

Análise comparativa do desempenho de diferentes *chatbots* de Inteligência Artificial generativa na resolução de questões do atendimento em farmácia comunitária

Adriana Ribeiro ^{1*}, Ângelo Jesus ², Agostinho Cruz ²

¹ Escola Superior de Saúde, Porto, Portugal,

² REQUIMTE/LAQV Escola Superior de Saúde, Instituto Politécnico do Porto, Porto, Portugal

* 10220385@ess.ipp.pt

Enquadramento: O uso da Inteligência Artificial generativa tem vindo a generalizar-se na área da saúde, apesar da ocorrência de alucinações e falhas éticas que podem comprometer as respostas dos *chatbots* [1,2].

Objetivo: Descrição do uso de *chatbots* na resolução de questões comuns em farmácia comunitária, através da análise comparativa do seu desempenho quanto às variáveis Precisão, Corroboração, Consistência, Clareza, Relevância e Reprodutibilidade. **Métodos:** Estudo observacional, no qual cinco investigadores independentes analisaram, de forma cega, as respostas fornecidas, em outubro de 2025 e fevereiro de 2026, pelos LLM (Grandes Modelos de Linguagem) disponibilizados gratuitamente pelo *ChatGPT*, *Microsoft Copilot*, *Google Gemini*, *CLAUDE*, *Komo*, *Deepseek*, *GROK* e *Meta AI*, sobre dois casos clínicos acerca de medicamentos e suplementação, recorrendo a metodologia modificada descrita por Daraquel et al. [3].

Resultados: As medianas das pontuações variaram entre 0 e 2, com Intervalo Interquartil (IQR) > 0 em 36 das 40 tabelas-resumo *chatbot*/variável/caso clínico/submissão analisadas. Nas restantes 4 tabelas, embora o IQR fosse sempre 0, a distribuição percentual das pontuações revelou 80% de concordância, demonstrando discrepâncias mesmo nas combinações com menor dispersão. Tal comprometeu a análise comparativa, evidenciando-se a influência da subjetividade como um fator limitante. Apesar disso, destacaram-se como principais erros cometidos: incorreta distinção entre medicamentos com/sem prescrição, recomendação de produtos não existentes em Portugal e indicação de posologias erradas. **Conclusões:** Os resultados indicam que fatores como a sensibilidade ao *prompt*, linguagem, localização, mecanismos de segurança e área em estudo podem influenciar as respostas dos *chatbots* que parecem carecer de julgamento clínico, cometendo erros eventualmente graves. Enquanto não forem criados métodos para a sua avaliação, o seu papel na saúde deve reduzir-se apenas a ferramentas complementares. São necessárias atualizações das bases de dados, consoante diretrizes em vigor e tendo em consideração práticas locais, de forma a garantir a fiabilidade das suas respostas.

Palavras-chave: *chatbots*; LLM; análise; comparação; farmácia.

Reconhecimentos

Financiamento: Esta pesquisa não recebeu financiamento externo

Referências

1. Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*. 2023;15(2):e35179. doi:10.7759/cureus.35179 PubMed PMID: 36811129; PubMed Central PMCID: PMC9939079.
2. Ning Y, Teixayavong S, Shang Y, Savulescu J, Nagaraj V, Miao D, et al. Generative artificial intelligence and ethical considerations in health care: a scoping review and ethics checklist. *Lancet Digit Health*. novembro de 2024;6(11):e848–56. doi:10.1016/S2589-7500(24)00143-2 PubMed PMID: 39294061; PubMed Central PMCID: PMC11542614.
3. Daraquel B, Wafaie K, Mohammed H, Cao L, Mheissen S, Liu Y, et al. The performance of artificial intelligence models in generating responses to general orthodontic questions: ChatGPT vs Google Bard. *American Journal of Orthodontics and Dentofacial Orthopedics*. 1 de junho de 2024;165(6):652–62. doi:10.1016/j.ajodo.2024.01.012