

COMPARAÇÃO ENTRE GEMINI 2.5 FLASH COM E SEM RAG EM APLICAÇÃO MÉDICA: USO DO HISTÓRICO CONVERSACIONAL NO MANEJO DE PACIENTES COM DIABETES TIPO 2

**COMPARATIVE EVALUATION OF GEMINI 2.5 FLASH WITH AND WITHOUT
RETRIEVAL-AUGMENTED GENERATION IN A CLINICAL SETTING:
CONVERSATIONAL HISTORY FOR TYPE 2 DIABETES MANAGEMENT**

Ciências Exatas e da Terra, Ciências da Saúde • 28/07/2026

REGISTRO DOI: [10.70773/revistatopicos/784954242](https://doi.org/10.70773/revistatopicos/784954242)

Bryan Neves Pinto
Francisco Ponciano Maciel Neto
Johnson Tavares Pinto
Juíle Yoshie Sarkis Hanada
Maria Giovanna Gonçalves Sales
Hylbert Bentes Rocha Rodrigues
Lilian Kalinka da Silva Carvalho
Ailon dos Santos Teixeira

RESUMO

Este estudo avalia comparativamente o modelo Gemini 2.5 Flash operando com e sem a técnica de Retrieval-Augmented Generation (RAG) em um agente conversacional voltado à assistência e educação de pacientes com diabetes mellitus tipo 2. Os resultados indicam que o uso de RAG favorece a assertividade e a fundamentação documental das respostas, enquanto o modo sem RAG apresenta maior empatia e clareza comunicacional, evidenciando um trade-off relevante para aplicações em saúde digital.

Palavras-chave: Gêmeos 2.5 Flash; Geração Aumentada por Recuperação; diabetes mellitus tipo 2; Linguagem de Grande Porte.

ABSTRACT

This study comparatively evaluates the Gemini 2.5 Flash model operating with and without the Retrieval-Augmented Generation (RAG) technique in a conversational agent designed to support the assistance and education of patients with type 2 diabetes mellitus. The results indicate that the use of RAG enhances the assertiveness and documentary grounding of the generated responses, whereas the non-RAG configuration exhibits higher levels of empathy and communicative clarity, highlighting a relevant trade-off for digital health applications.

Keywords: Gemini 2.5 Flash; Retrieval-Augmented Generation; diabetes mellitus tipo 2; Linguagem de Grande Porte.

1. INTRODUÇÃO

Deve-se fazer uma contextualização/apresentação breve do tema a ser estudado ou do tema O diabetes mellitus tipo 2 (DM2) é um fator de risco relevante para a insuficiência cardíaca (IC), atuando tanto de

forma direta, ao comprometer a função cardíaca, quanto indiretamente, por meio de comorbidades associadas. Entre essas condições destacam-se a hipertensão, a doença arterial coronariana, a disfunção renal e a obesidade, que ampliam o impacto do DM2 na saúde cardiovascular [Palazzuoli and Iacoviello, 2022].

Com a intensificação da transformação digital no cenário pós-pandemia, os pacientes passaram a monitorar sua saúde de maneira mais ativa. Esse movimento tem sido favorecido por inovações em sensores biomédicos e pelo uso crescente da Inteligência Artificial, especialmente dos Modelos de Linguagem de Grande Porte, que podem oferecer suporte à personalização do cuidado em doenças crônicas como o DM2 [Dao et al., 2024a].

Abbasian et al. [2024] afirma que tecnologias avançadas como as LLMs melhoram substancialmente a acessibilidade e a eficácia no autogerenciamento do diabetes, pois permitem que os pacientes adquiram uma melhor assistência com um conhecimento abrangente. porém, as LLMs podem apresentar problemas, como limitações em respostas muito específicas e alucinações [Johno et al., 2025].

Uma técnica que vem sendo muito utilizada nas LLM's é a RAG (Retrieval-Augmented Generation) que visa aprimorar a geração de respostas e a confiabilidade das LLMs, recuperando informações relevantes a partir de fontes externas de conhecimento confiável [Johno et al., 2025].

Lewis et al. [2020] afirmam que a técnica RAG surgiu como uma solução promissora para lidar com as limitações nas respostas das LLM's, reduzindo as alucinações e garantindo que os modelos

forneçam informações atualizadas por meio da integração de documentos confiáveis na base de conhecimento.

Este estudo tem como objetivo avaliar e comparar a eficácia da técnica de geração aumentada por recuperação RAG (Retrieval-Augmented Generation) aplicada ao modelo Gemini 2.5 Flash, no contexto do desenvolvimento de um agente conversacional voltado a assistência e educação de pacientes com diabetes mellitus tipo 2.

Nesse contexto, será realizada uma análise comparativa entre duas versões do modelo Gemini 2.5 Flash com RAG e Gemini 2.5 Flash sem RAG a fim de investigar em que medida a incorporação de mecanismos de recuperação de conhecimento pode aprimorar a clareza, empatia e assertividade das respostas geradas pelo agente em cenários clínicos e educacionais.

Com esse propósito, será desenvolvida uma interface web interativa de perguntas e respostas, capaz de analisar o histórico de conversas do paciente e, a partir dele, gerar *insights* clínicos e educacionais personalizados nos domínios do tratamento do diabetes tipo 2.

2. TRABALHOS RELACIONADOS

Esta seção apresenta estudos que utilizam Modelos de Linguagem de Grande Porte na assistência médica de pacientes com diabetes, as pesquisas visam apresentar novos horizontes da pesquisa e preencher lacunas do conhecimento.

De acordo com o estudo de [Panguraj, 2023], as LLMs possuem a capacidade de avaliar a probabilidade do surgimento de doenças como diabetes ou problemas cardíacos. Elas fazem isso ao analisar um conjunto abrangente de informações biomédicas, que inclui o

histórico médico do paciente, os resultados de exames laboratoriais e dados sobre o estilo de vida, como dieta e rotina de exercícios.

O mesmo estudo também aponta que os métodos de diagnóstico assistidos por IA não se limitam apenas à previsão de doenças, mas também são eficazes para aprimorar diversos sistemas de diagnóstico. Ao fornecer suporte adicional aos profissionais de saúde, essa tecnologia os auxilia na tomada de decisões mais precisas e eficientes em relação ao tratamento dos pacientes.

O estudo Diabetica [Wei et al., 2025] apresenta um LLM especializado em diabetes, derivado do Qwen2-7B-Instruct e refinado com LoRA. Seu pipeline inclui curadoria rigorosa de dados e um método de autodestilação que aprimora o alinhamento clínico e a consistência nas respostas. Essas técnicas reduzem o desvio de distribuição entre dados e modelo. Como resultado, o Diabetica demonstra desempenho superior e maior robustez em tarefas médicas especializadas.

Nas avaliações de bancada, o Diabetica apresentou desempenho superior em múltipla escolha, preenchimento de lacunas e diálogos clínicos. O modelo atingiu 87,2% de acurácia, superando tanto LLMs open-source quanto sistemas proprietários. Os resultados evidenciam a eficácia de modelos especializados e mostram que arquiteturas menores podem superar modelos generalistas quando bem treinadas. O estudo reforça a importância de pipelines rigorosos de curadoria e treinamento.

O estudo de [Dao et al., 2024b] apresenta uma solução digital abrangente que utiliza modelos GPT-3.5 da OpenAI para a prevenção proativa do diabetes. O sistema inclui uma interface amigável para

web e celular, um chatbot de IA, sistemas de lembretes personalizados, um agregador de recursos de saúde e um módulo de suporte emocional.

A pesquisa também avaliou a qualidade das respostas geradas pelo modelo GPT-3.5 em duas configurações: usando a API Assistants e um sistema de geração aumentada de recuperação (RAG) com FAISS. Os testes, baseados em um conjunto de dados estruturados de perguntas e respostas, revelaram que o modelo combinado com a API Assistants se mostrou superior, fornecendo respostas mais relevantes, detalhadas e personalizadas.

A abordagem adotada destaca o potencial da IA como uma ferramenta valiosa na prevenção do diabetes e reforça a importância da confiabilidade e consistência em soluções de saúde.

3. PROTOCOLO EXPERIMENTAL

Esta seção descreve a metodologia experimental empregada, incluindo o Modelo de Linguagem de Grande Porte selecionado — Gemini 2.5 Flash —, bem como a arquitetura de software desenvolvida e as tecnologias utilizadas. O estudo não compara modelos distintos, mas sim duas configurações do mesmo modelo:

3.1. Gemini-2.5-Flash

De acordo com [O'Shea et al., 2024], o Gemini 2.5-Flash constitui uma evolução da base do modelo 2.0-Flash. Essa versão incorpora melhorias significativas em recursos de raciocínio, mantendo a velocidade de execução e a eficiência de custo como prioridades centrais. Dessa forma, o modelo se posiciona como uma solução

avançada, capaz de equilibrar desempenho técnico com acessibilidade operacional.

Além disso, o Gemini 2.5-Flash é descrito como o primeiro modelo de raciocínio totalmente híbrido. Essa característica permite que os desenvolvedores ativem ou desativem o modo de pensamento, conforme a necessidade da aplicação. Tal flexibilidade possibilita a definição de orçamentos ajustados, equilibrando a qualidade das respostas com os recursos computacionais disponíveis [O'Shea et al., 2024].

3.2. Gemini-2.5-Flash com Técnica RAG

A técnica de Retrieval-Augmented Generation (RAG) foi aplicada para aumentar a precisão e a fundamentação das respostas geradas pelos LLMs. O processo inclui ingestão documental, extração textual e segmentação em blocos, que são convertidos em vetores densos via embeddings Gemini. Esses vetores compõem um corpus vetorial global utilizado para busca semântica eficiente.

Na etapa de consulta, a pergunta do usuário é transformada em embedding e comparada ao corpus por similaridade de cosseno, permitindo recuperar os trechos mais relevantes. Esses blocos são incorporados ao prompt do LLM, garantindo geração orientada por evidências externas verificáveis. Assim, o sistema reduz alucinações, aumenta a confiabilidade e assegura rastreabilidade das fontes.

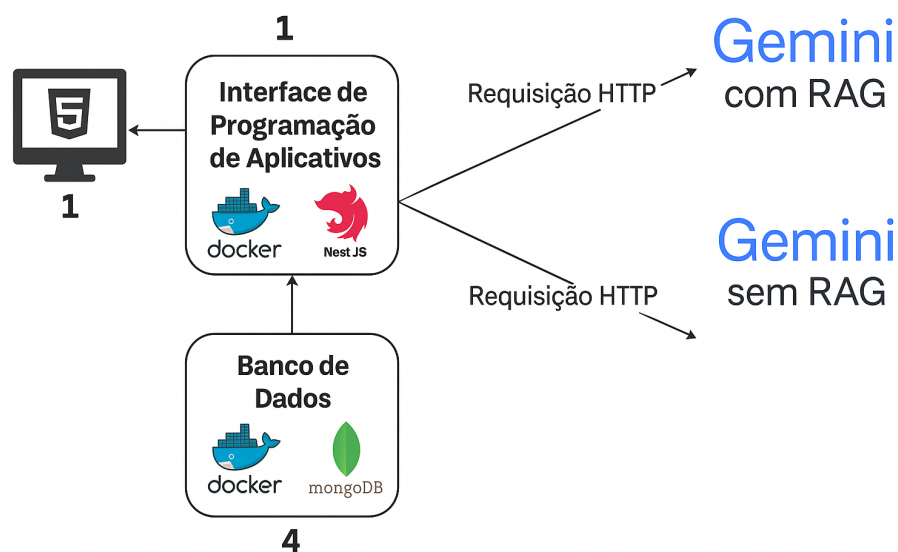
3.3. Arquitetura de Software

Esta seção apresenta a arquitetura de software definida para este projeto de pesquisa. De acordo com [Garlan, 2008], a arquitetura de software compreende o design, a especificação e a análise da

estrutura geral do sistema, considerando componentes, estruturas de controle, protocolos de comunicação e mecanismos de acesso a dados.

A arquitetura proposta integra o aplicativo móvel, a API (Application Programming Interface) e dois modos distintos de utilização do modelo Gemini 2.5: com a técnica de Retrieval-Augmented Generation (RAG) e sem RAG.

No modo sem RAG, o modelo responde apenas com base no conhecimento adquirido durante o treinamento. Já no modo com RAG, a API recupera trechos relevantes de um corpus vetorial para enriquecer o contexto fornecido ao modelo. Essa estrutura permite comparar o desempenho do Gemini 2.5 com e sem suporte documental, mantendo simplicidade de uso e alta robustez operacional.



3.4. Fluxo de Comunicação

O fluxo de comunicação foi estruturado para possibilitar a comparação entre o modelo Gemini 2.5 Flash operando em dois modos distintos: com a técnica de Retrieval-Augmented Generation

(RAG) e sem RAG. O usuário interage com o aplicativo móvel, que envia a consulta para a API desenvolvida em NestJS. Após validação e padronização da requisição, a API direciona a mesma pergunta tanto ao módulo de processamento com RAG quanto ao módulo sem RAG.

As respostas retornam à API, onde passam por verificação estrutural e são armazenadas para avaliação comparativa quanto à clareza, assertividade e empatia percebida. Em seguida, a resposta é entregue ao aplicativo de forma uniformizada, garantindo uma interação transparente com o usuário e permitindo a coleta sistemática de dados para o estudo experimental.

4. AVALIAÇÃO QUALITATIVA POR PROFISSIONAIS DE SAÚDE E PACIENTES

A avaliação qualitativa envolveu profissionais de saúde e usuários leigos, incluindo pacientes, permitindo uma análise geral das respostas. Os resultados indicaram diferenças claras entre o Gemini 2.5 Flash com e sem RAG. Observou-se variação no perfil linguístico e no público-alvo percebido. Além disso, o nível de empatia transmitido pelas respostas mostrou-se distinto entre as duas abordagens.

No cenário com RAG, tanto profissionais de saúde quanto participantes leigos relataram que as respostas apresentaram um caráter predominantemente técnico, sendo direcionadas principalmente a um público prescritor, como médicos e outros especialistas da área da saúde. Sob a ótica dos profissionais, esse perfil foi considerado adequado do ponto de vista da precisão

conceitual e do rigor informacional, refletindo maior alinhamento com práticas clínicas e terminologia especializada.

Entretanto, os participantes fora da área médica apontaram que a linguagem adotada nesse modo não se mostrou apropriada para usuários leigos ou pacientes sem formação na área da saúde, limitando sua aplicabilidade em contextos de comunicação direta com pacientes. Além disso, ambos os grupos destacaram uma limitação relevante relacionada à baixa empatia percebida nas respostas, que foram descritas como excessivamente objetivas e pouco acolhedoras.

Em contraste, o modelo operando sem a técnica de RAG foi avaliado de forma mais positiva por participantes leigos, que perceberam maior empatia, sensibilidade comunicacional e uso de linguagem mais acessível. Os profissionais de saúde, por sua vez, reconheceram que esse modo de operação produziu explicações mais claras e compreensíveis, adequadas tanto para o público leigo quanto para indivíduos com formação na área da saúde.

Essa característica foi considerada especialmente relevante em contextos de educação em saúde, orientação inicial ao paciente e esclarecimento de dúvidas relacionadas a condições clínicas. De modo geral, os resultados indicam que o uso da técnica de RAG tende a direcionar o modelo para um estilo de resposta mais técnico e especializado, favorecendo a assertividade e a fundamentação documental.

Por outro lado, essa abordagem ocorre em detrimento da empatia e da adaptação linguística ao público não especializado, o que pode impactar negativamente a experiência de pacientes e usuários

leigos. Em contrapartida, a ausência de RAG permite maior flexibilidade discursiva, resultando em respostas mais empáticas e comunicativamente acessíveis, ainda que potencialmente menos rigorosas quanto à rastreabilidade das informações.

Os resultados quantitativos obtidos a partir da avaliação das respostas geradas pelo modelo Gemini 2.5 Flash, nos cenários com e sem a técnica de Retrieval-Augmented Generation (RAG), são apresentados na Figura 2.

Sem RAG		
Empatia 4.47	Clareza 4.58	Assertividade 4.53
Com RAG		
Empatia 2.33	Clareza 3.44	Assertividade 4.28

Observa-se que, no cenário sem RAG, o modelo obteve médias superiores nos critérios de empatia (4.47), clareza (4.58) e assertividade (4.53). Em contrapartida, no cenário com RAG, os valores médios foram de 2.33 para empatia, 3.44 para clareza e 4.28 para assertividade.

5. CONCLUSÃO

Este estudo comparou o modelo Gemini 2.5 Flash com e sem a técnica de Retrieval-Augmented Generation (RAG) no apoio à assistência e educação de pacientes com diabetes tipo 2. Os resultados indicaram que o uso de RAG aumenta a precisão técnica, melhora a fundamentação documental e amplia a rastreabilidade das informações fornecidas ao usuário.

Do ponto de vista qualitativo, identificaram-se diferenças relevantes nos critérios de empatia, clareza e assertividade. No cenário sem

RAG, o modelo obteve médias superiores de empatia (4,47), clareza (4,58) e assertividade (4,53). Já no modo com RAG, as médias foram de 2,33 para empatia, 3,44 para clareza e 4,28 para assertividade. Esses valores indicam um discurso mais técnico e orientado a evidências no uso de RAG.

Em ambos os cenários, não foram identificadas alucinações, reforçando a confiabilidade geral do modelo. Assim, a escolha pelo uso ou não da técnica de RAG deve considerar o público-alvo e o objetivo da aplicação. O modo com RAG é mais adequado para cenários técnicos, acadêmicos e de apoio a profissionais de saúde. Por outro lado, o modo sem RAG mostra-se mais apropriado para interações diretas com pacientes.

Como perspectiva adicional, os resultados evidenciam a importância de avaliar agentes conversacionais também sob critérios humanos e comunicacionais. A empatia e a clareza mostraram-se fatores determinantes na percepção de utilidade do sistema pelos usuários leigos. Nesse contexto, abordagens híbridas e adaptativas surgem como alternativas promissoras. Essas estratégias podem equilibrar rigor técnico e empatia, ampliando a efetividade de LLMs na área da saúde.

REFERÊNCIAS BIBLIOGRÁFICAS

Aisling O'Shea, John C. D. White, Minsuk Cho, Aakash Prasad, D. J. Seo, M. G. Chung, and S. Y. Lee. Comece a desenvolver com gemini 1.5 flash: a nova versão do modelo multimodal mais rápida do google. Google Developers Blog, 2024. URL <https://developers.googleblog.com/pt-br/start-building-with-gemini-25-flash/>.

Alberto Palazzuoli and Massimo Iacoviello. Diabetes leading to heart failure and heart failure leading to diabetes: epidemiological and clinical evidence. *Heart Failure Reviews*, 27(5):1693–1709, 2022.

Antony Ronald Reagan Panguraj. Exploring how llms can analyse patient data, medical literature, and symptoms to assist in the early detection of chronic conditions like diabetes, cancer, or cardiovascular diseases. *International Journal of Scientific Research in Engineering and Management (IJSREM)*, 07(06), jun 2023. ISSN 2582-3930.

David Garlan. Software architecture. 2008. doi: 10.1002/9780470050118.ecse375. URL <https://doi.org/10.1002/9780470050118.ecse375>.

Dung Dao, Jun Yi Claire Teo, Wenru Wang, and Hoang D. Nguyen. Llm-powered multi-modal ai conversations for diabetes prevention. Association for Computing Machinery, New York, 2024a.

Dung Dao, Jun Yi Claire Teo, Wenru Wang, and Hoang D. Nguyen. Llm-powered multimodal ai conversations for diabetes prevention. In *The 1st ACM Workshop on AI-Powered QA Systems for Multimedia (AIQAM '24)*, AIQAM '24, page 6, New York, NY, USA, 2024b. ACM. ISBN 979-8-4007-0547-2/24/06. doi: 10.1145/3643479.3662049. URL <https://doi.org/10.1145/3643479.3662049>.

Hisashi Johno, Yuki Johno, Akitomo Amakawa, Junichi Sato, Ryota Tozuka, Atsushi Komaba, Hiroaki Watanabe, Hiroki Watanabe, Chihiro Goto, Hiroyuki Morisaka, Hiroshi Onishi, and Kazunori Nakamoto. Enhancing pancreatic cancer staging with large language models: The role of retrieval-augmented generation. *arXiv*

preprint arXiv:2503.15664, 2025. URL <https://arxiv.org/abs/2503.15664>.
Department of Radiology, University of Yamanashi, Japan.

Lai Wei, Zhen Ying, Muiyang He, Yutong Chen, Qian Yang, Yanzhe Hong, Jiaping Lu, Kaipeng Zheng, Shaoting Zhang, Xiaoying Li, Weiran Huang, and Ying Chen. Diabetica: Adapting large language model to enhance multiple medical tasks in diabetes care and management. DIABETICA, 2025.

Mahyar Abbasian, Zhongqi Yang, Elahe Khatibi, Pengfei Zhang, Nitish Nagesh, Iman Azimi, Ramesh Jain, and Amir M Rahmani. Knowledge-infused llm-powered conversational health agent: A case study for diabetes patients. arXiv preprint arXiv:2402.10153, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge intensive nlp tasks. In Advances in Neural Information Processing Systems (NeurIPS 2020), volume 33, pages 9459–9474. Curran Associates, Inc., 2020.
URL <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>.