57º SBPO
SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL
Gramado • RS
DE 5 A 9 DE OUTUBRO DE 2025

Uso de Large Language Models para Geração Automática de Sumários Clínicos em Sistemas de Informação Hospitalares

Everton Pereira Militão
Departamento de Informática - DI
Instituto Tecgraf PUC-Rio
evertonmilitao@tecgraf.puc-rio.br

Pedro Azambuja
INI - FIOCRUZ
pedroaza@gmail.com

Leonardo dos Santos Lourenço Bastos
Departamento de Engenharia Industrial
Instituto Tecgraf PUC-Rio
lslbastos@puc-rio.br

A maior complexidade e volume de dados em Prontuários Eletrônicos tornou desafiadora a tarefa de revisar e resumir informações clínicas pelos profissionais de saúde. Os *Large Language Models* (LLMs) têm demonstrando alta capacidade de síntese de informações não-estruturadas. Este trabalho buscou aplicar LLMs para a geração de resumos clínicos da condição de pacientes durante a internação. Desenvolveu-se uma aplicação em *Retrieval-Augmented Generation* (RAG) com os dados abertos do MIMIC-IV, e os modelos *Ollama-3.3-70b*, *DeepSeek-V3* e *Ascepilus-Ollama-8B*, com três estratégias de recuperação de texto: inserção integral no prompt, o RAG simples e o *Hypothetical Document Embeddings*. O uso do *DeepSeek* demonstrou melhor aderência ao benchmark gerado por um médico especialista (>60%), enquanto que o *Asclepilus*, mesmo de domínio especifico, apresentou piores métricas. Todos apresentaram baixa alucinação (Faithfuness>80%), e alta consistência com as notas clínicas (Answer Relevancy>80%). A solução demonstrou potencial de melhoria na rotina de prática clínica dos profissionais de saúde.

PALAVRAS CHAVE. Large Language Models, Retrieval Agumented Generation, Notas Clínicas, Prontuário Eletrônico, IA Generativa.

Tópicos: SA – PO na Área de Saúde, PO e IA

ABSTRACT

The increased complexity and volume of data in Electronic Health Records have challenged healthcare professionals to review and summarize patient clinical information. Large Language Models (LLMs) have shown an impressive ability to synthesize unstructured information. Here we aimed to apply LLMs to generate clinical summaries regarding patients' conditions during hospitalization. A Retrieval-Augmented Generation (RAG)-based solution was developed on the MIMIC-IV open dataset with the Ollama-3.3-70b, DeepSeek-V3, and Asclepilus-Ollama-8B models, and three information retrieval strategies: full insertion of clinical notes into the prompt, simple RAG, and Hypothetical Document Embeddings. Results showed the approach that best aligned with the benchmark created by a medical doctor used DeepSeek (>60%), while Asclepilus, despite being

https://proceedings.science/p/212188?lang=pt-br

domain-specific, showed poorer results. All models produced low leveos of hallucination (Faithfulness >80%) and showed high consistency with clinical notes (AnswerRelevancy>80%). Thus, the proposed solution showed potential for application in generating clinical summaries, supporting the clinical workflow of healthcare professionals.

KEYWORDS. Large Language Models, Retrieval Agumented Generation, Clinical notes, Electronic Health Records, Generative AI.

Paper topics: SA – PO na Área de Saúde, PO e IA

1. Introdução

No contexto da saúde, os Registros Eletrônicos de Saúde (EHRs) são elementos fundamentais para a tomada de decisão clínica, pois integram uma diversidade de dados multimodais — como sinais vitais, resultados laboratoriais, notas clínicas e evoluções — que documentam, em profundidade, a trajetória do cuidado prestado ao paciente [Johnson et al., 2016; Mohsen et al., 2022]. Em casos de pacientes com múltiplas internações e condições clínicas complexas, esses registros tornam-se volumosos, sendo muitas vezes não estruturados, como as notas clínicas, o que impõe uma carga cognitiva significativa aos profissionais de saúde encarregados de analisá-los. Esse cenário pode comprometer a qualidade da tomada de decisão clínica, impactando diretamente a segurança e os desfechos do paciente [Kruse et al., 2025].

Diante disso, a aplicação de IA generativa na geração de sumários clínicos surge como uma ferramenta potencial para apoiar a transição de cuidado entre diferentes níveis assistenciais e melhorar a acurácia das decisões clínicas. No entanto, a recuperação das informações mais relevantes, aliada à extensão e à complexidade dos registros clínicos, impõe desafios substanciais à tarefa de sumarização. Esses desafios evidenciam a necessidade de modelos capazes de lidar com dados longos, desestruturados e distribuídos temporalmente, bem como com o raciocínio clínico envolvido nesses contextos [Kruse et al., 2025; Lyu et al., 2024].

Embora modelos de linguagem de grande porte (*Large Language Models* — LLMs) tenham demonstrado desempenho promissor em tarefas de geração e compreensão de linguagem natural, sua aplicação direta em contextos clínicos longos e complexos ainda apresenta limitações. Em especial, esses modelos tendem a sofrer com perda de informação em entradas extensas, alucinações factuais e dificuldade em reter a coerência temporal ao longo de trajetórias clínicas [Liu et al. 2024], [Kruse et al. 2025]. Nesse cenário, a integração com técnicas de *Retrieval Augmented Generation* (RAG ou ”Geração Aumentada por Recuperação”) se torna especialmente relevante, pois permite que o modelo acesse dinamicamente informações contextuais relevantes a partir de bases estruturadas ou não estruturadas, mesmo quando fora da janela de atenção do modelo [Lewis et al. 2020].

Portanto, a combinação entre LLMs e RAG representa uma abordagem promissora para lidar com os desafios impostos pela sumarização automatizada de Registros Eletrônicos de Saúde (EHRs) em cenários reais de cuidado ao paciente. Dessa forma, este trabalho teve como objetivo aplicar LLMs, aliados a técnicas de RAG, para a realizar a sumarização automatizada de notas clínicas presentes em registros eletrônicos de saúde (EHRs) com foco na transição de cuidado.

2. Metodologia

A metodologia deste estudo foi estruturada de acordo com as cinco etapas para adoção de sistemas com IA generativa [Lewis et al., 2020]. Cada etapa utilizada neste trabalho é descrita a seguir.

2.1. Identificar:

O primeiro passo consistiu na definição do cenário de aplicação: a geração automatizada de sumários clínicos com o objetivo de apoiar a transição de cuidado entre diferentes níveis assistenciais. Esse desafio é intensificado pelo crescente volume e complexidade dos Registros Eletrônicos de Saúde (EHRs), que agregam dados clínicos multimodais — como sinais vitais, resultados laboratoriais e notas evolutivas — frequentemente armazenados de forma não estruturada [Johnson et al. 2016; Mohsen et al. 2022].

2.2. Determinar:

Em seguida, foram definidos os requisitos do sistema: os sumários clínicos deveriam ser concisos, informativos e temporalmente coerentes, a fim de apoiar uma transição de cuidado eficiente. Para isso, utilizou-se o *Medical Information Mart for Intensive Care* (MIMIC) IV, um banco de dados clínico aberto amplamente adotado em pesquisas de inteligência artificial em saúde.

O MIMIC-IV contém aproximadamente 2 milhões de notas clínicas de cerca de 60.000 pacientes internados em unidades de terapia intensiva e mais de 200.000 internações em departamentos de emergência [Johnson et al. 2016]. Os dados utilizados neste estudo foram extraídos especificamente da tabela *NOTEEVENTS*, que contém registros de notas clínicas geradas por diferentes profissionais de saúde.

2.3. Selecionar:

Nesta etapa foram definidos os modelos de linguagem e as tarefas clínicas a serem avaliadas. Com base em benchmarks recentes no domínio biomédico [Zhang et al. 2023; Kruse et al. 2025], foram selecionados (*LLMs*) de grande porte tais como o *Ollama-3* (Parâmetros: 70 bilhões) e o *DeepSeek-V3* (Parâmetros 671 bilhões), sendo ambos modelos domínio geral. Além disso, foi considerado um modelo especializado em domínio espefício biomédico, o *Asclepius* (Parâmetros 8 bilhões), que corresponde a um modelo *Ollama-3* refinado via (*fine-tuned*) com dados clínicos do MIMIC-III [Mehdi et al. 2024]. As tarefas definidas incluíram a sumarização de alta hospitalar. Além disso, junto aos modelos, utilizaram-se três estratégias de recuperação de texto: (i) inserção direta do conteúdo da nota clínica integralmente no prompt; uma (ii) aplicação de *Retrieval-Augmented Generation* (RAG) simples utilizando as notas clínicas como bases externas; e (iii) utilização da técnica *Hypothetical Document Embeddings* (HyDE), que é uma técnica de recupeação semântica que utiliza um modelo de linguagem para gerar um documento hipotético a partir de uma consulta, o que é então codificado em um vetor de *embedding* para recuperar documentos reais semanticamente semelhantes [Gao et al., 2023].

2.4. Adaptar:

Os modelos selecionados foram adaptados ao contexto clínico por meio de engenharia de *prompts* e aplicação da técnica de RAG, a fim de enriquecer o conteúdo inserido nos modelos com evidências extraídas semanticamente dos registros clínicos. Para isso, foram utilizados *embeddings* de domínio geral, extraídos do modelo *sentence-transformers/all-MPNet-base-v2*, disponível na plataforma *Hugging Face*, amplamente reconhecido por seu desempenho em tarefas de similaridade textual e recuperação semântica.

2.5. Implementar:

A implementação do sistema foi realizada em *Python*, utilizando a biblioteca *LangChain* para a orquestração e integração com *LLMs*, além de ferramentas complementares como o *FAISS* para vetorização e recuperação semântica de contexto. Três estratégias de recuperação de texto foram testadas: (i) a inserção integral da última nota clínica no prompt do modelo; (ii) a técnica tradicional de *Retrieval-Augmented Generation* (RAG), com busca semântica vetorial; e (iii) a abordagem baseada em *Hypothetical Document Embeddings* (HyDE), que explora documentos hipotéticos gerados a partir da consulta para refinar a recuperação de informações. A experimentação computacional foi conduzida com foco na avaliação da qualidade dos sumários gerados. Para isso, foram utilizadas métricas específicas: a *Answer Relevancy* (AR), para mensurar a relevância da resposta em relação à nota clínica original; a *Faithfulness*, para detectar a presença de alucinações; e uma métrica de aderência semântica baseada na comparação com um sumário clínico de referência gerado e validado por um médico especialista.

3. Resultados

A comparação entre as três estratégias de geração de respostas — *Prompt Injection*, *RAG* e *HyDE* — revela diferenças importantes quanto à aderência ao *benchmark*, alucinação e relevância das respostas. A análise dos resultados evidencia que a melhor aderência ao *benchmark* foi alcançada pela estratégia baseada no modelo *DeepSeek*, com desempenho superior a 60% na métrica *G-Eval*. Curiosamente, o modelo *Asclepius*, ainda que treinado com foco em dados médicos, apresentou os piores resultados nessa métrica, sugerindo que a especialização de domínio não garantiu

melhor alinhamento com o *benchmark*. Em contrapartida, todas as abordagens testadas demonstraram alta consistência no conteúdo gerado, com métricas de *Faithfulness* e *Answer Relevancy* acima de 80%, indicando baixo risco de alucinação e boa relevância clínica das respostas.

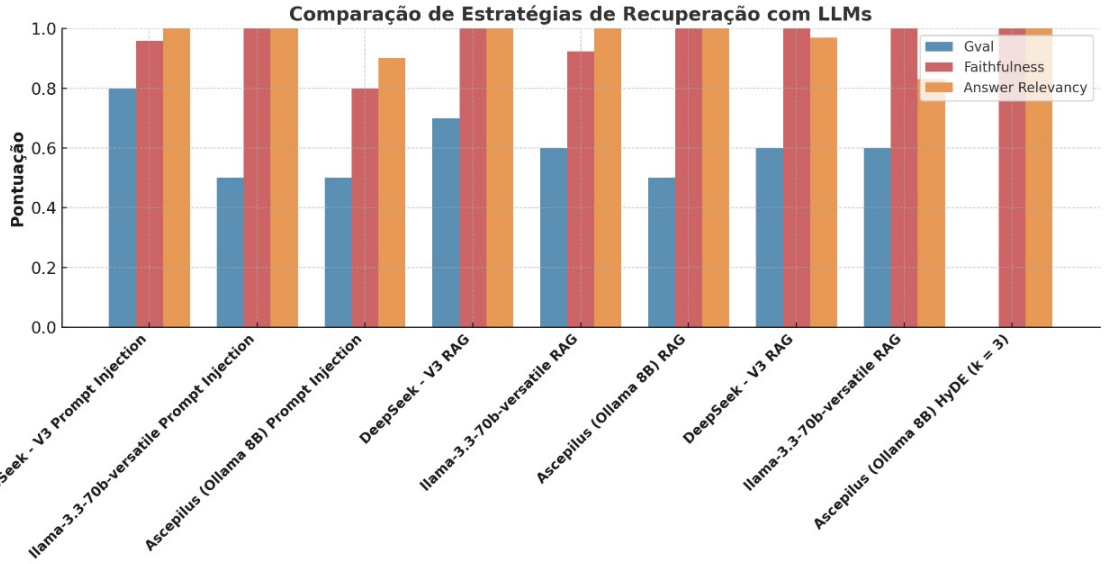


Figura 1: Comparação dos Resultados Obtidos pelas Métricas GEval, Faithfulness e Answer Relevancy.

4. Considerações Finais

A abordagem desenvolvida demonstrou viabilidade na geração automatizada de sumários clínicos, destacando-se pelo uso eficiente de linguagem de domínio geral aliados a estratégias de recuperação semântica. Potencialmente, o maior número de parâmetros e capacidade de processamento pela janela de contexto impactaram o desempenho mesmo dos modelos de domínio específico. De maneira geral, soluções baseadas em IA generativa, sobretudo os LLMs, podem contribuir significativamente para a rotina assistencial e a continuidade do cuidado.

Referências

Gao, L., Ma, X., Lin, J., e Callan, J. (2023). Precise zero-shot dense retrieval without relevance labels. p. 1762–1777. URL `https://aclanthology.org/2023.acl-long.99.pdf`.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., e Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035.

Kruse, M., Hu, S., Derby, N., e et al. (2025). Zero-shot large language models for long clinical text summarization with temporal reasoning. *arXiv preprint arXiv:2501.18724*.

Lewis, P., Perez, E., Piktus, A., e et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Lyu, C., Smith, J., e Yang, Y. (2024). Benchmarking llms on medical discharge summaries. In *Proceedings of the ACL*.

Mohsen, M. A., Zhang, W., Wu, Y., e et al. (2022). Clinical notes analysis using machine learning: A systematic review. *Journal of Biomedical Informatics*.