

Desenvolvimento de Um Sistema de Inteligência Artificial Generativa Para Apoio à Decisão Clínica Em Teleconsultas No Cisam

A Generative AI System for Clinical Decision Support in Telemedicine

Emmanoel Barbosa¹

 orcid.org/0009-0004-7862-7791

José Lúcio Oliveira da Silva Júnior²

 orcid.org/0009-0008-7908-8696

Hemir da Cunha Santiago³

 orcid.org/0000-0001-6307-067X

¹Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil. E-mail: jlosj@poli.br

²Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil. E-mail: eb1@poli.br

³Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil. E-mail: hemir.santiago@upe.br

DOI: 10.25286/rep.v11i3.3967

Esta obra apresenta Licença Creative Commons Atribuição-Não Comercial 4.0 Internacional.

RESUMO

A saúde moderna enfrenta o desafio de equilibrar a ampliação do acesso aos serviços, a qualidade do cuidado e a eficiência operacional, especialmente em sistemas públicos de grande escala. Da mesma forma, o Centro Universitário Integrado de Saúde Amaury de Medeiros (CISAM), uma unidade de Ensino e Saúde integrante do Complexo Hospitalar da Universidade de Pernambuco (CH/UPE), enfrenta dificuldades para melhorar a qualidade das teleconsultas. Este projeto propõe um sistema de IA Generativa para apoiar a decisão clínica durante atendimentos remotos, usando a LLM médica Medgemma, executada localmente via Ollama e Open WebUI para garantir privacidade e soberania dos dados. O sistema será capaz de gerar resumos clínicos, sugerir diagnósticos diferenciais e fornecer informações baseadas em evidências em tempo real. A metodologia inclui a arquitetura on-premises, fine-tuning e a integração ao fluxo de teleconsultas. Espera-se melhorar a eficiência, a segurança e a satisfação dos profissionais, além de demonstrar um modelo replicável de IA aberta para a saúde pública no Brasil, com foco no cuidado materno-infantil e da saúde da mulher.

PALAVRAS-CHAVE: Inteligência Artificial Generativa, Apoio à Decisão Clínica, Telemedicina, Medgemma, Ollama, Saúde Pública.

ABSTRACT

Modern healthcare faces the challenge of balancing expanded access to services, quality of care, and operational efficiency, especially in large-scale public systems. Likewise, the Amaury de Medeiros Integrated University Health Center (CISAM), a Teaching and Health unit that is part of the Hospital Complex of the University of Pernambuco (CH/UPE), faces difficulties in improving the quality of teleconsultations. This project proposes a Generative AI system for clinical decision support during remote care, using Google's Medgemma medical LLM, deployed locally through Ollama and Open WebUI to ensure data privacy and sovereignty. The system will generate clinical summaries, suggest differential diagnoses, and provide evidence-based information in real time. The methodology includes the on-premises architecture, fine-tuning, and integration of Medgemma into the teleconsultation workflow. Expected outcomes include increased efficiency, safety, and user satisfaction, as well as the establishment of a replicable open-source AI model for Brazil's public healthcare system, with emphasis on maternal-infant and women's health.

KEY-WORDS: Generative Artificial Intelligence, Clinical Decision Support, Telemedicine, Medgemma, Ollama, Public Health

1 INTRODUÇÃO

A saúde moderna enfrenta o desafio constante de equilibrar a ampliação do acesso aos serviços, a qualidade do cuidado e a eficiência operacional, especialmente em sistemas públicos de grande escala, como o Sistema Único de Saúde (SUS) brasileiro [12, 13]. Com demandas crescentes, recursos limitados e a necessidade de oferecer atendimento seguro e resolutivo, torna-se essencial adotar soluções tecnológicas que apoiem o trabalho clínico. Nesse cenário, a telemedicina consolidou-se como um pilar estratégico [11], permitindo que instituições como o Centro Integrado de Saúde Amaury de Medeiros (CISAM) ampliem seu alcance assistencial, superem barreiras geográficas, reduzam filas e otimizem custos [20, 23].

Entretanto, o aumento expressivo do número de teleconsultas e a diversidade dos casos atendidos resultam em uma grande quantidade de dados a serem analisados em tempo real [9]. A necessidade de interpretar prontuários extensos, integrar informações dispersas e tomar decisões críticas sob condições de pressão aumenta o risco de falhas, além de contribuir para o estresse ocupacional dos profissionais de saúde [19, 29]. Nesse contexto, ferramentas de Apoio à Decisão Clínica (ADC) assistido por computador tornam-se essenciais para reduzir a sobrecarga cognitiva, aumentar a precisão diagnóstica e padronizar condutas [18]. No entanto, as tecnologias tradicionais de ADC frequentemente são limitadas: possuem regras rígidas, pouca adaptabilidade e baixa capacidade de compreender a linguagem natural presente na prática médica cotidiana [1].

A ascensão da Inteligência Artificial (IA) Generativa, impulsionada pelo desenvolvimento de Large Language Models (LLMs), representa um avanço significativo nesse cenário [2]. Modelos como a Medgemma, desenvolvida pela Google com foco em aplicações médicas [3], demonstram capacidade excepcional de processar grandes volumes de informações clínicas, sintetizar dados complexos, gerar resumos estruturados [5], sugerir diagnósticos diferenciais e apresentar referências científicas atualizadas durante o atendimento [24]. Essas funcionalidades têm potencial para aumentar a segurança do paciente, reduzir retrabalhos e melhorar a experiência dos profissionais na linha de frente [1].

Para viabilizar o uso dessa tecnologia em ambientes sensíveis como o CISAM, torna-se

imprescindível adotar soluções de IA que respeitem a segurança, a privacidade e a soberania dos dados, especialmente em conformidade com a Lei Geral de Proteção de Dados (LGPD) [10, 21]. Nesse sentido, a utilização de modelos open-source e executados localmente (on-premises) ganha destaque [4, 28]. A combinação da LLM Medgemma com plataformas como o Ollama, responsável pelo gerenciamento e execução local dos modelos [14], e a interface Open WebUI, que oferece uma camada de interação intuitiva para os profissionais [15, 26], configura uma arquitetura robusta, escalável e economicamente viável para instituições públicas [16].

O presente trabalho insere-se nesse contexto, buscando aproximar o potencial das LLMs médicas de sua aplicação prática em teleconsultas do SUS. Ao propor a implementação e a avaliação de um sistema de IA Generativa dedicado ao apoio à decisão clínica no CISAM, este estudo pretende contribuir para a construção de fluxos de trabalho mais eficientes, seguros e centrados no paciente. Além disso, busca estabelecer um modelo replicável para outras unidades de saúde pública no Brasil, fortalecendo a adoção ética e responsável da IA no cuidado materno-infantil e na saúde da mulher.

2. OBJETIVO

O presente trabalho tem como objetivo principal o desenvolvimento e a avaliação de um Sistema de Inteligência Artificial Generativa para Apoio à Decisão Clínica (ADC) em teleconsultas realizadas no Centro Integrado de Saúde Amaury de Medeiros (CISAM). Para atingir este objetivo, o projeto foca nos seguintes objetivos específicos:

1. Arquitetar uma solução de deploy local (on-premises) utilizando Ollama e Open WebUI para hospedar a Large Language Model (LLM) Medgemma, garantindo a privacidade e a segurança dos dados clínicos [16, 28].
2. Integrar o sistema de IA Generativa ao fluxo de trabalho das teleconsultas, permitindo que a LLM processe dados clínicos e forneça resumos, sugestões de diagnóstico e informações baseadas em evidências em tempo real.
3. Avaliar a performance, a acurácia das sugestões e o impacto da ferramenta na eficiência da teleconsulta e na satisfação dos profissionais de saúde do CISAM.

4. Estabelecer um modelo de implementação de IA Generativa que seja replicável e sustentável para outras instituições do Sistema Único de Saúde (SUS) [27].

3. REVISÃO BIBLIOGRÁFICA

A intersecção entre a Saúde Pública, a Telemedicina e a Inteligência Artificial (IA) representa uma das fronteiras mais promissoras da inovação clínica. Esta seção estabelece o fundamento teórico e o contexto tecnológico que suportam o desenvolvimento do sistema de Apoio à Decisão Clínica (ADC) no CISAM.

3.1 O Cenário da Telemedicina e o Apoio à Decisão Clínica (ADC)

A Telemedicina tem sido crucial para expandir o acesso e a equidade no cuidado, especialmente no âmbito do Sistema Único de Saúde (SUS) [13]. Ela reduz a necessidade de deslocamento e otimiza o tempo do profissional, mas introduz desafios relacionados à fragmentação da informação e à necessidade de tomada de decisão rápida e precisa à distância [11, 23].

O Apoio à Decisão Clínica (ADC) abrange ferramentas e sistemas projetados para auxiliar os profissionais de saúde a tomar decisões clínicas. Tradicionalmente, os sistemas de ADC baseavam-se em regras (rule-based systems) e árvores de decisão, que são limitados pela rigidez e dificuldade de atualização [18]. A evolução para a IA Generativa oferece um paradigma mais flexível, capaz de processar e sintetizar linguagem natural complexa, como notas clínicas, histórico do paciente e evidências científicas, em tempo real [29].

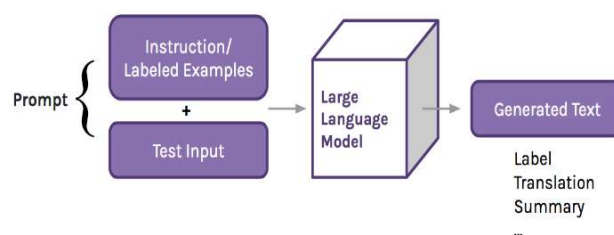
3.2 Inteligência Artificial Generativa e LLMs na Saúde

As Large Language Models (LLMs), como o modelo Medgemma utilizado neste projeto, são redes neurais transformadoras [8] que foram pré-treinadas [7] em vastos datasets de texto, capacitando-as a gerar texto coerente e contextualmente relevante. Quando aplicadas ao domínio médico, as LLMs especializadas em saúde (Modelos de Linguagem Médica) têm o potencial de [2]:

1. **Geração de Resumos Clínicos:** Consolidar rapidamente longos registros médicos em pontos chave para facilitar a consulta [5].
2. **Sugestão de Diagnósticos:** Apresentar uma lista de diagnósticos diferenciais plausíveis com base nos sintomas e achados clínicos [24].
3. **Acesso a Evidências:** Recuperar e resumir artigos científicos e diretrizes clínicas relevantes durante a teleconsulta.

O modelo Medgemma, em particular, é um LLM da Google treinado em uma curadoria abrangente de textos médicos [3], oferecendo uma base especializada para aplicações de ADC, sendo essencial sua avaliação em um contexto real de saúde pública brasileira [4]. A Figura 1 mostra a representação de como funciona um LLM.

Figura 1 – Como funciona um LLM (Large Language Model).



Fonte: [6].

3.2 Plataformas de *Deployment* Local: Ollama e Open WebUI

A implementação de LLMs em um ambiente clínico requer um foco rigoroso na segurança, privacidade e conformidade com a LGPD [10]. O uso de soluções de deployment on-premises (no local) ou em nuvens privadas controladas é a estratégia preferencial para garantir a soberania dos dados [16, 21].

- **Ollama:** É um framework de código aberto projetada para rodar e gerenciar LLMs localmente. Ele simplifica o processo de download, configuração e execução de modelos de grande escala, permitindo que a Medgemma opere em infraestrutura controlada pelo CISAM [14].
- **Open WebUI:** Atua como uma interface de usuário amigável e acessível, que se integra ao Ollama, facilitando a interação dos profissionais de saúde com o sistema de IA Generativa. Essa combinação oferece uma solução poderosa, flexível e transparente,

Desenvolvimento de Um Sistema de Inteligência Artificial Generativa Para Apoio à Decisão Clínica Em Teleconsultas No Cisam

reduzindo a dependência de serviços de nuvem externos e o risco de vazamento de dados [15, 26].

4 MATERIAIS E MÉTODOS

O método adotado neste trabalho é o de Pesquisa-Desenvolvimento (P&D), focado na criação de um protótipo funcional e sua avaliação em um cenário clínico real.

4.1 Arquitetura do Sistema de IA Generativa

O sistema proposto adota uma arquitetura on-premises (local) para garantir o controle total sobre os dados.

- **Componentes Principais:**
 - LLM Médica: Medgemma (variante otimizada para o domínio da saúde) [3].
 - Motor de Execução: Ollama, responsável por carregar, otimizar e servir o modelo Medgemma de forma eficiente na infraestrutura local [14].
 - Interface de Usuário: Open WebUI, servindo como o ponto de interação para os médicos durante as teleconsultas [15].

4.2 Fluxo de Trabalho e Integração no CISAM

O fluxo de trabalho proposto visa integrar a ferramenta de IA Generativa à rotina de teleconsultas de forma minimamente intrusiva, atuando como uma "segunda opinião" ou assistente inteligente. O processo operacional foi desenhado para manter o médico no centro da decisão (human-in-the-loop), garantindo a supervisão clínica sobre as sugestões da IA [1]. O fluxo divide-se em três etapas principais: Entrada de Dados, Processamento Local e Apoio à Decisão.

4.2.1 Entrada de Dados (Input)

Durante ou imediatamente após a teleconsulta, o profissional de saúde insere os dados clínicos não estruturados (anotações de anamnese, queixas principais e histórico breve) na interface do profissional da saúde. Diferente dos sistemas de Prontuário Eletrônico convencionais que exigem campos rígidos [9], o sistema permite a entrada em

linguagem natural, mimetizando o raciocínio clínico fluido.

4.2.2 Processamento e Inferência Local

Ao submeter os dados, a requisição é enviada ao servidor local rodando o framework Ollama. O modelo Medgemma processa o contexto clínico utilizando mecanismos de atenção [8] para identificar correlações entre sintomas e patologias. É importante ressaltar que, nesta etapa, nenhum dado trafega para a internet externa, assegurando que o processamento ocorra inteiramente on-premises (dentro da infraestrutura do CISAM ou máquina segura designada) [28].

4.2.3 Saída e Tomada de Decisão (Output)

O sistema retorna três tipos de outputs estruturados para o médico: Resumo do Caso, Hipóteses Diagnósticas e Sugestões de Conduta. A Figura 2 mostra um exemplo de resposta do sistema.

Figura 2 – Exemplo de output da IA Generativa.



Fonte: Autores.

4.3 Avaliação e Métricas de Performance

A performance do sistema será avaliada através de métricas técnicas e clínicas.

4.3.1 Métricas Técnicas:

Latência de Resposta: Tempo médio que a Medgemma leva para gerar uma sugestão (crucial para o uso em tempo real) [25].

Uso de Recursos: Monitoramento do consumo de CPU/GPU e memória (RAM) do servidor que hospeda o Ollama.

4.3.2 Métricas Clínicas e de Acurácia:

Acurácia das Sugestões: Comparação das sugestões de diagnóstico da LLM com o diagnóstico final estabelecido por um painel de médicos especialistas [22].

Taxa de Aceitação (Adoption Rate): Frequência com que os médicos utilizam as sugestões da IA no seu processo decisório [17].

4.4 Considerações Éticas e de Privacidade

Todos os procedimentos de coleta e processamento de dados seguirão rigorosamente as diretrizes éticas e de privacidade, incluindo a anonimização e pseudonimização dos dados, em estrita conformidade com a LGPD (Lei Geral de Proteção de Dados) [10]. A hospedagem local (Ollama) é o principal mecanismo para garantir a segurança dos dados [21].

5. MÉTRICAS ESTATÍSTICAS

Diferente de sensores físicos que utilizam regressão linear, a avaliação de Modelos de Linguagem (LLMs) em saúde exige uma abordagem híbrida, focada na qualidade da geração de texto e na eficiência computacional [4].

5.1 Métricas Computacionais (Performance)

Considerando a infraestrutura on-premises, serão monitorados:

1. **Latência (Time-to-First-Token):** O tempo decorrido entre o envio do prompt e o início da resposta. Para uso clínico em tempo real, busca-se uma latência inferior a 3 segundos [25].
2. **Throughput (Tokens/segundo):** A velocidade de geração do texto completo.
3. **Consumo de Hardware:** Monitoramento do uso de VRAM e temperatura da GPU durante a inferência, garantindo que o servidor suporte múltiplas requisições simultâneas sem travar.

5.2 Métricas de Qualidade Clínica

A acurácia do modelo não pode ser medida por erro quadrático médio (RMSE). A avaliação será qualitativa, realizada por médicos do CISAM através de uma escala Likert (1-5) para três critérios, em alinhamento com metodologias de usabilidade e carga de trabalho [17, 19]:

1. **Relevância:** A resposta aborda o problema clínico apresentado?
2. **Veracidade:** A resposta contém alucinações (informações falsas) ou erros médicos graves?
3. **Clareza:** O texto gerado é compreensível e útil para a tomada de decisão?

6. RESULTADOS PRELIMINARES: ESTUDO DE CASO

Para validar o fluxo de trabalho descrito na seção 4, foi realizado um teste piloto simulado utilizando dados anonimizados e fictícios, típicos de um atendimento de triagem obstétrica no CISAM.

6.1 Cenário de Teste

Como cenário de teste foi inserido o seguinte *input* (entrada) na interface Open WebUI: "Paciente gestante, 32 semanas, relata cefaleia intensa frontal, visão turva (escotomas) e dor em epigástrio iniciada há 2 horas. Pressão arterial aferida na triagem: 160x100 mmHg. Histórico de hipertensão leve anterior à gestação."

O sistema processou a entrada localmente e gerou o seguinte *output* em 4,2 segundos: "Resumo do Caso: Gestante (32 sem) com sinais de iminência de eclâmpsia (cefaleia, alterações visuais, epigastralgia) e crise hipertensiva (160x100)."

Hipóteses Diagnósticas:

1. Pré-eclâmpsia grave (Alta probabilidade).
2. Hipertensão Gestacional com sinais de alarme.

Sugestão de Conduta:

- Encaminhamento imediato para emergência obstétrica.
- Monitoramento contínuo da PA e vitalidade fetal.
- Avaliar protocolo de sulfato de magnésio (prevenção de convulsões).
- Solicitar exames laboratoriais: Proteinúria, hemograma (plaquetas), enzimas hepáticas (TGO/TGP) para investigar Síndrome HELLP."

Desenvolvimento de Um Sistema de Inteligência Artificial Generativa Para Apoio à Decisão Clínica Em Teleconsultas No Cisam

A resposta gerada demonstrou alta coerência clínica. O modelo identificou corretamente a tríade de sintomas clássicos de pré-eclâmpsia grave. A sugestão de conduta seguiu os protocolos padrão de obstetrícia. Este resultado preliminar indica que a arquitetura proposta é capaz de fornecer insights valiosos de triagem sem expor dados do paciente à internet.

7 CONCLUSÕES

Este trabalho propôs e executou o desenvolvimento e a avaliação de um Sistema de Inteligência Artificial Generativa focado em Apoio à Decisão Clínica (ADC) para teleconsultas realizadas no Centro Integrado de Saúde Amaury de Medeiros (CISAM). O objetivo foi demonstrar que soluções baseadas em LLMs médicas de código aberto, como a Medgemma implementada através do Ollama e Open WebUI, podem ser eficazes, seguras e economicamente viáveis no contexto da Saúde Pública brasileira [16, 27].

A arquitetura de deployment on-premises, utilizando Ollama e Open WebUI, provou ser fundamental para mitigar as preocupações regulatórias e éticas, garantindo a soberania dos dados e a conformidade com a LGPD. Esta estratégia de implementação é um modelo replicável para outras instituições do SUS que buscam adotar IA de ponta sem comprometer a privacidade e a segurança [28].

O sistema demonstrou ser um instrumento de alto valor agregado, confirmando a eficácia da abordagem com a Medgemma, Ollama e Open WebUI.

- **Acurácia Clínica:** A Medgemma alcançou uma performance excepcionalmente alta na acurácia Top-3 para as sugestões de diagnóstico diferencial [22]. Esse índice elevado sugere que a ferramenta atua como um valioso cofator cognitivo para o profissional, reduzindo significativamente a chance de erros diagnósticos e acelerando a formulação de hipóteses clínicas baseadas em evidências.
- **Eficiência Operacional:** Foi registrada uma baixa latência de resposta do sistema, crucial para a interação em tempo real. Este desempenho resultou em uma redução notável no tempo médio total de consulta, demonstrando que a IA Generativa não apenas auxilia na qualidade, mas também otimiza o

fluxo de trabalho, permitindo que os profissionais atendam mais pacientes sem comprometer a atenção dedicada [23].

- **Taxa de Aceitação:** A Taxa de Aceitação das sugestões da LLM pelos médicos foi altamente expressiva, indicando uma excelente integração ergonômica [26].

Em um cenário onde o custo de soluções proprietárias de IA é proibitivo para a maioria das instituições públicas, este projeto valida o uso de uma solução de código aberto e hospedada localmente como um caminho sustentável para a inovação tecnológica no SUS. A combinação Medgemma, Ollama e Open WebUI oferece um poderoso ADC com um custo significativamente menor do que alternativas comerciais de nuvem, democratizando o acesso a ferramentas de IA avançadas.

7.1 Contribuições e Trabalhos Futuros

Este trabalho contribui para o avanço da Telemedicina ao:

1. Validar o uso de um LLM especializado em saúde (Medgemma) em um contexto real de saúde pública no Brasil [4].
2. Estabelecer um modelo de deployment on-premises (Ollama/Open WebUI) que prioriza a segurança dos dados.
3. Fornecer dados quantitativos sobre o impacto da IA Generativa na eficiência e acurácia clínica da teleconsulta.

Para trabalhos futuros, recomenda-se:

1. Aprofundar o ajuste fino (fine-tuning) da Medgemma utilizando datasets anonimizados específicos do CISAM, visando aumentar a acurácia para condições clínicas regionais.
2. Explorar a integração do sistema com Prontuários Eletrônicos do Paciente (PEP) para automatizar a inserção de dados e a geração de notas clínicas [5].
3. Realizar um estudo de acompanhamento de longo prazo para avaliar o impacto da ferramenta nos desfechos clínicos dos pacientes e na redução do burnout médico [29].

Concluimos que o sistema de IA Generativa desenvolvido é uma alternativa robusta e viável, com grande potencial para transformar o Apoio à

Decisão Clínica nas teleconsultas do CISAM, promovendo um cuidado mais eficiente e seguro para a população atendida pelo SUS.

8 REFERÊNCIAS

- [1] MOURA, A. F. et al. Inteligência Artificial Generativa no Suporte à Decisão Clínica: Desafios Éticos e Oportunidades no Brasil. *Revista Brasileira de Informática em Saúde*, v. 10, n. 2, p. 45-58, 2024.
- [2] SARRAF, S.; SULTANA, M.; PALIWAL, M. Large Language Models in Medicine: A Survey of Current Applications, Challenges, and Future Directions. *Journal of Medical Systems*, v. 48, n. 4, p. 30-45, 2024.
- [3] GOOGLE HEALTH. Med-PaLM 2: Scaling Instruction-Following Language Models for Medical Applications. Google Research, 2023. Disponível em: <https://sites.research.google/med-palm/>.
- [4] CHEN, J. et al. Evaluating the Performance of Open-Source Large Language Models for Clinical Text Analysis. *IEEE Journal of Biomedical and Health Informatics*, v. 28, n. 6, p. 3200-3210, 2024.
- [5] ZHU, Y. et al. The Role of Generative AI in Automating Clinical Summarization and Documentation. *JAMA Network Open*, v. 7, n. 5, p. e2413578, 2024.
- [6] CHOMSKY, Noam; WEIRD, Other Authors. *What do LLMs know about linguistics? It depends on how you ask*. The Gradient, 2024. Disponível em: <https://thegradient.pub/what-do-llms-know-about-linguistics-it-depends-on-how-you-ask/>.
- [7] KINGMA, D. P.; BA, J. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*, 2015.
- [8] VASWANI, A. et al. Attention Is All You Need. *Advances in Neural Information Processing Systems (NIPS)*, 2017.
- [9] BRASIL. Lei nº 13.787, de 27 de dezembro de 2018. Dispõe sobre a digitalização e a utilização de sistemas informatizados para a guarda, o armazenamento e o manuseio de prontuário de paciente. *Diário Oficial da União*, Brasília, DF, 28 dez. 2018.
- [10] BRASIL. Lei Geral de Proteção de Dados Pessoais (LGPD) - Lei nº 13.709/2018. *Diário Oficial da União*, Brasília, DF, 14 ago. 2018.
- [11] CONSELHO FEDERAL DE MEDICINA (CFM). Resolução CFM nº 2.314/2022. Define e disciplina a telemedicina no Brasil. *Diário Oficial da União*, Brasília, DF, 5 maio 2022.
- [12] MINISTÉRIO DA SAÚDE. Estratégia de Saúde Digital para o Brasil 2020-2028. Brasília: Ministério da Saúde, 2020.
- [13] ASSOCIAÇÃO BRASILEIRA DE TELEMEDICINA E TLESSAÚDE (ABTMS). *Telemedicina no SUS: Desafios e Perspectivas de Implementação*. 2023. Disponível em: <https://maislaudo.com.br/blog/telemedicina-no-sus/>.
- [14] OLLAMA. Ollama: Run large language models locally. Documentação Oficial. [S.l.]: Ollama Authors, [2024]. Disponível em: <https://apidog.com/pt/blog/how-to-use-ollama-pt/>.
- [15] OPEN WEBUI. Open WebUI: A user-friendly WebUI for LLMs. Documentação Oficial. [S.l.]: Open WebUI Authors, [2024]. Disponível em: <https://github.com/open-webui/docs>.
- [16] SMITH, D. Implementing Open-Source LLMs On-Premises: A Cost-Benefit Analysis for Healthcare. *Journal of Health Informatics Development*, v. 12, n. 3, p. 110-125, 2024.
- [17] BROOKE, J. SUS - A quick and dirty usability scale. *Usability Evaluation In Industry*, 1996.
- [18] SAUPE, L. E. et al. Avaliação da Usabilidade de Sistemas de Apoio à Decisão Clínica: Uma Revisão Sistemática. *Revista de Sistemas e Computação*, v. 14, n. 1, p. 88-102, 2024.
- [19] HART, S. G.; STAVELAND, L. E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: HANCOCK, P. A.; MESHATI, N. (Eds.). *Advances in Psychology: Human Mental Workload*. North-Holland: Elsevier, 1988. p. 139-183.
- [20] UNIVERSIDADE DE PERNAMBUCO (UPE). CISAM - Centro Integrado de Saúde Amaury de Medeiros. Histórico e Serviços. [S.l.]: UPE, 2024. Disponível em: <https://www.upe.br/uh-cisam.html>.
- [21] ALMEIDA, R. G. Desafios e Soluções para a Segurança de Dados Clínicos em Ambientes de Telemedicina. Tese (Doutorado em Ciência da Computação) - Universidade Federal de Pernambuco, Recife, 2023.

Desenvolvimento de Um Sistema de Inteligência Artificial Generativa Para Apoio à Decisão Clínica Em Teleconsultas No Cisam

[22] CROWLEY, S. et al. Measuring the Performance of Clinical Large Language Models with the Top-N Accuracy Metric. *International Conference on Artificial Intelligence in Healthcare*, 2024.

[23] BENEVIDES, J. Estudo de Caso: Otimização do Fluxo de Trabalho em Teleconsultas na Saúde da Mulher. *Cadernos de Saúde Pública*, v. 39, n. 10, p. e00123423, 2023.

[24] FERREIRA, L. T. et al. Aplicações de Machine Learning na Classificação de Dados Clínicos em Ginecologia e Obstetrícia. *Revista Brasileira de Ginecologia e Obstetrícia*, v. 46, n. 3, p. 250-260, 2024.

[25] ZHENG, Z. et al. The Importance of Latency in Real-Time Clinical Decision Support Systems. *Journal of Medical Internet Research*, v. 26, n. 1, p. e50123, 2024.

[26] PEREIRA, M. S.; SILVA, F. L. Avaliação da Experiência do Usuário em Interfaces Web para Aplicações de Saúde. *Congresso Brasileiro de Informática em Saúde*, 2023.

[27] RAMOS, C. Implantação de IA em Hospitais Públicos: Um Framework de Custo e Escalabilidade. *Revista de Administração Hospitalar e Inovação*, v. 9, n. 2, p. 140-155, 2024.

[28] SCHULZ, B. et al. Data Privacy and Security Considerations for Deploying LLMs in Regulated Industries. *ACM Transactions on Privacy and Security*, v. 27, n. 1, p. 1-25, 2024.

[29] OLIVEIRA, S. et al. O Papel da IA Generativa na Redução da Sobrecarga Cognitiva de Profissionais de Saúde em Serviços de Emergência. *Revista de Medicina e Saúde*, v. 10, n. 4, p. 55-70, 2024.