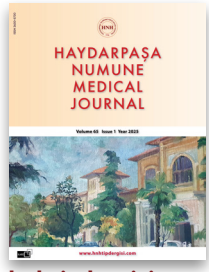
HAYDARPAŞA NUMUNE MEDICAL JOURNAL

DOI: 10.14744/hnhj.2025.48154
Haydarpasa Numune Med J 2025;65(2):165–173



hnhtipdergisi.com

ORIGINAL ARTICLE

Evaluating AI in Psychiatry Board Exams: A Comparative Study of ChatGPT-4 and Google Gemini

İpek Özönder Ünal[1], Miray Pirinççi Aytaç[2]

[1]Department of Psychiatry, Tuzla State Hospital, Istanbul, Türkiye
[2]Department of Psychiatry, Sancaktepe Şehit Prof.Dr. İlhan Varank Training and Research Hospital, Istanbul, Türkiye

Abstract

Introduction: Artificial intelligence (AI) is revolutionizing medical education, with large language models (LLMs) such as ChatGPT-4 (OpenAI) and Google Gemini (Google AI) increasingly used as learning tools. This study examines ChatGPT-4 and Google Gemini's accuracy in answering board-level psychiatry examination questions and classifying question difficulty.

Methods: This cross-sectional study evaluated ChatGPT-4 and Google Gemini using 993 validated board-style psychiatry questions from BoardVitals. AI models were tested using standardized prompts, and their responses were analyzed for accuracy and difficulty classification.

Results: Both ChatGPT-4 and Google Gemini demonstrated high accuracy, significantly surpassing the peer benchmark of 75.95% ($p<0.001$). No statistically significant difference was found between the models in overall accuracy (ChatGPT-4: 90.4%, Google Gemini: 90.8%; $p=0.658$). Both models exhibited only fair agreement with BoardVitals' difficulty categorizations, with ChatGPT-4 ($\kappa w=0.373$) and Gemini ($\kappa w=0.30$) frequently underestimating difficult questions.

Discussion and Conclusion: ChatGPT-4 and Google Gemini show high accuracy in answering psychiatry board-style questions, highlighting their potential as adjunctive tools in medical education. However, their limitations in higher-order reasoning and difficulty classification underscore the need for further refinement. Future research should explore AI integration into real-world clinical decision-making while ensuring human oversight to maintain reliability and ethical considerations.

Keywords: Academic performance; artificial intelligence; psychiatry.

The rapid evolution of artificial intelligence (AI) has introduced transformative possibilities across various domains, including healthcare and medical education.[1] Advanced AI models, such as ChatGPT-4 (OpenAI, San Francisco, USA) and Google Gemini, are increasingly utilized in academic and clinical settings to enhance efficiency, accuracy, and accessibility.[2] These generative models, built on natural language processing (NLP) technologies, are capable of interpreting complex information, providing human-like responses, and adapting to specialized tasks through iterative learning.[3]

In medical education, AI systems have shown significant potential in preparing students and professionals for board examinations by simulating real-world problem-solving scenarios.[4] Recent studies have demonstrated AI's ability to perform comparably to human learners in specialized board-style examinations.[5] For instance, ChatGPT has been assessed for its accuracy and iterative learning in answering neurology board-style questions, achieving accuracy levels akin to resident physicians.[6] Similarly, comparative evaluations of ChatGPT-4 and Google Bard in PMR board exams have highlighted their strengths in

Correspondence: İpek Özönder Ünal, M.D. Department of Psychiatry, Tuzla State Hospital, Istanbul, Türkiye
Phone: +90 536 573 91 94 **E-mail:** ipekozonder@gmail.com
Submitted Date: 03.02.2025 **Revised Date:** 04.03.2025 **Accepted Date:** 11.03.2025
Haydarpaşa Numune Medical Journal

answering discipline-specific questions while emphasizing the importance of oversight in clinical applications.[7]

Psychiatry is a critical medical specialty that focuses on the diagnosis, treatment, and prevention of mental, emotional, and behavioral disorders. As a discipline, it requires a profound understanding of the interplay between biological, psychological, and social factors that influence mental health. Psychiatrists address a wide spectrum of conditions, from mood and anxiety disorders to psychotic and neurodevelopmental conditions, employing a combination of psychopharmacological, psychotherapeutic, and holistic approaches.[8] Given the increasing prevalence of mental health issues globally, the field plays a pivotal role in improving quality of life and advancing public health.[9] Specialization in psychiatry requires not only a strong theoretical foundation but also clinical reasoning, empathy, and interdisciplinary collaboration, making board certification a rigorous process that ensures practitioners meet the highest professional standards.

Despite these advancements, the performance of AI in psychiatry-a field defined by its reliance on nuanced clinical reasoning and multifaceted diagnostic frameworks-remains largely unexplored. Psychiatry board examinations present a unique challenge, requiring comprehensive knowledge of diagnostic criteria, therapeutic strategies, and patient-centered care approaches.[10] As such, understanding how AI models like ChatGPT-4 and Google Gemini perform in this context is essential to evaluating their potential utility as adjunct tools in psychiatric training and assessment.

This study aims to address this gap by evaluating and comparing the performance of ChatGPT-4 and Google Gemini in answering psychiatry board-style questions sourced from the BoardVitals question bank.[11] Specifically, the study assesses the models' accuracy in identifying the correct answers and their ability to classify question difficulty (easy, moderate, or difficult) relative to established benchmarks. By examining these parameters, this research seeks to contribute to the growing body of evidence supporting AI integration into medical education and explore its potential application in psychiatry—a critical and complex medical specialty.[12]

Materials and Methods

Study Design

This study employed a comparative, cross-sectional design to evaluate the performance of two large language models (LLMs)—ChatGPT-4 (OpenAI, San Francisco, CA, USA) and Google Gemini (Google AI, Mountain View, CA, USA)—in answering board-level psychiatry examination questions. These LLMs were specifically selected as they represent the most advanced models currently available, ensuring that the study's findings reflect the cutting edge of AI capabilities in this domain. The primary objectives were to assess the accuracy of the models' responses and their ability to classify question difficulty (easy, moderate, or difficult) relative to a benchmark provided by BoardVitals, a widely recognized online medical question bank.

Data Source and Question Selection

A total of 1,000 psychiatry board-style multiple-choice questions were sourced from BoardVitals between January 10 and January 20, 2025. BoardVitals is a physician-authored question bank designed for medical specialty board certification preparation.[11] Each question consisted of one correct answer among four options and was pre-categorized by BoardVitals into one of three difficulty levels: easy, moderate, or difficult.

To ensure consistency in the evaluation process, seven questions requiring image-based interpretation were excluded from this study, as the focus was solely on text-based question comprehension and response accuracy.

AI Models and Testing Protocol

The AI models evaluated in this study were ChatGPT-4 and Google Gemini. Each model was presented with the same standardized prompt for every question: "The following is a board-level exam question for psychiatrists. Read the question and indicate the level of difficulty as easy, moderate, or difficult, then choose the correct option." For each question, two key parameters were recorded. First, the response accuracy of the AI models was documented, categorizing each answer as either correct or incorrect. Second, the difficulty classification assigned by the AI model was noted, identifying whether the model classified the question as easy, moderate, or difficult. To eliminate potential bias, each AI model's chat interface was cleared between each question before presenting the next item.

BoardVitals classifications served as the gold standard for AI difficulty assessment. To further validate the accuracy and consistency of BoardVitals' difficulty categorizations, two senior psychiatrists, blinded to AI responses and each other's assessments, independently reviewed all 1,000 questions, focusing solely on the assigned difficulty levels (easy, moderate, difficult). Any discrepancies between the psychiatrists' assessments and the BoardVitals classifications were resolved through consensus, ensuring a robust and reliable benchmark for comparison.

Data Collection

For each question, data were systematically recorded to evaluate the performance of the AI models. For each question, response accuracy was recorded (1=correct, 0=incorrect), and AI-assigned difficulty levels (easy, moderate, difficult) were documented. BoardVitals' classifications served as the gold standard for comparison. Performance metrics included overall accuracy (percentage of correct answers) and accuracy stratified by difficulty level. Concordance between AI-assigned and BoardVitals difficulty classifications was analyzed to assess agreement with human benchmarks.

Statistical Analysis

To compare the performance of ChatGPT-4 and Google Gemini, multiple statistical tests were employed. McNemar's test was used to analyze differences in the proportion of correct responses between the two models, providing insight into their comparative accuracy. Additionally, Chi-Square tests were applied to assess the accuracy of difficulty classifications across the three predefined levels (easy, moderate, and difficult) to determine any significant deviations between AI-generated and benchmark difficulty categorizations. To assess the level of agreement between ChatGPT-4 and Google Gemini in difficulty classification, Cohen's Kappa (κ) was used as a statistical measure of inter-rater reliability. A p-value<0.05 was considered statistically significant. Statistical analyses were performed using IBM SPSS Statistics, version 22.0 (IBM Corp., Armonk, NY, USA).

Ethical Considerations

This study involved no human participants or identifiable patient data and relied solely on publicly available, validated BoardVitals questions. AI models were used exclusively for analytical purposes within an educational research framework. The study was conducted in full compliance with the Declaration of Helsinki, upholding principles of data integrity, scientific rigor, and ethical use of AI in medical education. All procedures adhered to established standards for responsible research conduct.

Results

The performance of ChatGPT-4 and Google Gemini was evaluated against a peer performance benchmark derived from the BoardVitals database. This benchmark represents the compiled correct response rates for each individual question based on previous test-takers—psychiatrists preparing for their board exams who have utilized the BoardVitals question bank. The mean accuracy across all questions was 75.95%±19.19% (range: 14–98%). Both AI models significantly outperformed this benchmark, with ChatGPT-4 achieving 90.4% accuracy and Google Gemini achieving 90.8% accuracy (p<0.001 for both), indicating significantly higher success rates.

When directly comparing ChatGPT-4 and Google Gemini, no statistically significant difference was observed. These findings suggest that both models performed at a similarly high level relative to the peer benchmark.

The performance of ChatGPT-4 and Google Gemini was compared across three difficulty levels—easy, moderate, and difficult—as well as overall agreement. The degree of agreement between the two AI models in classifying question difficulty levels relative to the BoardVitals (BV) benchmark was measured using Cohen's Kappa (κ), while McNemar's test was applied to assess statistical differences (Fig. 1).

Figure 1 illustrates the accuracy of ChatGPT-4 and Google Gemini across different difficulty levels. While both models performed well on easy and moderate questions, their accuracy declined significantly for difficult questions, indicating challenges with more complex psychiatry board-style questions.

For easy questions (n=369), the agreement in correct responses between the models was substantial (κ=0.723), indicating high consistency. The McNemar test (p=0.250) showed no statistically significant difference, suggesting both models aligned closely with the BoardVitals (BV) benchmark.

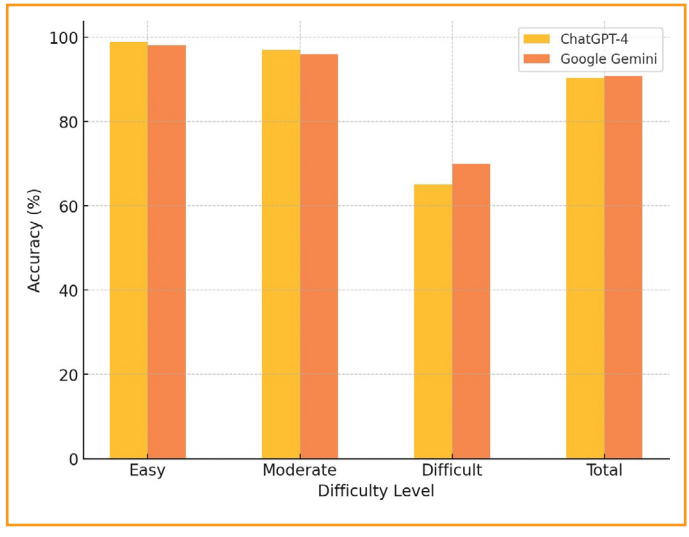


Figure 1. Accuracy Comparison Across Difficulty Levels.

For moderate questions (n=398), the agreement was moderate (κ=0.482), reflecting some variability. However, the McNemar test (p=0.424) showed no significant difference in accuracy between the two models, indicating comparable performance in handling these questions.

For difficult questions (n=226), the models exhibited substantial agreement (κ=0.708), demonstrating strong consistency in accuracy. The McNemar test (p=0.063) suggested a potential trend toward differences in handling the most challenging questions, but overall, both models remained aligned with the BV benchmark.

When analyzing all 993 questions combined, the models showed substantial agreement (κ=0.727), indicating consistent accuracy across difficulty levels. The McNemar test (p=0.658) confirmed no statistically significant difference between ChatGPT-4 and Google Gemini in overall accuracy (Fig. 2).

Figure 2 illustrates the Kappa values representing the agreement between ChatGPT-4 and Google Gemini across different difficulty levels. Higher Kappa values indicate stronger agreement, with the models demonstrating substantial agreement in easy and difficult questions but moderate agreement in moderate questions. The red dashed line marks the threshold for substantial agreement (κ=0.6), highlighting that while overall agreement was strong, moderate-level questions exhibited the most variability in classification between the two models (Table 1).

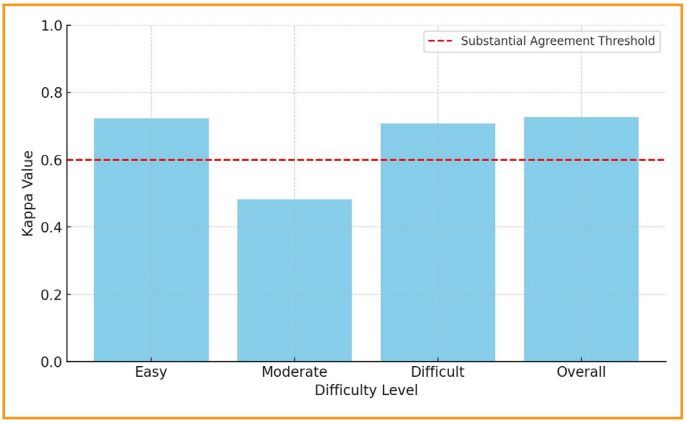


Figure 2. Agreement Across Difficulty Levels.

Performance Across Psychiatry Topics

The analysis of performance across 30 psychiatry topics revealed no statistically significant differences between ChatGPT-4 and Google Gemini, with p-values exceeding 0.05 for all comparisons (Table 2).

Performance Analysis Based on Difficulty Levels

When evaluated against the BV benchmark difficulty classifications, both ChatGPT-4 and Google Gemini demonstrated excellent accuracy for easy and moderate questions. For easy questions, ChatGPT-4 achieved an accuracy of 98.9%, while Google Gemini closely followed with 98.1%. Similarly, for moderate questions, ChatGPT-4 and Google Gemini maintained strong performance, achieving 97.0% and 96.0% accuracy, respectively. These findings indicate that both models performed exceptionally well in handling lower-complexity board-style questions.

However, for difficult questions, a significant drop in accuracy was observed. ChatGPT-4's accuracy declined to 65.0%, while Google Gemini performed slightly better with 69.9%. This decline in performance was statistically significant (p<0.001), confirming that both AI models struggled to maintain high accuracy when faced with complex, higher-order reasoning questions.

In addition to their performance relative to the BV benchmark, each AI model's self-assigned difficulty levels were analyzed in relation to its response correctness. Both models were significantly less accurate when answering questions they classified as difficult themselves.

For ChatGPT-4, questions it classified as easy had an accuracy of 99.0%, while those labeled as moderate had a slightly reduced accuracy of 87.5%. However, when ChatGPT-4 categorized a question as difficult, its accuracy dropped substantially to 62.5% (p<0.001).

Similarly, Google Gemini showed a progressive decline in accuracy as question difficulty increased. When the model classified a question as easy, it answered correctly 97.0% of the time. For moderate questions, accuracy decreased to 89.2%. However, for questions it identified as difficult,

Table 1. Comparison of ChatGPT-4 and Google Gemini Accuracy Across Difficulty Levels

Board Vitals Difficult Level	ChatGPT-4 Correct n (%)	ChatGPT-4 Incorrect n (%)	Google Gemini Correct n (%)	Google Gemini Incorrect n (%)	p
Easy (n=369)	365 (98.9)	4 (1.1)	362 (98.1)	7 (1.9)	0.250
Moderate (n=398)	386 (97.0)	12 (3.0)	382 (96.0)	16 (4.0)	0.424
Difficult (n=226)	147 (65.1)	79 (34.9)	158 (69.9)	68 (30.1)	0.063
Total	898 (90.4)	95 (9.6)	902 (90.8)	91 (9.2)	0.658

Table 2. Performance Comparison of ChatGPT-4 and Google Gemini Across Psychiatry Topics

	Number of Questions	Chat GPT-4 incorrect answer	Google Gemini incorrect answer	Chat GPT 4 score	Google Gemini score
Anxiety Disorders	73	9	8	87.67%	89.04%
Behavioral/Social Sciences and Psychosocial Mechanisms of Disease	20	1	1	95.00%	95.00%
Bipolar and Related Disorders	56	12	10	78.57%	82.14%
Clinical Aspects of Psychiatric and Neuropsychiatric Disorders	153	9	9	94.12%	94.12%
Depressive Disorders	82	7	8	91.46%	90.24%
Developmental Processes and Development Through the Life Cycle	36	1	1	97.22%	97.22%
Diagnostic Procedures	20	1	1	95.00%	95.00%
Disruptive, Impulse-Control, and Conduct Disorders	17	1	1	94.12%	94.12%
Dissociative Disorders	13	1	0	92.31%	100.00%
Eating Disorders	23	1	0	95.65%	100.00%
Elimination Disorders	10	1	0	90.00%	100.00%
Gender Dysphoria	6	0	1	100.00%	83.33%
Interpersonal and Communication Skills	8	0	1	100.00%	87.50%
Neurocognitive Disorders	49	9	9	81.63%	81.63%
Neurodevelopmental Disorders	76	12	12	84.21%	84.21%
Neurologic Disorders	56	13	13	76.79%	76.79%
Neuroscience and Mechanisms of Disease	50	8	8	84.00%	84.00%
Non-Pharmacological Treatments	43	2	2	95.35%	95.35%
Obsessive-Compulsive and Related Disorders	16	2	1	87.50%	93.75%
Other Conditions that may be Focus of Clinical Attention	36	10	10	72.22%	72.22%
Paraphilic Disorders	8	1	2	87.50%	75.00%
Personality Disorders	40	2	2	95.00%	95.00%
Practice-Based Learning and Improvementa	23	0	1	100.00%	95.65%
Professionalism, Ethics, and the Law	35	1	0	97.14%	100.00%
Psychopharmacology	225	25	24	88.89%	89.33%
Psychotherapy	86	1	1	98.84%	98.84%
Schizophrenia Spectrum and Other Psychotic Disorders	69	2	2	97.10%	97.10%
Sexual Dysfunction	5	0	1	100.00%	80.00%
Sleep-Wake Disorders	30	1	0	96.67%	100.00%
Somatic Symptom and Related Disorders	40	5	5	87.50%	87.50%
Substance-Related and Addictive Disorders	56	1	1	98.21%	98.21%
Trauma- and Stressor-Related Disorders	33	1	1	96.97%	96.97%

Please note that some questions cover more than one topic, which may influence the total distribution of questions across categories.

accuracy dropped sharply to 39.1%—the lowest observed among all classifications ($p<0.001$).

Table 3 presents the distribution of difficulty classifications assigned by ChatGPT-4 and Google Gemini in comparison to the BoardVitals (BV) benchmark. The number of questions classified as easy, moderate, or difficult by each AI model is displayed for each BV-defined difficulty level.

The agreement between BoardVitals (BV) difficulty classifications and ChatGPT-4's classifications was assessed using Cohen's Kappa ($\kappa=0.308$) and Weighted Kappa ($\kappa w=0.373$), indicating a fair level of agreement. ChatGPT-4 demonstrated higher alignment with BV's classifications

Table 3. Comparison of ChatGPT-4 and Google Gemini in Difficulty Classification Relative to BoardVitals Benchmark

	ChatGPT-4 Difficult Level			Google Gemini Difficult Level		
	Easy, n (%)	Moderate, n (%)	Difficult, n (%)	Easy, n (%)	Moderate, n (%)	Difficult, n (%)
Board Vitals Difficult Level						
Easy (n=369)	214 (60.0)	155 (40.0)	0	226 (61.2)	143 (38.8)	0
Moderate (n=398)	56 (14.1)	338 (84.9)	4 (1.0)	73 (18.3)	318 (79.9)	7 (1.8)
Difficult (n=226)	36 (15.9)	170 (75.2)	20 (8.9)	62 (27.4)	148 (65.5)	16 (7.1)
Total (n=993)	306 (30.8)	663 (66.8)	24 (2.4)	361 (36.4)	609 (61.3)	23 (2.3)

Table 4. Agreement Between ChatGPT-4 and Google Gemini in Self-Assigned Difficulty Classification

	Google Gemini Difficult Level n (%)			
	Easy	Moderate	Difficult	Total
ChatGPT-4 Difficult Level				
Easy	280 (91.5)	26 (8.5)	0	306
Moderate	75 (11.3)	565 (85.2)	23 (3.5)	663
Difficult	6 (25.0)	18 (75.0)	0	24
Total (n=993)	361 (36.4)	609 (61.3)	23 (2.3)	993

The percentages in parentheses indicate the proportion of questions within each row (ChatGPT-4 difficulty classification) that were classified at each corresponding difficulty level by Google Gemini.

for easy and moderate questions but exhibited greater discrepancies in the difficult category. Specifically, 58.0% of BV-classified easy questions were also categorized as easy by ChatGPT-4, while 42.0% were misclassified as moderate. For moderate questions, ChatGPT-4 correctly identified 84.9%, but 14.1% were misclassified as easy and 1.0% as difficult. In the difficult category, ChatGPT-4 correctly classified only 8.8% of questions, while 75.2% were downgraded to moderate and 15.9% to easy, reflecting a tendency to underestimate question difficulty.

The agreement between BV's difficulty classifications and Google Gemini's classifications was assessed using Cohen's Kappa (κ=0.29) and Weighted Kappa (κw=0.30), both indicating fair agreement. While some consistency was observed, notable variation existed in how Gemini classified question difficulty compared to the BV benchmark. Only 7.1% of BV-classified difficult questions were correctly categorized as difficult by Gemini, with the majority (65.5%) being downgraded to moderate and 27.4% to easy. Similarly, a substantial proportion of moderate questions were misclassified as easy. These findings suggest that while Gemini effectively distinguishes easy questions, it struggles with more complex items, frequently underestimating difficulty.

The agreement between ChatGPT-4 and Google Gemini in difficulty classification was substantial, with Cohen's Kappa (κ=0.688). The Weighted Kappa (κw=0.686) further indicated strong alignment when accounting for degrees of disagreement. Despite minor variations, particularly in the moderate category, the high level of agreement suggests that ChatGPT-4 and Gemini classify difficulty levels similarly, reinforcing their reliability in assessing question complexity (Table 4).

Discussion

This study provides a comparative analysis of ChatGPT-4 and Google Gemini in answering psychiatry board-style questions, emphasizing their potential as adjunctive tools in psychiatric education. Both models demonstrated high accuracy, significantly surpassing the 75.95% peer benchmark and exhibiting strong reliability across varying question difficulty levels. While minor variations in accuracy across psychiatric subdomains were observed, they were not statistically significant, reinforcing the robustness of both models. However, both AI models struggled with difficulty classification, showing only fair agreement with BoardVitals and frequently underestimating difficult questions. Although accuracy remained high for easy and moderate questions, performance declined significantly for difficult ones, which were often misclassified as moderate. Despite this, ChatGPT-4 and Google Gemini exhibited substantial internal agreement, indicating consistency in their difficulty assessments, even when misaligned

with the BoardVitals benchmark. These findings highlight the strengths of AI in psychiatry board-style question answering while underscoring the need for improvements in difficulty classification, particularly for complex cases.

These findings align with broader trends in AI applications across medical fields, where large language models excel in text-based assessments. Studies in ophthalmology, rheumatology, pulmonology, neurology, radiology, and rehabilitation medicine support AI's strong performance across specialties.[6,13–17] The high accuracy in psychiatry suggests AI models are particularly effective in text-heavy disciplines requiring nuanced reasoning and conceptual synthesis, reinforcing their potential in fields that prioritize clinical reasoning over factual recall.

The increasing integration of artificial intelligence (AI) into medical education and clinical decision-making has prompted numerous studies assessing the performance of large language models (LLMs) like ChatGPT-4 and Google Gemini on board and certification exams.[18] Across multiple medical disciplines, ChatGPT-4 has consistently outperformed Google Gemini, particularly in text-based multiple-choice questions. However, both models exhibit notable limitations in image-based assessments, highlighting the current deficiencies of multimodal AI in medical diagnostics.[19,20]

For instance, a study assessing AI performance on the Ophthalmic Knowledge Assessment Program (OKAP) examination found that ChatGPT-4 correctly answered 57.14% of text-based questions, significantly outperforming Gemini, which achieved 46.72% ($p<0.018$).[13] Similarly, in rheumatology, ChatGPT-4 achieved an 86.9% accuracy rate, compared to Gemini's 60.2% ($p<0.001$), with statistically significant differences across multiple subspecialties, particularly in basic and clinical science, osteoarthritis, and rheumatoid arthritis.[14] These trends were further corroborated in other disciplines, such as sleep medicine, where ChatGPT-4 achieved a 68.1% accuracy rate, outperforming both Gemini (45.5%) and its predecessor GPT-3.5 (46.8%).[15]

In oral and maxillofacial surgery (OMS), ChatGPT-4 demonstrated a significantly higher accuracy rate (83.69%) than Gemini (66.85%, $p=0.002$).[16] Additionally, ChatGPT-4 exhibited a 98.2% error correction rate upon multiple attempts, compared to Gemini's 70.71%, underscoring its capacity for iterative learning and reliability.

Despite strong performance in text-based exams, ChatGPT-4 and Gemini struggle with image-based medical questions. While this study did not assess image interpretation, prior research in ophthalmology found ChatGPT-4 (39.58%) and Gemini (33.33%) performed poorly on image-based OKAP questions ($p=0.530$).[13] Similarly, in radiology, GPT-4V, Gemini 1.5 Pro, and Claude 3.5 Sonnet failed to leverage visual inputs for improved accuracy, highlighting current AI limitations in integrating diagnostic imaging.[20]

While ChatGPT-4 generally outperformed Gemini, the magnitude of its superiority varied across different medical domains. The model excelled in fields requiring high levels of medical reasoning and textual comprehension, such as rheumatology, neurology, ophthalmology, and oral and maxillofacial surgery.[14,17] However, Gemini demonstrated relative strength in niche areas such as oculoplastics and refractive surgery, suggesting that differences in training datasets and optimization techniques may influence AI performance in specific subspecialties.[13]

Both models struggled in pathology and refractive surgery, likely due to limited access to specialized datasets and the need for deeper domain-specific knowledge.[19] Additionally, in pediatric urology, both AI models performed similarly, with ChatGPT-4 scoring 66.7% and Gemini achieving 68.6%, indicating that in some specialties, neither model holds a definitive advantage.[21]

Despite ChatGPT-4's edge over Google Gemini in most medical fields, key areas for improvement remain. Enhancing multimodal capabilities to interpret MRIs, X-rays, and fundus images is essential for AI-driven diagnostics. Fine-tuning with curated, peer-reviewed medical datasets would improve reliability, especially in pathology, radiology, and surgical subspecialties. Beyond exam-based assessments, real-world clinical testing is necessary to ensure AI recommendations are practically applicable. Addressing response inconsistencies is also critical—future AI models must deliver predictable, accurate, and consistent guidance to become reliable clinical tools.

Unlike other medical fields, psychiatry fundamentally relies on the therapeutic alliance—a collaborative, trusting relationship between clinician and patient built on empathy, mutual respect, and shared understanding.[22] This alliance is essential for effective treatment, providing a safe space for patients to explore their vulnerabilities, develop insights, and work towards recovery.[22] While AI can process vast amounts of medical data and identify patterns, it cannot foster the genuine human connection that forms the foundation of the therapeutic alliance.[23] AI cannot truly empathize with a patient's distress, provide nuanced emotional support, or build the trust necessary for open and honest communication. For instance, treating

major depressive disorder requires not only accurate diagnosis and appropriate interventions but also active listening, validation of the patient's emotional experience, and the cultivation of a strong therapeutic relationship—elements that AI, in its current form, cannot replicate.[24] Over-reliance on AI in psychiatry risks undermining the therapeutic alliance, potentially reducing the practice to mere symptom management and neglecting the essential relational aspects of care.[25] Therefore, AI should be viewed as an adjunct to, not a replacement for, human clinicians. Ethical integration of AI in psychiatry must prioritize patient-centered care, leverage AI's strengths in data analysis and pattern recognition, while preserving and supporting the development of strong therapeutic alliances, which remain the cornerstone of effective psychiatric treatment.[26]

While promising, this study has several limitations. First, reliance on a predefined question bank may not fully capture the complexity of real-world board exams or clinical reasoning. The exclusion of image-based questions limits applicability, particularly in neuroimaging-dependent assessments. Future research should integrate multimodal AI capabilities to analyze both text and visual data for broader clinical relevance. Second, while AI models performed well in lower-order reasoning, their ability to engage in higher-order cognitive tasks—such as differential diagnosis and holistic treatment planning—was not assessed. Psychiatry often requires critical thinking beyond factual recall, and future studies should evaluate AI's role in complex clinical reasoning. Third, the impact of prompt engineering on AI performance warrants further exploration. Refining prompt structures and considering language influences could enhance AI accuracy and applicability across diverse contexts. Finally, AI-human performance comparisons must be contextualized. Unlike clinicians who rely on experience and intuition, AI synthesizes probabilities from vast datasets, achieving high accuracy but lacking cognitive and emotional depth. AI should be viewed as a complementary tool, not a replacement, ensuring its ethical and effective integration into psychiatry.

Conclusion

ChatGPT-4 and Google Gemini exhibited strong performance in answering psychiatry board-style questions, significantly surpassing the peer benchmark and demonstrating consistent accuracy across varying difficulty levels and topics. These findings suggest that AI technologies can effectively process and respond to complex clinical questions in psychiatry, reinforcing their potential role as educational adjuncts. However, both models displayed only fair agreement with BoardVitals' difficulty classifications, often underestimating difficult questions. Despite these discrepancies, ChatGPT-4 and Google Gemini exhibited substantial internal agreement (κw=0.686), suggesting a degree of consistency in their difficulty assessment. Given AI's limitations in handling higher-order reasoning and question complexity, ongoing refinements in model training and multimodal integration are necessary. Future research should explore AI's role in real-world psychiatric education and clinical decision-making while ensuring human oversight to maintain reliability and ethical considerations.

Data Availability: The data supporting the findings of this study are not publicly available but can be obtained from the corresponding author upon reasonable request.

Ethics Committee Approval: This study did not involve human or animal subjects and relied solely on AI-generated responses to publicly available question bank data. Consequently, no formal ethical approval was required. However, the research was conducted in accordance with established principles of research integrity and data confidentiality.

Peer-review: Externally peer-reviewed.

Use of AI for Writing Assistance: Not declared.

Authorship Contributions: Concept – İ.Ö.Ü., M.P.A.; Design – İ.Ö.Ü., M.P.A. ; Supervision – İ.Ö.Ü., M.P.A.; Fundings – İ.Ö.Ü.; Materials – İ.Ö.Ü., M.P.A.; Data collection &/or processing – İ.Ö.Ü., M.P.A.; Analysis and/or interpretation – İ.Ö.Ü., M.P.A.; Literature search – İ.Ö.Ü., M.P.A.; Writing – İ.Ö.Ü., M.P.A.; Critical review – İ.Ö.Ü.

Conflict of Interest: The authors declare that there is no conflict of interest.

Financial Disclosure: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019;6:94–8. [CrossRef]
2. Lund BD, Ting W, Mannuru NR, Nie B, Shimray S, & Wang Z. ChatGPT and a new academic reality: Artificial ıntelligence-written research papers and the ethics of the large language models in scholarly publishing. J Assoc Inf Sci Technol 2023;74:570–81. [CrossRef]
3. Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst 2020;33:1877–901.
4. 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:e0000198. [CrossRef]
5. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, 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. JMIR Med Educ 2023;9:e45312. [CrossRef]
6. Chen TC, Multala E, Kearns P, Delashaw J, Dumont A, Maraganore D, et al. Assessment of ChatGPT's performance on neurology written board examination questions. BMJ Neurol Open 2023;5:e000530. [CrossRef]
7. Botross M, Mohammadi SO, Montgomery K, Crawford C. Performance of Google's artificial intelligence chatbot "Bard" (Now "Gemini") on ophthalmology board exam practice questions. Cureus 2024;16:e57348. [CrossRef]
8. American Psychiatric Association. What is psychiatry? Available at: https://www.psychiatry.org/patients-families/what-is-psychiatry. Accessed Apr 15, 2025.
9. Bhugra D, Liebrenz M, Ventriglio A, Ng R, Javed A, Kar A, et al. World psychiatric association-asian journal of psychiatry commission on public mental Health. Asian J Psychiatr 2024;98:104105. [CrossRef]
10. Johnson T, John NJ, Lang M, Shelton PG. Accrediting graduate medical education in psychiatry: Past, present, and future. Psychiatr Q 2017;88:235–47. [CrossRef]
11. BoardVitals. BoardVitals question bank dashboard. 2025. Available at: https://www.boardvitals.com/dashboard. Accessed Jan 10, 2025.
12. Panch T, Mattie H, Celi LA. The "inconvenient truth" about AI in healthcare. NPJ Digit Med 2019;2:77. [CrossRef]
13. Gill GS, Tsai J, Moxam J, Sanghvi HA, Gupta S. Comparison of gemini advanced and ChatGPT 4.0's performances on the ophthalmology resident ophthalmic knowledge assessment program (OKAP) examination review question banks. Cureus 2024;16:e69612. [CrossRef]
14. Is EE, Menekseoglu AK. Comparative performance of artificial intelligence models in rheumatology board-level questions: Evaluating Google Gemini and ChatGPT-4o. Clin Rheumatol 2024;43:3507–13. [CrossRef]
15. Cheong RCT, Pang KP, Unadkat S, Mcneillis V, Williamson A, Joseph J, et al. Performance of artificial intelligence chatbots in sleep medicine certification board exams: ChatGPT versus Google Bard. Eur Arch Otorhinolaryngol 2024;281:2137–43. [CrossRef]
16. Mahmoud R, Shuster A, Kleinman S, Arbel S, Ianculovici C, Peleg O. Evaluating artificial intelligence chatbots in oral and maxillofacial surgery board exams: Performance and potential. J Oral Maxillofac Surg 2025;83:382–9. [CrossRef]
17. Chen CH, Hsieh KY, Huang KE, Lai HY. Comparing vision-capable models, GPT-4 and Gemini, with GPT-3.5 on Taiwan's pulmonologist exam. Cureus 2024;16:e67641. [CrossRef]
18. Rossettini G, Rodeghiero L, Corradi F, Cook C, Pillastrini P, Turolla A, et al. Comparative accuracy of ChatGPT-4, Microsoft Copilot and Google Gemini in the Italian entrance test for healthcare sciences degrees: A cross-sectional study. BMC Med Educ 2024;24:694. [CrossRef]
19. Irmici G, Cozzi A, Della Pepa G, De Berardinis C, D'Ascoli E, Cellina M, et al. How do large language models answer breast cancer quiz questions? A comparative study of GPT-3.5, GPT-4 and Google Gemini. Radiol Med 2024;129:1463–7. [CrossRef]
20. Sun SH, Chen K, Anavim S, Phillipi M, Yeh L, Huynh K, et al. Large language models with vision on diagnostic radiology board exam style questions. Acad Radiol 2024:S1076-6332(24)00873-0.
21. Azizoğlu M, Klyuev S. A comparative study on the question-answering proficiency of artificial intelligence models in bladder-related conditions: An evaluation of Gemini and ChatGPT 4. o. Med Records 2025;7:201–5. [CrossRef]
22. Opland C, Torrico TJ. Psychotherapy and therapeutic relationship. StatPearls Publishing. Available at: https://www.ncbi.nlm.nih.gov/books/NBK608012/. Accessed Jan 6, 2024.
23. Lopes E, Jain G, Carlbring P, Pareek S. Talking mental health: A battle of wits between humans and AI. J Technol Behav Sci 2024;9:628–38. [CrossRef]
24. Zhang Z, Wang J. Can AI replace psychotherapists? Exploring the future of mental health care. Front Psychiatry 2024;15:1444382. [CrossRef]
25. Babu A, Joseph AP. Artificial intelligence in mental healthcare: Transformative potential vs. the necessity of human interaction. Front Psychol 2024;15:1378904. [CrossRef]
26. Ayhan Y. The impact of artificial intelligence on psychiatry: Benefits and concerns-An essay from a disputed 'author'. Turk Psikiyatri Derg [Article in English, Turkish] 2023;34:65–7. [CrossRef]