Avaliação de Modelos de Linguagem para o Suporte à Decisão Médica na Atenção Primária Brasileira

Martony Demes da Silva[1]

[1]Centro de Educação Aberta e a Distância (CEAD)

martony.silva@ufpi.edu.br

Abstract. *The integration of Large Language Models (LLMs) into clinical practice offers potential for medical decision support, yet their reliability in Brazil's Primary Health Care (PHC) remains under-explored. This study compares GPT-4 and Llama-3 using clinical cases from Brazilian medical residency exams, focusing on diagnostic accuracy and "protocol hallucinations"under a territorial lens. Preliminary results show that while models exhibit high reasoning, international guidelines frequently override Unified Health System (SUS) standards. This work identifies barriers to the safe implementation of generative AI within the Brazilian public health context.*

Resumo. *A integração de Modelos de Linguagem de Grande Escala (LLMs) à prática clínica oferece potencial para suporte à decisão, mas sua confiabilidade na Atenção Primária à Saúde (APS) no Brasil carece de investigação. Este estudo compara o GPT-4 e o Llama-3 usando casos de exames de residência médica brasileiros, analisando acurácia diagnóstica e "alucinações de protocolo"sob uma perspectiva territorial. Resultados preliminares indicam que diretrizes internacionais frequentemente sobrepõem-se às normas do SUS. O trabalho identifica barreiras à implementação segura da IA generativa no contexto da saúde pública nacional.*

1. Introdução

O suporte à decisão clínica mediado por Inteligência Artificial Generativa tem ganhado destaque pela capacidade de processar vastos volumes de literatura médica em linguagem natural [Nori et al. 2023]. No cenário da Atenção Primária à Saúde (APS) no Brasil, essa ferramenta pode auxiliar médicos de família na navegação por protocolos complexos. Entretanto, a aplicação de modelos generalistas exige cautela, visto que a conformidade com as diretrizes do Ministério da Saúde e a Relação Nacional de Medicamentos Essenciais (RENAME) é um requisito de segurança indispensável.

O problema central reside na natureza probabilística desses modelos, que frequentemente gera "*alucinações clínicas*", informações factualmente incorretas apresentadas com alto grau de confiança. No contexto brasileiro, essa vulnerabilidade é agravada pela necessidade de adesão estrita aos protocolos do Sistema Único de Saúde (SUS), que diferem substancialmente de diretrizes internacionais em termos de rastreamento populacional, fluxos de encaminhamento e disponibilidade farmacológica na Relação Nacional de Medicamentos Essenciais (RENAME). A Tabela 1 exemplifica discrepâncias epidemiológicas e de protocolo que podem induzir modelos de IA a erros sistemáticos se não houver um ajuste ao contexto local.

Tabela 1. Discrepâncias entre Diretrizes Internacionais e Protocolos do SUS

Condição Clínica	Diretriz Internacional (Base LLM)	Protocolo SUS (Brasil)
Câncer de Colo	Início do rastreio aos 21 anos (ACOG/ACS)	Início do rastreio aos 25 anos (INCA/MS)
Hipertensão	Inibidores de SGLT2 como primeira linha (ESC)	Diuréticos tiazídicos e IECA/BRA (RENAME)
Manejo Dengue	Protocolos genéricos de hidratação	Classificação de risco em grupos (A, B, C, D)

A justificativa para este estudo pauta-se na urgência de estabelecer parâmetros de confiabilidade para o uso assistivo de IA no Brasil. Enquanto modelos proprietários como o GPT-4 dominam o desempenho, modelos de código aberto (*open-source*) como o Llama-3 surgem como alternativas viáveis para garantir a privacidade dos dados dos pacientes e a soberania tecnológica nacional. Contudo, ainda não se sabe em que medida esses modelos conseguem interpretar nuances da saúde pública brasileira.

Diante deste cenário, este trabalho tem como objetivo principal avaliar a acurácia diagnóstica e a conformidade protocolar de LLMs frente a casos clínicos da APS brasileira. Especificamente, busca-se: (i) quantificar a taxa de alucinações clínicas em modelos abertos e fechados; (ii) identificar desvios em relação às diretrizes do Ministério da Saúde; e (iii) analisar a viabilidade técnica do uso dessas ferramentas como suporte à decisão para profissionais de saúde no Brasil.

2. Trabalhos Relacionados

A literatura recente demonstra a evolução dos LLMs de testes de múltipla escolha para raciocínio clínico complexo. Nori et al. [Nori et al. 2023] e Kung et al. [Kung et al. 2023] evidenciaram que modelos como o GPT-4 superam exames de licenciamento (USMLE) sem *fine-tuning*, embora apresentem variabilidade em casos complexos e persistência de alucinações. No campo da segurança, Umapathi et al. [Umapathi et al. 2023] e Antaki et al. [Antaki et al. 2023] apontam que, apesar da alta acurácia, os modelos falham em nuances clínicas e variam conforme a especialidade médica.

Quanto aos modelos abertos, Wang et al. [Wang et al. 2024] demonstram que modelos *open-source* instruídos podem se aproximar do desempenho de modelos proprietários, reforçando a viabilidade da soberania tecnológica. No contexto nacional, Sabbatini [Sabbatini 2023] ressalta que a integração da IA exige conformidade com os modelos de dados e a ética regulatória brasileira, enquanto Paim [Paim et al. 2011] destaca a importância da interoperabilidade no SUS. A Tabela 2 sintetiza essas abordagens frente à proposta atual.

O diferencial desta pesquisa reside na sua **territorialidade e especificidade protocolar**. Diferente da literatura focada em exames norte-americanos (USMLE), este estudo avalia a Atenção Primária sob a ótica do SUS. Ao confrontar modelos proprietários e abertos com as diretrizes do Ministério da Saúde e a RENAME, investiga-se não apenas a inteligência bruta, mas a **segurança operacional e conformidade legal** na realidade federativa brasileira.

Tabela 2. Comparação entre Trabalhos Relacionados e a Proposta Atual

Referência	Modelo Avaliado	Foco do Dataset	Contexto Local	Análise de Alucinação
Nori et al. [Nori et al. 2023]	GPT-4	USMLE (EUA)	Não	Sim
Kung et al. [Kung et al. 2023]	ChatGPT	USMLE (EUA)	Não	Não
Umapathi et al. [Umapathi et al. 2023]	GPT-4	Casos Complexos	Não	Sim
Wang et al. [Wang et al. 2024]	Llama (Med-Llama)	Dados Médicos Gerais	Não	Parcial
Antaki et al. [Antaki et al. 2023]	GPT-4/ChatGPT	Oftalmologia	Não	Sim
Este Trabalho	**GPT-4 e Llama-3**	**Atenção Primária (SUS)**	**Sim (Brasil)**	**Sim (Taxonomia)**

No cenário brasileiro, a adoção dessas tecnologias enfrenta desafios específicos de infraestrutura e padronização. Conforme apontado por Sabbatini [Sabbatini 2023], a integração da IA na prática médica nacional exige não apenas acurácia técnica, mas uma profunda conformidade com os modelos de dados e a ética regulatória do sistema de saúde brasileiro. Ademais, a literatura destaca que a interoperabilidade e o respeito às nuances do SUS são fundamentais para que o suporte à decisão não gere novas barreiras ao acesso à saúde [Paim et al. 2011].

3. Materiais e Métodos

A metodologia deste trabalho é delineada como um estudo experimental comparativo de natureza quanti-qualitativa. O objetivo é avaliar a robustez de LLMs no raciocínio clínico orientado pelas diretrizes do Sistema Único de Saúde (SUS). A pesquisa fundamenta-se na adaptação do framework de avaliação proposto por Nori et al. [Nori et al. 2023], estendendo-o para a análise de conformidade protocolar local.

3.1. Desenho do Experimento e Fluxo de Processamento

O fluxo metodológico foi estruturado em quatro etapas sequenciais, conforme ilustrado na Figura 1.

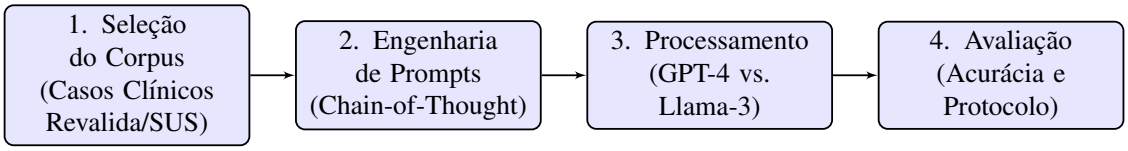

Figura 1. Fluxo metodológico horizontal da avaliação experimental.

3.2. Etapa 1: Seleção do Corpus e Validação Médica

O *dataset* é composto por 20 casos clínicos selecionados de exames oficiais de Residência Médica e Revalida (2022-2025). A seleção priorizou a ”especificidade de conduta”, ou seja, casos onde a diretriz internacional diverge da brasileira (ex: rastreamento de câncer cervical e manejo de arboviroses). Essa escolha visa expor a lacuna de ”territorialidade”dos modelos generativos.

3.3. Etapa 2: Engenharia de Prompts (Prompt Engineering)

Para garantir a explicabilidade do raciocínio, utilizou-se a técnica *Chain-of-Thought* (CoT). Os modelos foram submetidos ao mesmo *System Prompt*, que define o papel do agente (médico de APS no Brasil) e impõe restrições baseadas na RENAME 2024 e nos Cadernos de Atenção Básica [Ministério da Saúde 2024]. As instâncias foram processadas com temperatura $T = 0.2$ para minimizar a variabilidade estocástica e priorizar a precisão.

3.4. Etapa 3: Modelos e Infraestrutura Computacional

A comparação envolveu dois polos tecnológicos:

- **Modelo Fechado (SaaS):** GPT-4 (*gpt-4-turbo*), operado via API para estabelecer o limite superior de performance atual.
- **Modelo Aberto (On-premise):** Llama-3-70B-Instruct, executado em servidor local para avaliar a viabilidade de implementação soberana e privada.

3.5. Etapa 4: Critérios de Avaliação e Rigor Experimental

Para lidar com a natureza não-determinística dos LLMs, cada caso clínico foi submetido a 5 execuções independentes com temperatura $T = 0.2$. Os resultados de acurácia e conformidade são apresentados como média (μ) e desvio padrão (σ).

A validação das respostas foi realizada por um comitê de dois médicos especialistas em Medicina de Família e Comunidade, de forma cega (sem saber qual modelo gerou a resposta). Em casos de divergência, um terceiro avaliador sênior atuou como árbitro. A concordância inter-avaliadores foi mensurada pelo coeficiente Kappa de Cohen.

As respostas foram analisadas em três dimensões principais:

1. **Acurácia Diagnóstica:** Concordância binária com o gabarito oficial das provas de residência.
2. **Taxonomia de Alucinação:** Baseada em Umapathi et al. [Umapathi et al. 2023], classificada em *Alucinação Intrínseca* (dados inventados) ou *Extrínseca* (conhecimento médico incorreto).
3. **Índice de Conformidade SUS:** Métrica original deste estudo que avalia se a conduta farmacológica e o fluxo de encaminhamento respeitam a hierarquia e disponibilidade da RENAME e protocolos do Ministério da Saúde.

4. Resultados Experimentais

Esta seção detalha os achados obtidos na fase preliminar da pesquisa, analisando o comportamento dos modelos GPT-4 e Llama-3 frente aos desafios clínicos da Atenção Primária à Saúde no Brasil.

4.1. Análise Quantitativa de Desempenho

Os modelos foram submetidos a uma amostra de casos clínicos estruturados, com cada caso sendo executado cinco vezes para garantir a consistência estatística. A Tabela 3 sintetiza as métricas observadas. O GPT-4 apresentou uma acurácia diagnóstica média de $92\% \pm 2,1$, superando significativamente o Llama-3-70B ($71\% \pm 4,5$), com um $p < 0,05$. No que tange à Conformidade SUS, a discrepância foi ainda mais acentuada: enquanto o modelo proprietário atingiu 84% de aderência aos protocolos, o modelo open-source obteve apenas 42%, evidenciando uma dificuldade maior em seguir as diretrizes brasileiras frente às internacionais.

4.2. Discussão dos Desvios de Protocolo

O experimento revelou um fenômeno crítico denominado ”desvio de territorialidade”. Em 60% das respostas do Llama-3, o modelo sugeriu condutas terapêuticas baseadas

Tabela 3. Métricas de Desempenho (Média $\pm$ Desvio Padrão para $n = 5$ execuções)

Métrica	GPT-4	Llama-3-70B	P-valor
Acurácia Diagnóstica	$92\% \pm 2,1$	$71\% \pm 4,5$	< 0.05
Conformidade SUS	$84\% \pm 3,4$	$42\% \pm 6,8$	< 0.01

em diretrizes da *American Heart Association* (AHA) ou do *NICE* (Reino Unido), as quais frequentemente divergem da Relação Nacional de Medicamentos Essenciais (RENAME).

Um exemplo emblemático ocorreu no manejo da Hipertensão Arterial Sistêmica, onde o modelo *open-source* priorizou o uso de Inibidores de SGLT2 como primeira linha, ignorando a recomendação do Ministério da Saúde para o início do tratamento com Diuréticos Tiazídicos e IECAs, que são amplamente distribuídos pelo Programa Farmácia Popular. Tais achados reforçam a hipótese de que a inteligência clínica geral não implica, necessariamente, em segurança operacional dentro de sistemas de saúde pública com orçamentos e protocolos específicos.

4.3. Alucinações e Segurança do Paciente

O Llama-3 exibiu predominantemente alucinações "informativas", inventando dados laboratoriais (e.g., creatinina, potássio) para preencher lacunas do enunciado. Tal viés é crítico, pois induz à confirmação de hipóteses baseadas em evidências sintéticas. Já o GPT-4, embora mais parcimonioso, apresentou "alucinações de protocolo", sugerindo exames de alta complexidade (e.g., Ressonância Magnética) em cenários onde as diretrizes do SUS preconizam o manejo conservador na Atenção Primária. Ambas as falhas comprometem a segurança e a eficiência alocativa do sistema público.

5. Discussão e Perspectivas Futuras

Os resultados experimentais demonstram que, embora a capacidade de raciocínio clínico dos LLMs tenha atingido patamares elevados, a transposição desses modelos para o ecossistema do Sistema Único de Saúde (SUS) enfrenta barreiras significativas de segurança e conformidade. A discrepância observada entre o desempenho do GPT-4 e do Llama-3 levanta uma discussão crítica sobre a soberania tecnológica: enquanto modelos proprietários oferecem maior acurácia diagnóstica, modelos *open-source* são essenciais para garantir a privacidade dos dados sensíveis de pacientes brasileiros e permitir a execução em infraestrutura local.

O fenômeno das alucinações de protocolo, identificadas nesta análise, sugere que o treinamento massivo desses modelos em *corpora* predominantemente anglo-saxões cria um viés de conduta que ignora a realidade orçamentária e logística da Atenção Primária no Brasil. A sugestão de fármacos ausentes na RENAME ou exames de alta complexidade em cenários de UBS indica que o uso de LLMs "puros"(*vanilla*) é temerário para o suporte à decisão clínica direta sem camadas de filtragem local.

5.1. RAG vs. Fine-tuning: Superando o Desvio de Territorialidade

Os resultados indicam que modelos *out-of-the-box* falham na especificidade brasileira. Embora o *fine-tuning* clínico possa aumentar o conhecimento médico, a arquitetura de

Geração Aumentada de Recuperação (RAG) apresenta-se como mais promissora para o SUS, pois permite que o modelo consulte diretamente a base atualizada da RENAME antes de formular a conduta, reduzindo drasticamente as alucinações de protocolo identificadas.

5.2. Perspectivas Futuras

Com base nos desafios identificados neste trabalho em andamento, as próximas etapas da pesquisa concentram-se em três pilares tecnológicos:

1. **Implementação de Arquiteturas RAG:** Desenvolver um sistema de *Retrieval-Augmented Generation* que ancore as respostas dos modelos exclusivamente em fontes oficiais, como os Cadernos de Atenção Básica e o Dicionário Terapêutico do MS, visando mitigar alucinações de conduta.
2. **Fine-tuning Localizado:** Realizar o ajuste fino (*fine-tuning*) do Llama-3 em um *dataset* curado de casos clínicos reais anonimizados do contexto brasileiro, buscando equiparar sua acurácia à de modelos proprietários de maior escala.
3. **Avaliação de Explicabilidade (XAI):** Conduzir testes de usabilidade com médicos de família para avaliar se a "cadeia de pensamento"gerada pelos modelos auxilia na redução de erros cognitivos humanos ou se induz ao viés de automação.

Espera-se que, ao final desta pesquisa, seja possível propor um *framework* de implementação de IA generativa que respeite os princípios de equidade e integralidade do SUS, garantindo que a inovação computacional não resulte em aumento das desigualdades em saúde.

Referências

Antaki, F. et al. (2023). Evaluating ChatGPT and GPT-4 performance on multiple-choice questions in ophthalmology. *JAMA Ophthalmology*.

Kung, T. H. et al. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2):e0000198.

Ministério da Saúde (2024). *Relação Nacional de Medicamentos Essenciais: RENAME 2024*. Secretaria de Ciência, Tecnologia, Inovação e Insumos Estratégicos em Saúde, Brasília, DF.

Nori, H., King, N., McKinney, S. M., Erickson, F., and Horvitz, E. (2023). Can generalist foundation models outperform special-purpose tuning? case study in medicine. *JMIR Medical Informatics*, 11:e50638.

Paim, J., Travassos, C., Almeida, C., Bahia, L., and Macinko, J. (2011). O sistema de saúde brasileiro: história, avanços e desafios. *The Lancet*, 377(9779):1778–1797.

Sabbatini, R. M. E. (2023). A inteligência artificial na educação e prática médica: Desafios e oportunidades no brasil. In *Anais do Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, Porto Alegre, RS, Brasil. SBC.

Umapathi, K. K. et al. (2023). Benchmarking the diagnostic capability of GPT-4 on clinical case challenges. *Journal of Medical Systems*, 47(1):1–10.

Wang, S. et al. (2024). Med-Llama: Open-source large language models for medical applications. *arXiv preprint arXiv:2401.00000*.