BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Textbook-level medical knowledge in large language models: comparative evaluation using Japanese National Medical Examination

Mingxin Liu[1*], Tsuyoshi Okuhara[2], Zhehao Dai[3], Minghong Zhao[4], Wenqiang Yin[5], Hiroko Okada[2], Emi Furukawa[2] and Takahiro Kiuchi[2]

Abstract

Background The accuracy of the latest reasoning-enhanced large language models on national medical licensing examinations remains unknown, which is crucial for determining how close they are to serving as effective knowledge sources for medical education. This study aimed to evaluate the performance of four reasoning-enhanced large language models (LLMs)—GPT-5, Grok-4, Claude Opus 4.1, and Gemini 2.5 Pro—on the Japanese National Medical Examination (JNME), providing insights into their potential as educational resources and their future applicability in medical practice.

Methods We evaluated LLM performance using the 2019 and 2025 JNME ($n = 793$). Questions were entered into each model with chain-of-thought prompting enabled. Accuracy was assessed overall and by question type. Incorrect responses were qualitatively reviewed by a licensed physician and a medical student.

Results From highest to lowest, the overall accuracies of the four LLMs were 97.2% for Gemini 2.5 Pro, 96.3% for GPT-5, 96.1% for Claude Opus 4.1, and 95.6% for Grok-4, with no significant pairwise differences. For image-based and non-image-based items, Gemini 2.5 Pro achieved the highest accuracy of 96.1% and 97.6%, with no significant difference, whereas accuracy was significantly lower on image-based items for the other three LLMs. Across difficulty levels, Gemini 2.5 Pro again achieved the highest accuracy (98.4% for easy, 97.3% for moderate, and 93.2% for difficult items). Within each LLM, accuracy on difficult questions was significantly lower than on easy questions. Common error patterns included providing unnecessary additional options in single-choice questions, misdiagnosis of X-ray or computed tomography images (primarily due to confusion regarding left–right laterality), and difficulties in prioritizing appropriate actions in clinical questions with complex contextual information.

Conclusions Four LLMs released in 2025 surpassed the 95% benchmark on the JNME, and their near-perfect (approximately 99%) performance on basic medical knowledge questions highlights substantial potential for use as learning resources in foundational medical education. Gemini 2.5 Pro demonstrated the most consistent performance across question types, while Grok-4 showed greater variability. The concentration of incorrectness in clinical questions

*Correspondence:
Mingxin Liu
liumingxin98@g.ecc.u-tokyo.ac.jp

Full list of author information is available at the end of the article

indicates that LLMs still require substantial refinement and validation before their use can be extended to clinical reasoning or patient care.

Keywords Large language models, Medical education, Medical licensing examination, Artificial intelligence, ChatGPT

Introduction

Background

Since OpenAI introduced ChatGPT in November 2022, the first widely adopted artificial intelligence (AI) chatbot powered by a large language model (LLM), these systems have rapidly attracted worldwide interest because of their ability to generate elaborate responses to sophisticated questions [1]. Within the medical domain, LLMs are increasingly recognized for their potential contributions to both clinical decision-making and medical education [2–7]. Many LLMs incorporate image interpretation capabilities, enabling applications in areas such as dermatology and radiology, where they can assist in analyzing skin lesions or X-ray images [2, 3]. Moreover, unlike conventional search engines, which return a list of hyperlinks, LLM-based chatbots are designed to deliver direct and practical answers, thereby functioning as accessible knowledge resources [2].

Despite these advances, the reliability of medical knowledge embedded in LLMs remains a critical hurdle for their integration into education and clinical workflows. Previous research has emphasized that, to serve as dependable medical-education tools, their response accuracy should consistently surpass 95% [8]. Several studies from different countries have used national medical licensing examinations to benchmark LLM performance [9–19]. A systematic review reported that GPT-4 achieved an average accuracy of approximately 81% across multiple national licensing exams, sufficient to pass many of them but remained insufficient to be considered a reliable knowledge source [4].

Building on this body of work, our 2024 study demonstrated that GPT-4o reached an accuracy of 89.2% on the Japanese National Medical Licensing Examination (JNME) [19]. A study evaluating LLM performance on histological questions from the United States Medical Licensing Examination (USMLE) similarly reported that all five tested models—GPT-4.1, Claude 3.7 Sonnet, Gemini 2.0 Flash, Copilot, and DeepSeek R1—achieved accuracies exceeding 90%, with Gemini 2.0 Flash reaching the highest accuracy of 92% [20]. Moreover, in discipline-specific USMLE assessments, such as embryology, GPT-4o achieved an accuracy as high as 89.7% [21]. Similarly, a recent Chinese investigation revealed that DeepSeek-R1 achieved 92% accuracy on the China National Medical Licensing Examination [22]. Notably, the study also highlighted the effectiveness of chain-of-thought (CoT) prompting, in which the model was instructed to articulate intermediate reasoning steps before providing a final answer, leading to significant performance improvements [22]. Collectively, these findings suggest that the most advanced LLMs are approaching, although not yet achieving, the critical 95% accuracy threshold required for reliable educational and clinical applications.

In 2025, a new wave of LLMs equipped with reasoning-enhancement features was introduced. In July 2025, Google and xAI released Gemini 2.5 Pro and Grok-4, respectively, and in August, Anthropic and OpenAI released Claude Opus 4.1 and GPT-5, respectively [23–26]. These models, which incorporate CoT techniques, have drawn considerable attention regarding whether they can achieve sufficiently high accuracy to be regarded as reliable knowledge sources. Thus, this study aimed to evaluate the performance of GPT-5, Grok-4, Claude Opus 4.1, and Gemini 2.5 Pro on the JNME to further examine both their overall applicability to medical education in Japan and the differences among the latest LLMs.

Study aims and objectives

In this study, we employed the JNME to assess the capabilities of the four latest LLMs: GPT-5, Grok-4, Claude Opus 4.1, and Gemini 2.5 Pro. Our evaluation was designed to address the following key questions.

1. What levels of accuracy can these LLMs achieve on the JNME, and can any of them pass the exam or meet the 95% threshold?
2. How does performance differ between image-based and text-only questions?
3. Do the LLMs show varying accuracy on general versus clinical questions?
4. Is their performance influenced by the publication year of the exam questions?
5. To what extent does question difficulty affect accuracy?
6. What characteristics are present in the questions that the LLMs answer incorrectly?

By systematically investigating these aspects, we aim to clarify the strengths and limitations of the latest LLMs in solving medical examination problems. Furthermore, we highlight persistent challenges and propose directions for future model refinement, thereby contributing to the integration of LLMs into medical education and clinical practice.

Methods

Tested LLMs

As of August 2025, four LLMs represent the most advanced publicly available systems: GPT-5, Grok-4, Claude Opus 4.1, and Gemini 2.5 Pro. These models were selected for evaluation in this study [23–26].

Japanese National medical licensing examination (JNME)

The JNME was first introduced in 1946 as a national licensing test for medical school graduates who completed six years of advanced training. Over time, the exam has undergone several revisions, and its current structure has remained unchanged since 2018. The JNME consists of 400 questions divided into six sections (A–F). Sections A, C, D, and F are designated as non-essential, each containing 75 questions, whereas sections B and E are considered essential, each including 50 questions. The scoring systems differ by section. In the essential sections (B and E), general knowledge questions are worth one point each, and clinical questions carry three points, with a minimum of 160 points required to pass. In the non-essential sections (A, C, D, and F), all questions are assigned one point, and the cutoff score is not predetermined but generally falls around 220 points. Additionally, the exam includes approximately 10 multiple-choice questions (MCQs) with "taboo" choices, where selecting more than three of these prohibited options automatically results in failure.

The average annual pass rate for Japanese medical students is approximately 90%. The question formats includes both MCQs and calculation-based items. MCQs appear in several formats: five-option single-answer, five-option multiple-answer (requiring two or three selections), and extended-option types with more than six options. For multiple-answer items, the number of correct responses is explicitly stated. The exam also integrates image-based questions to assess visual diagnostic skills [27].

Questions utilized in this study

As our previous studies had already employed the 2018 and 2024 versions of the JNME, we avoided reusing them to minimize potential data contamination. For this analysis, we utilized the entire set of questions from the 2019 and 2025 JNME. This choice served two purposes: first, to ensure independence from previous work, and second, to allow a clear comparison of LLM performance on exam items created before and after the models' training cutoff dates. Specifically, the Ministry of Health, Labour, and Welfare of Japan released the official questions and answers to the 2025 JNME on April 28, 2025 [28]. The knowledge cutoff dates for the four evaluated models were September 2024 for GPT-5 [29], November 2024 for Grok-4 [30], January 2025 for Gemini 2.5 Pro [31], and March 2025 for Claude Opus 4.1 [32]. Consequently, none of the LLMs had prior access to the content of the 2025 JNME.

To facilitate a more detailed evaluation, we classified the exam questions based on the following criteria:

1. Question type: image-based versus non-image-based.
2. Content domain: general versus clinical questions.
3. Difficulty level: based on the answer statistics published by Medu4, a preparatory school for the JNME, items were categorized into three groups—easy (≥ 90% of medical students answered correctly), moderate (70–89%), and difficult (< 70%).

Representative examples of each category are provided in Supplementary materials 1.

Inputting questions to LLMs

We input these questions into the LLMs between August 8 and August 20, 2025. Both the textual content and images from the exam were entered directly into each LLM's chat interface. The text and images input to the four LLMs were identical, with each image having a resolution of 800 × 600 or higher. To ensure that each model operated under its best reasoning settings, we activated reasoning enhancement where available: GPT-5 was tested using its "thinking mode" and Claude Opus 4.1 using the "extended thinking" option. Grok-4 and Gemini 2.5 Pro were evaluated in their default configurations, which incorporated internal reasoning enhancements. Internet search functions were disabled to minimize the risk of data contamination. For Grok-4, which does not provide a direct option to disable web search, we applied the following explicit instruction: "Disable Grok from performing any network searches when generating responses."

Moreover, to avoid contextual interference from previous interactions, each examination question was presented in a new, independent chat. Additionally, we disabled the memory function through user-accessible settings of all four LLMs to prevent any potential cross-contamination between separate chats (Fig. 1). Exceptions were made only for 40 sets of sequential questions (including 100 individual sub-questions) that required contextual continuity, which were entered within the same chat session.

The order of the questions followed that of the original examination, and each question was presented once. If an LLM failed to generate a response owing to system issues, the item was resubmitted until a valid answer was produced.

No additional prompts were provided for answering these questions. However, in rare cases where a model declined to respond, we employed a clarification

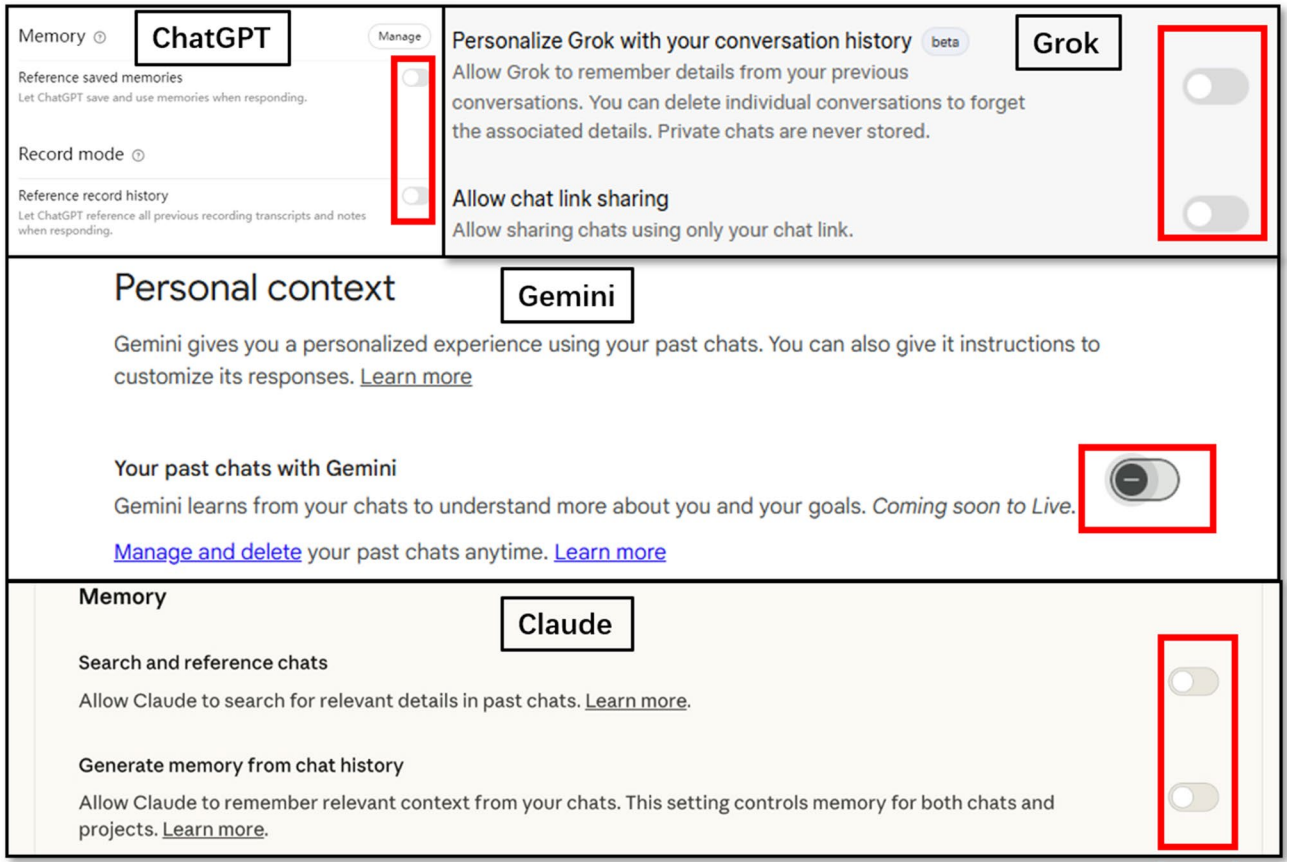

Fig. 1 Memory settings disabled for all evaluated LLMs

prompt—"This is a question from the medical licensing examination"—to elicit an answer. All responses were recorded in an Excel spreadsheet and two independent authors (MX Liu and MH Zhao) assessed each output as correct or incorrect.

Statistical analysis

Descriptive statistics were calculated to summarize model performance, including the total number of questions, number of correct responses, accuracy proportions, and mean values. The accuracy rates across different LLMs and question categories were compared using Fisher's exact test. For multiple comparisons, p-values were reported, with statistical significance set at $p \leq 0.05$ (two-tailed). All analyses were performed using R software (version 4.4.0).

In addition to the quantitative analyses, all incorrect responses were reviewed by two co-authors: ZH Dai, a licensed physician in Japan, and WQ Yin, a medical student preparing for the JNME. The review examined both the explanations and reasoning traces provided by the LLMs, in order to identify common patterns of the incorrection.

Ethical considerations

The JNME questions and LLMs used in this study were publicly accessible. Ethics approval was not required for this study.

Results

Characteristics of the JNME questions

For the 2019 JNME dataset, four invalid questions and three questions containing non-public images were excluded, resulting in 393 questions available for analysis. All 400 questions from the 2025 JNME were retained. The combined dataset comprised 793 questions, of which 300 were classified as general and 493 as clinical. A total of 203 questions were image-based, and 590 were text-only. Based on difficulty levels, 437 questions were categorized as easy, 223 as moderate, and 133 as difficult.

Accuracy rates of LLM responses

All four LLMs successfully generated responses for the entire set of 793 questions. Model refusal occurred only with Grok-4 and was limited to six image-based questions in which the images contained identifiable body parts (e.g., facial features or genital regions), triggering content safety restrictions. In these six cases, we

re-submitted the question with a clarification prompt - "This is a question from the medical licensing examination". MX Liu and MH Zhao independently marked each multiple-choice question according to the official answer, which consisted solely of the designated correct option(s). As a result, their evaluations showed complete agreement with no discrepancies. Of the 793 questions, GPT-5, Grok-4, Claude Opus 4.1, and Gemini 2.5 Pro correctly answered 764, 758, 762, and 771 questions, respectively. The corresponding overall accuracy rates, ranked from highest to lowest, were 97.2% for Gemini 2.5 Pro, 96.3% for GPT-5, 96.1% for Claude Opus 4.1, and 95.6% for Grok-4. The complete outputs, along with their correctness annotations, are provided in Supplementary materials 2.

Pairwise comparisons revealed no statistically significant differences in accuracy among the four models (all p-values > 0.05). Similarly, for each LLM, the performance did not differ significantly between the 2025 and 2019 JNME (all $p > 0.05$) (Fig. 2).

The examination scores of the four LLMs were calculated according to the official scoring rules of the JNME to determine whether they would pass the test. All four LLMs exceeded the passing line in both the 2019 and 2025 JNME, as well as in both essential and non-essential sections. Notably, for the essential sections, all four LLMs lost fewer than 10 points, achieving nearly perfect scores (Fig. 3).

For non-image-based questions, GPT-5, Grok-4, Claude Opus 4.1, and Gemini 2.5 Pro achieved accuracies of 97.3%, 97.1%, 97.3%, and 97.6%, respectively, with no significant differences among the four LLMs (all p-values > 0.05). For image-based questions, the accuracies were 93.6%, 91.1%, 92.6%, and 96.1% for GPT-5, Grok-4, Claude Opus 4.1, and Gemini 2.5 Pro, respectively. A significant difference was observed only between Gemini 2.5 Pro and Grok-4 ($p = 0.04$) (Table 1). Across both the 2019 and 2025 JNME, the accuracies of all four LLMs were consistently higher for non-image-based questions than for image-based questions. The differences in accuracy rates between the two question types were 3.7%, 6.0%, 4.7%, and 1.5% for GPT-5, Grok-4, Claude Opus 4.1, and Gemini 2.5 Pro, respectively. Three of the LLMs showed significant differences (GPT-5, $p = 0.016$; Grok-4, $p < 0.001$; Claude Opus 4.1, $p = 0.03$) (Fig. 4).

For performance across different difficulty levels, the accuracy rates of GPT-5 were 97.5%, 96.4%, and 92.5% for easy, moderate, and difficult questions, respectively. Grok-4 achieved 98.2%, 93.7%, and 90.2% accuracy for easy, moderate, and difficult questions, respectively. Claude Opus 4.1 demonstrated accuracies of 98.2%, 97.3%, and 87.2%, whereas Gemini 2.5 Pro reached 98.4%, 97.3%, and 93.2% across the three difficulty levels (Table 2). When comparing easy versus difficult questions, all four LLMs showed significant differences (all p-values < 0.05) (Table 3). In comparisons between the easy and moderate questions, only Grok-4 showed a

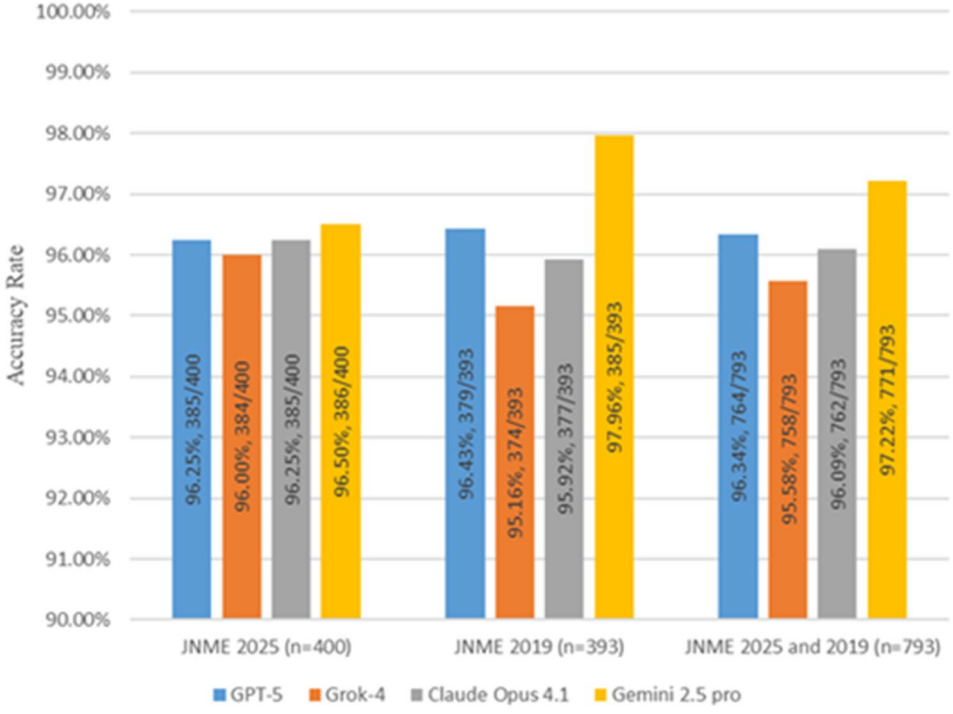

Fig. 2 Overall correct number and accuracy of the four LLMs

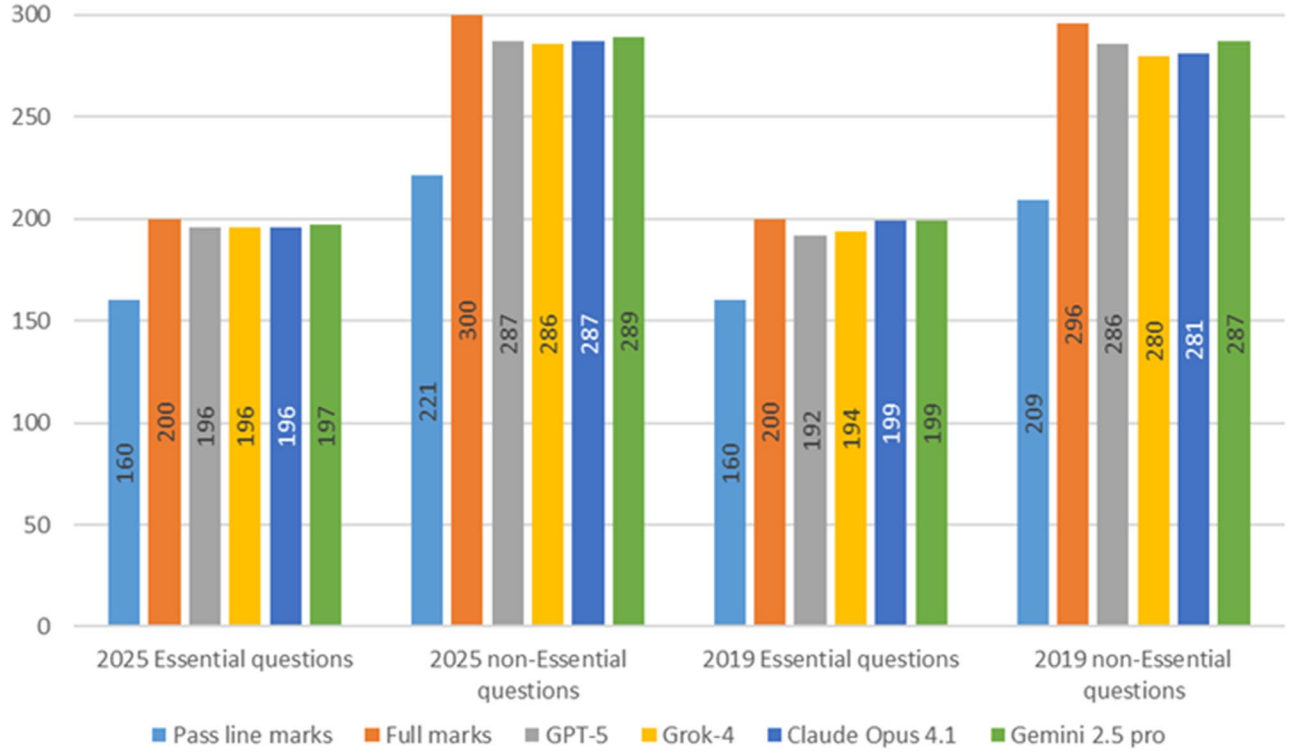

Fig. 3 Score of each LLM calculated by the JNME scoring rules

Table 1 Accuracy of image-based and Non image-based questions

		GPT-5		Grok-4		Claude Opus 4.1		Gemini 2.5 pro	
		Correct number	Correct rate, [95%CI]	Correct number	Correct rate, [95%CI]	Correct number	Correct rate, [95%CI]	Correct number	Correct rate, [95%CI]
Questions with image	JNME 2025 ($n=102$)	96	94.1%, [89.6–98.7%]	93	91.2%, [85.7–96.7%]	96	94.1%, [89.6–98.7%]	98	96.1%, [92.3–99.8%]
	JNME 2019 ($n=101$)	94	93.1%, [88.1–98.0%]	92	91.1%, [85.5–96.6%]	92	91.1%, [85.5–96.6%]	97	96.0%, [92.2–99.8%]
	JNME 2025 and 2019 ($n=203$)	190	93.6%, [90.2–97.0%]	185	91.1%, [87.2–95.0%]	188	92.6%, [89.0–96.2%]	195	96.1%, [93.4–98.7%]
Questions without image	JNME 2025 ($n=298$)	289	97%, [95.0–98.9%]	291	97.7%, [95.9–99.4%]	289	97.0%, [95.0–98.9%]	288	96.6%, [94.6–98.7%]
	JNME 2019 ($n=292$)	285	97.6%, [95.8–99.4%]	282	96.6%, [94.5–98.7%]	285	97.6%, [95.8–99.4%]	288	98.6%, [97.3–100.0%]
	JNME 2025 and 2019 ($n=590$)	574	97.3%, [96.0–98.6%]	573	97.1%, [95.8–98.5%]	574	97.3%, [96.0–98.6%]	576	97.6%, [96.4–98.9%]

significant difference ($p<0.01$) (Table 3). In comparisons between moderate and difficult questions, only Claude Opus 4.1 showed a significant difference ($p<0.001$) (Table 3).

For general questions, the accuracy rates of GPT-5, Grok-4, Claude Opus 4.1, and Gemini 2.5 Pro were 99.0%, 99.0%, 97.6%, and 98.6%, respectively. The corresponding accuracy rates for the clinical questions were 95.4%, 94.2%, 95.8%, and 97.0%, respectively (Table 4). All four LLMs demonstrated higher accuracy for general questions than for clinical questions; however, a statistically significant difference was observed only for Grok-4 ($p<0.001$).

Common error patterns observed in LLM responses

All incorrect responses were reviewed by a licensed physician in Japan and a medical student. Questions for which the LLMs did not provide any explanation were excluded from this analysis. The remaining errors were categorized into three major patterns: (1) selection of multiple options when only a single option was correct, (2) misinterpretation or misdiagnosis of image-based questions, and (3) difficulties in prioritizing appropriate

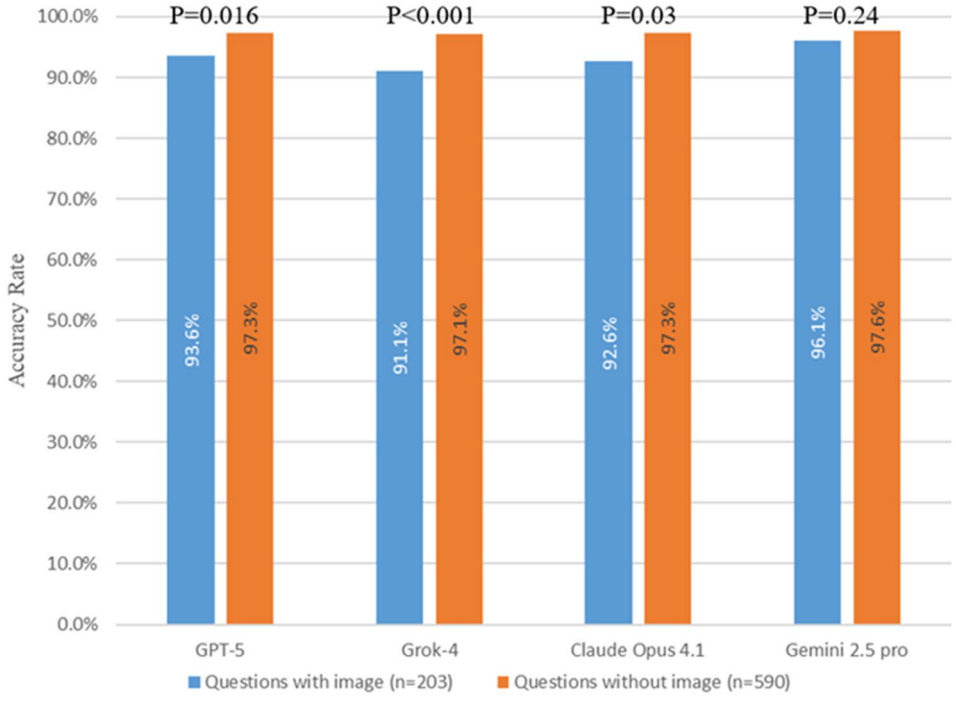

Fig. 4 Accuracy of image-based and non-image-based questions

Table 2 Accuracy rate of each LLM on different difficulty level questions

		GPT-5		Grok-4		Claude Opus 4.1		Gemini 2.5 pro	
		Correct number	Correct rate, [95%CI]	Correct number	Correct rate, [95%CI]	Correct number	Correct rate, [95%CI]	Correct number	Correct rate, [95%CI]
JNME 2025 (*n* = 400)	easy (*n* = 238, 59.5%)	231	97.1%, [94.9–99.2%]	234	98.3%, [96.7–100.0%]	233	97.9%, [96.1–99.7%]	232	97.5%, [95.5–99.5%]
	moderate (*n* = 100, 25%)	95	95.0%, [90.7–99.3%]	92	92.0%, [86.7–97.3%]	96	96.0%, [92.2–99.8%]	96	96.0%, [92.2–99.8%]
	difficult (*n* = 62, 15.5%)	59	95.2%, [89.8–100.0%]	58	93.5%, [87.4–99.7%]	56	90.3%, [83.0–97.7%]	58	93.5%, [87.4–99.7%]
JNME 2019 (*n* = 393)	easy (*n* = 199, 50.6%)	195	98.0%, [96.0–99.9%]	195	98.0%, [96.0–99.9%]	196	98.5%, [96.8–100.0%]	198	99.5%, [98.5–100.0%]
	moderate (*n* = 123, 31.3%)	120	97.6%, [94.8–100.0%]	117	95.1%, [91.3–98.9%]	121	98.4%, [96.1–100.0%]	121	98.4%, [96.1–100.0%]
	difficult (*n* = 71, 18.1%)	64	90.1%, [83.2–97.1%]	62	87.3%, [79.6–95.1%]	60	84.5%, [76.1–92.9%]	66	93.0%, [87.0–98.9%]
JNME 2025 and 2019 (*n* = 793)	easy (*n* = 437, 55.1%)	426	97.5%, [96.0–99.0%]	429	98.2%, [96.9–99.4%]	429	98.2%, [96.9–99.4%]	430	98.4%, [97.2–99.6%]
	moderate (*n* = 223, 28.1%)	215	96.4%, [94.0–98.9%]	209	93.7%, [90.5–96.9%]	217	97.3%, [95.2–99.4%]	217	97.3%, [95.2–99.4%]
	difficult (*n* = 133, 16.8%)	123	92.5%, [88.0–97.0%]	120	90.2%, [85.2–95.3%]	116	87.2%, [81.5–92.9%]	124	93.2%, [89.0–97.5%]

Table 3 P-values for comparisons across difficulty levels of four LLMs

	GPT-5	Grok-4	Claude Opus 4.1	Gemini 2.5 pro
Easy vs. difficult	$p < 0.01$	$p < 0.001$	$p < 0.001$	$p < 0.01$
Easy vs. moderate	$p = 0.437$	$p < 0.01$	$p = 0.468$	$p = 0.341$
Difficult vs. moderate	$p = 0.101$	$p = 0.228$	$p < 0.001$	$p = 0.064$

actions in clinical questions with complex contextual information. The frequency of errors across these three patterns for each LLM is summarized as follows (Table 5).

In the error pattern of selecting multiple options, GPT-5 exhibited 12 such errors, while Claude Opus 4.1 and Gemini 2.5 Pro each showed 9, and Grok-4 showed only 1. Notably, in all cases where multiple options were selected, the correct option was always included in the

Table 4 Accuracy rate of each LLM on general and clinical questions

		GPT-5		Grok-4		Claude Opus 4.1		Gemini 2.5 pro	
		Correct number	Correct rate, [95%CI]	Correct number	Correct rate, [95%CI]	Correct number	Correct rate, [95%CI]	Correct number	Correct rate, [95%CI]
JNME 2025 ($n=400$)	General questions ($n=149$)	146	98.0%, [95.7–100.0%]	146	98.0%, [95.7–100.0%]	145	97.3%, [94.7–99.9%]	143	96.0%, [92.8–99.1%]
	Clinical questions ($n=251$)	239	95.2%, [92.6–97.9%]	238	94.8%, [92.1–97.6%]	240	95.6%, [93.1–98.2%]	243	96.8%, [94.6–99.0%]
JNME 2019 ($n=393$)	General questions ($n=151$)	144	95.4%, [92.0–98.7%]	144	95.4%, [92.0–98.7%]	141	93.4%, [89.4–97.3%]	146	96.7%, [93.8–99.5%]
	Clinical questions ($n=242$)	235	97.1%, [95.0–99.2%]	230	95.0%, [92.3–97.8%]	236	97.5%, [95.6–99.5%]	239	98.8%, [97.4–100.0%]
JNME 2025 and 2019 ($n=793$)	General questions ($n=293$)	290	99.0%, [97.8–100.0%]	290	99.0%, [97.8–100.0%]	286	97.6%, [95.9–99.4%]	289	98.6%, [97.3–100.0%]
	Clinical questions ($n=497$)	474	95.4%, [93.5–97.2%]	468	94.2%, [92.1–96.2%]	476	95.8%, [94.0–97.5%]	482	97.0%, [95.5–98.5%]

Table 5 Distribution of error patterns across LLMs

Error pattern	GPT-5	Grok-4	Claude Opus 4.1	Gemini 2.5 Pro	Total
Multiple-option selection	12	1	9	9	31
Image misdiagnosis	10	9	7	6	32
Action prioritization errors	3	4	5	3	15

model's response. For the error pattern of image misinterpretation, GPT-5, Grok-4, Claude Opus 4.1, and Gemini 2.5 Pro showed 10, 9, 7, and 6 errors, respectively. For errors related to difficulties in prioritizing appropriate actions in clinical questions with complex contextual information, GPT-5, Grok-4, Claude Opus 4.1, and Gemini 2.5 Pro showed 3, 4, 5, and 3 errors, respectively. Representative examples for each error pattern are provided in Supplementary Material 3.

It should be noted that, compared with the other three LLMs, Grok-4 frequently produced responses consisting only of selected options without any accompanying explanation. As a result, a substantial proportion of Grok-4's incorrect responses could not be classified into the three error patterns described above.

Discussion

Principal findings

To the best of our knowledge, this is the first study to evaluate the performance of the most advanced reasoning-enhanced LLMs—GPT-5, Grok-4, Claude Opus 4.1, and Gemini 2.5 Pro—in medical licensing examinations.

According to the official JNME scoring rules, all four models passed both the 2019 and 2025 examinations, achieving near-perfect scores in the essential sections. None of the LLMs selected any taboo choices. In terms of overall accuracy (correct question numbers/total question numbers), the LLMs ranked as follows: Gemini 2.5 Pro (97.2%), GPT-5 (96.3%), Claude Opus 4.1 (96.1%), and Grok-4 (95.6%). All exceeded the 95% threshold, which has been proposed as a benchmark for considering LLMs as reliable sources of medical knowledge. Additionally, although the Japanese government does not publicly release the average scores of medical students on the National Medical Licensing Examination, we identified reports from a Japanese exam-preparation institution (Ishinkai) indicating that students' average scores tend to be close to the passing threshold [33]. For example, in 2025, the passing cutoff for the non-essential section was 221 out of 300, whereas the reported average student score was approximately 230 [33]. In contrast, the four LLMs evaluated in this study achieved scores of around 290, suggesting substantially higher performance relative to the reported student average. This suggests that the medical knowledge encoded in these LLMs approaches textbook level, representing a milestone in their practical application as educational tools in medicine.

Notably, the accuracies achieved in this study are the highest reported to date among all evaluations of LLMs for medical licensing examinations worldwide, significantly surpassing the performance of GPT-4o (89.2% accuracy) in previous studies [19]. Furthermore, unlike earlier findings in which GPT-4o significantly outperformed Gemini 1.5 Pro and Claude 3 Opus [19], the present evaluation revealed no significant differences in accuracy among the four cutting-edge LLMs. We believe that these four pragmatic advances best explain this improvement over prior models. Firstly, reasoning enhancements that make models "think before answering"—including CoT, Self-Consistency, and search-style Tree-of-Thoughts—are known to improve performance on multistep problems typical of clinical reasoning and computation. Moreover, "process supervision" trains models to produce correct intermediate steps, not just correct final answers, which aligns well with medical exam questions requiring stepwise justification [20, 34,

35]. Second, the latest models plausibly benefit from broader and more recent corpora with improved coverage of medical content (e.g., guideline-like text and exam-style phrasing) and better Japanese biomedical material, narrowing the historical gap seen in earlier English-skewed systems [36]. Third, no instances of AI hallucinations were observed in any of the LLMs examined. Compared with earlier LLMs (e.g., GPT-4o and Gemini 1.5 pro), current LLMs more often ground their answers and self-check their drafts. Techniques like retrieval-augmented generation supply external evidence, while "chain-of-verification" style decoding has been shown to lower hallucinations by planning and answering verification questions before finalizing a response—both of which are valuable for factual, guideline-consistent medical Q&A [37].

All four LLMs achieved accuracies exceeding 97% for non-image-based questions, with Gemini 2.5 Pro demonstrating near-perfect performance by attaining the highest accuracy of 98.6%. Consistent with previous studies [19, 20, 38, 39], their accuracies on image-based questions were lower than on non-image-based questions. GPT-5, Claude 4.1 Opus, and Grok 4 showed a statistically significant difference between image and non-image bases questions, whereas the Gemini 2.5 Pro did not. However, unlike previous studies, in which LLMs typically achieved less than 80% accuracy on image-based questions [4], the present study found that all four models exceeded 90% accuracy even on image-based questions, with Gemini 2.5 Pro reaching the highest accuracy of 96.1%. These findings suggest that the image interpretation capabilities of the latest generation of LLMs have substantially improved compared with earlier models. Notably, Gemini 2.5 Pro performed exceptionally well, achieving an accuracy above the 95% benchmark even for image-based questions, which supports its potential applicability as a medical imaging diagnostic tool. In contrast, the other three LLMs may pose risks if applied prematurely in medical imaging analyses or clinical diagnostic support.

We categorized all questions into three levels of difficulty (easy, moderate, and difficult) based on human medical students' accuracy and evaluated the LLMs accordingly. All four models demonstrated significantly higher accuracy for easy questions than for difficult questions. However, no significant differences were observed between easy and moderate questions or between moderate and difficult questions. In the easy, moderate, and difficult categories, Gemini 2.5 Pro achieved the highest accuracy across the board. Importantly, except for Grok-4, the remaining three LLMs maintained accuracies above 90% even for difficult questions. These findings suggest that although previous studies have consistently reported that LLMs perform better on easier items [38, 40–44], the performance gap attributable to difficulty narrows as LLM capabilities advance. Collectively, these patterns indicate enhanced stability across difficulty strata, stronger internal consistency in reasoning processes, and greater robustness to item complexity and distributional shifts. Consequently, the classical "difficulty effect" is attenuated, with model accuracies approaching a ceiling and variance across categories being substantially reduced. This enhanced uniformity across difficulty levels suggests that advanced LLMs may serve as more consistent and equitable tools for medical education and assessment, thereby reducing biases introduced by item complexity.

Typically, the performance of LLMs on general questions reflects their accuracy as knowledge sources in medical education, whereas their performance on clinical questions reflects their capabilities in clinical reasoning and diagnostic decision-making. In this study, we found that all four LLMs achieved higher accuracy on general questions than on clinical questions, with the difference reaching statistical significance only for Grok-4. Notably, GPT-5, Grok-4, and Gemini 2.5 Pro achieved 99% accuracy on general questions. Because they generally require factual recall rather than complex reasoning, this near-perfect accuracy further demonstrates that, with training on increasingly large-scale datasets, the latest LLMs have achieved almost textbook-level mastery of fundamental medical knowledge, the latest LLMs have achieved almost textbook-level mastery of fundamental medical knowledge, consistent with findings from prior studies [20]. With appropriate ethical oversight and regulatory safeguards, these models may hold considerable promise as educational tools to support basic medical training. In clinical questions, Gemini 2.5 Pro still achieved the highest accuracy rate of 97.0%, with GPT-5 and Claude 4.1 Opus also exceeding 95%. However, unlike performance on general knowledge items, the Japanese National Medical Examination (JNME), while standardized and transparent, remains an artificial testing environment. The dataset consists of exam-style questions with predefined correct answers and minimal ambiguity, whereas real-world clinical practice rarely offers such clarity. Patients frequently present with multiple comorbidities, incomplete histories, and overlapping imaging findings that require nuanced interpretation. At the same time, LLM capabilities for contextual reasoning, uncertainty management, and longitudinal decision-making required for clinical practice or electronic health record workflows remain insufficient [45, 46]. Therefore, higher accuracy thresholds and greater safety margins are required when evaluating LLM performance on clinical questions, as even small errors in medical education could propagate into clinical practice [21]. By relying solely on licensing exam questions, our study likely overestimates model

performance, as these items emphasize factual recall and pattern recognition rather than situational awareness and clinical judgment. More realistic evaluations—such as handling equivocal imaging findings, prioritizing between competing diagnoses, or managing incidental findings—will be necessary to uncover clinically relevant weaknesses that licensing-style assessments cannot capture.

After reviewing the explanation of incorrect questions. We found the most common cause for incorrect answers was misunderstanding of the number of selected options. As instructed in the beginning of each section, the default number of correct choices was one. However, LLMs chose more than one in some questions, leading to incorrect results. It is critical to retain the instruction contents while analyzing the questions until the end of each session. The second most common cause for incorrect answers was misdiagnosis from clinical images such as CT or misunderstanding of illustrations. Gemini 2.5 pro performed better in answering questions with images. Nevertheless, improvement in comprehending images, especially those images that directly contributes to a final diagnosis, remains a common challenged shared by LLMs. In addition, LLMs lack the common knowledge in perceiving the directions of clinical images: they always got the left and right directions wrong on X-ray images or CT. Third, although LLMs presented with high performance in collecting evidence from publicly available information such as clinical guidelines and textbooks, they seem to have a long way to go when it comes to prioritizing clinical actions. They tend to choose incorrect options or falsely choose multiple options in questions asking "the most appropriate action as the immediate next step".

Finally, we believe that the present findings highlight both the immediate promise and the remaining challenges of applying LLMs in medicine. While the results demonstrate that advanced LLMs have achieved near–textbook-level mastery of medical knowledge, their particularly strong performance in basic medical knowledge highlights substantial potential for use as learning resources in foundational medical education. However, the transition from knowledge recall to real-world clinical reasoning remains a substantial challenge. Clinical decision-making requires not only factual accuracy but also contextual interpretation, prioritization among competing diagnoses, sensitivity to uncertainty, and accountability for patient outcomes. Current LLMs lack these dimensions of professional judgment. Thus, whereas their role in foundational medical training appears increasingly feasible, their translation into frontline diagnostic tools will require further advances in reasoning, reliability, and regulatory oversight.

Limitation

Firstly, although we disabled all internet-search functionalities of the LLMs to prevent reliance on external information, we did not incorporate memorization diagnostics. Therefore, the possibility that some 2019 examination questions were encountered during model pretraining cannot be fully excluded. Importantly, one of the aims of this study was to examine whether publication year of the questions before versus after the cutoff dates would affect LLMs' performance. In this regard, we observed that Grok-4 and Claude Opus 4.1 performed even worse on the 2019 examination than on the 2025 examination. For these reasons, we clarify that the 2025 examination results should be interpreted as the primary findings of this study, whereas the 2019 results are intended as a supplementary comparison to contextualize performance across examination years.

Secondly, regarding statistical methodology, multiple pairwise comparisons were conducted across models, question types, difficulty levels, and examination years. Although p-values were reported using a conventional threshold ($p \leq 0.05$), the risk of Type I error may be inflated. Therefore, some statistically significant findings should be interpreted with caution. Additionally, to ensure reliability, this study evaluated a large sample of 793 questions, which increases statistical power such that very small absolute differences may reach statistical significance. However, such differences may not necessarily reflect meaningful educational or clinical relevance. Future studies should incorporate multiple-testing corrections and effect size measures to provide a more comprehensive interpretation.

Finally, this study exclusively evaluated the performance of LLMs on the JNME, which is written in Japanese. Therefore, these findings may not be generalizable to medical licensing examinations in other countries or languages. However, previous studies indicated that LLMs tend to perform better on examinations written in English than those presented in other languages [4]. Based on this evidence, it is reasonable to hypothesize that the four LLMs tested in the present study might achieve even higher and potentially near-perfect scores on medical examinations delivered in English. In contrast, their performance may be lower in examinations that incorporate traditional medicine domains (e.g., Traditional Chinese Medicine or Korean Traditional Medicine), where relevant training data are likely limited [47]. Therefore, future studies should examine the performance of LLMs across diverse linguistic and cultural contexts to better assess their global applicability in medical education and licensing.

Conclusion

This study evaluated the performance of GPT-5, Grok-4, Claude Opus 4.1, and Gemini 2.5 Pro on the JNME. All four LLMs achieved accuracies exceeding 95%, markedly higher than those reported in previous studies. To the best of our knowledge, this is the first study to demonstrate that LLMs can surpass the accuracy threshold across an entire set of medical licensing examination questions, representing a milestone in their potential application in medical education.

Although image-based questions, clinical questions, and higher-difficulty questions negatively affected the performance of these models, the magnitude of these effects was substantially smaller than that observed in earlier-generation LLMs. This reflects the enhanced stability, internal consistency, and robustness of the latest models in handling complex question formats and reasoning challenges. In this study, the main patterns of errors in LLMs were selecting extra options and failing to correctly identify left and right during X-ray and CT image recognition.

In particular, Gemini 2.5 Pro achieved the highest overall accuracy (97.2%) and consistently maintained a performance above 95%, even in subgroups traditionally disadvantageous for LLMs (e.g., image-based and clinical questions), demonstrating exceptional robustness and reliability. In contrast, Grok-4 showed more pronounced performance gaps between image-based and non-image-based questions and between general and clinical questions.

Abbreviations

LLM	Large language model
JNME	Japanese national medical examination
CT	Computed Tomography
MCQs	Multiple-choice questions
CoT	Chain-of-thought

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-026-03370-y.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Acknowledgements

Not applicable.

Author contributions

Conceptualization: Mingxin Liu, Tsuyoshi Okuhara; Methodology: Mingxin Liu, Tsuyoshi Okuhara, Zhehao Dai, Wenqiang Yin; Formal analysis and investigation: Mingxin Liu, Tsuyoshi Okuhara, Wenqiang Yin, Hiroko Okada; Writing – original draft preparation: Mingxin Liu, Tsuyoshi Okuhara; Writing – review and editing: Mingxin Liu, Tsuyoshi Okuhara, Zhehao Dai, Minghong Zhao, Wenqiang Yin, Hiroko Okada, Emi Furukawa, Takahiro Kiuchi; Funding acquisition: Mingxin Liu; Resources: Takahiro Kiuchi; Supervision: Tsuyoshi Okuhara.

Funding

This work was supported by JSPS KAKENHI Grant [Number 24KJ0830].

Data availability

All Japanese National Medical Examination (JNME) questions used in this study are publicly available from the Ministry of Health, Labour and Welfare of Japan website (https://www.mhlw.go.jp/kouseiroudoushou/shikaku_shiken/ishi/) and from authorized reproduction materials. The large language models (LLMs) evaluated in this study (GPT-5, Claude 4.1 Opus, Gemini 2.5 Pro, Grok-4) were accessed via their publicly available web interfaces.

Declarations

Ethics approval and consent to participate

The ethics approval of this study was waived by the Research Ethics Committee of the Graduate School of Medicine, The University of Tokyo, in accordance with the Ethical Guidelines for Medical and Biological Research Involving Human Subjects (MEXT/MHLW/METI, 2021). The study did not involve human participants, identifiable personal data, or human tissue. The study adhered to the principles of the Declaration of Helsinki. Additionally, informed consent was obtained from both the licensed physician and medical student prior to their participation in the evaluation process.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Health Communication, Graduate School of Medicine, The University of Tokyo, Hongo 7-3-1, Bunkyo, Tokyo, Japan
[2]Department of Health Communication, School of Public Health, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
[3]Department of Cardiovascular Medicine, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
[4]Department of engineering, University of Cambridge, Cambridge, UK
[5]Faculty of Medicine, The University of Tokyo, Tokyo, Japan

Received: 23 October 2025 / Accepted: 30 January 2026
Published online: 03 February 2026

References

1. OpenAI. ChatGPT. https://chat.openai.com/chat. Accessed 8 Aug 2025.
2. Tsang R. Practical Applications of ChatGPT in undergraduate medical education. J Med Educ Curric Dev. 2023;10:23821205231178449. Published 24 May 2023. https://doi.org/10.1177/23821205231178449. PMID: 37255525.
3. Hristidis V, Ruggiano N, Brown EL, Ganta SRR, Stewart S. ChatGPT vs Google for queries related to dementia and other cognitive decline: comparison of results. J Med Internet Res. 2023;25:e48966. Published 25 Jul 2023. https://doi.org/10.2196/48966. PMID: 37490317.
4. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, Kiuchi T. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res. 2024;26:e60807. https://doi.org/10.2196/60807.
5. Liu M, Okuhara T, Huang W, Ogihara A, Nagao HS, Okada H, Kiuchi T. Large Language models in dental licensing examinations: systematic review and Meta-Analysis. Int Dent J. 2025;75(1):213–22. https://doi.org/10.1016/j.identj.2024.10.014.
6. Liu M, Okuhara T, Chang X, Okada H, Kiuchi T. Performance of ChatGPT in medical licensing examinations in countries worldwide: A systematic review and meta-analysis protocol. PLoS ONE. 2024;19(10):e0312771. https://doi.org/10.1371/journal.pone.0312771.
7. Bolgova O, Shypilova I, Mavrych V. Large Language models in biochemistry education: comparative evaluation of performance. JMIR Med Educ. 2025;11:e67244. https://doi.org/10.2196/67244. PMID: 40209205; PMCID: PMC12005600.

8. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: a systematic review and a meta-analysis. BJOG. 2024;131(3):378–80. https://doi.org/10.1111/1471-0528.17641.
9. Alessandri Bonetti M, Giorgino R, Gallo Afflitto G, De Lorenzi F, Egro FM. How does ChatGPT perform on the Italian residency admission National exam compared to 15,869 medical graduates? Ann Biomed Eng. 2024;52(4):745–9. https://doi.org/10.1007/s10439-023-03318-7.
10. Aljindan FK, Al Qurashi AA, Albalawi IAS et al. ChatGPT Conquers the saudi medical licensing exam: exploring the accuracy of artificial intelligence in medical knowledge assessment and implications for modern medical education. Cureus. 2023;15(9):e45043. Published 11 Sep 2023. https://doi.org/10.7759/cureus.45043. PMID: 37829968.
11. Armitage RC. Performance of generative pre-trained transformer-4 (GPT-4) in membership of the royal college of general practitioners (MRCGP)-style examination questions. Postgrad Med J., 2024;100(1182): 274–275. https://doi.org/10.1093/postmj/qgad128. PMID: 38142282.
12. Ebrahimian M, Behnam B, Ghayebi N, Sobhrakhshankhah E. ChatGPT in Iranian medical licensing examination: evaluating the diagnostic accuracy and decision-making capabilities of an AI-based model. BMJ Health Care Inform. 2023;30(1):e100815. Published 11 Dec 2023. https://doi.org/10.1136/bmjhci-2023-100815. PMID: 38081765.
13. Fang C, Wu Y, Fu W et al. How does GPT-4 preform on non-English national medical licensing examination? an evaluation in Chinese language. PLOS Digit Health. 2023;2(12):e0000397. Published 1 Dec 2023. https://doi.org/10.1371/journal.pdig.0000397. PMID: 38039286.
14. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, et al. Performance of ChatGPT on the peruvian national licensing medical examination: cross-sectional study. JMIR Med Educ. 2023;9:e48039. Published 28 Sep 2023. https://doi.org/10.2196/48039. PMID: 37768724.
15. Garabet R, Mackey BP, Cross J, Weingarten M. GPT-4 Performance on USMLE Step 1 Style Questions and Its Implications for Medical Education: A Comparative Study Across Systems and Disciplines. Med Sci Educ. 2023;34(1):145–152. Published 27 Dec 2023. https://doi.org/10.1007/s40670-023-01956-z. PMID: 38510401.
16. Gilson A, Safranek CW, Huang T et al. How Does ChatGPT perform on the united states medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment [published correction appears in JMIR Med Educ. 2024;10:e57594]. JMIR Med Educ. 2023;9:e45312. Published 8 Feb 2023. https://doi.org/10.2196/45312. PMID: 36753318.
17. Gobira M, Nakayama LF, Moreira R, Andrade E, Regatieri CVS, R Belfort Jr. Performance of GPT-4 in answering questions from the Brazilian National examination for medical degree revalidation. Rev Assoc Med Bras (1992). 2023;69(10):e20230848. https://doi.org/10.1590/1806-9282.20230848. Published 2023 Sep 25. PMID: 37792871.
18. Yaneva V, Baldwin P, Jurich DP, Swygert K, Clauser BE. Examining ChatGPT performance on USMLE sample items and implications for assessment. Acad Med. 2024;99(2):192–7. https://doi.org/10.1097/ACM.0000000000005549.
19. Liu, M., Okuhara, T., Dai, Z., Huang, W., Gu, L., Okada, H., … Kiuchi, T. (2025). Evaluating the Effectiveness of advanced large language models in medical Knowledge:A Comparative study using Japanese national medical examination. Int J Med Inform. 193;105673. https://doi.org/10.1016/j.ijmedinf.2024.105673
20. Mavrych V, Yousef EM, Yaqinuddin A, Bolgova O. Large Language models in medical education: a comparative cross-platform evaluation in answering histological questions. Med Educ Online. 2025;30(1):2534065. https://doi.org/10.1080/10872981.2025.2534065. Epub 2025 Jul 12. PMID: 40651009; PMCID: PMC12258195.
21. Bolgova O, Ganguly P, Mavrych V. Comparative analysis of LLMs performance in medical embryology: A cross-platform study of ChatGPT, Claude, Gemini, and Copilot. Anat Sci Educ. 2025;18(7):718–726. https://doi.org/10.1002/ase.70044. Epub 2025 May 11. PMID: 40350555.
22. Wu J, Wang Z, Qin Y. Performance of DeepSeek-R1 and ChatGPT-4o on the Chinese National medical licensing examination: a comparative study. J Med Syst. 2025;49(1):74. https://doi.org/10.1007/s10916-025-02213-z.
23. Gemini 2.5 pro. Google Cloud. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro. Accessed 24 Aug 2025.
24. Grok 4. xAI. https://x.ai/news/grok-4. Accessed August 24, 2025.
25. Claude Opus 4. 1. Anthropic. https://www.anthropic.com/news/claude-opus-4-1. Accessed 24 Aug 2025.
26. Introducing, GPT-5. OpenAI. https://openai.com/ja-JP/index/introducing-gpt-5/. Accessed 24 Aug 2025.
27. Ministry of Health, Labour and Welfare. Japanese National Medical Examination. https://www.mhlw.go.jp/kouseiroudoushou/shikaku_shiken/ishi/. Accessed 8 Aug 2025.
28. Information Available in English. Ministry of Health, Labour and Welfare. https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/topics_150873_139_140.html. Accessed 8 Aug 2025.
29. GPT-5 Chat. OpenAI Platform. https://platform.openai.com/docs/models/gpt-5-chat-latest. Accessed 24 Aug 2025.
30. Models and Pricing. xAI. https://docs.x.ai/docs/models. Accessed August 24, 2025.
31. Gemini. GoogleDeepMind. https://deepmind.google/models/gemini/pro/. Accessed 24 Aug 2025.
32. Models overview. Anthropic. https://docs.anthropic.com/en/docs/about-claude/models/overview. Accessed 24 Aug 2025.
33. Trends in Average Scores for the national medical licensing examination. (Past 5 Years) explanation of passing standards and regional trends. Ishin-Kai. https://ishin-kai.info/column/exam/4706?utm_source=chatgpt.com. Accessed 21 Dec 2025.
34. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Zhou D. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inform Proc Syst. 2022;35:24824–24837.
35. Lightman H, Kosaraju V, Burda Y, Edwards H, Baker B, Lee T, Cobbe K. Let's verify step by step. In the twelfth international conference on learning representations. 2023.
36. Jiang J, Huang J, Aizawa A. JMedBench: a benchmark for evaluating Japanese biomedical large language models. 2024.
37. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Kiela D. Retrieval-augmented generation for knowledge-intensive nlp tasks. Adv Neural Inform Proc Syst. 2020;33:9459–9474.
38. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L et al. Evaluating the efficacy of ChatGPT in navigating the Spanish medical residency entrance examination (MIR): Promising horizons for ai in clinical medicine. Clin Pract. 2023;13(6):1460–1487. Published 20 Nov 2023. https://doi.org/10.3390/clinpract13060130. PMID: 37987431.
39. Nakao T, Miki S, Nakamura Y et al. Capability of GPT-4V(ision) in the Japanese national medical licensing examination: evaluation study. JMIR Med Educ. 2024;10:e54393. Published 12 Mar 2024. https://doi.org/10.2196/54393. PMID: 38470459.
40. Khorshidi H, Mohammadi A, Yousem DM, Abolghasemi J, Ansari G, Mirza-Aghazadeh-Attari M, Ardakani AA. Application of ChatGPT in multilingual medical education: How does ChatGPT fare in 2023's Iranian residency entrance examination. Inform Med Unl. 2023;41:101314. https://doi.org/10.1016/j.imu.2023.101314
41. Knoedler L, Alfertshofer M, Knoedler S, Hoch CC, Funk PF, Cotofana S, Lamby P. Pure wisdom or Potemkin villages? a comparison of ChatGPT 3.5 and ChatGPT 4 on USMLE step 3 style questions: quantitative analysis. JMIR Med Edu. 2024:10(1):e51148. https://doi.org/10.2196/51148
42. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish medical final examination. Sci Rep. 2023;13(1):20512. https://doi.org/10.1038/s41598-023-46995-z.
43. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med Educ. 2023;9(1):e48002. https://doi.org/10.2196/48002.
44. Watari T, Takagi S, Sakaguchi K, Nishizaki Y, Shimizu T, Yamamoto Y, Tokuda Y. Performance comparison of ChatGPT-4 and Japanese medical residents in the general medicine in-training examination: comparison study. JMIR Med Educ. 2023;9:e52202. https://doi.org/10.2196/52202.
45. Du X, Zhou Z, Wang Y, Chuang YW, Li Y, Yang R, Zhang W, Wang X, Chen X, Guan H, Lian J, Hong P, Bates DW, Zhou L. Testing and evaluation of generative large language models in electronic health record applications: a systematic review. medRxiv [Preprint]. Int J Med Inform. 2026;205:106091. https://doi.org/10.1016/j.ijmedinf.2025.106091. PMID: 39228726; PMCID: PMC11370524.
46. Baxter SL, Longhurst CA, Millen M, Sitapati AM, Tai-Seale M. Generative artificial intelligence responses to patient messages in the electronic health record: early lessons learned. JAMIA Open. 2024;7(2):ooae028. https://doi.org/10.1093/jamiaopen/ooae028. PMID: 38601475; PMCID: PMC11006101.

47. Jang D, Yun TR, Lee CY, Kwon YK, Kim CE. GPT-4 can pass the Korean National licensing examination for Korean medicine Doctors. PLOS Digit Health. 2023;2(12):e0000416. https://doi.org/10.1371/journal.pdig.0000416.