Psychiatric Quarterly
https://doi.org/10.1007/s11126-026-10275-6

RESEARCH

Artificial Intelligence in Psychiatry Training: Comparative Insights from Nine Large Language Models Across Cultural and Exam Contexts

Yusuf Selman Çelik[1] · Nagihan Özer[2] · Makbule Esen Öksüzoğlu[3] · Şeyma Selcen Macit[1] · Hande Günal Okumuş[4] · Meryem Kaşak[1] · Ayşegül Efe[1] · Yusuf Öztürk[1]

Received: 26 October 2025 / Accepted: 16 March 2026

Abstract

In recent years, large language models (LLMs) have demonstrated rapidly improving performance in medical knowledge tasks; however, most comparative evaluations have focused on general medical domains, leaving psychiatry—where contextual reasoning and nuance are critical—relatively underexplored. This study systematically compared the performance of nine LLMs on psychiatry-focused medical examination questions to evaluate their accuracy, reliability, and educational utility.

The models tested included ChatGPT-5, ChatGPT-4, Claude Sonnet-4 (free) and Sonnet-4.5 (Pro), Gemini-2.5 Flash and Gemini-2.5 Pro, Grok-3 and Grok-4, and DeepSeek-v3. A total of 100 multiple-choice psychiatry questions were administered, comprising 25 USMLE-type, 25 TUS-type, and 50 expert-authored items. Each model completed five independent testing sessions (4,500 responses total). Statistical analyses assessed overall accuracy, test–retest reliability, and performance differences by exam type.

Results showed significant overall variability among models ($\chi^2(8, N=4{,}500)=42.45$, $p<.001$). Claude Sonnet-4.5 Pro achieved the highest accuracy (94%), followed by Gemini-2.5 Pro (92.8%) and GPT-5 (92.6%). DeepSeek-v3 and GPT-4 demonstrated excellent reliability (ICC>0.90), whereas Gemini-2.5 Flash and Grok-4 exhibited only moderate stability (ICC≈0.65). Question format significantly influenced performance ($F(2,24)=16.19$, $p<.001$, $\eta^2=0.57$): accuracy was lower for USMLE-type items (83.8%) than for TUS-type (95.6%) or expert-authored questions (92.1%). Free and premium models performed comparably on factual tasks, though premium systems showed higher temporal consistency.

These findings indicate that current LLMs can achieve high accuracy on psychiatry-focused, exam-style multiple-choice questions, reflecting strong performance in structured factual knowledge tasks rather than clinical competence. High performance on multiple-choice questions should not be interpreted as equivalence to clinical expertise, which requires integrative reasoning, contextual judgment, and interpersonal skills beyond the scope of standardized examinations. Accordingly, free models may be valuable for foundational learning and examination preparation, while premium systems offer greater consistency for repeated educational assessments, without implying readiness for independent

Extended author information available on the last page of the article

Published online: 20 March 2026

clinical application. Psychiatry, requiring empathy and nuanced reasoning, remains an essential domain for testing AI's progression from factual mastery toward human-centered understanding.

Keywords Artificial intelligence · Psychiatry training · Large language models · Cross-cultural assessment · Exam

Introduction

In recent years, the integration of artificial intelligence (AI) into medical education and clinical practice has accelerated remarkably. During this evolution, large language models (LLMs) have emerged as transformative tools capable of processing and analyzing complex medical information. LLMs are deep learning–based AI systems trained on extensive textual datasets that can generate human-like language, perform logical reasoning, and synthesize knowledge across domains. Contemporary systematic reviews have demonstrated that these models can achieve near-human accuracy on standardized medical examinations and generate clinically coherent explanations across multiple disciplines [1–3]. Beyond information retrieval, LLMs have shown potential to support educational and decision-support contexts, including medical documentation and simulation-based learning [3, 4]. However, variations in architecture, training data, and reasoning mechanisms among major model families—such as ChatGPT, Claude, Gemini, Grok, and DeepSeek—have led to substantial performance heterogeneity, underscoring the need for domain-specific benchmarking studies [5, 6].

Standardized medical examinations such as the United States Medical Licensing Examination (USMLE) and the Turkish Medical Specialization Examination (Tıpta Uzmanlık Sınavı – TUS) serve as rigorous benchmarks for assessing medical knowledge and clinical reasoning skills. Although differing in structure and national context, both examinations rely heavily on multiple-choice questions to assess the ability to integrate diagnostic information, therapeutic principles, and contextual clinical cues, including within psychiatry-related content. As such, psychiatry-focused MCQs from these examinations provide a standardized and comparable framework for benchmarking LLM performance across different educational systems. Previous studies have shown that LLM accuracy declines as question difficulty increases [7, 8], suggesting that advanced cognitive reasoning tasks remain challenging for AI in medical contexts.

Recent evidence indicates that the performance of LLMs is both model- and task-dependent. Certain models have achieved accuracy rates of up to 96% on multiple-choice medical examinations, whereas others demonstrate superior pedagogical quality in generating detailed clinical scenarios [5]. Comparative analyses have also revealed a performance gap of approximately 10% points between ChatGPT-4o and DeepSeek-R1 on USMLE-style questions [9]. Beyond model architecture, linguistic and cultural context appear to influence accuracy. For example, ChatGPT-4 achieved a 91% success rate on the French version of the European Board of Ophthalmology examination [10], but only 73.6% on the Chinese National Medical Licensing Examination [11], underscoring that performance varies considerably across non-English settings. Importantly, emerging evidence suggests that question structure and cognitive demand may exert a stronger influence on LLM performance

than language alone, with item complexity and reasoning depth accounting for a substantial proportion of observed variability [12]. These findings highlight the need to evaluate LLMs not only in terms of accuracy but also with respect to explanatory depth, contextual coherence, linguistic adaptability, and cultural alignment.

Among all medical specialties, psychiatry stands out due to its unique cognitive and affective complexity. Unlike disciplines grounded in laboratory diagnostics or imaging data, psychiatry requires multilayered interpretation of human behavior, cultural context, and emotional nuance. Psychiatric assessment inherently involves empathy, ethical sensitivity, and theory-of-mind reasoning—features that remain particularly difficult for AI systems to emulate [4]. Accordingly, psychiatry-focused multiple-choice questions do not operationalize these capacities themselves; instead, they approximate selected cognitive components relevant to psychiatric knowledge, including diagnostic reasoning, treatment selection, and contextual interpretation within constrained formats. In this context, a study conducted in Türkiye in December 2025 evaluated ChatGPT-4–generated psychiatry MCQs and clinical vignettes across common diagnoses through structured prompting and expert psychiatrist review, demonstrating that such outputs can approximate clinical reasoning within educational settings while also revealing risks of ambiguity and the need for expert oversight [13]. Despite these advances, most comparative studies have focused on general medical or anatomical domains [1, 4, 14], leaving psychiatry markedly underrepresented. Given the field's cognitive and contextual demands, the absence of systematic head-to-head benchmarking studies in psychiatry constitutes an important gap, limiting evidence-informed guidance for the responsible educational use of AI-based tools.

Over the past few years, AI technologies—particularly LLMs—have rapidly advanced, transforming multiple areas of medicine, including psychiatry [4, 15–19]. Early evaluations suggested that LLMs could approximate passing-level performance on standardized medical examinations [20]; however, subsequent studies demonstrated substantial variability across models, examination formats, and linguistic contexts. For example, performance differences have been reported between GPT-3.5 and GPT-4 across national licensing examinations conducted in different languages [21, 22], a pattern further supported by meta-analytic evidence showing consistent but heterogeneous superiority of newer-generation models [23]. Similar variability has also been observed in domain-specific MCQ evaluations, including embryology-focused assessments [24]. Together, these findings underscore pronounced performance heterogeneity and reinforce the need for systematic, domain-specific benchmarking rather than reliance on isolated accuracy reports.

Evaluating LLM performance in psychiatry is not only a theoretical but also a practical concern. For medical students preparing for board examinations and residents in specialty training as well as clinicians using AI tools in educational support and exam preparation contexts, understanding relative model performance is important for educational validity and responsible use. Moreover, socioeconomic disparities and unequal access to premium AI systems mean that many learners depend on free platforms, potentially widening educational gaps. As highlighted in prior research, financial barriers can restrict access to advanced AI tools, contributing to disparities in learning opportunities [15]. Accordingly, comparing both free and paid versions of LLMs is essential to characterize performance differences and to inform discussions of fairness and accessibility in medical training. As AI systems become more widely used in educational settings, the development of standardized

evaluation frameworks remains important to support consistent and equitable use across diverse educational and socioeconomic contexts [2, 16].

Against this background, the present study conducts a systematic comparative analysis of nine distinct AI configurations: ChatGPT-5 (OpenAI), ChatGPT-4(OpenAI), Grok-3 (xAI) and Grok-4 (xAI), DeepSeek R3 (Hangzhou DeepSeek AI Basic Technology Research Co.), Claude Sonnet-4 free version (Anthropic), Claude Sonnet-4.5 professional version (Anthropic), Gemini 2.5-Flash free version (Google), and Gemini-2.5 Pro version (Google). By focusing specifically on psychiatry questions from both the USMLE and TUS examinations, this research aims to evaluate the relative strengths and limitations of current AI models in handling complex psychiatry-focused MCQs. The findings will contribute to the growing body of evidence on AI capabilities in medical domains while providing practical guidance for medical educators, students, and healthcare professionals regarding the optimal utilization of AI tools in psychiatric education and practice.

This study aims to bridge the gap between AI capabilities and psychiatric education through three primary inquiries. First, we establish a benchmark for model accuracy and reliability using a rigorous set of multiple-choice questions. We then expand our analysis to compare the efficacy of free and paid AI versions, alongside a comparison of public and expert-validated question sets. By considering the nuances of different examination systems and languages, such as the USMLE and TUS, we provide a structured framework to guide the educational integration of these models without overstepping into clinical guidance.

Materials and Methods

Study Design

This study compared the capabilities of nine publicly available LLMs in answering psychiatry-focused medical examination questions: ChatGPT-4 (OpenAI) and ChatGPT-5 (OpenAI), Grok-3 and Grok-4 (xAI), DeepSeek-v3 (Hangzhou DeepSeek AI Basic Technology Research Co.), Claude Sonnet-4 free version (Anthropic), Claude Sonnet-4.5 Pro version (Anthropic), Gemini-2.5 Flash free version (Google), and Gemini-2.5 Pro version (Google). The accuracy of these AI models in responding to multiple-choice psychiatry questions from standardized medical examinations was systematically assessed.

Question Selection

The question pool was designed to ensure representation of both internationally standardized and locally relevant medical examination formats. To facilitate interpretation for readers less familiar with these examination formats, the three question categories differed primarily in their cognitive focus and clinical framing. USMLE-type questions typically presented brief clinical vignettes emphasizing diagnostic reasoning and core pathophysiology, often requiring the selection of the most likely diagnosis or next best step. TUS-type questions were generally more concise and knowledge-oriented, focusing on factual recall and guideline-based information commonly tested in national examinations. In contrast, expert-authored questions were designed to be more integrative, incorporating clinical nuance, differential diagnosis, and treatment decision-making reflective of real-world psychiatric practice.

Specifically, 25 USMLE-type questions were directly drawn from the officially released USMLE Sample Questions database, publicly available through the National Board of Medical Examiners (NBME). For the TUS-type questions, items were constructed to closely mirror the content, structure, and difficulty of the psychiatry questions published by the Student Selection and Placement Center (ÖSYM) between 2015 and 2021. To supplement these, an additional 50 expert-generated unpublished questions were developed by the study authors themselves through a consensus process. To reduce potential author-related bias, expert-authored questions were developed without reference to any specific LLM architecture or known model strengths (model-blind authoring). Each expert-authored question was designed to reflect clinical reasoning, psychopathology, and psychopharmacology domains commonly encountered in medical examinations, while ensuring originality and alignment with contemporary psychiatric diagnostic frameworks (DSM-5). All items were reviewed independently by three senior psychiatrists to confirm clarity, content validity, and equivalence in cognitive demand across question categories. A subset of expert-authored questions was additionally pilot-tested on psychiatry residents to confirm clarity and expected difficulty range. In total, 100 psychiatry-focused multiple-choice questions were included (25 USMLE-type, 25 TUS-type, and 50 expert-authored items). USMLE-type, TUS-type, and expert-authored questions are presented in the Supplementary Materials.

Data Collection

Each AI model was required to answer all 100 psychiatry questions during the testing phase. The performance assessment was conducted in October 2025, with each model tested independently. Due to input limitations and conversation length restrictions of different LLM systems, the questions were divided into 4 sets of 25 multiple-choice questions each. Each set was presented with the standardized prompt:

> *"Respond to the following multiple-choice psychiatry questions by choosing the one best answer for each item. For every numbered question, indicate only the letter (A, B, C, D, E, or F) corresponding to your selected option."*

To assess response consistency and reliability, this process was replicated across five independent testing sessions, separated by 24-hour intervals. Each testing round was conducted as a fresh conversation session without access to previous responses, ensuring independence between attempts. For premium versions (GPT-5, Claude Sonnet 4.5 Pro and Gemini 2.5 Pro), identical testing protocols were followed using the respective paid subscription services.

The outputs from each AI model were systematically recorded in a Microsoft Excel (Microsoft 365) spreadsheet and evaluated for accuracy against the established correct answers. A total of 4500 responses (9 AI models × 5 rounds × 100 questions) were collected and analyzed.

Statistical Analyses

All statistical analyses were conducted to evaluate both the accuracy and reliability of LLMs across different medical question types. Descriptive and inferential statistics were used to

compare overall performance levels, examine consistency across repeated sessions, and identify significant differences between models and question formats. Although individual responses were treated as independent observations for categorical accuracy analyses, we acknowledge that responses are inherently nested within models and question sets. This approach was adopted to enable large-scale benchmarking and direct comparison with prior LLM evaluation studies using similar response-level aggregation. Importantly, model-level performance summaries (mean accuracy, confidence intervals, and ICC estimates) were emphasized to mitigate potential inflation of statistical significance arising from dependency structures. Crucially, at the time the study was conducted, the LLM versions evaluated did not retain or reference prior conversation history across sessions; each testing round was executed in a newly initialized, isolated session, thereby minimizing carryover effects and supporting the practical independence of responses.

First, descriptive statistics were computed for each model, including mean accuracy percentages, standard deviations, and 95% confidence intervals (CI). The intraclass correlation coefficient (ICC, two-way mixed model, absolute agreement) was calculated to assess test–retest reliability across five testing sessions, providing an estimate of each model's internal consistency and temporal stability. Next, model-level performance comparisons were performed using chi-square (χ^2) tests of independence based on the number of correct and incorrect responses across 4,500 total items (500 per model). This approach allowed examination of categorical differences in accuracy distributions among models. When the overall χ^2 test was significant, pairwise chi-square analyses were applied to detect specific model-to-model differences, with Bonferroni correction applied to control for Type I error in multiple comparisons. To evaluate differences in accuracy across question types (Expert, TUS, and USMLE), a one-way analysis of variance (ANOVA) was conducted. Prior to the ANOVA, Levene's test assessed the assumption of homogeneity of variances. Given the violation of this assumption, the Games–Howell post hoc test was employed for pairwise comparisons. Effect sizes were reported using eta squared (η^2) for the overall ANOVA and Cohen's d for post hoc contrasts to quantify the magnitude of observed differences. Finally, all analyses were performed using IBM SPSS Statistics version 29 and Python version 3.14, with the level of statistical significance set at $p < .05$ (two-tailed). Effect size thresholds were interpreted according to conventional criteria: $d = 0.2$ small, 0.5 medium, and 0.8 large; $\eta^2 = 0.01$ small, 0.06 medium, 0.14 large. Given the near-ceiling performance observed across several models, effect sizes and confidence intervals were prioritized alongside p-values to contextualize the practical significance of observed differences. Statistical significance was therefore interpreted cautiously, with particular attention to magnitude and overlap of performance estimates rather than sole reliance on null-hypothesis testing. Accordingly, statistical analyses were intended to support comparative benchmarking rather than infer individual-level decision-making validity.

Results

Figure 1 illustrates the distribution of correct answers across different exam formats, highlighting notable performance variations among models. GPT-5 and Gemini-2.5 Pro produced the highest number of correct responses overall, demonstrating consistent performance across expert-level and TUS-type questions. Claude Sonnet-4.5 Pro also performed

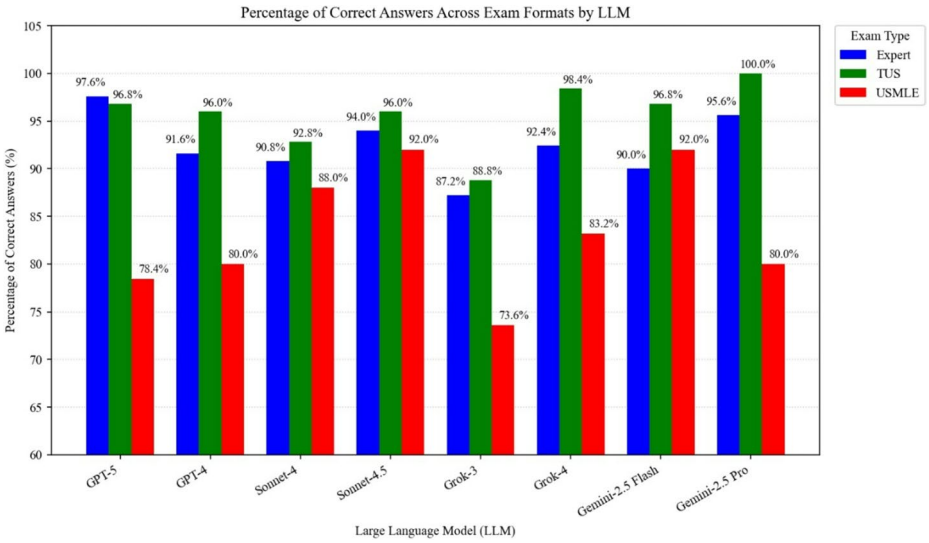

Fig. 1 Comparison of percentage of correct answers across exam formats by LLMs

Table 1 Accuracy comparison of LLMs across different medical exam formats

LLM	Expert (250 Q)		TUS (125 Q)		USMLE (125 Q)		Total (500 Q)
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Total Correct
GPT-5	244	6	121	4	98	27	**463 / 500**
GPT-4	229	21	120	5	100	25	**449 / 500**
Sonnet-4	227	23	116	9	110	15	**453 / 500**
Sonnet-4.5	235	15	120	5	115	10	**470 / 500**
Grok-3	218	32	111	14	92	33	**421 / 500**
Grok-4	231	19	123	2	104	21	**458 / 500**
Gemini-2.5 Flash	225	25	121	4	115	10	**461 / 500**
Gemini-2.5 Pro	239	11	125	0	100	25	**464 / 500**
DeepSeek	225	25	119	6	109	16	**453 / 500**

LLM Large Language Model, *TUS* Medical Specialty Examination in Turkey (Tıpta Uzmanlık Sınavı), *USMLE* United States Medical Licensing Examination. The table displays the number of correct and incorrect responses generated by each model across Expert-level (250 questions), TUS-type (125 questions), and USMLE-type (125 questions) exams. "Total Correct" indicates the cumulative correct answers out of 500 total items

comparably well, whereas Grok-3 yielded the fewest correct answers, particularly in the USMLE-type section.

Across all 4,500 items administered (500 per model), a chi-square test of independence revealed a significant overall difference in accuracy among LLMs ($\chi^2(8, N=4{,}500)=42.45$, $p < .001$). Table 1 provides a summary of the correct and incorrect responses for each model across the three question categories. (Expert-type, TUS-type, and USMLE-type). Claude Sonnet-4.5 Pro achieved the highest total accuracy (470/500; 94%), followed by Gemini-2.5 Pro (464/500; 92.8%) and GPT-5 (463/500; 92.6%). In contrast, Grok-3 exhibited the lowest overall accuracy (421/500; 84.2%), showing marked difficulty with USMLE-type items (92/125 correct). While Claude Sonnet-4.5 Pro and Gemini-2.5 Flash maintained near-

Table 2 Pairwise comparison of LLM accuracy rates using chi-square tests

Model 1	Model 2	χ^2	*p*	Significance (after Bonferroni)
GPT-5	Sonnet-4.5 Pro	0.66	0.416	Non significant
GPT-5	Grok-3	6.12	0.013	Non significant after correction
GPT-5	Gemini-2.5 Pro	0.96	0.327	Non significant
Sonnet-4.5 Pro	Grok-3	11.25	**0.0008**	Significant after correction
Sonnet-4.5 Pro	Gemini-2.5 Flash	0.01	0.933	Non significant
Grok-3	Gemini-2.5 Flash	12.38	**0.0004**	Significant after correction
Gemini-2.5 Flash	DeepSeek-v3	0.83	0.364	Non significant

LLM Large Language Model. The table presents pairwise chi-square (χ^2) comparisons of accuracy rates between models, with corresponding p-values and Bonferroni-adjusted significance results. Comparisons that remained significant after correction are indicated as "Significant after correction."

Table 3 Descriptive and reliability statistics of LLMs across exam types

LLM	Mean Accuracy (%)				Overall SD	ICC	95% CI	
	Expert	TUS	USMLE	Overall			Lower bound	Upper bound
GPT-5 Pro	97.6	96.8	78.4	90.9	10.9	0.826	0.777	0.870
GPT-4 Free	91.6	96.0	80.0	89.2	8.3	0.935	0.914	0.953
Sonnet-4 Free	90.8	92.8	88.0	90.5	2.4	0.814	0.761	0.860
Sonnet-4.5 Pro	94.0	96.0	92.0	94.0	2.0	0.859	0.818	0.895
Grok-3 Free	87.2	88.8	73.6	83.2	8.4	0.844	0.798	0.883
Grok-4 Pro	92.4	98.4	83.2	91.3	7.7	0.664	0.586	0.739
Gemini-2.5 Flash	90.0	96.8	92.0	92.9	3.5	0.655	0.576	0.731
Gemini-2.5 Pro	95.6	100.0	80.0	91.9	10.5	0.718	0.648	0.783
DeepSeek-v3	90.0	95.2	87.2	90.8	4.1	0.919	0.893	0.940

LLM Large Language Model, *TUS* Medical Specialty Examination in Turkey (Tıpta Uzmanlık Sınavı), *USMLE* United States Medical Licensing Examination, *SD* Standard Deviation, *ICC* Intraclass Correlation Coefficient, *CI* Confidence Interval. The table displays mean accuracy percentages for each LLM across three exam types (Expert-level, TUS-type, and USMLE-type), along with overall mean accuracy, variability, and reliability indices. Higher ICC values indicate stronger consistency in model performance across exam types

ceiling accuracy in TUS-style questions, performance variability increased substantially in USMLE-type items across models. It should be noted that near-ceiling accuracy was observed for a subset of high-performing models, particularly on TUS-type items, which may limit the interpretability of fine-grained between-model differences despite statistical significance. Accordingly, non-significant pairwise contrasts among top-performing models should be interpreted as reflecting performance convergence rather than absence of meaningful competence.

Following the overall significant chi-square test, pairwise comparisons were conducted to identify specific differences in accuracy rates between models (Table 2). After Bonferroni correction for multiple testing, Claude Sonnet-4.5 Pro demonstrated significantly higher accuracy than Grok-3 ($\chi^2 = 11.25$, $p = .001$), and Gemini-2.5 Flash also outperformed Grok-3 ($\chi^2 = 12.38$, $p < .001$). All other pairwise comparisons, including contrasts between GPT-5, Claude Sonnet 4.5 Pro, Gemini 2.5 Pro, and DeepSeek-v3, did not reach statistical significance after correction ($p > .05$).

Descriptive statistics and reliability indices for each LLM are presented in Table 3. Mean accuracy rates across all question types ranged from 83.2% (Grok-3) to 94.0% (Claude Son-

net-4.5 Pro). Among the models, Claude Sonnet-4.5 Pro demonstrated the highest overall accuracy (M = 94.0%, SD = 2.0) with excellent internal consistency (ICC = 0.859, 95% CI [0.818, 0.895]). GPT-5 (M = 90.9%, ICC = 0.826) and DeepSeek-v3 (M = 90.8%, ICC = 0.919) also exhibited strong reliability across testing sessions. In contrast, Grok-4 (ICC = 0.664) and Gemini-2.5 Flash (ICC = 0.655) showed only moderate test–retest consistency, indicating greater fluctuation in performance over time. Importantly, variability in ICC values revealed meaningful differences in temporal stability across models, with several systems demonstrating only moderate test–retest reliability despite relatively high mean accuracy (Table 4).

Accuracy by question type revealed systematic differences across domains (see Fig. 2). Most models achieved near-ceiling performance on TUS-type questions, whereas accuracy declined in USMLE-type questions, particularly for GPT-based models (GPT-5: 78.4%, GPT-4: 80.0%). Conversely, Claude Sonnet-4.5 Pro and Gemini-2.5 Flash maintained higher stability and generalization across all categories.

The ANOVA was conducted to evaluate differences in accuracy across question types (Expert, TUS, and USMLE). Descriptive statistics indicated mean accuracies of 92.13% (SD = 3.18) for expert-type, 95.64% (SD = 3.25) for TUS-type, and 83.82% (SD = 6.38) for USMLE-type questions. Levene's test of homogeneity was significant ($F(2,24) = 4.60$, $p = .020$), indicating unequal variances across groups; therefore, the Games–Howell post hoc procedure was employed. The ANOVA revealed a significant main effect of question type on model accuracy, $F(2,24) = 16.19$, $p < .001$, with a large effect size ($\eta^2 = 0.57$). Post hoc comparisons showed that USMLE-type questions yielded significantly lower accuracy than both Expert-type (mean difference = 8.31, $p = .012$, Cohen's $d = 1.66$) and TUS-type questions (mean difference = 11.82, $p = .001$, Cohen's $d = 2.32$). The difference between Expert-type and TUS-type questions was not statistically significant ($p = .082$).

The table presents descriptive statistics and inferential analyses comparing mean accuracy scores across Expert-type, TUS-type, and USMLE-type exams. Levene's test indicated non-homogeneous variances ($p = .020$). A one-way ANOVA revealed a significant main effect of exam type ($F(2,24) = 16.19$, $p < .001$, $\eta^2 = 0.57$). Post hoc Games–Howell comparisons showed that accuracy was significantly higher in the TUS-type exam compared to the USMLE-type exam ($p = .001$, d = 2.32), while the difference between Expert and TUS exam types was not statistically significant.

Table 4 Comparison of mean accuracy scores across exam types

	Variable / Comparison	n	Mean	SD	F / Mean Diff.	df	P	Effect size
Descriptive Statistics	Expert-type exam	9	92.13	3.18	—	—	—	—
	TUS-type exam	9	95.64	3.25	—	—	—	—
	USMLE-type exam	9	83.82	6.38	—	—	—	—
	Total	27	90.53	6.67	—	—	—	—
Levene's Test of Homogeneity	—	—	—	—	**F = 4.60**	(2,24)	**.020**	—
One-Way ANOVA	—	—	—	—	**F = 16.19**	(2,24)	**< .001**	$\eta^2 = .57$
Post Hoc (Games–Howell)	Expert – TUS	—	—	—	–3.51	—	.082	—
	Expert– USMLE	—	—	—	**+8.31**	—	**.012**	—
	TUS – USMLE	—	—	—	**+11.82**	—	**.001**	d = 2.32

SD Standard Deviation, *df* Degrees of Freedom, η^2 Eta Squared, *d* Cohen's d

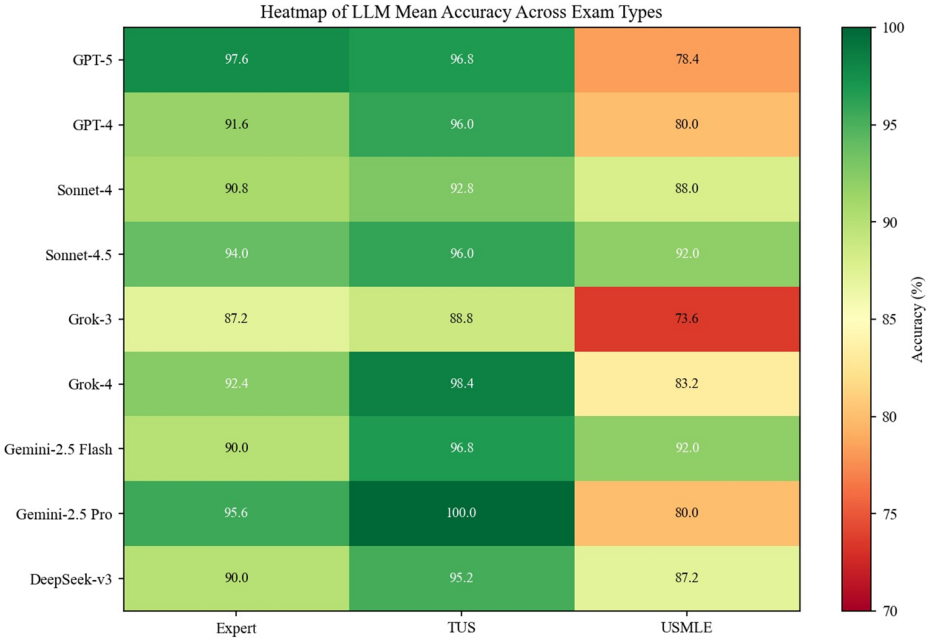

Fig. 2 Heatmap of LLMs mean accuracy across medical exam types

Discussion

Our systematic evaluation of nine LLMs across 4,500 psychiatry examination items yielded four principal findings that delineate the current capabilities and boundaries of artificial intelligence in psychiatric reasoning. First, premium models achieved near-ceiling accuracy (Claude Sonnet-4.5 Pro: 94%; Gemini-2.5 Pro: 92.8%; GPT-5: 92.6%), with differences failing to reach statistical significance after correction. Second, temporal stability varied substantially, ranging from excellent (DeepSeek-v3, GPT-4: ICC > 0.90) to moderate (Gemini-2.5 Flash, Grok-4: ICC ≈ 0.65). Third, question type exerted a significant effect on performance ($F(2, 24) = 16.19$, $p < .001$, $\eta^2 = 0.57$), with USMLE-style items producing markedly lower accuracy (83.8%) than TUS-type (95.6%) or Expert-type (92.1%) questions. Fourth, free models demonstrated accuracy comparable to premium versions on structured factual tasks, suggesting that subscription-based systems may not be essential for foundational educational use. Collectively, these patterns indicate that while current LLMs approximate expert-level performance on structured factual examination items in psychiatry, limitations persist in reasoning depth and response consistency. Importantly, high accuracy on multiple-choice psychiatry questions reflects proficiency in answer selection under constrained conditions and should not be interpreted as evidence of clinical judgment, therapeutic decision-making, or patient-centered reasoning.

When examined within the context of psychiatry, the current landscape of research on LLMs remains notably underdeveloped. Despite the growing number of evaluations across general medical domains, only a limited subset has systematically investigated psychiatric reasoning and diagnostic understanding. Most prior studies have concentrated on ChatGPT-

based systems [25], providing little insight into how newer architectures—such as Claude, Gemini, Grok, and DeepSeek—perform in this cognitively and affectively complex field. Recent large-scale investigations have begun to address this gap. Xu et al. (2025) systematically benchmarked 15 LLMs and reported that DeepSeek-R1-671B (86.83%) and DeepSeek-V3 (84.68%) achieved the highest accuracies, followed by QwQ-32B (84.27%), GLM-Z1 (78.90%), and GPT-4o (72.72%) on mental-health knowledge testing in the Chinese context [26]. Similarly, Hou et al. (2025) demonstrated that GPT-5 attained 94.1% accuracy in the MedQA benchmark—nearly 50% points above the best supervised baseline—highlighting rapid advances in biomedical reasoning and factual retrieval [27]. Consistent with these global findings, the present study revealed that Claude Sonnet-4.5 Pro achieved the highest total accuracy (94%), closely followed by Gemini-2.5 Pro (92.8%) and GPT-5 (92.6%). Although these differences did not reach statistical significance after correction, their convergence at a near-ceiling level suggests that premium-tier models now demonstrate comparable mastery in structured psychiatric question answering. Conversely, Grok-3 showed significantly lower performance than Claude Sonnet-4.5 Pro and Gemini-2.5 Flash, underscoring persistent variability among model families. Together, these data suggest the presence of a potential ceiling effect inherent to multiple-choice–based evaluation frameworks, rather than evidence of saturation in psychiatric reasoning capacity. As accuracy approaches the upper bounds of MCQ performance, future differentiation among models is more likely to depend on evaluation paradigms that probe higher-order reasoning, contextual integration, and explanatory coherence.

Beyond accuracy, our test–retest analyses highlighted notable differences in temporal stability across models. DeepSeek-v3 and GPT-4 demonstrated excellent reliability (ICC > 0.90), whereas Claude Sonnet-4.5 Pro and GPT-5 maintained high consistency (ICC $\approx$ 0.83–0.86). By contrast, Gemini-2.5 Flash and Grok-4 exhibited only moderate stability (ICC ≈ 0.65), indicating fluctuating accuracy across repeated sessions. From a psychometric perspective, such variability has direct practical implications: models with ICC values below 0.70 risk generating inconsistent answers to identical psychiatric vignettes. This inconsistency likely stems from stochastic sampling, temperature parameters, or architecture-specific randomness. The importance of temporal stability is underscored in psychiatry, where diagnostic judgments often integrate information across multiple encounters. Compared with earlier domain-specific analyses—such as Mavrych et al. (2025), who reported mean accuracies around 67% in neuroscience questions [28]—our psychiatry-focused evaluation achieved a mean accuracy approaching 90%, reflecting both technological maturation and broader assessment scope. These results indicate that contemporary premium-tier models combine high accuracy with substantial reliability, reinforcing their potential role as educational adjuncts and as exploratory tools for hypothesis generation, rather than autonomous clinical decision support systems. Future research should examine the sources of instability in low-ICC models, including context-window length, sampling randomness, and fine-tuning methods, to optimize their consistency for high-stakes educational or clinical applications.

A major determinant of performance was question format. Models attained near-ceiling scores on TUS-type questions (>95%), but accuracy declined sharply on USMLE-type items—especially for GPT-based systems (GPT-5: 78.4%; GPT-4: 80.0%). Statistical analyses confirmed a strong main effect of question type ($F(2, 24) = 16.19$, $p < .001$, $\eta^2 = 0.57$). Games–Howell post hoc tests revealed that USMLE-type accuracy was significantly

lower than both Expert-type (p = .012, d = 1.66) and TUS-type (p = .001, d = 2.32) questions, whereas the difference between Expert and TUS items was nonsignificant (p = .082). This difference may be related to features of USMLE-style items, such as longer clinical vignettes and the need for multi-step integration of information. However, as item-level cognitive complexity was not directly measured in the present study, these interpretations should be considered exploratory rather than definitive. In contrast, the concise and monolingual structure of TUS questions favors factual recall, an area where LLMs excel due to extensive pretraining on structured textual corpora.

Previous studies examining linguistic factors have suggested that translation or non-English contexts exert only minimal influence on accuracy [10, 11, 21, 22]. Instead, cognitive and structural properties—such as reasoning depth, distractor plausibility, and contextual ambiguity—better explain performance variability [29, 30]. Our findings align with this interpretation: models performed strongly on short, fact-based TUS-type questions but faltered on reasoning-intensive USMLE items requiring differential diagnosis among plausible alternatives. Thus, linguistic familiarity alone does not appear to account for observed differences; the intrinsic complexity of the task itself remains the dominant determinant of LLM success in psychiatry.

From an educational perspective, the near-expert accuracy achieved in TUS-type and expert-authored questions suggests that LLMs can serve as valuable adjuncts in psychiatric training. They can assist in generating multiple-choice questions, simulating patient scenarios, and supporting formative assessments in residency programs. For clinicians, these systems may facilitate rapid review of diagnostic criteria, psychopharmacologic algorithms, and differential diagnoses. However, their reduced accuracy and lower reliability on complex USMLE-type reasoning tasks warrant caution when deploying them in high-stakes contexts such as board examinations or formal competency evaluations. Moreover, differences in test–retest reliability imply that not all LLMs are equally appropriate for longitudinal educational or research applications.

The present findings also have implications for model accessibility and cost efficiency. Although premium systems (Claude Sonnet-4.5 Pro, GPT-5, Gemini-2.5 Pro) achieved slightly higher mean accuracies than their free counterparts, these differences were not statistically significant after correction. This parallels the report by Jiang et al. (2025), who found that five free Chinese generative AI systems surpassed medical students on molecular biology examinations [31]. Similarly, our data show that free or lightweight models appear sufficient for low-stakes, factual educational tasks such as exam familiarization, concept review, and item generation, but should not be extrapolated to complex clinical reasoning or decision-making contexts. Nevertheless, premium models consistently demonstrated superior reliability and temporal stability. In line with this, Wu et al. (2023) observed that proprietary systems (GPT-4, Claude Sonnet-2) outperformed open-source or free models on nephrology assessments, while a meta-analysis in *Nature Medicine* (2024) confirmed that premium models (e.g., Med-PaLM 2, GPT-4) exceeded open-access alternatives across standardized clinical examinations [32, 33]. These converging findings suggest that while free models suffice for foundational learning and exam preparation, premium LLMs remain preferable for tasks demanding high reproducibility, contextual reasoning, or evaluative rigor. Overall, model selection should remain purpose-driven. Free and publicly accessible versions can effectively support routine study, question generation, and low-stakes educational applications. In contrast, premium systems appear better suited for advanced

reasoning simulations, supervisory feedback, and integration into clinical decision-support frameworks where consistency and interpretive depth are paramount. As LLMs technology continues to evolve, the field must also consider ethical and pedagogical implications—ensuring transparency, minimizing bias, and maintaining clinician oversight in AI-assisted training.

Limitations and Future Directions

Although this study offers one of the earliest large-scale evaluations of nine major LLMs on 4,500 psychiatry-focused exam items, several limitations should be acknowledged. Most importantly, the present evaluation focuses on answer selection accuracy rather than the underlying reasoning process. As such, it does not assess explanation coherence, clinical justification, or ethical judgment—dimensions that are fundamental to psychiatric practice and clinical decision-making. Relatedly, the binary scoring system (correct/incorrect) omits qualitative analysis of explanation quality, reasoning depth, and transparency, which are critical for assessing educational usefulness and clinical plausibility. First, subscription-based access to premium systems (e.g., GPT-5, Claude Sonnet-4.5 Pro, Gemini-2.5 Pro) limits its reproducibility in low-resource settings. Second, reliance on text-based multiple-choice questions may favor factual recall over the nuanced reasoning and contextual understanding required in real clinical decision-making, reducing ecological validity. Third, possible training data overlap between test items and model corpora could inflate accuracy despite efforts to include author-generated, unpublished questions. Fourth, the study's linguistic and cultural scope—largely reflecting Western psychiatry—may restrict generalizability. Future studies should incorporate structured rubric-based evaluations, error taxonomy, or expert ratings to capture these dimensions. Fifth, ethical issues—including algorithmic bias propagation, potential breaches of data privacy, and the reinforcement of psychiatric stigma through oversimplified or context-insensitive outputs—require ongoing scrutiny. Finally, the near-ceiling performance observed in several models may have limited the sensitivity of traditional accuracy-based comparisons. When performance distributions cluster toward the upper boundary, subtle differences among high-performing systems may be difficult to detect using classical test theory metrics alone. Future studies could therefore consider incorporating item-level psychometric approaches, such as item response theory modeling, to enable more refined differentiation through difficulty calibration and latent performance estimation. Addressing these concerns will necessitate interdisciplinary collaboration, transparent reporting standards, and the development of domain-specific ethical frameworks for the use of LLMs in psychiatry.

Future research should move beyond static accuracy benchmarks toward evaluating dynamic reasoning, empathy, and ethical sensitivity in psychiatry-specific contexts. Longitudinal studies incorporating real-world clinical vignettes, multimodal data (e.g., text + speech + emotion cues), and bilingual comparisons could provide a more comprehensive understanding of model performance. Additionally, developing standardized psychometric frameworks for AI evaluation—analogous to inter-rater reliability or construct validity in human testing—will be crucial for establishing these systems as trustworthy educational and clinical tools. As LLMs approach the upper bound of performance observed in expert-authored MCQ settings in factual knowledge, the next frontier lies not in accu-

racy but in judgment, empathy, and contextual intelligence—the very qualities that define psychiatry itself.

Conclusion

This study represents one of the first systematic, head-to-head evaluations of contemporary large language models on psychiatry-focused licensing content, demonstrating overall high performance yet revealing substantial variability across model families and exam formats. While premium systems achieved slightly higher accuracy than free counterparts, most pairwise differences were not statistically significant, and lightweight configurations reached near-expert proficiency on structured factual tasks. Nonetheless, premium models displayed superior temporal stability, suggesting greater suitability for structured educational simulations and supervised analytic tasks, rather than autonomous use in high-stakes clinical decision-making. Importantly, high reproducibility reflects statistical consistency across repeated administrations and should not be conflated with clinical robustness. Clinical robustness entails reliable performance across heterogeneous, context-rich, and unpredictable real-world scenarios, which extend beyond standardized examination formats. Distinguishing these constructs will be essential in future research evaluating the clinical applicability of LLM systems.

The sharp decline in accuracy on USMLE-style items underscores that multi-step reasoning, distractor density, and contextual ambiguity remain unresolved challenges for current architectures. Taken together, these findings support a purpose-driven deployment strategy in which large language models are primarily positioned as educational adjuncts, with future research needed to determine whether and how such systems might safely contribute to clinically relevant reasoning under appropriate human oversight. Future research should move beyond static multiple-choice contexts to retrieval-augmented and multimodal evaluation frameworks, incorporating measures of reasoning coherence, empathy, and explanatory quality. Ultimately, psychiatry provides a critical proving ground for assessing whether artificial intelligence can evolve from factual mastery toward truly human-centered understanding.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11126-026-10275-6.

Authors' Contributions Y.S.Ç. designed the study, coordinated data generation, performed the analyses, interpreted the findings, and drafted the initial version of the manuscript. N.Ö. contributed to data analysis, assisted in developing the methodology, and drafted the initial version of the manuscript. M.E.Ö. performed statistical analyses, contributed to data interpretation, and assisted in the preparation of figures and tables. Ş.S.M. participated in data generation, conducted the literature review, and contributed to the preparation of the method sections. H.G.O. verified the accuracy and integrity of the data, ensured quality control, and contributed to manuscript revisions. M.K. contributed to the development of technical and software components, supported data processing, and assisted in data visualization. A.E. supervised the overall research process, contributed to the conceptual framework, and reviewed all manuscript versions. Y.Ö. supervised the overall research process, contributed to the conceptual framework, and reviewed all manuscript versions. All authors read and approved the final version of the manuscript.

Funding No grant was obtained for this research.

Data Availability The datasets analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Ethics Approval and Consent to Participate This study did not involve human participants, human subjects, patient data, or human-derived materials.

Competing Interests The authors declare no competing interests.

References

1. Sumbal A, Sumbal R, Amir A. Can ChatGPT-3.5 Pass a Medical Exam? A Systematic Review of ChatGPT's Performance in Academic Testing. J Med Educ Curric Dev. 2024;11:23821205241238641. PMID: 38487300; PMCID: PMC10938614.
2. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Med Inf Decis Mak. 2025;25(1):117. https://doi.org/10.1186/s12911-025-02954-4. PMID: 40055694; PMCID: PMC11889796.
3. Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A Review of Large Language Models in Medical Education, Clinical Decision Support, and, Administration H. Healthcare (Basel). 2025;13(6):603. https://doi.org/10.3390/healthcare13060603. PMID: 40150453; PMCID: PMC11942098.
4. Cheng SW, Chang CW, Chang WJ, Wang HW, Liang CS, Kishimoto T, et al. The now and future of ChatGPT and GPT in psychiatry. Psychiatry Clin Neurosci. 2023;77(11):592–6. https://doi.org/10.1111/pcn.13588. Epub 2023 Sep 11. PMID: 37612880; PMCID: PMC10952959.
5. Szabó A, Laein GD. Comparative evaluation of large language models performance in medical education using urinary system histology assessment. Sci Rep. 2025;15(1):31933. https://doi.org/10.1038/s41598-025-17571-4. PMID: 40883524; PMCID: PMC12397380.
6. Kang J, Ahn J. Technologies, opportunities, challenges, and future directions for integrating generative artificial intelligence into medical education: a narrative review. Ewha medical journal. 2025 Oct 2. https://doi.org/10.12771/emj.2025.00787. Epub ahead of print. PMID: 41035210.
7. Alfertshofer M, Knoedler S, Hoch CC, Cotofana S, Panayi AC, Kauke-Navarro M, Tullius SG, Orgill DP, Austen WG Jr, Pomahac B, Knoedler L. Analyzing Question Characteristics Influencing ChatGPT's Performance in 3000 USMLE®-Style Questions. Med Sci Educ. 2024;35(1):257–67. https://doi.org/10.1007/s40670-024-02176-9. PMID: 40144074; PMCID: PMC11933601.
8. Takahashi H, Shikino K, Kondo T, Komori A, Yamada Y, Saita M, Naito T. Educational Utility of Clinical Vignettes Generated in Japanese by ChatGPT-4: Mixed Methods Study. JMIR Med Educ. 2024;10:e59133. https://doi.org/10.2196/59133. PMID: 39137031; PMCID: PMC11350316.
9. Zhang R, Cai Q, Sartori A, Gayed N, Collette H. Comparing Artificial Intelligence Large Language Models in Medical Training: A Performance Analysis of ChatGPT and DeepSeek on United States Medical Licensing Examination (USMLE) Style Questions. Cureus. 2025;17(8):e90212. https://doi.org/10.7759/cureus.90212. PMID: 40970008; PMCID: PMC12442328.
10. Panthier C, Gatinel D. Success of ChatGPT, an AI language model, in taking the French language version of the European Board of Ophthalmology examination: A novel approach to medical knowledge assessment. J Fr Ophtalmol. 2023;46(7):706–11. https://doi.org/10.1016/j.jfo.2023.05.006. Epub 2023 Aug 1. PMID: 37537126.
11. Fang C, Wu Y, Fu W, Ling J, Wang Y, Liu X, Jiang Y, Wu Y, Chen Y, Zhou J, Zhu Z, Yan Z, Yu P, Liu X. How does ChatGPT-4 preform on non-English national medical licensing examination? An evaluation in Chinese language. PLOS Digit Health. 2023;2(12):e0000397. https://doi.org/10.1371/journal.pdig.0000397. PMID: 38039286; PMCID: PMC10691691.
12. Lee J, Park S, Shin J, Cho B. Analyzing evaluation methods for large language models in the medical field: a scoping review. BMC Med Inf Decis Mak. 2024;24(1):366. https://doi.org/10.1186/s12911-024-02709-7. PMID: 39614219; PMCID: PMC11606129.
13. Emekli E, Emekli E, Özel B. Artificial Intelligence-Assisted Generation of Case Scenarios and Multiple-Choice Questions in Psychiatry: A Pilot Study. Acad Psychiatry. 2025 Dec 18. https://doi.org/10.1007/s40596-025-02298-1. Epub ahead of print. PMID: 41413359.

14. Rajan A, Alexander SM, Shenvi CL. Can AI grade like a professor? comparing artificial intelligence and faculty scoring of medical student short-answer clinical reasoning exams. Adv Health Sci Educ Theory Pract. 2025 Aug 6. https://doi.org/10.1007/s10459-025-10462-3. Epub ahead of print. PMID: 40767985.
15. Singal A, Goyal S. Comparative evaluation of AI platforms Google Gemini 2.5 Flash, Google Gemini 2.0 Flash, DeepSeek V3 and ChatGPT 4o in solving multiple-choice questions from different subtopics of anatomy. Surg Radiol Anat. 2025;47:193. https://doi.org/10.1007/s00276-025-03707-8.
16. Lee QY, Chen M, Ong CW, Ho CSH. The role of generative artificial intelligence in psychiatric education- a scoping review. BMC Med Educ. 2025;25(1):438. https://doi.org/10.1186/s12909-025-07026-9. PMID: 40133891; PMCID: PMC11938615.
17. Li H, Zhang R, Lee YC, et al. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. npj Digit Med. 2023;6:236. https://doi.org/10.1038/s41746-023-00979-5.
18. Fiske A, Henningsen P, Buyx A. Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy. J Med Internet Res. 2019;21(5):e13216. https://doi.org/10.2196/13216. PMID: 31094356; PMCID: PMC6532335.
19. Lu K, Sun S, Liu W, Jiang J, Yan Z. Mapping Key Nodes and Global Trends in AI and Large Language Models for Medical Education: A Bibliometric Study. Adv Med Educ Pract. 2025;16:1421–38. PMID: 40832621; PMCID: PMC12360372.
20. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198. PMID: 36812645; PMCID: PMC9931230.
21. Nakao T, Miki S, Nakamura Y, Kikuchi T, Nomura Y, Hanaoka S, et al. Capability of GPT-4V(ision) in the Japanese National Medical Licensing Examination: Evaluation Study. JMIR Med Educ. 2024;10:e54393. https://doi.org/10.2196/54393. PMID: 38470459; PMCID: PMC10966435.
22. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ. 2024;24:143. https://doi.org/10.1186/s12909-024-05125-7.
23. Jaleel A, Aziz U, Farid G, Zahid Bashir M, Mirza TR, Khizar Abbas SM, et al. Evaluating the Potential and Accuracy of ChatGPT-3.5 and 4.0 in Medical Licensing and In-Training Examinations: Systematic Review and Meta-Analysis. JMIR Med Educ. 2025;11:e68070. https://doi.org/10.2196/68070. PMID: 40973108; PMCID: PMC12495368.
24. 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.
25. Zong H, Wu R, Cha J, Wang J, Wu E, Li J, et al. Large Language Models in Worldwide Medical Exams: Platform Development and Comprehensive Analysis. J Med Internet Res. 2024;26:e66114. https://doi.org/10.2196/66114. PMID: 39729356; PMCID: PMC11724220.
26. Xu Y, Fang Z, Lin W, Jiang Y, Jin W, Balaji P, Wang J, Xia T. Evaluation of large language models on mental health: from knowledge test to illness diagnosis. Front Psychiatry. 2025;16:1646974. PMID: 40842952; PMCID: PMC12365771.
27. Hou Y, Zhan Z, Zhang R. Benchmarking GPT-5 for biomedical natural language processing. 2025. arXiv preprint arXiv:2509.04462.
28. Mavrych V, Yaqinuddin A, Bolgova O, Claude. ChatGPT, Copilot, and Gemini performance versus students in different topics of neuroscience. Adv Physiol Educ. 2025;49(2):430–7. https://doi.org/10.1152/advan.00093.2024. Epub 2025 Jan 17. PMID: 39824512.
29. Lee KH, Lee RW. ChatGPT's Accuracy on Magnetic Resonance Imaging Basics: Characteristics and Limitations Depending on the Question Type. Diagnostics (Basel). 2024;14(2):171. https://doi.org/10.3390/diagnostics14020171. PMID: 38248048; PMCID: PMC10814518.
30. Knoedler L, Knoedler S, Hoch CC, Prantl L, Frank K, Soiderer L, Cotofana S, Dorafshar AH, Schenck T, Vollbach F, Sofo G, Alfertshofer M. In-depth analysis of ChatGPT's performance based on specific signaling words and phrases in the question stem of 2377 USMLE step 1 style questions. Sci Rep. 2024;14(1):13553. https://doi.org/10.1038/s41598-024-63997-7. PMID: 38866891; PMCID: PMC11169536.
31. Jiang X. Comparison of the Performance of Five Generative Artificial Intelligence Models on a Medical Molecular Biology Examination. Cureus. 2025;17(7):e88480. https://doi.org/10.7759/cureus.88480. PMID: 40851717; PMCID: PMC12368473.

32. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, Hou L, Clark K, Pfohl SR, Cole-Lewis H, Neal D, Rashid QM, Schaekermann M, Wang A, Dash D, Chen JH, Shah NH, Lachgar S, Mansfield PA, Prakash S, Green B, Dominowska E, Agüera Y, Arcas B, Tomašev N, Liu Y, Wong R, Semturs C, Mahdavi SS, Barral JK, Webster DR, Corrado GS, Matias Y, Azizi S, Karthikesalingam A, Natarajan V. Toward expert-level medical question answering with large language models. Nat Med. 2025;31(3):943–50. https://doi.org/10.1038/s41591-024-03423-7. Epub 2025 Jan 8. PMID: 39779926; PMCID: PMC11922739.
33. Wu S, Koo M, Blum L, et al. Benchmarking open-source large language models, GPT-4 and Claude 2 on multiple-choice questions in nephrology. NEJM AI. 2024;1(2):AIdbp2300092. https://doi.org/10.1056/AIdbp2300092.

Authors and Affiliations

Yusuf Selman Çelik[1] · Nagihan Özer[2] · Makbule Esen Öksüzoğlu[3] · Şeyma Selcen Macit[1] · Hande Günal Okumuş[4] · Meryem Kaşak[1] · Ayşegül Efe[1] · Yusuf Öztürk[1]

✉ Hande Günal Okumuş
drhandegunal@gmail.com

Yusuf Selman Çelik
yusufselmancelik@gmail.com

Nagihan Özer
nagihan.ozerrr@gmail.com

Makbule Esen Öksüzoğlu
mesenoksuzoglu@kastamonu.edu.tr

Şeyma Selcen Macit
selcen.macit@hotmail.com

Meryem Kaşak
meryemkasak90@gmail.com

Ayşegül Efe
aysegulboreas@gmail.com

Yusuf Öztürk
yusuf26es@hotmail.com

[1] Department of Child and Adolescent Psychiatry, Ankara Etlik City Hospital, Ankara, Turkey

[2] Department of Child and Adolescent Psychiatry, Şırnak State Hospital, Şırnak, Turkey

[3] Department of Child and Adolescent Psychiatry, Kastamonu University, Kastamonu, Turkey

[4] Department of Child and Adolescent Psychiatry, Uşak Training and Research Hospital, Uşak, Turkey