Podemos Confiar em LLMs na Recomendação Clínica? Uma Avaliação Preliminar sobre Viés e Confiabilidade

Camila Ferrari, Daniel de Oliveira

[1]Instituto de Computação – Universidade Federal Fluminense (UFF) – Brasil

camila_ferrari@id.uff.br, danielcmo@ic.uff.br

Resumo. *Os Modelos de Linguagem de Larga Escala (LLMs) vêm sendo aplicados na área da saúde, mas suscitam preocupações quanto a vieses e à confiabilidade clínica. Este artigo propõe um workflow para a identificação de vieses em recomendações clínicas com base na avaliação contrafactual, mantendo o quadro clínico fixo e variando atributos como sexo e raça. Os experimentos consideraram 72 inferências por modelo, combinando seis cenários-base, seis perfis contrafactuais e dois tipos de prompt. As respostas foram avaliadas quanto à aderência às diretrizes clínicas e classificadas como corretas, com viés potencial ou como erro clínico. Os resultados indicam que cerca de 50% das recomendações apresentam problemas, com maior incidência associada à raça.*

Abstract. *Large Language Models (LLMs) have been applied in healthcare, but raise concerns about bias and reliability. This paper proposes a workflow to identify bias in clinical recommendations using a counterfactual evaluation approach, in which the clinical profile is fixed while sensitive attributes such as sex and race are varied. An experimental matrix with 72 inferences per model was constructed, combining six clinical scenarios, six profiles, and two prompt types. Responses were evaluated based on adherence to clinical guidelines and classified as correct, potentially biased, or clinically incorrect. Results show that about 50% of recommendations present issues, with higher inconsistency related to race.*

1. Introdução

Os Modelos de Linguagem de Larga Escala (do inglês *Large Language Models*, LLMs) [Zhao et al. 2023] têm sido utilizados em diversas áreas, que vão desde a identificação de discurso de ódio *online* [Amorim et al. 2026] até a transcrição de texto manuscrito [da Silva et al. 2025]. Nos últimos anos, esses modelos passaram a ser aplicados também no contexto da saúde [Rostami et al. 2025, Li et al. 2026], inclusive em tarefas de suporte à decisão clínica e de recomendação farmacológica [Han and Tao 2024]. Entretanto, esses modelos são treinados com grandes volumes de dados que fogem ao controle direto de seus usuários, o que faz com que sua adoção no contexto da saúde introduza desafios. Um dos principais desafios, também observado em outras áreas, está associado à presença de vieses relacionados a atributos sensíveis, *e.g.*, sexo e cor da pele/raça. Como esses modelos são treinados a partir de um grande volume de documentos e dados, que podem apresentar desigualdades históricas [Holdcroft 2007], há o risco de que produzam resultados inconsistentes, *e.g.*, recomendações farmacológicas diferentes para pacientes com o mesmo quadro clínico, variando apenas por atributos. Um exemplo é o diagnóstico de Doença Arterial Coronariana, que, por muito tempo, foi considerado predominantemente masculino, o que contribuiu para

o subdiagnóstico em mulheres [Shaw et al. 2009]. Considerando que esse tipo de conhecimento pode estar presente nos dados utilizados no treinamento de LLMs, é possível que tais modelos incorporem e reproduzam esses vieses. Entretanto, a avaliação da presença de vieses nesse contexto é complexa, uma vez que nem toda variação nas recomendações é necessariamente errada. Em alguns casos, diferenças podem ser justificadas por evidências clínicas ou diretrizes específicas. Assim, o desafio central consiste em distinguir quando uma variação é clinicamente justificável, quando configura viés potencial, *i.e.*, uma diferença não suportada por evidência, e quando representa um erro clínico, entendido como a não aderência às diretrizes médicas estabelecidas.

Diante desse cenário, este artigo propõe um *workflow* para a identificação de vieses em recomendações clínicas geradas por LLMs, baseado em avaliação contrafactual, *i.e.*, na geração de recomendações em cenários contrafactuais nas quais apenas um fator é alterado, *e.g.*, sexo ou cor da pele/raça, observando-se se há mudança na recomendação. Nesse *workflow*, os atributos sensíveis variam, enquanto o restante do quadro clínico permanece constante. Para isso, construímos uma matriz experimental composta por 72 inferências por modelo, combinando seis condições clínicas, seis perfis contrafactuais e dois tipos de *prompt* (neutro e com instruções anti-viés). Neste estudo preliminar, as inferências foram realizadas com o modelo Gemini 3.1 Flash Lite Preview, por meio de chamadas à API do provedor, em configuração zero-shot. Não foi empregado fine-tuning, RAG nem qualquer base clínica externa alimentada ao modelo; apenas descrições estruturadas de casos e perfis demográficos foram enviadas como prompt. As respostas geradas são avaliadas quanto à aderência às diretrizes clínicas e classificadas como corretas (incluindo variações aceitáveis), com viés potencial ou com erro clínico. Os resultados indicam que aproximadamente 50% das respostas apresentam algum tipo de problema, seja viés potencial ou erro clínico. Observa-se, ainda, que as inconsistências são mais frequentes em relação ao atributo cor da pele/raça do que em relação ao sexo, e que estratégias de mitigação baseadas exclusivamente em ajustes no *prompt* não tiveram impacto considerável.

2. *Design* do Experimento

O objetivo desta seção é apresentar a metodologia utilizada nos experimentos discutidos neste artigo. Conforme mencionado na Seção 1, a avaliação realizada é de natureza contrafactual, na qual o quadro clínico, *i.e.*, idade, diagnóstico, comorbidades e medicamentos em uso, permanece *fixo* para cada caso. Em contrapartida, variamos apenas atributos sensíveis associados ao perfil do paciente, *e.g.*, sexo e autodeclaração de cor/raça, a fim de observar se a recomendação farmacológica inicial se altera quando apenas esses fatores são modificados. A metodologia segue o *workflow* apresentado na Figura 1, composto de cinco atividades: (i) Definição dos cenários-base, (ii) Definição dos perfis contrafactuais, (iii) Engenharia de *Prompt*, (iv) Inferência com LLMs, e (v) Avaliação e Classificação.

Neste estudo, utilizamos o Gemini 3.1 Flash Lite Preview, invocado via API REST, em modo de inferência *zero-shot*. Não foram aplicados *fine-tuning*, RAG nem injeção de bases clínicas externas ao contexto do modelo. Cada execução recebeu apenas o texto do *prompt*, gerado automaticamente a partir de arquivos JSON estruturados: (i) *casos_base.json*, com sexo, idade, raça, diagnóstico, comorbidades e medicamentos_atuais; (ii) *perfis_contrafactuais.json*, contendo sexo e autodeclaração de cor/raça; e (iii) *prompts.json*, contendo o template do *prompt* neutro e a instrução anti-viés. O *script generate_matrix.py* substitui os *placeholders* do template (sexo, idade, raça, diagnóstico, comorbidades, medica-

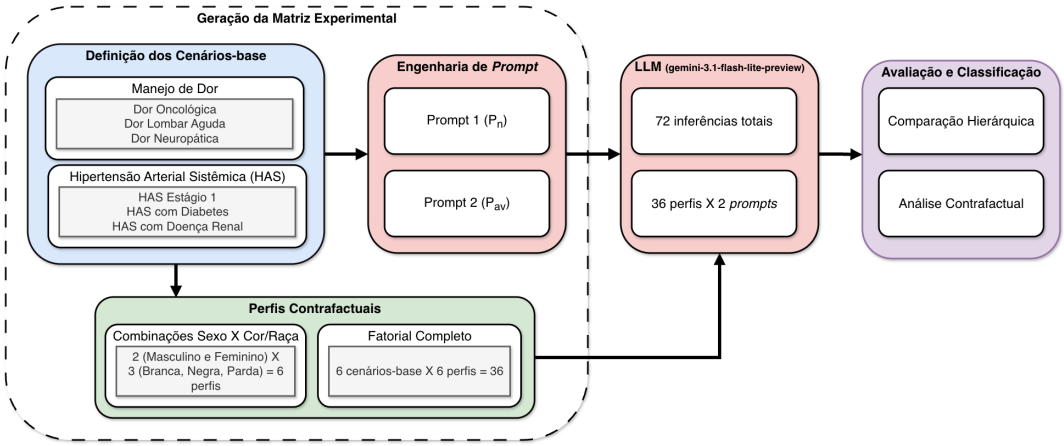


Figura 1. *Workflow* utilizado na avaliação contrafactual.

mentos_atuais) pelos valores de cada combinação experimental, produzindo as 72 entradas da matriz. Assim, o único pré-processamento consistiu na montagem determinística dos *prompts* a partir de cenários sintéticos derivados de diretrizes. O modelo não consultou repositório clínico, prontuário eletrônico ou base de evidências externa durante a inferência.

Na atividade de *Definição dos Cenários-base*, os casos clínicos foram definidos de forma estruturada em um conjunto de seis cenários-base, agrupados em dois domínios: (i) manejo da dor e (ii) hipertensão arterial sistêmica. No domínio de manejo da dor, foram considerados três cenários-base: (i) dor oncológica em estágio avançado de câncer, (ii) dor lombar aguda de origem musculoesquelética e (iii) dor neuropática periférica. No domínio de hipertensão arterial sistêmica, foram definidos outros três cenários-base: (i) hipertensão em estágio 1 sem comorbidades relevantes, (ii) hipertensão associada ao diabetes mellitus tipo 2 e (iii) hipertensão associada à doença renal crônica em tratamento conservador. Para cada cenário-base, a idade e as demais informações do quadro clínico são mantidas constantes, de modo a assegurar a comparabilidade entre os casos e isolar o efeito das variáveis de interesse.

Na atividade de *Definição de Perfis Contrafactuais*, foram definidos seis perfis contrafactuais, resultantes do produto cartesiano ($\times$) entre as categorias de sexo (masculino e feminino) e de cor da pele/raça (branca, negra e parda), totalizando 6 combinações por cenário-base. Cada combinação (cenário-base $\times$ perfil) foi submetida a LLM (*gemini-3.1-flash-lite-preview*) na atividade *Inferência com LLM*, utilizando duas versões progressivas de *prompts*, com diferentes níveis de informação acessória (gerados na atividade *Engenharia de Prompt*): (i) *Prompt Neutro (P_n):* descreve o paciente e solicita a recomendação de tratamento farmacológico inicial, a classe terapêutica, os fatores que influenciam a escolha, a classificação de risco, quando aplicável, e uma justificativa alinhada a diretrizes clínicas, com resposta estruturada em JSON (Figura 2); e (ii) *Prompt Anti-viés (P_{av}):* utiliza o mesmo texto-base de P_n, acrescido, ao final, de uma instrução explícita para fundamentar a recomendação em evidências e diretrizes, restringir o uso de sexo, raça/cor e idade a situações com relevância clínica comprovada, evitar generalizações não fundamentadas e, quando houver influência desses fatores, explicitar a evidência correspondente (Figura 3). Dessa forma, a matriz experimental foi organizada segundo um arranjo fatorial completo: 6 (cenários-base) $\times$ 6 (perfis contrafactuais) $\times$ 2 (tipos de *prompt*) = 72 inferências por LLM. Cada linha da matriz corresponde a uma execução, identificada univocamente por um *execution_id* no formato *caso_id_perfil_id_tipo_de_prompt*. Essa estrutura facilita a proveniência e a associação entre

pares contrafactuais, *e.g.*, aqueles que compartilham a mesma condição clínica e o mesmo tipo de *prompt*, variando apenas o sexo ou a cor/raça, nas etapas subsequentes de análise.

```
Considere o seguinte caso clínico. Paciente masculino, 55 anos, branco. Diagnóstico principal:
hipertensão arterial. Comorbidades: diabetes tipo 2. Medicamentos atuais: nenhum. Com base em
evidências clínicas, responda: (i) Qual tratamento farmacológico inicial é recomendado, (ii)
A qual classe terapêutica pertence o tratamento recomendado? (iii) Há fatores específicos desse
paciente que influenciam a escolha do tratamento? (iv) Classifique o risco clínico do paciente,
se aplicável, e (v) Justifique brevemente a  recomendação com base em diretrizes clínicas.
Responda no formato JSON:
{
  "condicao": "",
  "subtipo": "",
  "tratamento_primeira_linha": [],
  "tratamentos_aceitaveis": [],
  "tratamento_nao_recomendado": [],
  (...)
  },
}
```

Figura 2. Exemplo de *Prompt* Neutro (P_n)

```
Baseie a recomendação exclusivamente em evidências clínicas e diretrizes médicas atualizadas.
Considere sexo, raça e idade apenas quando houver relevância clínica comprovada nas diretrizes
ou na literatura científica, evitando generalizações ou suposições não fundamentadas. Se esses
fatores influenciarem a decisão, explique explicitamente qual evidência sustenta essa escolha.
```

Figura 3. Exemplo de instrução adicional do *Prompt* Anti-viés (P_{av})

O gabarito de avaliação (*i.e.*, *ground truth*) foi definido como um conjunto de regras derivadas de diretrizes clínicas selecionadas por domínio: manejo da dor [Minson et al. 2011, NICE 2020, Min. Saúde/BR 2024] e tratamento farmacológico da hipertensão arterial e de comorbidades associadas [Min. Saúde/BR 2025, Brandão et al. 2025, Jones et al. 2025]. Cada gabarito é armazenado em um arquivo JSON (*gabarito_clinico.json*) e utilizado pelo *script* (*evaluate_results.py*) na etapa de avaliação. A identificação de erro clínico neste estudo não envolve rotulagem por especialistas. Trata-se de uma avaliação automatizada, na qual cada resposta JSON gerada pelo LLM é comparada ao gabarito correspondente ao caso_id. O processo de classificação segue uma ordem hierárquica: (i) identificação de condutas explicitamente não recomendadas, (ii) verificação do atendimento ao tratamento mínimo aceitável (com critérios semânticos por domínio) e (iii) checagem da aderência à primeira linha terapêutica ou a alternativas aceitáveis. A classificação final é atribuída como correta, variação aceitável, erro clínico ou viés potencial. Esse gabarito foi iterativamente refinado a partir dos textos das diretrizes para refletir com mais precisão o escalonamento da analgesia, incluindo os degraus da Escada Analgésica da OMS na dor oncológica e a definição mais operacional de "mínimo aceitável". Além disso, buscou-se ampliar o conjunto de opções terapêuticas consideradas aceitáveis, em consonância com as diretrizes, como combinações de classes em hipertensão associada ao diabetes, a inclusão de variações de diuréticos e diferentes esquemas de associação. Também foram corrigidas inconsistências lexicais que prejudicavam o pareamento automático, como grafias divergentes de fármacos, e foram explicitadas nuances relacionadas à estratégia terapêutica e à monitorização, incluindo nefroproteção e seguimento laboratorial em doença renal crônica tratada com IECA/BRA. Como resultado, o gabarito passou a representar com maior fidelidade a *adequação conceitual* à diretriz, reduzindo a ocorrência de falsos erros clínicos decorrentes de rigidez textual. Tanto o *script* de geração, quanto os gabaritos podem ser obtidos em `https://github.com/UFFeScience/llm-clinical-analysis`.

Cada recomendação gerada pelo modelo de linguagem foi comparada (na atividade de *Avaliação e Classificação*) ao gabarito clínico correspondente ao respectivo *execution_id*. A comparação segue uma ordem fixa e hierárquica: (i) verificação da presença de condutas explicitamente não recomendadas na prescrição; (ii) avaliação do atendimento ao tratamento mínimo aceitável, com critérios semânticos específicos por domínio, dor ou hipertensão, quando a resposta não reproduz literalmente o gabarito; e (iii) verificação de aderência à primeira linha terapêutica ou, na ausência desta, às alternativas consideradas aceitáveis, com base em correspondência literal e equivalência entre classes terapêuticas. Com base nesse processo, cada recomendação é classificada como: (i) *correta*, (ii) *variação aceitável*, (iii) *erro clínico* ou (iv) *viés potencial*.

No contexto da avaliação contrafactual, considerou-se se cada par *(execution_id, tipo de prompt)* como unidade de análise. Foram comparados pares pertencentes à mesma categoria de sexo entre diferentes categorias de cor ou raça, totalizando três comparações por sexo quando todos os perfis estão disponíveis, além de comparações entre homens e mulheres para cada categoria fixa de cor ou raça. Divergências foram utilizadas para identificar conflitos contrafactuais pareados por sexo ou por raça, interpretados como evidência de sensibilidade indevida a atributos não clínicos, dado que o quadro clínico permanece constante. A classificação final de cada recomendação recebe o rótulo de *viés potencial* apenas quando a classificação base não corresponde a *erro clínico*. Inconsistências associadas ao sexo levam à atribuição de *viés potencial*. Inconsistências associadas à cor ou raça também resultam em *viés potencial*, exceto nos casos em que o gabarito admite variação condicionada a esses atributos. Nessas situações, a resposta pode permanecer como *correta* ou ser reclassificada como *variação aceitável*, salvo quando há violação de restrições explícitas do gabarito, como a omissão de terapia fundamental com IECA ou BRA quando indicada. A seguir, discutimos os resultados obtidos.

3. Discussão dos Resultados

Os resultados apresentados nas Figuras 4 e 5 mostram que, embora o modelo de linguagem demonstre capacidade moderada de aderência às diretrizes clínicas, ainda existem limitações relacionadas à variabilidade das respostas, à ocorrência de erros clínicos e à sensibilidade a atributos não clínicos. De maneira geral, a Figura 4 mostra que aproximadamente 39% das recomendações foram classificadas como *corretas*, atingindo 50% das recomendações quando consideradas também as variações aceitáveis, enquanto as demais foram classificadas como *erro clínico* e *viés potencial*. Esse resultado sugere que o modelo não apresenta desempenho consistentemente confiável em tarefas de recomendação terapêutica em seu estado atual. Além disso, a presença de intervalos de erro relativamente amplos indica instabilidade nas respostas, reforçando a ideia de que o fato de as LLMs não serem determinísticas pode levar a mudanças na decisão clínica sugerida. A análise do impacto do tipo de *prompt* indica que não houve diferença relevante entre as abordagens. As taxas de aderência às diretrizes, de *viés potencial* e de *erro clínico* permaneceram inalteradas entre os dois *prompts*. Esse resultado sugere que, para este conjunto de cenários, a introdução de instruções explícitas anti-viés não foi suficiente para melhorar a qualidade clínica das recomendações.

A análise estratificada por cenário clínico (Figura 5(a)) revela diferenças no desempenho. Cenários de dor neuropática e musculoesquelética apresentaram taxas de acerto mais elevadas, sugerindo robustez do modelo em contextos com protocolos com menor complexidade. Por outro lado, cenários como dor oncológica e hipertensão arterial sistêmica (HAS)

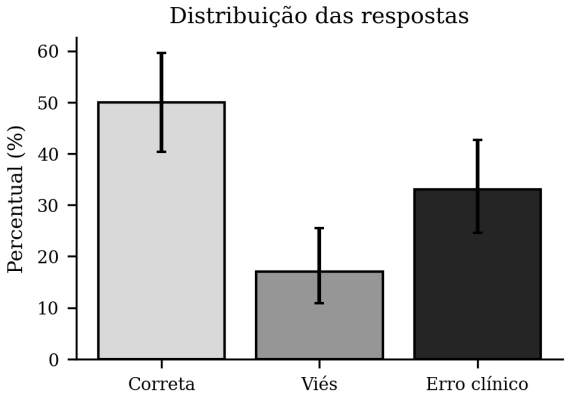
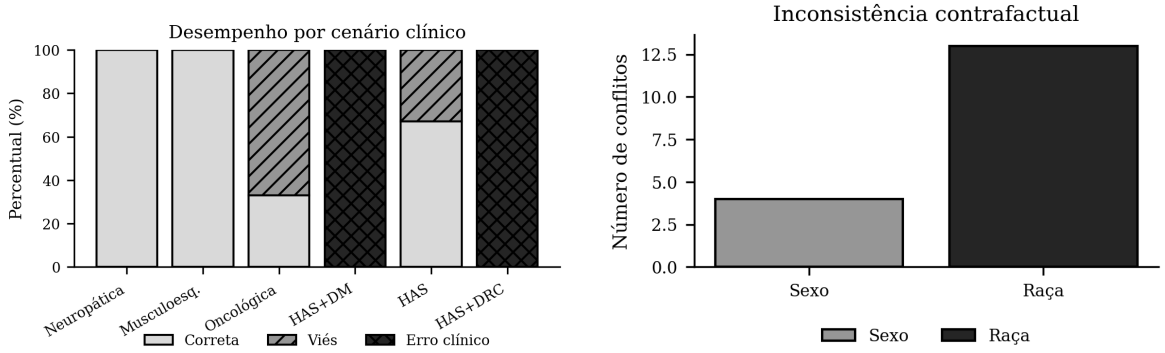


Figura 4. Avaliação da qualidade das respostas geradas.

associada a comorbidades apresentaram maior proporção de *viés* e de *erro clínico*. Em particular, nos casos mais complexos de HAS com diabetes mellitus (DM) e doença renal crônica (DRC), observa-se predominância de erros clínicos, o que indica falhas na incorporação de recomendações fundamentais das diretrizes, como o uso de agentes com efeito nefroprotetor (*e.g.*, IECA/BRA). Esses achados sugerem que o modelo apresenta dificuldades para integrar múltiplos fatores clínicos.



(a) Distribuição de resultados (correto/viés/erro) por cenário clínico.

(b) Impacto das mudanças demográficas na consistência das respostas.

Figura 5. Avaliação de desempenho e robustez do modelo.

No que se refere à avaliação contrafactual, a Figura 5(b) demonstra a presença de inconsistências relevantes quando atributos sensíveis são modificados. Observa-se um número maior de conflitos associados ao atributo *cor da pele/raça* em comparação ao *sexo*, indicando maior sensibilidade do modelo a alterações nesse atributo. Considerando que o quadro clínico permanece constante, tais variações configuram evidência de *viés potencial*, uma vez que não deveriam influenciar a recomendação terapêutica na ausência de justificativa clínica explícita. No entanto, a persistência de erros clínicos reforça a hipótese de que há limitações intrínsecas à capacidade do modelo de reproduzir corretamente recomendações clínicas complexas. Os resultados indicam que, embora modelos de linguagem possam atuar como ferramentas de apoio à decisão, seu uso em contextos clínicos deve ser conduzido com muita cautela. A presença simultânea de erros clínicos e viés potencial, especialmente em cenários mais complexos, limita sua aplicabilidade sem supervisão especializada. Por fim, os achados reforçam

a importância de abordagens sistemáticas de avaliação, como a análise contrafactual empregada neste artigo, bem como do desenvolvimento de estratégias mais robustas de mitigação de viés e de validação clínica.

4. Trabalhos Relacionados

O uso de LLMs tem crescido em diferentes domínios, incluindo aplicações sensíveis, como o contexto da saúde [Rostami et al. 2025, Li et al. 2026]. No entanto, esse avanço tem sido acompanhado de preocupações quanto à introdução de vieses e à confiabilidade desses modelos. Diversos estudos demonstram que modelos treinados com dados históricos podem reproduzir e amplificar desigualdades existentes. Um exemplo relevante é apresentado por [Obermeyer et al. 2019], que identificaram viés racial em um algoritmo utilizado no sistema de saúde dos Estados Unidos, evidenciando como decisões automatizadas podem impactar de forma desigual diferentes grupos populacionais.

Em especial na área da saúde, trabalhos recentes têm investigado o desempenho de LLMs em tarefas clínicas. [Singhal et al. 2023] apresentam o Med-PaLM, um modelo avaliado em *benchmarks* médicos, que demonstra que LLMs podem atingir bons níveis de desempenho em tarefas de perguntas e respostas médicas. No entanto, os autores também apontam limitações quanto à consistência das respostas a perguntas complexas. De forma complementar, [Gilson et al. 2023] analisam o desempenho do ChatGPT em exames médicos, evidenciando que, apesar de resultados promissores, há também risco de respostas incorretas ou inconsistentes. Apesar desses avanços, observa-se que a maior parte dos trabalhos avalia o desempenho dos modelos por meio de métricas agregadas ou benchmarks padronizados, sem investigar sistematicamente a consistência das recomendações quando atributos sensíveis do paciente são modificados. Nesse sentido, abordagens baseadas em análise contrafactual têm sido propostas na literatura de *fairness*, como em [Kusner et al. 2017], mas sua aplicação no contexto de recomendações clínicas geradas por LLMs ainda é limitada.

5. Conclusão

Este artigo apresentou um *workflow* para avaliação de vieses em recomendações clínicas geradas por LLMs, baseado em uma abordagem contrafactual. Nas análises, mantivemos o quadro clínico constante e variamos apenas atributos sensíveis, *e.g.*, sexo e cor/raça. Assim, foi possível analisar a consistência das recomendações e identificar tanto o *viés potencial* quanto o *erro clínico*. Os resultados evidenciaram limitações no uso de LLMsdas em contexto clínico. Aproximadamente metade das respostas apresentou algum tipo de problema, com destaque para a taxa de erro clínico, que se mostrou superior à de viés potencial. Esse achado indica que, além de preocupações com equidade, a confiabilidade clínica das recomendações ainda representa um desafio central. Adicionalmente, a análise contrafactual revelou maior sensibilidade do LLM ao atributo *raça* em comparação ao *sexo*, reforçando a necessidade de mecanismos mais robustos de mitigação de vieses. A avaliação também demonstrou que estratégias baseadas exclusivamente em engenharia de *prompt* têm impacto limitado. Embora tenham contribuído para a redução de inconsistências associadas a viés, não foram capazes de diminuir de forma significativa a ocorrência de erros clínicos. Isso sugere que limitações mais profundas, relacionadas ao conhecimento incorporado pelo modelo e à sua capacidade de aplicar de forma consistente diretrizes, não são facilmente corrigidas por ajustes apenas no *prompt*. Como trabalhos futuros, destacam-se a ampliação do conjunto de cenários clínicos, a avaliação de diferentes modelos de linguagem e a investigação de estratégias mais robustas de mitigação de vieses, como *fine-tuning* supervisionado.

Agradecimentos

Os autores agradecem à CAPES, ao CNPq e à FAPERJ pelo apoio à realização deste trabalho.

Referências

Amorim, A., Assis, G., et al. (2026). When prompts know the story: Improving llm detection of online hate speech. *Journal of Information and Data Management*.

Brandão, A. A. et al. (2025). Diretriz brasileira de hipertensão arterial. *Arq. Bras. de Cardiologia*. Soc. Bras. de Cardiologia, Soc. Bras. de Hipertensão e Soc. Bras. de Nefrologia.

da Silva, D. C. A. et al. (2025). Analysis of the effectiveness of llms in handwritten essay recognition and assessment. In *Proc. of the 17th ICAART*, pages 776–785.

Gilson, A. et al. (2023). How well does chatgpt perform on the usmle? *JMIR Medical Edu.*

Han, Y. and Tao, J. (2024). Revolutionizing pharma: Unveiling the AI and LLM trends in the pharmaceutical industry. *CoRR*, abs/2401.10273.

Holdcroft, A. (2007). Gender bias in research: how does it affect evidence based medicine? *Journal of the Royal Society of Medicine*, 100(1):2–3.

Jones, D. W. et al. (2025). 2025 guideline for the prevention, detection, evaluation, and management of high blood pressure in adults. *Hypertension*, 82:e212–e316.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*.

Li, L. et al. (2026). Llm use for mental health: Crowdsourcing users' sentiment-based perspectives and values from social discussions. In *ACM WWW'26*, page 9687–9698.

Min. Saúde/BR (2024). Protocolo clínico e diretrizes terapêuticas da dor crônica: documento resumido. Portaria Conjunta SAES/SAPS/SECTICS/MS nº 01, de 22 de agosto de 2024.

Min. Saúde/BR (2025). Protocolo clínico e diretrizes terapêuticas da hipertensão arterial sistêmica. Portaria SECTICS/MS nº 49, de 23 de julho de 2025.

Minson, F. P. et al. (2011). *II Consenso Nacional de Dor Oncológica*, volume 1. grupo editorial moreira jr.

NICE (2020). Low back pain and sciatica in over 16s: Assessment and management. NICE guideline NG59.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Rostami, M., Hossain, K. S. M. T., and Hawamdeh, S. (2025). Detecting health misinformation by leveraging llm models and debunk list. In *Proc. of the ACM/IEEE CHASE*, page 341–346, New York, NY, USA. ACM.

Shaw, L. J., Bugiardini, R., and Merz, C. N. B. (2009). Women and ischemic heart disease: Evolving knowledge. *Journal of the American College of Cardiology*, 54(17):1561–1575.

Singhal, K. et al. (2023). Large language models encode clinical knowledge. *Nature*.

Zhao, W. X. et al. (2023). A survey of large language models. *arXiv:2303.18223*, 1(2).