Barw Medical Journal | 2024;2(2):3-6
Journal home page: https://barw.krd/index.php/BMJ

Original Article

Assessment of Nursing Skill and Knowledge of ChatGPT, Gemini, Microsoft Copilot, and Llama: A Comparative Study

Dilan S. Hiwa[1*], Sarhang Sedeeq Abdalla[2], Aso S. Muhialdeen[2], Hussein M. Hamasalih[3], Sanaa O. Karim[6]

1. *College of Medicine, University of Sulaimani, Sulaymaniyah, Kurdistan, Iraq*
2. *Smart Health Tower, Madam Mitterrand Street, Sulaymaniyah, Kurdistan, Iraq*
3. *College of Nursing, University of Sulaimani, Sulaymaniyah, Kurdistan, Iraq*

[*] ***Corresponding author***: dilan.sarmad.hiwa@gmail.com *(D.S. Hiwa). Ashty Street 30 - Zone 1 - house number 6, Zip code: 46001, Sulaymaniyah, Iraq*

Keywords:
MCQ
Artificial intelligence
Nursing
AI

Received: April 2, 2024
Revised: April 15, 2024
Accepted: April 27, 2024
Published: May 7, 2024

Citation: Hiwa DS, Abdalla SS, Muhialdeen AS, Hamasalih HM, Karim SO. Assessment of Nursing Skill and Knowledge of ChatGPT, Gemini, Microsoft Copilot, and Llama: A Comparative Study. Barw Medical Journal. 2024 May 7;2(2):3-6.
https://doi.org/10.58742/bmj.v2i2.87

Abstract

Introduction

Artificial intelligence (AI) has emerged as a transformative force in healthcare. This study assesses the performance of advanced AI systems—ChatGPT-3.5, Gemini, Microsoft Copilot, and Llama 2—in a comprehensive 100-question nursing competency examination. The objective is to gauge their potential contributions to nursing healthcare education and future potential implications.

Methods

The study tested four AI systems (ChatGPT 3.5, Gemini, Microsoft Copilot, Llama 2) with a 100-question nursing exam in February of 2024. A standardized protocol was employed to administer the examination, covering diverse nursing competencies. Questions derived from reputable clinical manuals ensured content reliability. The AI systems underwent evaluation based on accuracy rates.

Results

Microsoft Copilot demonstrated the highest accuracy at 84%, followed by ChatGPT 3.5 (77%), Gemini (75%), and Llama 2 (68%). None achieved complete accuracy on all questions. Each of the AI systems has answered at least one question that only they got correctly.

Conclusion

The variations in AI answers underscore the significance of selecting appropriate AI systems based on specific application requirements and domains, as no singular AI system consistently surpassed others in every aspect of nursing knowledge.

1. Introduction

Artificial intelligence (AI) has surfaced as an innovative technology with the capacity to transform numerous sectors, such as healthcare. The domain of AI has experienced significant progress in recent times, especially in the field of chatbot technology. There is a growing belief that AI, having outperformed humans in various areas, can bring about significant improvements in healthcare, AI has the potential to enhance disease prevention, detection, diagnosis, and treatment [1,2].

The utilization of AI in the healthcare industry has garnered considerable interest due to its vast potential to enhance healthcare provision and patient results. The field of nursing is one area that could experience a significant transformation through the implementation of AI technology. Nursing examinations play a crucial role in assessing the competency and knowledge of nursing professionals. Nurse examinations cover a wide range of topics, including procedures such as nasogastric tube insertion, urinary catheterization, administration of drugs, and their knowledge in surgical settings. With the rise of AI systems, it is important to evaluate their performance in nurse

examinations. In recent years, several AI tools have become widely available, offering a range of services and capabilities. One such AI system is ChatGPT 3.5, an advanced language model created by OpenAI that was trained using an expansive collection of textual content derived from websites, literature, and diverse sources via language modeling tasks. This feature distinguishes it as one of the most expansive and resilient language models ever created, integrating an astonishing 175 billion parameters. ChatGPT version 3.5 has been introduced with wide-ranging applicability throughout multiple sectors, including healthcare [1]. Another AI system that has gained attention is Gemini, formerly known as Google Bard, is an AI-powered information retrieval tool is an advanced chatbot that utilizes a "native multimodal" model to efficiently analyze and adapt to a wide range of data formats such as text, audio, and video [3,4]. Furthermore, Microsoft Copilot, another AI system, is an AI tool created by Microsoft that integrates language models with organizational data to amplify productivity and creativity. It is engineered to support users in diverse tasks by offering recommendations, code excerpts, and additional forms of aid tailored to the user's work context [5]. Additionally, Llama 2, a series of pre-trained and fine-tuned large language models created and launched by GenAI, Meta, encompasses models ranging from 7 billion to 70 billion parameters. Among these variants are models specifically tailored for dialogue applications, referred to as Llama 2-Chat [6]. The performance of these AI systems in nurse examinations is an important area of study. Understanding their ability to provide accurate answers to nursing-related questions can provide insights into their potential use in healthcare settings.

The study aims to compare the performance of advanced AI systems – namely, ChatGPT-3.5, Gemini, Microsoft Copilot, and Llama 2 – when applied to an examination focused on essential nursing competencies. By conducting this comparative analysis, we seek to shed light on the current state of AI integration within nursing education and identify opportunities for further development.

2. Methods

In this comparative study, four different AI systems (ChatGPT 3.5, Gemini, Microsoft copilot, Llama 2) were tested through an examination consisting of 100 multiple-choice questions, with each having five options (A-E). The examination was tailored for nurses, and it incorporated 21 nasogastric intubation questions, 14 urinary catheterization questions, 21 surgery-related questions, 44 other questions about anesthesia drugs, and other general nursing questions. The questions were derived from The Royal Marsden Manual of Clinical Nursing Procedures, tenth Student Edition [7], and Oxford Handbook of Anesthesia, 5th Edition [8]. The questions and answers were reviewed and analyzed by a board-certified anesthesiologist and a senior surgeon separately. The questions were entered on the 17th through 18th of February 2024 into each of the AI systems in the same manner: by first greeting the AI systems by entering a prompt writing “Hello,” and secondly asking them this inquiry “Please, choose a single correct answer for the following multiple-choice questions.” Then, the multiple-choice questions were copy-pasted from a prepared Word document, and the answers were recorded in a table. While conducting a literature review for the current study, papers were specifically included from reputable journals and excluded those published in predatory journals, following the criteria outlined in Kscien’s list [9].

3. Results

In a comparative assessment between ChatGPT 3.5, Gemini, Microsoft copilot, and Llama 2 on 100 nursing multiple-choice questions. ChatGPT 3.5 had a correct answer percentage of 77%, Gemini scored 75%, Microsoft copilot scored 84%, and Llama 2 scored 68%. All of the AI systems showed complete agreeability on only 51% of the questions collectively that they got correctly. About 5% of the questions were answered incorrectly by all the AI systems (Figure 1).

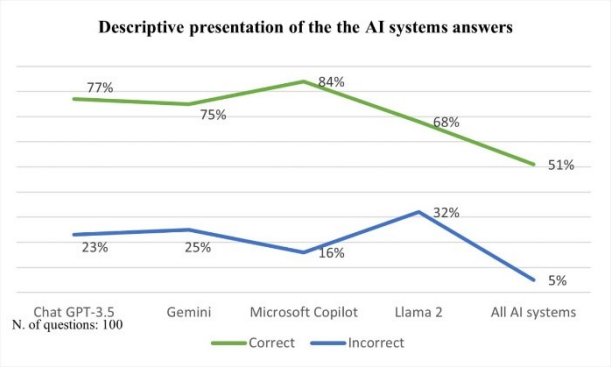

Figure 1: Descriptive presentation of the AI systems answers.

All the AI systems answered correctly on what to do when resistance is felt during urinary catheter insertion and the importance of correct positioning of the nasogastric tube as well as the uses of atropine. However, they all answered incorrectly about the associated risk factors regarding postoperative nausea and vomiting, as well as another question about the optimal position during nasogastric tube insertion. ChatGPT was the only AI system to correctly answer a question regarding prophylaxis of postoperative nausea and vomiting. But was the only one to answer a question about pheochromocytoma incorrectly. Gemini was the sole AI system to provide accurate information regarding pre-procedure steps for nasogastric intubation. But also, the sole system to incorrectly answer a question about risk factors of urinary tract infection. A question concerning the first-line method for confirming nasogastric tube placement was accurately answered by only Microsoft copilot. And Llama 2 was the only AI system to answer that obtaining consent should be the first step for urinary catheter insertion. All the questions and the corresponding AI answers are provided in an Excel sheet (Supplementary 1).

4. Discussion

Nurses play a crucial function in the healthcare sector, utilizing their unique abilities and specialized knowledge to provide

patient-focused healthcare services. Nurses demonstrate proficiency in various areas that complement AI capabilities. Primarily, they prioritize patient-centered care by customizing treatment plans to individual requirements, while AI contributes by furnishing evidence-based information. Moreover, nurses have clinical judgment abilities and extensive medical expertise, enabling them to evaluate patients and make well-informed decisions. AI can also offer valuable prompts to augment the decision-making process. Nurses possess strong communication skills, which enable them to establish rapport with patients and facilitate information exchange. AI can assist in this area to some degree by providing chatbots that can answer patient queries [10-12]. One of the advantages of AI systems is the potential to automate routine tasks, freeing up nurses' time to focus on higher-level activities and personalized patient care [13]. Furthermore, AI-driven monitoring systems can perhaps track patient vital signs and alert nurses to any deviations from normal parameters, facilitating timely interventions [14].

The present investigation explores the performance of advanced AI systems in addressing fundamental nursing competencies, shedding light on the prospective impact of AI technologies on nursing education and practice. Our findings reveal intriguing distinctions among ChatGPT 3.5, Gemini, Microsoft Copilot, and Llama 2 when confronted with a comprehensive set of 100 multiple-choice nursing questions. Our results indicate that while none of the AI systems achieved full agreement across all questions, there were notable differences in their respective accuracies. For instance, Microsoft Copilot performed best among the AI systems when it came to questions related to urinary catheterization, and Llama 2 performed the worst. In addition, Gemini and Microsoft Copilot performed better than the other two AI systems and had the same score when it came to nasogastric intubation. These variations underscore the significance of selecting appropriate AI systems based on specific application requirements and domains. Moreover, our findings suggest that no singular AI system consistently surpassed others in every aspect of nursing knowledge. Instead, each AI system displayed strengths and weaknesses in distinct areas of nursing competence. For example, ChatGPT 3.5 excelled in providing information about prophylactic measures against postoperative nausea and vomiting but struggled with a pheochromocytoma question. Similarly, Gemini proved adept at describing pre-procedure steps for nasogastric intubation yet faltered in recognizing risk factors for urinary tract infection. Meanwhile, Microsoft Copilot distinguished itself by accurately responding to a question about the first-line method for confirming nasogastric tube placement but failed to recognize the necessity of obtaining informed consent prior to urinary catheter insertion.

In a study conducted by Hirosawa et al., physicians consistently demonstrated superior accuracy rates compared to Google Bard in various categories, including the top 10, the top 5, and the top differential diagnosis [15]. Furthermore, in a study by Taira et al. They evaluated the performance of ChatGPT on the Japanese National Nurse Examinations and found that ChatGPT met the passing criteria for the 2019 examination and performed close to the passing level in the 2020-2023 examinations. While ChatGPT did not pass all the examinations, it showed promising results, with only a few more correct answers needed to pass [16]. In an investigation by Rick et al. Google Bard surpassed ChatGPT in accurately executing mass casualty incident triage, achieving a 60% accuracy compared to ChatGPT's 26.67%. This dissimilarity was statistically significant [17].

One of the challenges that face the integration of AI systems into healthcare is the lack of contextual understanding; while AI systems can provide accurate information, they may lack the ability to understand the nuanced context of patient care, potentially leading to inappropriate recommendations or responses. Furthermore, human nurses retain indispensable critical thinking skills and emotional intelligence necessary for comprehensive patient care, qualities that AI may not completely emulate. [14,18]. It's legitimate to worry about AI systems developing and reinforcing prejudices present within their learning materials or input datasets. This apprehension has been recognized as a significant issue due to research indicating that such bias-prone behavior may arise when AI models absorb and reproduce existing discriminatory patterns. It is crucial to develop effective approaches to address the issue of biases in AI models and integrate algorithms that prioritize equitableness and rightfulness during the developmental phase. This is particularly significant in the medical field, where it is essential to prevent any biases in patient care and the decision-making capabilities of AI systems. By doing so, these AI models can become a more dependable resource for medical professionals in the future. Moreover, advanced language models possess the ability to produce compelling yet emotionally persuasive material even when they are incorrect. As such, it becomes indispensable to actively confront the associated dangers and maintain accountable and moral utilization of these intricate systems. Another ethical concern is discerning the ownership and authorship attribution of content produced by AI systems [1].

5. Conclusion

Variations in the different AI systems underscore the significance of selecting appropriate AI systems based on specific application requirements and domains, as no singular AI system consistently surpassed others in every aspect of nursing knowledge.

Declarations

Conflicts of interest: The author(s) have no conflicts of interest to disclose.

Ethical approval: Not applicable.

Patient consent (participation and publication)**:** Not applicable.

Funding: The present study received no financial support.

Acknowledgments: None to be declared.

Authors' contributions: DSH was a major contributor to the conception of the study, as well as to the literature search for related studies. SSA, ASM, and HMHS were involved in the literature review, the design of the study, the critical revision of the manuscript, and participated in data collection. SOK and

DSH were involved in the literature review, study design, and writing the manuscript. SSA and DSH confirm the authenticity of all the raw data. All authors approved the final version of the manuscript.

Use of AI: AI was not used in the drafting of the manuscript, the production of graphical elements, or the collection and analysis of data.

Data availability statement: Not applicable.

References

1. Kuzucu 1. Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems. 2023. doi:10.1016/j.iotcps.2023.04.003
2. 2. Ahamed ZM, Dhahir HM, Mohammed MM, Ali R, Hassan SH, Muhialdeen AS, Saeed YA, Fatah ML, Qaradakhy AJ, Ali RM, Ahmed SF. Comparative Analysis of ChatGPT and Human Decision-Making in Thyroid and Neck Swellings: A Case-Based Study. Barw Medical Journal. 2023;1(4):2-6. doi:10.58742/bmj.v1i2.43
3. 3. Masalkhi M, Ong J, Waisberg E, Lee AG. Google DeepMind's gemini AI versus ChatGPT: a comparative analysis in ophthalmology. Eye. 2024 14:1-6. doi:10.1038/s41433-024-02958-w
4. 4. Abbas YN, Hassan HA, Hamad DQ, Hasan SJ, Omer DA, Kakamad SH, et al. Role of ChatGPT and Google Bard in the Diagnosis of Psychiatric Disorders: A Cross Sectional Study. Barw Medical Journal. 2023;1(4):14-19. doi:10.58742/4vd6h741
5. 5. Semeraro F, Gamberini L, Carmona F, Monsieurs KG. Clinical questions on advanced life support answered by artificial intelligence. A comparison between ChatGPT, Google Bard and Microsoft Copilot. Resuscitation. 2024 1;195. doi:10.1016/j.resuscitation.2024.110114
6. 6. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288. 2023 18. doi:10.48550/arXiv.2307.09288
7. 7. Lister S, Hofland J, Grafton H, Wilson C. The Royal Marsden Manual of Clinical Nursing Procedures, Student Edition. Google Books. John Wiley & Sons; 2021. https://books.google.iq/books?
8. 8. Freedman R, Herbert L, O'Donnell A, Ross N. Oxford Handbook of Anaesthesia. Oxford University Press; 2022. doi:10.1177/0310057X221134636
9. 9. Muhialdeen AS, Ahmed JO, Baba HO, Abdullah IY, Hassan HA, Najar KA, Mikael TM, Mustafa MQ, Mohammed DA, Omer DA, Bapir R. Kscien's List; A New Strategy to Discourage Predatory Journals and Publishers (Second Version). Barw Medical Journal. 2023 1. doi:10.58742/bmj.v1i1.14
10. 10. Javaid M, Haleem A, Singh RP. ChatGPT for healthcare services: An emerging stage for an innovative perspective. BenchCouncil Transactions on Benchmarks, Standards and Evaluations. 2023 1;3(1):100105. doi:10.1016/j.tbench.2023.100105
11. 11. Miao H, Ahn H. Impact of ChatGPT on interdisciplinary nursing education and research. Asian/Pacific Island Nursing Journal. 2023 24;7(1): e48136. doi:10.2196/48136
12. 12. Sauerbrei A, Kerasidou A, Lucivero F, Hallowell N. The impact of artificial intelligence on the person-centred, doctor-patient relationship: some problems and solutions. BMC Medical Informatics and Decision Making. 2023;23(1):1-4. doi:10.1186/s12911-023-02162-y
13. 13. De Gagne JC. The State of Artificial Intelligence in Nursing Education: Past, Present, and Future Directions. International Journal of Environmental Research and Public Health. 2023 10;20(6):4884. doi:10.3390/ijerph20064884
14. 14. Yelne S, Chaudhary M, Dod K, Sayyad A, Sharma R. Harnessing the Power of AI: A Comprehensive Review of Its Impact and Challenges in Nursing Science and Healthcare. Cureus. 2023 22;15(11). doi:10.7759/cureus.49252
15. 15. Hirosawa T, Mizuta K, Harada Y, Shimizu T. Comparative evaluation of diagnostic accuracy between Google Bard and physicians. The American Journal of Medicine. 2023 1;136(11):1119-23. doi:10.1016/j.amjmed.2023.08.003
16. 16. Taira K, Itaya T, Hanada A. Performance of the large language model ChatGPT on the National Nurse Examinations in Japan: evaluation study. JMIR nursing. 2023; 6: e47305. doi:10.2196/47305
17. 17. Gan RK, Ogbodo JC, Wee YZ, Gan AZ, González PA. Performance of Google bard and ChatGPT in mass casualty incidents triage. The American journal of emergency medicine. 2024 1; 75:72-8. doi:10.1016/j.ajem.2023.10.034
18. 18. Muhialdeen AS, Mohammed SA, Ahmed NH, Ahmed SF, Hassan WN, Asaad HR, et al. Artificial Intelligence in Medicine: A Comparative Study of ChatGPT and Google Bard in Clinical Diagnostics. Barw Medical Journal. 2023;1(4):7-13. doi:10.58742/pry94q89