The rise of AI in healthcare education: DeepSeek and GPT-4o take on the 2024 Taiwan Pharmacist Exam

Ying-Mei Wang[a,b,c], Tzeng-Ji Chen[d,e,f,*]

[a]Department of Medical Education and Research, Taipei Veterans General Hospital Hsinchu Branch, Hsinchu, Taiwan, ROC; [b]Department of Pharmacy, Taipei Veterans General Hospital Hsinchu Branch, Hsinchu, Taiwan, ROC; [c]School of Medicine, National Tsing Hua University, Hsinchu, Taiwan, ROC; [d]Department of Family Medicine, Taipei Veterans General Hospital Hsinchu Branch, Hsinchu County, Taiwan, ROC; [e]Department of Family Medicine, Taipei Veterans General Hospital, Taipei, Taiwan, ROC; [f]Department of Post-Baccalaureate Medicine, National Chung Hsing University, Taichung, Taiwan, ROC

DEAR EDITOR,

In recent years, the rapid development of large language model (LLM) technology has not only changed the landscape of artificial intelligence (AI) research but has also begun to have a profound impact on the fields of medicine and pharmacy. In the past, LLMs were primarily dominated by technology giants operating under closed commercial models, such as OpenAI's GPT series and Google DeepMind's Gemini. However, the rise of open-source LLMs is reshaping this competitive landscape and opening up new opportunities for the research community and industry.

DeepSeek is one of the most popular open-source LLMs in recent times, not only for its strong technical performance but also for its strengths in multilingual processing, cost-effectiveness, and openness.[1] This model has made waves in the LLM market with its groundbreaking Mixture-of-Experts (MoE) architecture and reinforcement learning techniques, enabling it to compete with OpenAI's GPT-4o despite limited computational resources, and even outperform it in mathematics, programming, and reasoning capabilities.[2] Furthermore, DeepSeek's open-source strategy and lower operating costs have quickly drawn global attention, attracting numerous scientists to explore its applications in various fields.[3,4] This technological evolution is poised to drive considerable advances in both academia and industry.

DeepSeek has demonstrated potential in scientific areas such as bioinformatics, computational chemistry, and neuroscience, achieving performance comparable to OpenAI's GPT-4o in certain data analysis and diagnostic tasks.[5] However, its application in the medical domain remains limited by factors such as the quality of annotated data, excessive transparency of reasoning leading to verbose output, and ethical and regulatory challenges, including data privacy and liability issues.[1] Additionally, although DeepSeek is freely accessible, its reliability in clinical decision-making requires further validation. Moreover, it still exhibits errors in some basic tasks, such as simple calculations.[6] Comparing the applications of DeepSeek and ChatGPT in clinical research, both models can accelerate scientific research, streamline clinical data analysis and diagnostics, and support manuscript writing. In academic publishing, they help researchers refine manuscripts and reduce repetitive tasks.[7]

The impact of LLMs on medical education and professional examinations is a growing area of interest. For example, the applications of ChatGPT in pharmacy have been extensively studied. The *Journal of the Chinese Medical Association* previously published a study on the performance of ChatGPT on the pharmacist licensing examination in Taiwan.[8] The authors also published a study in another journal comparing ChatGPT-3.5 with ChatGPT-4 on the same examination.[9] Most studies indicate that ChatGPT's performance in non-English contexts needs improvement.[8–11] DeepSeek, trained on extensive Chinese and English datasets, claims to have a competitive advantage over other mainstream LLMs (such as GPT-4 and Gemini) on tasks involving Chinese text generation, question answering, and reading comprehension.[12] Given this advantage, how does DeepSeek perform in professional pharmacist examinations in Taiwan?

We attempted a similar test with DeepSeek, using questions from the 2024 second session of the Taiwan national pharmacist licensing examination, which covered three basic subjects (pharmacology and pharmaceutical chemistry, pharmaceutical analysis and pharmacognosy, pharmaceutics and biopharmaceutics) and three clinical subjects (dispensing and clinical pharmacy, therapeutics, pharmacy administration and law).

A 25% random sample of examination questions was taken for each subject. If a subject included two subfields, a separate 25% sample was taken. For example, in "pharmaceutics and biopharmaceutics," which contains 80 questions (40 per subfield), 10 questions were sampled from pharmaceutics and another 10 from biopharmaceutics. The sampled questions were entered directly into DeepSeek and ChatGPT, respectively, without additional prompts or instructions, and the responses were recorded.

The performance of both AI models was analyzed in three dimensions: overall accuracy, subject-specific performance, and performance by question type. Examination questions were classified into four categories based on reasoning complexity: memory-based questions (single correct answer), judgment questions (choosing the correct statement), reverse questions

*Address correspondence. Dr. Tzeng-Ji Chen, Department of Family Medicine, Taipei Veterans General Hospital Hsinchu Branch, 81, Section 1, Zhongfeng Road, Zhudong, Hsinchu 310, Taiwan, ROC. E-mail address: tjchen@vhct.gov.tw (T.-J. Chen).

Conflicts of interest: Dr. Tzeng-Ji Chen, an editorial board member at Journal of the Chinese Medical Association, had no role in the peer review process of or decision to publish this article. The authors declare that they have no conflicts of interest related to the subject matter or materials discussed in this article.

Journal of Chinese Medical Association. (2025) 88: 338-339.

doi: 10.1097/JCMA.0000000000001220

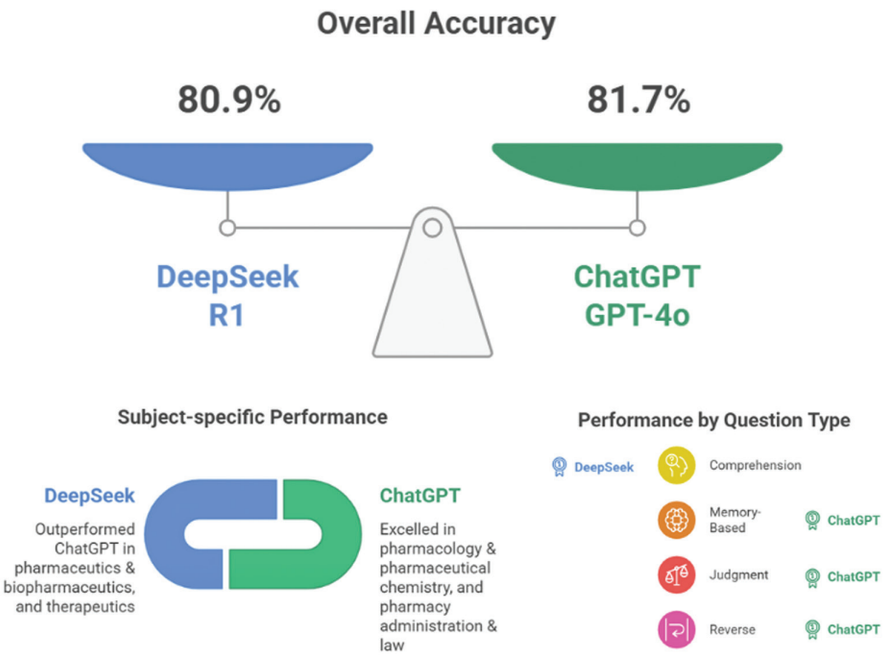

Fig. 1 Performance comparison of DeepSeek and ChatGPT in answering randomly sampled questions from the 2024 Taiwan national pharmacist licensing examination.

(choosing an incorrect option or statement), and comprehension questions (including multiple-choice, matching, calculation, and case-based questions). The study used the free versions of DeepSeek R1 and ChatGPT GPT-4o, with testing conducted between February 2 and February 9, 2025.

The results demonstrate that DeepSeek and ChatGPT have similar overall accuracy on the 2024 Taiwan national pharmacist licensing examination, with DeepSeek scoring 80.9% and ChatGPT 81.7%. However, subject-specific analysis revealed distinct performance differences: DeepSeek outperformed ChatGPT in pharmaceutics and biopharmaceutics, as well as therapeutics, whereas ChatGPT performed better in pharmacology and pharmaceutical chemistry, as well as pharmacy administration and law. In terms of question types, DeepSeek excelled in comprehension questions (85.7%), surpassing ChatGPT (75.0%), while underperforming in other categories (Fig. 1).

These findings suggest that different AI models exhibit variations in language comprehension, reasoning ability, and the application of domain knowledge, likely influenced by differences in training data and model architectures. Although AI can serve as a valuable supplementary tool in pharmacy education, its accuracy and reliability must be critically assessed, particularly in non-English settings. Future research should explore the broader applicability of generative AI in professional examinations, strategies for integrating the strengths of different models to enhance the role of AI in medical and pharmaceutical research, and methods for refining AI-driven tools to improve accuracy in specialized domains. By deepening our understanding of AI's capabilities and limitations, we can better harness these technologies to support education, research, and professional practice in the medical and pharmaceutical fields.

REFERENCES

1. Peng Y, Malin BA, Rousseau JF, Wang Y, Xu Z, Xu X, et al. From GPT to DeepSeek: Significant gaps remain in realizing AI in healthcare [online ahead of print]. *J Biomed Inform* 2025;163:104791.
2. Normile D. Chinese firm's large language model makes a splash. *Science* 2025;387:238.
3. Conroy G, Mallapaty S. How China created AI model DeepSeek and shocked the world. *Nature* 2025;638:300–1.
4. Gibney E. China's cheap, open AI model DeepSeek thrills scientists. *Nature* 2025;638:13–4.
5. Gibney E. Scientists flock to DeepSeek: how they're using the blockbuster AI model [online ahead of print]. *Nature* 2025. doi: 10.1038/d41586-025-00275-0.
6. Smith J. Daily briefing: the pros and cons of DeepSeek [online ahead of print]. *Nature* 2025. doi: 10.1038/d41586-025-00330-w.
7. Kayaalp ME, Prill R, Sezgin EA, Cong T, Królikowska A, Hirschmann MT. DeepSeek versus ChatGPT: multimodal artificial intelligence revolutionizing scientific discovery. From language editing to autonomous content generation—redefining innovation in research and practice [online ahead of print]. *Knee Surg Sports Traumatol Arthrosc* 2025. doi: 10.1002/ksa.12628.
8. Wang YM, Shen HW, Chen TJ. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. *J Chin Med Assoc* 2023;86:653–8.
9. Wang YM, Shen HW, Chen TJ, Chiang SC, Lin TG. Performance of ChatGPT-3.5 and ChatGPT-4 in the Taiwan national pharmacist licensing examination: comparative evaluation study. *JMIR Med Educ* 2025;11:e56850.
10. Jin HK, Lee HE, Kim E. Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: a systematic review and meta-analysis. *BMC Med Educ* 2024;24:1013.
11. Jin HK, Kim E. Performance of GPT-3.5 and GPT-4 on the Korean pharmacist licensing examination: comparison study. *JMIR Med Educ* 2024;10:e57451.
12. DeepSeek. Available at https://www.deepseek.com/. Accessed February 15, 2025.