RESUMO - III. PESQUISA CIENTÍFICA RELACIONADA ÀS TECNOLOGIAS EDUCACIONAIS E ASSISTENCIAIS

COMPARAÇÃO DO DESEMPENHO DO CHATGPT E DO DEEPSEEK NO TESTE DE PROGRESSO DE ENFERMAGEM NO BRASIL

Carmen Silvia Gabriel (cgabriel@eerp.usp.br)

Maíra De Oliveira Azevedo (m.deoliveiraazevedo@erasmusmc.nl)

Letícia Jesus (leticiajesus@usp.br)

Dario Cecilio Fernandes (d.fernandes@erasmusmc.nl)

Renato Ferretti (r.ferretti@unesp.br)

Pedro Tadao Hamamoto Filho (pedro.hamamoto@unesp.br)

Introdução: A inteligência artificial tem sido utilizada para responder a questões no contexto do ensino médico, mas existem poucos estudos demonstrando seu desempenho na educação em enfermagem. Este estudo teve como objetivo avaliar o desempenho, a confiabilidade e a precisão do ChatGPT e do DeepSeek ao responder a questões de múltipla escolha (QMEs) aplicadas a estudantes de enfermagem. Métodos: Realizamos um estudo observacional transversal para avaliar o desempenho do ChatGPT-5.0 e do DeepSeek (DeepThink-R1) em 605 QMEs do Teste de Progresso Interinstitucional de

Enfermagem brasileiro (2021–2025). A análise focou na precisão dos modelos em seis domínios: Saúde Pública, Gestão, Saúde da Mulher, Saúde da Criança, Saúde do Adulto e Saúde Mental. Para ambos os modelos, realizamos três tentativas. Após cada sessão, os resultados foram integrados a um arquivo mestre e o histórico do chat excluído. Resultados: Análises indicam variação no desempenho entre os chatbots nas áreas. O ChatGPT mostrou maior precisão em Saúde do Adulto e Mental, enquanto o DeepSeek teve melhor desempenho em Saúde Pública e da Criança. Nas três tentativas, os modelos mostraram padrões de desempenho distintos. Na primeira tentativa, o DeepSeek-R1 alcançou a maior precisão com 91,07%, caindo para 70,74% na segunda e 42,98% na terceira. A precisão do ChatGPT-5 na primeira tentativa foi de 77,02%, e alucinou, recusou-se a responder ou falhou em seguir o comando em 30 questões. Na segunda tentativa, a precisão do ChatGPT-5 melhorou para 89,92%, caindo para 77,52% na terceira. Conclusão: Os achados sugerem que LLMs podem replicar níveis moderados de desempenho na avaliação do conhecimento em enfermagem, embora sua eficácia seja inconsistente entre os domínios. As discrepâncias sugerem diferenças nos dados de treinamento, arquitetura do modelo e raciocínio contextual. Embora as ferramentas de IA prometam suporte educacional, elas complementam, e não substituem, o ensino e a avaliação liderados por humanos.

Palavras chave: Inteligência Artificial, ChatGPT, DeepSeek, Educação em Enfermagem, Avaliação Educacional, Teste de Progresso

Performance Comparison of ChatGPT and DeepSeek in a Brazilian Nursing Progress Test

Introduction: Artificial intelligence has been used to answer questions in the medical education context, but there are few studies demonstrating its performance in nursing education. This study aimed to evaluate the performance, reliability, and precision of ChatGPT and DeepSeek in responding to multiple-choice questions (MCQs) administered to nursing students.Methods: We conducted a cross-sectional observational study to assess the performance of ChatGPT-5.0 and DeepSeek (DeepThink-R1) on 605 MCQs from the

Brazilian Interinstitutional Nursing Progress Test (2021–2025). The analysis focused on the models' accuracy across six domains: Public Health, Management, Woman's Health, Child Health, Adult Health, and Mental Health. For both models, we performed three attempts. After each session, the results were integrated into a master file and the chat history was deleted.Results: Analyses indicate variation in performance between the chatbots across fields. ChatGPT showed higher accuracy in Adult Health and Mental Health, while DeepSeek performed better in Public Health and Child Health. Across the three attempts, the models displayed distinct performance patterns. In the first attempt, DeepSeek-R1 achieved the highest accuracy at 91.07%, dropping to 70.74% in the second and 42.98% in the third. ChatGPT-5's accuracy in the first attempt was 77.02%, and it hallucinated, refused to answer, or failed to follow the prompt in 30 questions. In the second attempt, ChatGPT-5's accuracy improved to 89.92%, falling to 77.52% in the third.Conclusion: The findings suggest that LLMs can replicate moderate performance levels in nursing knowledge assessment, although their efficacy is inconsistent across domains. The discrepancies suggest differences in training data, model architecture, and contextual reasoning. While AI tools show promise for educational support, they complement rather than replace human-led assessment.

Key words: Artificial Intelligence, ChatGPT, DeepSeek, Nursing Education, Educational Measurement, Progress Test.

Comparación del Rendimiento de ChatGPT y DeepSeek en el Examen de Progreso de Enfermería en Brasil

Introducción: La inteligencia artificial se ha utilizado para responder preguntas en la educación médica, pero existen pocos estudios sobre su rendimiento en la educación en enfermería. Este estudio evaluó el rendimiento, confiabilidad y precisión de ChatGPT y DeepSeek al responder preguntas de opción múltiple (POM) aplicadas a estudiantes de enfermería.Métodos: Estudio observacional transversal para evaluar rendimiento de ChatGPT-5.0 y DeepSeek (DeepThink-R1) en 605 POM del Examen de Progreso Interinstitucional de Enfermería de Brasil (2021–2025). Se analizó la precisión de los modelos en seis dominios: Salud Pública, Gestión, Salud de la Mujer, Salud del Niño, Salud del Adulto y

Salud Mental. Para ambos se realizaron tres intentos. Tras cada sesión, los resultados se integraron en un archivo maestro y se eliminó el historial del chat.Resultados: Los análisis indican variación en el rendimiento entre chatbots según el área. ChatGPT mostró mayor precisión en Salud del Adulto y Salud Mental; DeepSeek obtuvo mejor rendimiento en Salud Pública y Salud del Niño. En los tres intentos, los modelos mostraron patrones distintos. En el primer intento, DeepSeek logró la mayor precisión con 91.07%, disminuyendo al 70.74% en el segundo y al 42.98% en el tercero. La precisión de ChatGPT en el primer intento fue del 77.02%, presentando alucinaciones, negativas a responder o fallos al seguir la instrucción en 30 preguntas. En el segundo intento, la precisión de ChatGPT mejoró al 89.92%, cayendo al 77.52% en el tercero.Conclusión: Los hallazgos sugieren que los LLM pueden replicar niveles moderados de rendimiento en la evaluación del conocimiento en enfermería, aunque su eficacia es inconsistente entre dominios. Las discrepancias sugieren diferencias en los datos de entrenamiento, la arquitectura del modelo y el razonamiento contextual. Aunque las herramientas de IA prometen un apoyo educativo, deben complementar, y no reemplazar, la enseñanza y la evaluación guiadas por docentes humanos.

Palabras clave:Inteligencia Artificial, ChatGPT, DeepSeek, Educación em Enfermería, Evaluación Educacional, Prueba de Progreso.

Palavras-chave: inteligência artificial; chatgpt; deepseek; educação em enfermagem; avaliação educacional; teste de progresso.