Comparative Evaluation of Advanced AI Reasoning Models in Korean National Licensing Examination OpenAI vs DeepSeek

Jin-Gyu Lee, Gyeong Hoon Kim, Jiahn Chae, Jun Suh Lee, Hyun-Young Shin

Submitted to: JMIR Medical Education
on: March 27, 2025

Table of Contents

Comparative Evaluation of Advanced AI Reasoning Models in Korean National Licensing Examination OpenAI vs DeepSeek

Jin-Gyu Lee[1*] MD, MSc; Gyeong Hoon Kim[1*] MD; Jiahn Chae[2]; Jun Suh Lee[3*] MD, phD; Hyun-Young Shin[4*] MD, MPH, phD

[1] Department of Dermatology, Asan Medical Center, University of Ulsan College of Medicine Seoul KR
[2] Kyungpook National University School of Medicine Daegu KR
[3] Bucheon Sejong Hospital Bucheon KR
[4] Seoul St. Mary's Hospital Seoul KR
[*]these authors contributed equally

Abstract

Artificial intelligence (AI) has advanced in natural language processing and reasoning, with large language models (LLMs) increasingly assessed for medical education and licensing exams. Given the growing use of AI in medical licensing examinations, evaluating their performance on non-Western, region-specific tests like the Korean Medical Licensing Examination (KMLE) is crucial for assessing their real-world applicability. This study compared five LLMs—GPT-4o, o1, o3-mini (OpenAI), DeepSeek-V3, and DeepSeek-R1 (DeepSeek)—on the KMLE.

A total of 150 multiple-choice questions from the 2024 KMLE were extracted and categorized into three domains: Local Health & Medical Laws, Preventive Medicine, and Clinical Medicine. Graph-based questions were excluded. Each model completed five independent runs via API, with accuracy assessed against official answers. Statistical differences were analyzed using ANOVA, and consistency was measured using Fleiss' kappa coefficient.

o1 achieved the highest overall accuracy (94.3%), excelling in Clinical Medicine (97.5%) and Local Health & Medical Law (81.0%), while DeepSeek-R1 led in Preventive Medicine (92.6%). Despite domain-specific variations, all models surpassed passing criteria. For consistency, o1 ranked highest (97.1%), with DeepSeek-V3 excelling in Local Health & Medical Law (97.5%). Performance declined in Local Health & Medical Law, likely due to legal complexities and limited Korean-language training data.

This is the first study to compare OpenAI and DeepSeek models on medical licensing exam, demonstrating their strong performance, with o1 and DeepSeek-R1 ranking within the top 10% of human candidates. While o1 was the most accurate, DeepSeek-R1 provided a cost-effective alternative. Future research should optimize LLMs for non-English exams and develop Korea-specific AI models to improve accuracy in legal domains.

(JMIR Preprints 27/03/2025:75032)
DOI: https://doi.org/10.2196/preprints.75032

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Comparative Evaluation of Advanced AI Reasoning Models in Korean National Licensing Examination OpenAI vs DeepSeek

Jin-Gyu Lee[1†], Gyeong Hoon Kim[2†], Jiahn Chae[3], Hyun-Young Shin[4*], Jun Suh Lee[5*]

1 Yeungnam Sok Union Internal Medicine Clinic, Daegu, Republic of Korea

2 Department of Dermatology, Asan Medical Center, University of Ulsan College of Medicine

3 Kyungpook National University School of Medicine, Daegu, Republic of Korea

4 Department of Family Medicine, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea

5 Department of Surgery, Bucheon Sejong Hospital, Bucheon, Korea

† * These authors contributed equally.

*Corresponding author

Hyun-Young Shin, MD, MPH, PhD

Department of Family Medicine, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, 222, Banpo-daero, Seocho-gu, Seoul, 06591, Republic of Korea

TEL: +82-10-9050-1098

E-mail: SHYDEBORAH@CATHOLIC.AC.KR

https://orcid.org/0000-0001-7261-3365

Jun Suh Lee, MD, PhD

Department of Surgery, Bucheon Sejong Hospital, Bucheon, Korea

28, Hohyeon-ro 489 beon-gil, Bucheon, Republic of Korea

Tel: +82-10-2747-6320

E-mail: rudestock@gmail.com

https://orcid.org/0000-0001-9487-9826

Running Title

OpenAI versus DeepSeek in Korean Medical License Examination

Keywords

Artificial Intelligence; GPT; DeepSeek; Machine Learning; Licensure, Medical; Educational Measurement; Natural Language Processing

Abstract

Artificial intelligence (AI) has advanced in natural language processing and reasoning, with large language models (LLMs) increasingly assessed for medical education and licensing exams. Given the growing use of AI in medical licensing examinations, evaluating their performance on non-Western, region-specific tests like the Korean Medical Licensing Examination (KMLE) is crucial for assessing their real-world applicability. This study compared five LLMs—GPT-4o, o1, o3-mini (OpenAI), DeepSeek-V3, and DeepSeek-R1 (DeepSeek)—on the KMLE.

A total of 150 multiple-choice questions from the 2024 KMLE were extracted and categorized into three domains: Local Health & Medical Laws, Preventive Medicine, and Clinical Medicine. Graph-based questions were excluded. Each model completed five independent runs via API, with accuracy assessed against official answers. Statistical differences were analyzed using ANOVA, and consistency was measured using Fleiss' kappa coefficient.

o1 achieved the highest overall accuracy (94.3%), excelling in Clinical Medicine (97.5%) and Local Health & Medical Law (81.0%), while DeepSeek-R1 led in Preventive Medicine (92.6%). Despite domain-specific variations, all models surpassed passing criteria. For consistency, o1 ranked highest (97.1%), with DeepSeek-V3 excelling in Local Health & Medical Law (97.5%). Performance declined in Local Health & Medical Law, likely due to legal complexities and limited Korean-language training data.

This is the first study to compare OpenAI and DeepSeek models on medical licensing exam, demonstrating their strong performance, with o1 and DeepSeek-R1 ranking within the top 10% of human candidates. While o1 was the most accurate, DeepSeek-R1 provided a cost-effective alternative. Future research should optimize LLMs for non-English exams and develop Korea-specific AI models to improve accuracy in legal domains.

Introduction

Artificial intelligence (AI) has demonstrated significant advancements in natural language processing (NLP) and reasoning, particularly in medical education and assessment. Large language models (LLMs), such as OpenAI's ChatGPT, are being increasingly evaluated for their ability to assist in clinical reasoning, medical decision-making, and standardized testing performance.[1,2] Given the rigorous nature of medical licensing examinations, these AI models have the potential to serve as supplementary learning tools, automated tutors, or even decision-support systems in clinical settings. While OpenAI's ChatGPT has been extensively studied, emerging alternatives like DeepSeek-R1 offer new opportunities for comparative evaluation.

DeepSeek-R1 is a large language model (LLM) developed by DeepSeek, a Chinese artificial intelligence research company founded in July 2023. Launched in January 2025, DeepSeek-R1 became the most-downloaded freeware application on the iOS App Store in the United States, surpassing ChatGPT in popularity.[3] Designed for various NLP tasks, including question answering, text generation, and knowledge synthesis, it is trained on a diverse multilingual corpus, including medical literature, and is optimized for reasoning and professional applications such as medicine, law, and engineering.

Understanding AI performance on licensing examinations like the Korean Medical Licensing Examination (KMLE) is essential in determining their real-world applicability. The KMLE, managed by the Korea Health Personnel Licensing Examination Institute, ensures standardized medical competence and has evolved from a government-controlled system to a privatized and specialized process, influenced by historical events.[4] This exam presents unique challenges,

including language barriers, culturally specific guidelines, and regionally relevant clinical scenarios. Evaluating AI models on the KMLE provides insights into their ability to navigate complex medical reasoning beyond Western contexts.

Previous studies have examined AI models in medical education, with many focusing on their ability to pass licensing examinations such as the United States Medical Licensing Examination (USMLE).[5,6] However, most evaluations have centered on Western medical exams, leaving a gap in understanding how these models perform in non-English, region-specific tests like the KMLE. Additionally, while earlier versions of AI models have been assessed in broad medical contexts, there remains a lack of direct comparative studies between different advanced AI reasoning models on licensing examinations, particularly those trained with region-specific medical knowledge.

As AI models continue to evolve, there is a pressing need to assess their applicability in diverse medical education settings. The introduction of DeepSeek-R1 necessitates an objective comparison with established models such as OpenAI's o1. Given the differences in their training methodologies, data sources, and underlying architectures, a direct performance comparison on the KMLE can help determine their relative strengths and weaknesses in medical reasoning. This study aims to systematically compare flagship versions of ChatGPT and DeepSeek, such as o1 and DeepSeek-R1 on the KMLE, highlighting their strengths, limitations, and implications for medical education. By addressing the gap in non-English AI evaluation, this research contributes to understanding the feasibility of AI as a supplementary tool for medical professionals and students preparing for licensing examinations.

Materials & methods

Input Source

KMLE is an annual exam taken by all medical school graduates in the Republic of Korea to become licensed medical doctors. It comprises three subjects— Local Health & Medical Law, Preventive Medicine, and Clinical Medicine—and consists of four sessions of 80 questions each, yielding a total of 320 questions. Every question provides five answer options, with only one correct answer. Some questions include images or graphs, while others are text-only. The entire exam is written in Korean. The scanned images of the exam papers are publicly available online, but the images and graphs included in each question are obscured due to copyright issues[7].

We obtained the publicly available scanned images of 2024 KMLE exam papers and extracted the questions as text using the Gemini 2.0 flash model (Google LLC, Mountain View, CA, USA) with the prompt: "Extract the questions as text from this exam paper image". All extracted questions were manually reviewed and corrected for any errors. Questions containing graphs were excluded, and a final selection of 150 questions was included in this study, comprising 20 questions from Medical Laws and Regulations, 19 from Preventive Medicine, and 111 from Clinical Medicine. Since this study uses publicly available data, it is exempt from IRB review.

Data Collection

In this study, we utilized a total of five models: GPT-4o, o1, o3-mini (OpenAI,

Inc., San Francisco, CA, USA), DeepSeek-V3, and DeepSeek-R1 (DeepSeek, Inc., Hangzhou, Zhejiang, China). 150 questions were individually submitted to each LLM along with the prompt: 'Choose the most appropriate answer to the above question,' to obtain the LLM's response. For each LLM model, this process was executed in five runs. The LLMs were accessed via Application Programming Interfaces (APIs), specifically calling the '*gpt-4o-2024-11-20*', '*o1-2024-12-17*', and '*o3-mini*' models from OpenAI, and the '*deepseek-chat*' and '*deepseek-reasoner*' models from DeepSeek. All LLM analyses were conducted within a single day (February 26, 2025), using the most up-to-date API versions of the models available on that date, so that multiple versions of the models were not used concurrently.

Statistical Analysis

First, the analysis of 150 KMLE questions was conducted using API-based evaluations across five LLMs. The accuracy of each model was assessed by comparing its responses with the official answers provided by the Korea Health Personnel Licensing Examination Institute (KHPLEI) for the 2024 KMLE. Each model underwent five independent evaluation runs, and the accuracy from each run was recorded. The mean accuracy across the five runs was then computed, and statistical significance in differences among the three or more groups was assessed using analysis of variance (ANOVA).

Furthermore, to evaluate the reliability of each LLM across independent runs, consistency was measured using Fleiss' kappa coefficient. This method was employed to verify the statistical significance of consistency.

All statistical analyses were performed using Microsoft Excel and IBM SPSS

statistics 26. Additionally, to prevent textual errors (e.g., typographical mistakes) during the conversion of KMLE questions into text format, a rigorous manual review process was repeatedly conducted.

Results

Figure 1 presents a schematic overview of our conceptual evaluation comparing five models—three variants of ChatGPT (o4, o1, o3-mini) and two variants of DeepSeek (DeepSeek-V3, DeepSeek-R1)—on KMLE. As depicted in Figure 2, out of the initial 320 questions, 150 were selected for analysis after excluding those containing images, which neither ChatGPT nor DeepSeek supports. The final dataset comprised 20 questions on Local Health & Medical Law, 19 on Preventive Medicine, and 111 on Clinical Medicine.

1) Comparison of the accuracy between OpenAI models (4o, o1, o3-mini) and DeepSeek models (DeepSeek-V3, DeepSeek-R1)

For each model, the raw text file of these 150 questions was processed in five independent runs to compute both accuracy and consistency for each subject area. Notably, the exam passing criteria require an overall accuracy of at least 60% and subject-specific accuracies of at least 40%, and all models surpassed these thresholds.

Comparison of accuracy between OpenAI models and DeepSeek models in solving KMLE across each subject area (Figure 3, Table 1) revealed that o1 achieved the highest overall accuracy at 94.3%, followed by DeepSeek-R1 at 91.6%, o3-mini at 88.7%, GPT-4o at 82.0%, and DeepSeek-V3 at 74.9% (Figure 3a). The statistical significance between each LLM is indicated by the *P*-value

(Figure 3a, P <.001). In the Local Health & Medical Law domain, o1 led with an accuracy of 81.0%, while DeepSeek-V3 recorded 75.0%, DeepSeek-R1 69.0%, GPT-4o 60.0%, and o3-mini 56.0% (Figure 3b). This trend indicates that all models encountered challenges with questions from this subject area.

In the domain of Preventive Medicine, DeepSeek-R1 excelled with an accuracy of 92.6%, slightly outperforming o1 at 89.5%. o3-mini, GPT-4o, and DeepSeek-V3 achieved accuracies of 87.4%, 85.3%, and 84.2%, respectively (Figure 3c). Meanwhile, in Clinical Medicine, o1 demonstrated exceptional performance with an accuracy of 97.5%, followed closely by DeepSeek-R1 at 95.5% and o3-mini at 94.8%. GPT-4o and DeepSeek-V3 trailed in this domain, with accuracies of 85.4% and 73.3%, respectively (Figure 3d).

Overall, while o1 consistently provided superior performance across most domains, DeepSeek-R1 showed notable strength in Preventive Medicine. Importantly, despite the variations observed across different subject areas, all evaluated models exceeded the minimum criteria required for passing the national exam, underscoring their potential utility in high-stakes assessments such as the KMLE.

2) Comparison of the consistency between OpenAI models (4o, o1, o3-mini) and DeepSeek models (DeepSeek-V3, DeepSeek-R1)

Next, we compared the consistency of the five model variants—three from ChatGPT (GPT-4o, o1, o3-mini) and two from DeepSeek (V3, R1)—across the three medical domains (Local Health & Medical Law, Preventive Medicine, and Clinical Medicine). Comparison of consistency between GPTs and DeepSeeks in solving KMLE across each subject area (Figure 4, Table 2) revealed that GPT-o1 displayed the highest overall consistency at 97.1%, followed by DeepSeek-V3

and DeepSeek-R1 (both 94.5%), o3-mini (94.0%), and GPT-4o (92.3%) (Figure 4a) The statistical significance test results for repeated evaluations of each LLM are represented by the *P*-value.

Domain-specific analysis revealed that DeepSeek-V3 achieved the highest consistency in Local Health & Medical Law (97.5%), whereas o1 recorded 84.9%, o3-mini 85.7%, GPT-4o 89.6%, and DeepSeek-R1 81.5% (Figure 4b). In the Preventive Medicine domain, o1 and DeepSeek-V3 both attained 100.0% consistency, while o3-mini, GPT-4o, and DeepSeek-R1 reached 95.1%, 93.4%, and 94.7%, respectively (Figure 4c). Finally, in Clinical Medicine, o1 again outperformed the others with a consistency of 98.9%, followed by DeepSeek-R1 at 96.8%, o3-mini at 95.2%, GPT-4o at 92.5%, and DeepSeek-V3 at 92.9% (Figure 4d).

Overall, o1 demonstrated the most stable performance across all domains, though DeepSeek-V3 surpassed other models in Local Health & Medical Law and tied with o1 in Preventive Medicine. These findings highlight a degree of domain-dependent variation in consistency, underscoring the importance of evaluating AI models across multiple subject areas.

3) Comparison of the accuracy and consistency on the questions that showed the poorest performance within the Local Health & Medical Law domain

Finally, we focused on the Local Health & Medical Law domain, which showed significantly lower accuracy across all models. In this domain, we identified the five questions with the lowest average accuracy across the five variants and compared their performance in terms of accuracy (Figure 5, Table 3). The results

reveal a stark contrast in model performance. o1 demonstrated the highest accuracy at 52%, significantly outperforming the other models. o3-mini followed at 16%, while DeepSeek-R1, DeepSeek-V3, and GPT-4o struggled, achieving only 4%, 0%, and 0% accuracy, respectively (Figure 5a). These findings indicate that even the most advanced models faced substantial difficulty with these specific questions, reinforcing the notion that the Local Health & Medical Law domain presents unique challenges for AI models (Figure 5b) The statistical significance test results for comparisons between different LLMs and repeated evaluations within each LLM are represented by the *P*-value.

Despite the low accuracy, consistency varied notably among models. GPT-4o exhibited perfect consistency at 100%, indicating that while it consistently provided incorrect answers, its responses remained stable across multiple runs. DeepSeek-V3 and o3-mini both maintained relatively high consistency at 86%, suggesting a degree of reliability despite their lower accuracy. In contrast, o1 and DeepSeek-R1 displayed lower consistency, at 68.3% and 67.3%, respectively, implying greater variability in their responses. This inconsistency, particularly in the case of o1, may be attributed to its higher accuracy—fluctuating between correct and incorrect answers rather than repeatedly making the same errors.

Discussion

This study is the first to compare the performance of GPT and DeepSeek using medical licensing exam questions. The results confirm that both GPT and DeepSeek passed the KMLE. Notably, all tools, except for DeepSeek-V3, passed with scores higher than the human average. o1 and DeepSeek-R1 showed

superior performance by passing within the top 10%. In the domain-specific evaluation, a performance decline was observed in the Local Health & Medical Law exam, which reflects the regional characteristics of the Republic of Korea, compared to globally recognized preventive medicine and clinical medicine questions.

The performance differences between OpenAI and DeepSeek models were minimal, given that all models surpassed 60% accuracy. The best-performing model in terms of accuracy was o1, whereas DeepSeek-R1 emerged as the most cost-effective option. Regarding the performance differences across question categories, the lowest accuracy was observed in the domain of Local Health & Medical Law. Further analysis of the questions with the lowest average accuracy across the five variants revealed that o1 exhibited the highest accuracy among all models in this category.

Research on the application of LLMs in medical education and assessment has been continuously evolving. Notably, MedExamLLM, a web-based platform summarizing LLM performance across various medical examinations worldwide, reported that GPT-4, achieved the highest performance [7]. This meta-analysis identified language-based performance discrepancies among LLMs, with superior results observed in English-language assessments due to the larger volume of training data available. This underscores the need for studies evaluating LLM performance on Korean-language medical examinations, highlighting the significance of the present study.

A recent study evaluating LLM performance on the United States Medical Licensing Examination (USMLE) reported an accuracy of 46.23% for GPT-3.5, which increased to 71.33% for GPT-4, surpassing the average human test-taker

accuracy of 54.38%[6]. These findings align with the trends observed in our study, demonstrating that LLM performance improves with model advancements (Figure 1, Table 1).

Similarly, a comparative study of LLM performance on the Japanese Medical Licensing Examination (which uses Japanese rather than English) found that GPT-4o achieved the highest accuracy (89.2%) and reliability (>95%) [9]. This is consistent with our findings for the Korean-language KMLE, which also showed high accuracy and consistency on the latest GPT model (Figure 3,4, Table 1,2).

In the domain of traditional Korean medicine, where both Korean and Classical Chinese are used, an LLM-based study reported that GPT-4 achieved an overall accuracy of 66.18%, surpassing the 60% passing threshold [10]. However, while pharmacology (87.5%) and neuropsychiatry (81.2%) exhibited high accuracy, public health and legal regulations (40%) and traditional Korean medicine (55.4%) demonstrated lower performance. These language-based discrepancies align with our study's findings, where the lowest accuracy was observed in the Local Health & Medical Law category (Figure 3).

Additionally, a study comparing advanced reasoning AI models (o1 and DeepSeek-R1) in pediatric clinical decision support analyzed 500 multiple-choice pediatric-related questions [11]. It concluded that GPT o1 (92.8%) outperformed DeepSeek-R1 (87.0%) in diagnostic accuracy. This corresponds to our study's findings, where o1 (94.3%) surpassed DeepSeek-R1 (91.6%) (Table 1).

In the 2024 KMLE, 3,231 candidates participated, with 3,084 passing (scoring above 60%). The average score of all candidates was 256.2 out of 320 (80.1%), with a standard deviation of 25.3. Excluding DeepSeek-V3 (74.9%), all LLM models in our study outperformed this average, demonstrating superior

accuracy compared to human test-takers.

Among the top 10% of candidates, the average score was 286 out of 320 (89.4%). Given that o1 achieved 94.3% accuracy and DeepSeek-R1 achieved 91.6%, both models exceeded the performance of top human candidates (Table 1).

Among the various OpenAI models used, o1 demonstrated the highest accuracy (94.3%) and consistency (97.1%), likely due to its enhanced deep reasoning capabilities and proficiency in solving complex problems. Additionally, DeepSeek-R1 significantly outperformed its predecessor (accuracy: 91.6% vs. 74.9%), which can be attributed to its optimization for handling more sophisticated tasks.

Notably, despite minor variations in overall performance, all models exhibited relatively similar pass rates and consistently high reliability (>90%). This finding supports the feasibility of utilizing LLMs in medical education and assessment.

Furthermore, o1 showed the best performance, while DeepSeek-R1 demonstrated a slight performance gap. Given that DeepSeek-R1 is available for free, while GPT requires a monthly subscription fee of $200, this suggests that users may have different preferences when choosing between the two, considering the respective advantages and disadvantages.

Accuracy for Local Health & Medical Law questions was notably lower across all models. Additional analysis of the five most frequently misclassified questions revealed consistently poor performance across multiple LLMs. Several factors could explain this trend:

- Localization of Legal Regulations: Local Health & Medical Law questions reflect Korea’s unique legal framework, and the limited availability of

relevant localized data may hinder LLM performance. While legal texts are available in both Korean and English on the National Law Information Center, LLMs may struggle to efficiently extract and apply relevant regulations.

- Insufficient Korean-Language Training Data: OpenAI's training dataset consists of over 90% English-language data, with Korean comprising only 0.017%. This imbalance may contribute to lower accuracy in Local Health & Medical Law questions compared to clinical medicine questions.
- Frequent Legal Revisions: Medical regulations evolve based on societal needs and policy changes. Given that the latest OpenAI models have a knowledge cutoff of April 2023, they may lack updates on recent legislative amendments, further limiting their performance in this domain.

This study holds significant novelty as the first to evaluate LLM performance on the Korean Medical Licensing Examination. Moreover, by streamlining the analysis process through API-based evaluations, our study provides a framework for future research utilizing different AI models. Additionally, presenting results in terms of accuracy and consistency enhances accessibility for medical students, practitioners, and the general public, promoting awareness of LLM capabilities in the KMLE and encouraging further research.

However, our study also has limitations. Due to copyright concerns related to personal information, we excluded multimedia-based questions from the 2024 KMLE dataset. As a result, we could not assess LLMs' image interpretation abilities, limiting our evaluation of visual reasoning and clinical judgment skills. Nevertheless, this study is the first research to evaluate medical licensing exam

using GPT and DeepSeek.

Furthermore, while we attempted to enhance reliability by having each model answer the same question multiple times, our assessment of LLMs' reasoning processes remained limited. Future studies could address this by comparing AI-driven reasoning with expert clinical judgment.

LLMs continue to evolve, with recent developments such as Grok (by X) and Manus AI (from China) garnering attention. Future studies should extend our research by evaluating these emerging models for their medical applications. Additionally, the development of a Korea-specific AI model may overcome the limitations observed in Local Health & Medical Law-related questions, enhancing overall performance in localized assessments.

Conclusion

This study evaluated the performance of OpenAI and DeepSeek models on the KMLE, demonstrating that all models achieved accuracy above the passing threshold and exhibited high consistency. While o1 requires a $200 monthly subscription, DeepSeek-R1 is available for free and still demonstrates a similar level of accuracy As the first study to assess the four most advanced LLMs—GPT-4 and DeepSeek—on medical licensing exam, our research provides valuable insights into their capabilities and limitations. Given the increasing role of AI in healthcare, further efforts are needed to optimize LLMs for medical applications, enhancing physician efficiency and accuracy in both patient care and examination preparation. As LLM models continue to evolve, further extensive research is needed on the vulnerabilities of LLMs in various healthcare

national exams, such as the KMLE, the potential to overcome limitations in languages other than English, and the evaluation of their economic effectiveness.

Funding Declaration

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(Ministry of Science and ICT, Grant number: RS-2022-NR071874).

Data availability

The materials published by the National Health Personnel Licensing Examination Institute of Korea are available for download from the official website and can be used for research purposes if needed. (https://www.kuksiwon.or.kr/EngHome/index.do)

References

X1. Savage, Thomas, et al. "Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine." NPJ Digital Medicine 7.1 (2024): 20.

2. Goh, Ethan, et al. "Large language model influence on diagnostic reasoning: a randomized clinical trial." JAMA Network Open 7.10 (2024): e2440969-e2440969.

3. Metz, C., and M. Tobin. "How Chinese AI Start-Up DeepSeek Is Competing With Silicon Valley Giants." The New York Times. https://www. nytimes. com/2025/01/23/technology/deepseek-china-ai-chips. html (2025).

4. Hwang, Sang-Ik. "History of the medical licensure system in Korea from the late 1800s to 1992." Journal of Educational Evaluation for Health Professions 21 (2024): 36.

5. Meyer, Annika, Janik Riese, and Thomas Streichert. "Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the written German medical licensing examination: observational study." JMIR Medical Education 10 (2024): e50965.

6. Penny, Parker, Riley Bane, and Valerie Riddle. "Advancements in AI Medical Education: Assessing ChatGPT's Performance on USMLE-Style Questions Across Topics and Difficulty Levels." Cureus 16.12 (2024).

7. URL link: https://www.kuksiwon.or.kr/EngHome/index.do
8. Zong, Hui, et al. "Large language models in worldwide medical exams: Platform development and comprehensive analysis." Journal of Medical Internet Research 26 (2024): e66114.

9. Liu, Mingxin, et al. "Evaluating the Effectiveness of advanced large language models in medical Knowledge: A Comparative study using Japanese national medical examination." International Journal of Medical Informatics 193 (2025): 105673.
10. Jang, Dongyeop, et al. "GPT-4 can pass the Korean national licensing examination for Korean medicine doctors." PLOS Digital Health 2.12 (2023): e0000416.

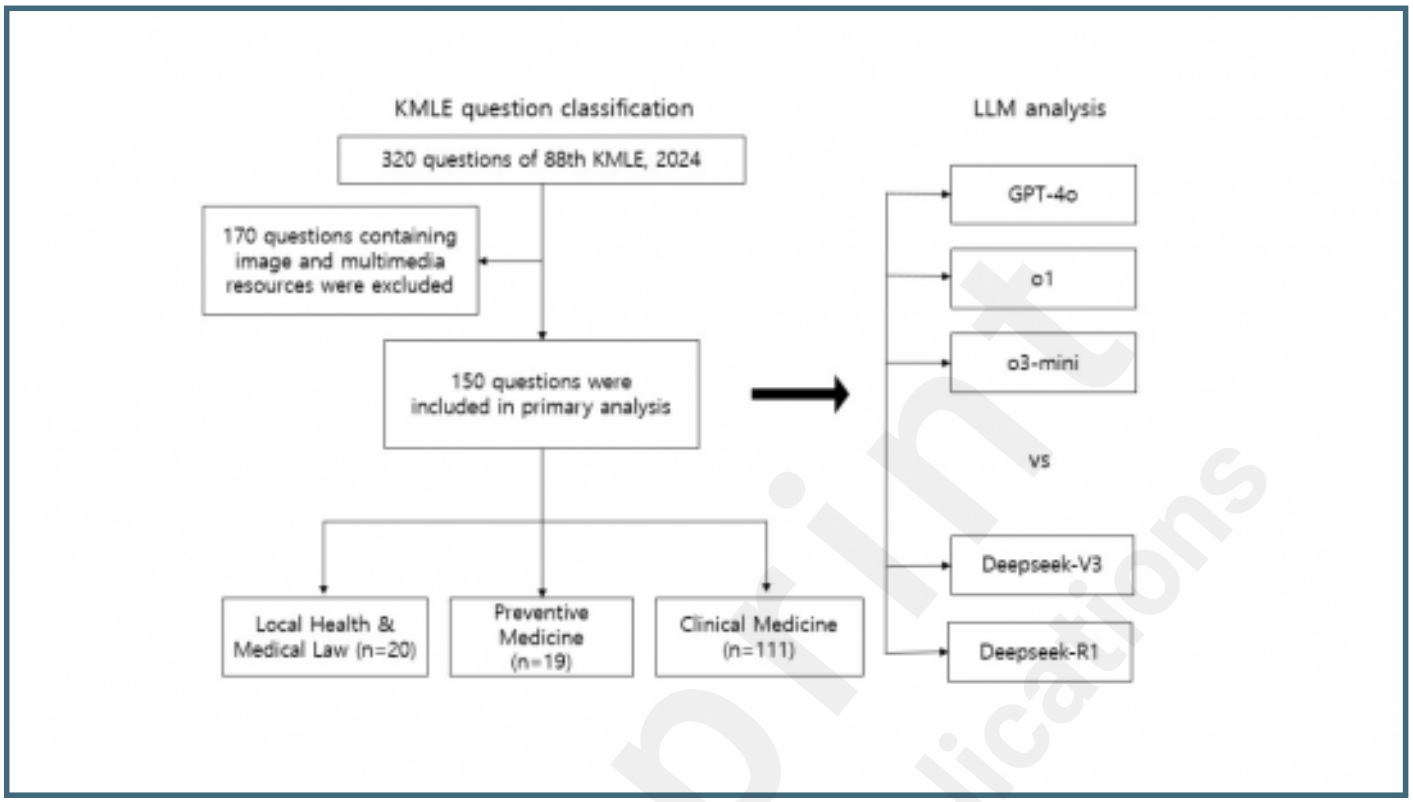
11. Mondillo, Gianluca, et al. "Comparative Evaluation of Advanced AI Reasoning Models in Pediatric Clinical Decision Support: ChatGPT O1 vs. DeepSeek-R1." medRxiv (2025): 2025-01.

Supplementary Files

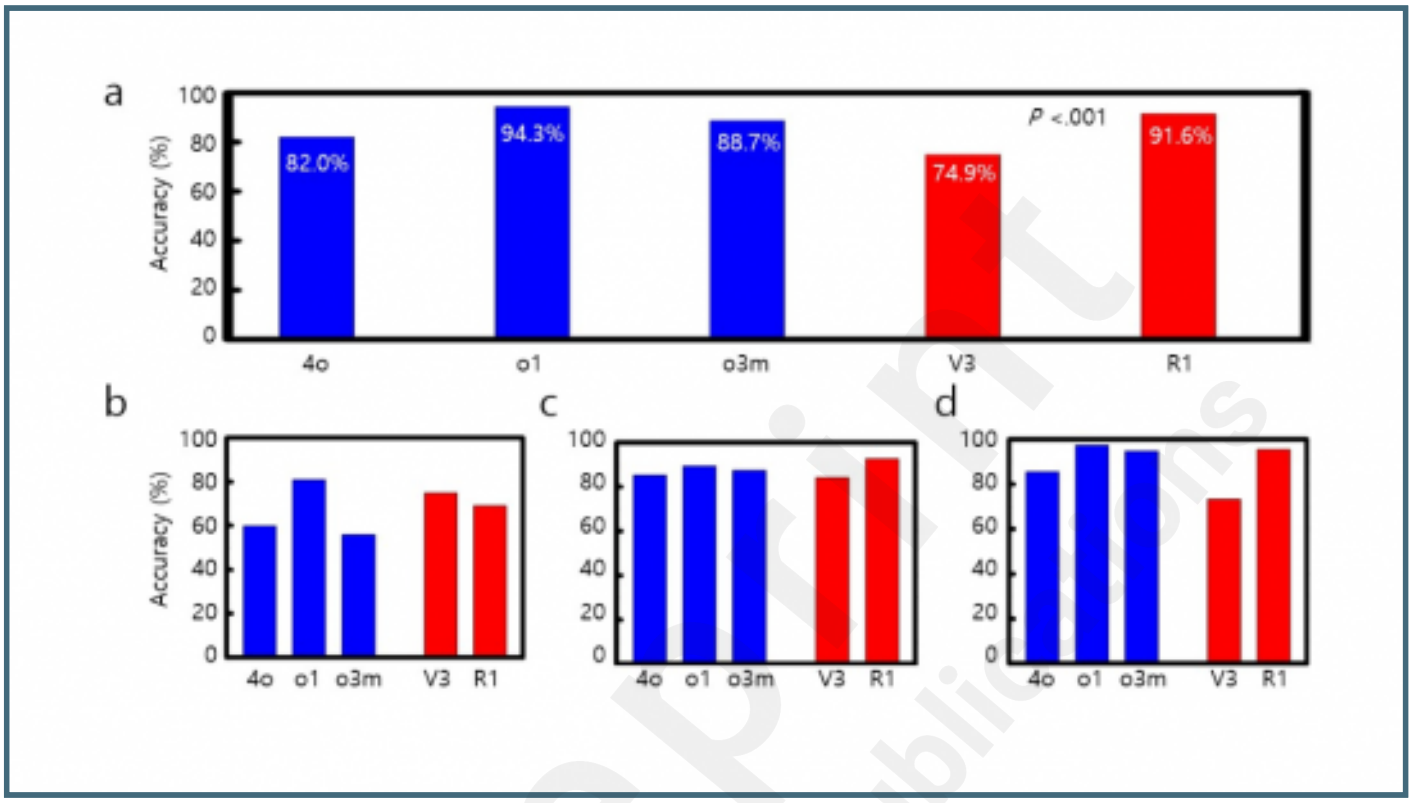
Figures

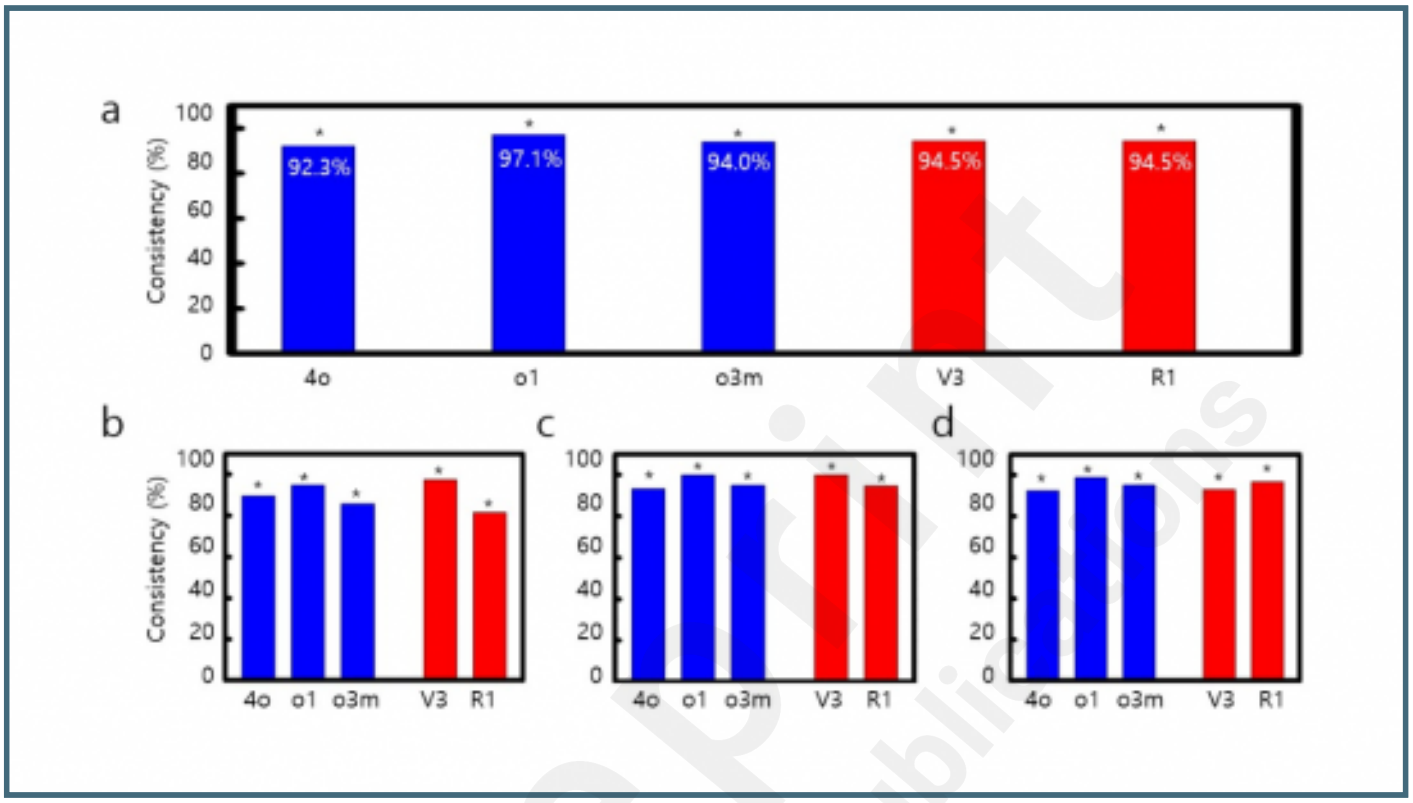
Schematic overview of conceptual illustration of ChatGPT(4o, o1, o3-mini) and DeepSeek (V3, R1) engaging in a comparative evaluation on the Korean Medical Licensing Examination(KMLE). The diagram visualizes the comparison of outputs generated by both models in response to KMLE questions.

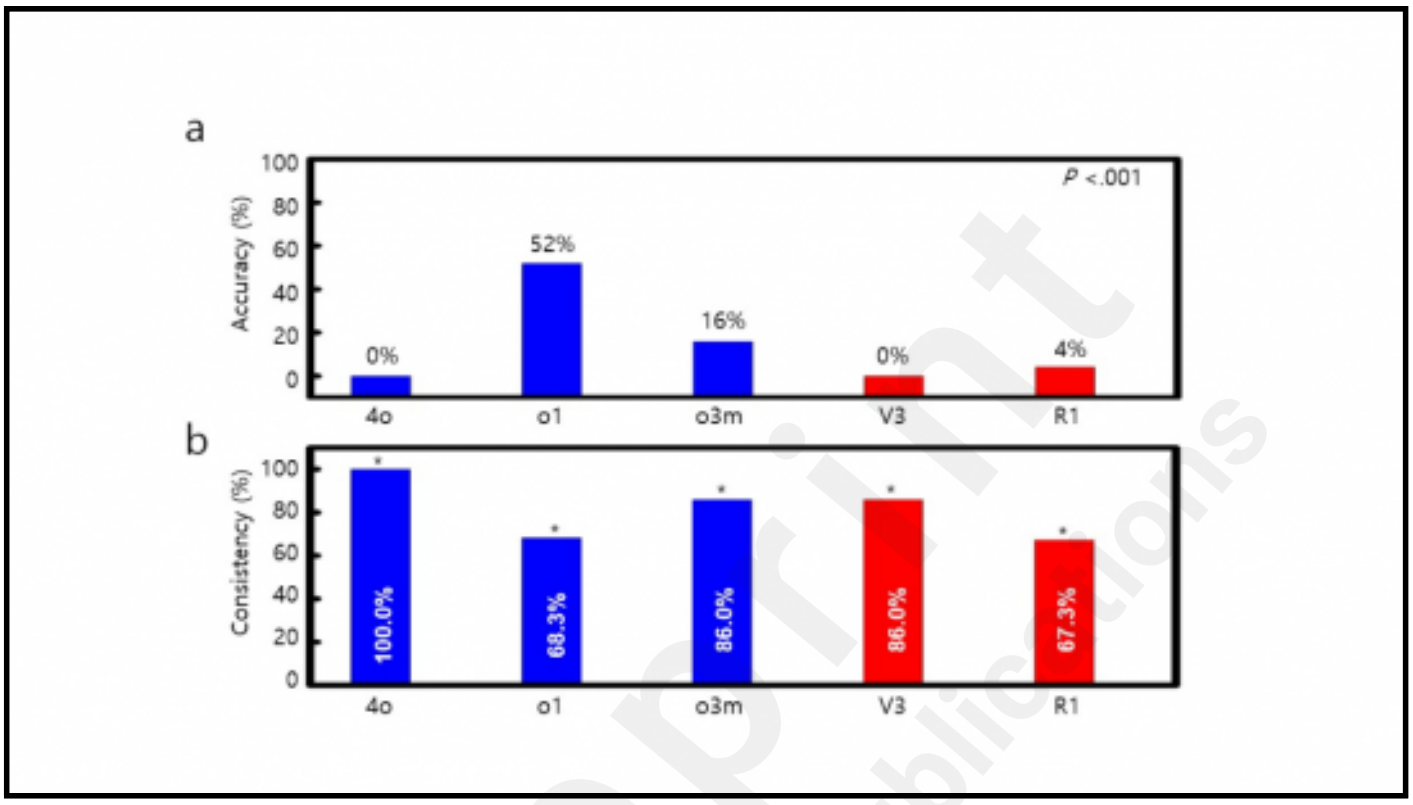
Summary of KMLE question enrollment, classification and LLM analysis process. This figure shows the algorithm for question enrollment, classification of KMLE before LLM analysis process.



Comparison of accuracy between GPTs and DeepSeeks in solving KMLE across each subject area. (a) comparison of accuracy of all included questions between GPT 4o, o1, o3-mini(o3m) and Deepseek V3(V3), R1(R1). (n=150) (b) of Local Health & Medical Law (n=20) (c) of Preventive Medicine (n=19) (d) of Clinical Medicine (n=111).







Multimedia Appendixes

Table numbers are indicated in the file.

URL: http://asset.jmir.pub/assets/6a1516917110a7920b8f251e80ddb50b.docx