições de medicamentos que fossem claras e adequadas, na tentativa de torná-las facilmente compreensíveis para os pacientes. O prompt resultante desse processo é apresentado a seguir.

1. Você é um médico que deve fornecer instruções diretas e objetivas.
2. Comece se dirigindo à pessoa pelo nome dela.
3. Apresente o medicamento exatamente como consta na entrada, incluindo a forma de apresentação, usando números para indicar dose, concentração e duração do uso.
4. Se houver alguma preparação incluída nas informações do medicamento, mencione-a antes das instruções de uso do medicamento.
5. Substitua a palavra genérica "tomar" por uma palavra que seja consistente com a via de administração do medicamento.
6. Não instrua a pessoa a remover a embalagem se o medicamento não for uma cápsula ou comprimido.
7. Se houver algum tipo especial de armazenamento nas informações do medicamento, mencione-o.
8. Apresente as instruções no modo imperativo e não reforce a necessidade de usar o medicamento.
9. Evite jargões técnicos.
10. Apresente cada ação que o paciente deve realizar em ordem cronológica.
11. Não mencione o médico ou terceiros ou interações medicamentosas.
12. Forneça instruções de uso de acordo com a via de administração, por exemplo, para uso oral, indique líquido suficiente para engolir, avise para não abrir a cápsula ou dividir o comprimido; se o medicamento for líquido, indique a quantidade da dose e como medi-la; para cremes, pomadas e géis, indique o tamanho da área de aplicação.
13. Não repita palavras.
14. Não use excessivamente a palavra "comece".
15. Especifique o nome completo da via de administração, por exemplo, inalação oral.
16. Não mencione validade ou prazo de validade.
17. Sempre informe o nome completo do medicamento.
18. Exemplo: Ana Beatriz, tome 1 comprimido de medicamento de 300 mg a cada 8 horas. Retire o comprimido da embalagem somente quando for tomá-lo e tome-o com água para que você engula o comprimido inteiro, sem esmagá-lo ou mastigá-lo. Continue por 7 dias.

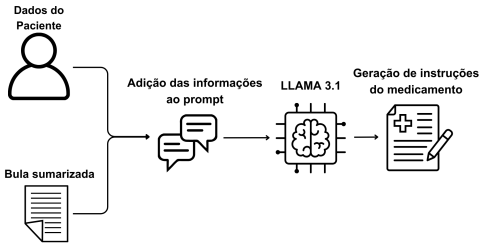
As linhas 2 a 7 do prompt atendem aos seis primeiros requisitos definidos. O prompt também inclui solicitações que não estão diretamente relacionadas a nenhum dos requisitos, progressivamente adicionados durante o processo iterativo para atenuar os vieses identificados nos testes. Um exemplo foi a necessidade de evitar instruções repetitivas como "Comece a tomar...", ajustado na linha 14 do prompt. Para cumprir o requisito 7, foi criada uma tabela de sugestões de cronograma baseada na dosagem prescrita, superando limitações matemáticas dos LLMs. O requisito 8 foi tratado com uma condição para inserir um alerta sobre renovação da prescrição, caso a duração do uso seja indeterminada. O requisito 9, aplicável a todas as prescrições, foi incorporado da mesma forma.

3.1.1 Retrieval-Augmented Generation (RAG)

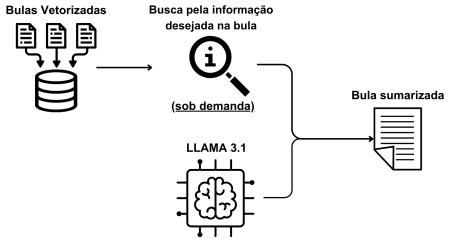
Procuramos reduzir a complexidade do prompt, aumentando a probabilidade de o modelo atender a todos os requisitos estabelecidos com maior precisão e consistência. Apesar desses esforços, vários medicamentos têm prescrições mais complexas devido às suas características particulares, que podem não estar totalmente representadas nos dados usados no treinamento original do LLM. Para superar isso, propomos o uso da técnica RAG, que visa melhorar o desempenho do modelo ao incorporar informações de fontes externas, enriquecendo significativamente os dados contextuais disponíveis para o LLM durante a geração de instruções. A **Figura 1a** apresenta uma representação visual do processo.

Primeiramente realizamos a coleta no site da Anvisa[2] das bulas de informações ao paciente para todos os medicamentos disponibilizados e armazenamos localmente. Realizar um único RAG com todo o conteúdo para o LLM é um processo computacional caro, podendo aumentar significativamente o tempo de resposta e afetar a qualidade dos resultados ao incluir informações irrelevantes para a medicação. Para resolver esse problema, adotamos uma estratégia de RAG sob demanda. Nesse método, quando um medicamento é prescrito, procuramos exclusivamente a bula do paciente desse medicamento específico em nosso banco de dados. Ainda sim, essa bula pode conter uma quantidade substancial de informações que não são relevantes para o cenário específico. A inserção dessas informações na íntegra poderia

[2] `https://consultas.anvisa.gov.br/#/bulario/)`



(a) Visão geral do RAG



(b) Sumarização da Bula

Figura 1. Processo Completo de Retrieval-Augmented Generation (RAG)

afetar negativamente a precisão do modelo ou resultar em um excesso de detalhes técnicos complexos, o que poderia prejudicar a compreensão do paciente. Outro problema é o risco de o paciente não aderir ao tratamento devido à apresentação excessiva dos efeitos colaterais, que, embora relatados, geralmente são raros. Portanto, aprimoramos nossa abordagem com uma etapa adicional. Em vez de incorporar toda a bula ao LLM, fornecemos ao LLM apenas a seção específica relacionada à administração. Para isso, extraímos informações textuais da bula, segmentando-a em partes de igual tamanho. Cada bloco foi posteriormente convertido em uma representação vetorial. Para garantir que apenas a seção correspondente às instruções de uso fosse selecionada, aplicamos a métrica de distância de cosseno para identificar os dois pedaços cuja representação vetorial estava mais próxima da frase "6. Como devo usar este medicamento?". Mesmo essa seção específica que trata de como o medicamento deve ser usado ainda é bastante densa e grande. A execução do RAG com essas informações ainda poderia impedir que o LLM extraísse e usasse informações essenciais para gerar as instruções. Assim, nossa proposta introduziu uma nova etapa antes da RAG propriamente dita, na qual as informações extraídas da bula foram resumidas pelo LLM, conforme ilustrado na **Figura 1b**. As informações extraídas dessa etapa são adicionadas ao prompt que solicita a geração de instruções. Essa abordagem busca equilibrar a necessidade de detalhes com a clareza e a facilidade de uso das informações fornecidas. A seguir, detalhamos como avaliamos nossa proposta.

4 Avaliação Experimental

4.1 Ambiente Experimental

4.1.1 Casos Clínicos

Para avaliar a aplicabilidade das nossas propostas, criamos um banco de dados com casos clínicos hipotéticos, elaborados com auxílio de cinco profissionais de saúde altamente experientes, com autoridade prescritiva e mais de 20 anos de experiência em prescrições médicas, especialistas em informática em saúde, farmacêuticos e linguistas aplicados. As prescrições de medicamentos simuladas foram modeladas com base no padrão do e-SUS nacional[3]. O sistema fornece aos prescritores acesso completo ao histórico de medicamentos de um paciente, permitindo que eles prescrevam medicamentos específicos usando menus estruturados para selecionar o nome da marca, a concentração, a dosagem, a via de administração, a frequência de uso e a quantidade. A base de dados contém 83 prescrições, 39 (47%) para pacientes do sexo feminino e 44 (53%) para pacientes do sexo masculino, com 83 princípios ativos em 12 formas de apresentação diferentes indicadas para oito vias de administração distintas. Essas prescrições foram selecionadas para abranger várias condições médicas e opções de tratamento, proporcionando uma avaliação abrangente do protótipo do software. Das prescrições, 61 (73,5%) eram para uso temporário e 22 (26,5%) para uso prolongado.

4.1.2 Repositório de Bulas

Realizamos uma coleta, no site da Anvisa, das bulas dos medicamentos correspondentes aos 83 princípios ativos. Cada bula foi processada e armazenada em sua forma vetorizada. A vetorização consiste em transformar o texto original da bula, segmentado em partes, em representações vetoriais (numéricas), utilizando o modelo multilingual M3-Embedding Chen *et al*. [2024]. Esse processo busca capturar as informações mais relevantes de cada seção e suas características, facilitando seu uso em métodos de aprendizado de máquina e recuperação de informações.

4.1.3 Metodologia de Avaliação

Para cada uma dos 83 casos clínicos criados, as instruções elaboradas pelos especialistas, consideradas nosso padrão de referência, foram comparadas com instruções produzidas por três modelos baseados em LLMs, todos instanciados utilizando o LLM Llama 3.1 com 8b de parâmetros: (a) abordagem com o melhor resultado relatado em Reis *et al*. [2024] ("Llama3.1 8b Modelo 1"); (b) abordagem utilizando o LLM com prompt apresentado na seção 3.1 ("Llama3.1 8b Modelo 2"); e abordagem utilizando o LLM com nossa proposta de RAG ("Llama3.1 8b Modelo RAG"). A escolha pelo Llama 3.1 8b se deu por ser um LLM de código aberto, e portanto reprodutível, e com resultados muito próximos de opções proprietárias, como GPT 3.4, conforme relatamos em Reis *et al*. [2024]. As instruções geradas foram avaliadas por cinco médicos experientes, chamados de juízes, de forma justa e imparcial. O Conselho Nacional de Pesquisa do Brasil aprovou o estudo (**CAAE: 78883924.7.0000.5149**). No total, foram realizadas 357 análises, 119 para cada modelo mencionado. A distribuição das avaliações seguiu um esquema em que quatro juízes realizaram 24 análises por modelo, enquanto um juiz realizou 23 análises. Para garantir uma base consistente e comparativa, nove avaliações foram comuns a todos os juízes.

4.1.4 Critérios de Avaliação

Os juízes avaliaram quatro critérios: (1) **Adequação**: refere-se à capacidade do modelo de gerar respostas precisas e adequadas ao objetivo, ao contexto e às necessidades do usuá-

[3] `https://sisaps.saude.gov.br/esus/`

rio; (2) **Clareza**: está relacionada à capacidade do modelo de apresentar informações de forma compreensível, organizada e sem ambiguidades; (3) **Personalização**: está relacionada à capacidade do modelo de gerar linguagem relevante e eficaz para tornar a interação natural, prática e alinhada com o objetivo do usuário; e (4) **Comparação**: avalia até que ponto as instruções geradas pelos modelos são melhores do que o padrão ideal na visão do avaliador. Para todos eles, foram considerados valores entre 0 e 100, onde quanto mais próximo de 100, melhor a avaliação. Uma exceção é o critério Comparação em que o valor 50 significa que a saída do modelo é equiparável ao padrão de referência. Os juízes também relataram seu grau de **Confiança**, que corresponde à capacidade autopercebida de analisar criticamente as instruções geradas. O quinto critério (similaridade) foi avaliado de forma automática aplicando-se a **Similaridade** de Cosseno para medir a proximidade entre o texto gerado por um dos modelos e o padrão de referência, resultando em um valor entre -1 e 1. Por fim, os juízes também puderam classificar os erros encontrados nas instruções em sete tipos, com base em Tam *et al*. [2024]: **Tipo 1** - instruções podem levar ao uso incorreto deste medicamento; **Tipo 2** - instruções de uso são contraditórias ou vagas; **Tipo 3** - falta de instruções essencias para uso; **Tipo 4** - há erros factuais (não médicos); **Tipo 5** - há instruções que não são apoiadas por evidências científicas; **Tipo 6** - há instruções não solicitadas; e **Tipo 7** - há alucinações (informações fabricadas/inventadas). Os testes pareados de Friedman compararam as medianas e o teste post-hoc de Bonferroni forneceu comparações entre pares. Com relação aos tipos de erro do LLM, os juízes atribuíram sim/não para sete tipos de erros, de acordo com a taxonomia de Roy *et al*. [2024]. O teste McNemar-Bowker avaliou a frequência de homogeneidade marginal dos erros. As proporções e diferenças medianas são relatadas juntamente com seus intervalos de confiança de 95% (IC 95%) e Intervalo Interquartil (IQR).

4.2 Análise dos Resultados

Primeiramente, focamos nossas análises nos resultados relacionados aos critérios de avaliação, apresentados na **Tabela 1**. Os resultados indicaram que, para dois dos cinco aspectos analisados (similaridade e personalização), não houve diferenças estatisticamente significativas entre o desempenho dos modelos avaliados. Entretanto, destaca-se o desempenho do Modelo 1, que, apesar de utilizar um prompt menos elaborado e também não utilizar técnicas adicionais (RAG), conseguiu resultados comparáveis aos outros modelos. Para os outros três aspectos restantes (adequação, clareza e comparação), foram observados diferenças relevantes entre os modelos, em que, para quantificar essas diferenças, foram realizados testes estatísticos de comparação par-a-par, possibilitando uma análise mais aprofundada do desempenho relativo de cada abordagem nos critérios avaliados. Com essas análises, foi possível observar as melhorias provenientes do modelo com uso de RAG, principalmente se comparado ao Modelo 1. O uso de RAG resultou no melhor desempenho ao analisar o aspecto Adequação, considerado crucial por estar relacionado à adequação dos resultados do modelo para a tarefa em questão, analisando detalhes essenciais na prescrição gerada, como instruções corretas para uso do medicamento. Na análise que corresponde à comparação entre as instruções geradas pelo modelo e o padrão de referência (Comparação), nossa proposta baseada em RAG apresenta instruções equiparados às geradas por especialistas, além de também serem as mais claras.

Nossa segunda análise focou na classificação dos erros identificados pelos juízes. Um erro foi considerado presente em uma prescrição apenas quando apontado pela maioria dos avaliadores, ou seja, por pelo menos três dos cinco especialistas envolvidos na avaliação. A distribuição do número de erros por modelo está presente na **Tabela 2**.

O **erro tipo** 1 se destaca como o mais preocupante do ponto de vista médico. Esse tipo de erro ocorre quando o modelo gera informações que podem induzir o paciente ao uso incorreto do medicamento, potencialmente comprometendo a eficácia do tratamento ou causando complicações à saúde. Os resultados relacionados ao erro tipo 1 demonstraram uma melhora significativa com o uso de nossa abordagem com LLM (Llama 3.1) com RAG quando comparado aos outros modelos, reduzindo significativamente a quantidade de erros. O **erro tipo 2** está relacionado à capacidade dos modelos em gerar uma conclusão vaga ou incorreta. É possível ver uma melhora, principalmente, da efetividade de um prompt bem elaborado no processo, onde o quantitativo de erros diminuiu drasticamente do Modelo 1 para o Modelo 2. O uso de nossa abordagem RAG consegue reduzir ainda mais as ocorrências de erros tipo 2. O **erro tipo 3** refere-se à ausência de informações essenciais para o uso correto da medicação. Nos resultados obtidos, é possível observar que há um empate estatístico entre os Modelos 1 e 2, ou seja, apenas a engenharia de prompt não foi suficiente para amenizar esse tipo de erro. Entretanto, é notável a diferença significativa da quantidade desse erro entre o Modelo 2 e o Llama 3.1 com RAG, demonstrando, novamente, como a técnica contribui positivamente para aperfeiçoar as saídas geradas. O **erro tipo 4** corresponde a falhas pontuais, como erros de escrita ou operações matemáticas incorretas. Apesar de haver um empate estatístico entre o Modelo 1 e o Modelo 2, há uma diferença estatística entre os resultados do Modelo 1 e do Llama 3.1 com RAG, reforçando a efetividade das técnicas empregadas. O **erro tipo 5** diz respeito a falhas metodológicas específicas ao medicamento, caracterizadas por instruções inadequadas ao contexto da prescrição. Um exemplo seria orientar o paciente a tomar o medicamento ”em caso de melhora”, mesmo quando o uso correto seria em uma situação específica, como em caso de febre. Como evidenciado na tabela, o quantitativo deste erro foi extremamente baixo, além de não haver diferença estatística significativa entre os resultados dos modelos. O **erro tipo 6** retrata situações em que o modelo retornou informações não solicitadas, como dados sobre validade ou interações medicamentosas. Para esse tipo de erro, a engenharia de prompt resultou em uma melhora significativa, de acordo com os testes estatísticos, em que o desempenho do Modelo 2 foi superior ao do Modelo 1. Entretanto, com o uso de RAG, houve empate estatístico entre o Modelo 1. É possível que a falta de melhora seja devido à dificuldade do modelo de lidar com uma quantidade maior de informações,, fazendo-o agregar informações não desejadas nas saídas geradas. O **erro tipo 7** refere-se a casos em que o modelo gerou alucinações, apresentando informações completamente

Tabela 1. Avaliação dos modelos sob diferentes critérios. *Teste pareado de Friedman. Comparação par-a-par com teste Bonferroni entre: [1]Llama3.1 8b Modelo 1 e Llama3.1 8b Modelo 2; [2]Llama3.1 8b Modelo 1 e Llama3.1 8b Modelo RAG; e [3]Llama3.1 8b Modelo 2 e Llama3.1 8b Modelo RAG.

	Llama3.1 8b Modelo 1 (n=83)	Llama3.1 8b Modelo 2 (n=83)	Llama3.1 8b Modelo 3 (n=83)	Comparação entre modelos*	P-valor[1]	P-valor[2]	P-valor[3]
Similaridade, mediana (IQR)	91.4 (5.4)	93.8 (5.2)	94.3 (5.3)	0.064			
Adequação, mediana (IQR)	93.0 (40.0)	94.0 (53.3)	100 (15.0)	0.001	0.815	<0.001	<0.001
Clareza, mediana (IQR)	90.0 (60.0)	92.0 (55.0)	95 (24.0)	0.012	0.326	0.008	0.091
Personalização, mediana (IQR)	95.0 (20.0)	95 (35.0)	100 (10.0)	0.052			
Comparação, mediana (IQR)	44 (35)	47 (27.0)	50 (21.0)	0.026	0.239	0.010	0.157

Tabela 2. Erros nos resultados gerados pelos diferentes modelos. Teste de McNemar-Bowker para: [1] Modelo 1 vs. Modelo 2; [2] Modelo 1 vs. Modelo RAG. [3] Modelo 2 vs. Modelo RAG; ... Imensurável

	Llama3.1 8b Modelo 1 (n=83)	Llama3.1 8b Modelo 2 (n=83)	Llama3.1 8b Modelo 3 (n=83)	P-valor[1]	P-valor[2]	P-valor[3]
Erro tipo 1: Uso incorreto do medicamento	31 (37.3%) [27.4–48.0]	24 (28.9%) [19.9 - 39.2]	14 (16.9%) [9.9 - 25.9]	0.143	<0.001	0.013
Erro tipo 2: Instruções contraditórias ou vagas	31 (37.3%)[27.4 - 48.0]	18 (21.7%)[13.8 - 31.3]	16 (19.3%)[11.8 - 28.6]	0.019	0.014	0.832
Erro tipo 3: Falta de instruções essenciais	25 (30.1%)[21.0 - 40.5]	28 (33.7%)[24.2 - 44.3]	16 (19.3%)[11.8 - 28.6]	0.678	0.078	0.017
Erro tipo 4: Erros factuais (não médicos)	17 (20.5%)[12.8 - 30.0]	8 (9.6%) [4.5 - 17.2]	3 (3.6%) [0.9 - 9.1]	0.078	0.001	0.180
Erro tipo 5: Instruções sem evidências científicas	0 (0%) [0.0 - 0.23]	0 (0%) [0.0 - 0.23]	2 (2.4%) [0.4 - 7.3]	...	...	...
Erro tipo 6: Instruções não solicitadas	9 (10.8%) [5.4 - 18.7]	1 (1.2%) [0.1 - 5.2]	6 (7.2%) [2.9 - 14.1]	0.021	0.549	0.125
Erro tipo 7: Alucinações (informações inventadas/fabricadas)	10 (12.0%) [6.2 - 20.2]	6 (7.2%) [2.9 - 14.1]	4 (4.8%) [1.5 - 10.8]	0.388	0.180	0.687

descorrelacionadas com o pedido original. Novamente, para esse tipo de erro, não foram observadas diferenças estatisticamente significativas entre os modelos, todavia, em termos absolutos, o modelo com RAG teve apenas 4 ocorrência dentre todas as 83 analisadas, o que pode ser considerado um resultado muito robusto.

5 Conclusões e Trabalhos Futuros

Este artigo visa promover o uso seguro e eficaz de medicamentos por meio de uma solução baseada em LLMs, integradas com uma estratégia de Retrieval-Augmented Generation (RAG) que melhora significativamente o contexto fornecido aos LLMs durante a preparação da prescrição utilizando informações das bulas fornecidas pela Anvisa. Nossa solução identifica o trecho mais relevante de uma bula referente ao contexto específico da prescrição para realizar o enriquecimento. Avaliamos a aplicabilidade da proposta considerando um conjunto de prescrições de casos clínicos hipotéticos, criados por experientes profissionais de saúde. Cinco avaliadores médicos realizaram uma comparação dessas prescrições manuais com as geradas por nossa proposta. Os resultados dessa avaliação destacam o potencial dos LLMs como ferramentas capazes de auxiliar significativamente os processos de comunicação na área da saúde. Este estudo demonstra que mesmo os modelos de código aberto com número menor de parâmetros menores podem obter resultados satisfatórios na geração de prescrições, especialmente com a integração de RAGs. Como trabalhos futuros, pretendemos ampliar a proposta considerando e comparando outros LLMs de código aberto recentes, como o DeepSeek, bem como avaliar outras informações das bulas que podem ser usadas pelo RAG. Por fim, pretendemos realizar uma avaliação muito mais ampla, considerando vários profissionais de saúde.

Declarações complementares

Financiamento

Este trabalho foi financiado por Grant Challenges Brazil, financiado por Fundação Gates e Fiocruz (número de subsídio INV-009289), CNPq (números de subsídio 400758/2024-5, 313103/2021-6, 307229/2021-1) e FAPEMIG.

Contribuições dos autores

As implementações e execuções dos experimentos foram realizadas pelo aluno Davi, sob a orientação do professor Leonardo. A concepção do projeto e as análises de resultados foram feitas em conjunto, aluno e professor. Esse trabalho está inserido em um projeto mais amplo, financiado pela Fundação Gates, coordenado pela professora Zilma.

Conflitos de interesse

Os autores declaram que não têm nenhum conflito de interesse.

Disponibilidade de dados e materiais

Os conjuntos de dados gerados e/ou analisados durante o estudo atual estão disponíveis em `https://ftp.medicina.ufmg.br/cins/pesquisa/iapolis/`.

Referências

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. (2024). M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the ACL*. DOI: 10.18653/v1/2024.findings-acl.137.

Gosselin, L., Thibault, M., Lebel, D., and Bussières, J.-F. (2021). Utilisation de l'intelligence artificielle en pharmacie : une revue narrative. *CJHP*, 74. DOI: 10.4212/cjhp.v74i2.3100.

Goyal, S., Rastogi, E., Rajagopal, S. P., Yuan, D., Zhao, F., Chintagunta, J., Naik, G., and Ward, J. (2024). HealAI: A healthcare LLM for effective medical documentation. In *Proceedings of the 17th WSDM*. ACM. DOI: 10.1145/3616855.3635739.

Klepser, D., Lanham, A., and Cochran, G. (2016). Electronic prescriptions: opportunities and challenges for the patient and pharmacist. *AHCT*. DOI: 10.2147/AHCT.S64477.

Lu, Z., Tian, J., Wei, W., Qu, X., Cheng, Y., Xie, W., and Chen, D. (2024). Mitigating boundary ambiguity and inherent bias for text classification in the era of large language models. In *Findings of ACL*. DOI: 10.18653/v1/2024.findings-acl.467.

Reis, Z. S. N., Pagano, A. S., Ramos De Oliveira, I. J., Dias, C. D. S., Lage, E. M., Mineiro, E. F., Varella Pereira, G. M., De Carvalho Gomes, I., Basilio, V. A., Cruz-Correia, R. J., De Jesus, D. D. R., De Souza Júnior, A. P., and Rocha, L. (2024). Evaluating large language model–supported instructions for medication use: First steps toward a comprehensive model. *Mayo Clinic Proceedings: Digital Health*, 2. DOI: 10.1016/j.mcpdig.2024.09.006.

Rezende, L. H. O., Gehrke, F. D. S., Silva, M. A., Carneiro, A. M. F., Abreu, R. M., Monteiro, C. N., Leão, K. A., and Takei, K. (2020). Prescrição de medicamentos: uma análise para a implantação da prescrição eletrônica ambulatorial. *Acervo Saúde*, 12. DOI: 10.25248/reas.e3638.2020.

Roy, S., Khatua, A., Ghoochani, F., Hadler, U., Nejdl, W., and Ganguly, N. (2024). Beyond accuracy: Investigating error types in GPT-4 responses to USMLE questions. In *Proceedings of the 47th ACM SIGIR*. ACM. DOI: 10.1145/3626772.3657882.

Shiffman, S., Gerlach, K. K., Sembower, M. A., and Rohay, J. M. (2011). Consumer understanding of prescription drug information: An illustration using an antidepressant medication. *Ann Pharmacother*, 45. DOI: 10.1345/aph.1P477.

Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., Osterhoudt, H., Wu, X., Visweswaran, S., Fu, S., Mathur, P., Cacciamani, G. E., Sun, C., Peng, Y., and Wang, Y. (2024). A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digit. Med.*, 7. DOI: 10.1038/s41746-024-01258-7.

Xiong, G., Jin, Q., Lu, Z., and Zhang, A. (2024). Benchmarking retrieval-augmented generation for medicine. In *Findings of ACL*. Association for Computational Linguistics. DOI: 10.18653/v1/2024.findings-acl.372.

Yang, B., Jiang, S., Xu, L., Liu, K., Li, H., Xing, G., Chen, H., Jiang, X., and Yan, Z. (2024). DrHouse: An LLM-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8. DOI: 10.1145/3699765.

Zakka, C., Shad, R., Chaurasia, A., Dalal, A. R., Kim, J. L., Moor, M., Fong, R., Phillips, C., Alexander, K., Ashley, E., Boyd, J., Boyd, K., Hirsch, S., Vogelsong, M. A., Cunningham, J. P., and Hiesinger, W. (2024). Almanac — retrieval-augmented language models for clinical medicine. *NEJM AI*, 1. DOI: 10.1056/AIoa2300068.