INTERNATIONAL JOURNAL of MEDICAL STUDENTS

ORIGINAL RESEARCH

52. Cross-Sectional Descriptive Study of Comparative Accuracy of ChatGPT, Google Gemini, And Microsoft Copilot in Solving NEET PG Medical Entrance Test

Manik Bhise [1], Kale Sachin Sadanand [1], Patil Anuradha Vishwanath [1], Jali Nandita Vivekanand [1],

[1] MGM Institute of Health Science / MGM Medical College and Hospital, Chhatrapati Sambhajinagar (Aurangabad), Maharashtra, India – 431001

Background: Artificial Intelligence (AI) is increasingly applied in healthcare and medical education, with tools capable of assisting in diagnosis, treatment planning, and exam preparation. The NEET-PG is India's national entrance examination for postgraduate medical training, with case vignettes forming a major component of assessment. AI chatbots therefore hold potential as aids in exam preparation. Previous studies have reported variable accuracy of AI tools in medical licensing exams, but head-to-head comparisons across question types, subjects, and platforms are scarce. Given their rapidly growing use by students and educators, establishing the reliability of these tools is critical. This study directly compares three leading AI chatbots.

The objective was to assess and compare the accuracy of ChatGPT-4, Google Gemini, and Microsoft Copilot in solving the NEET-PG 2023 examination and to evaluate their performance across different question types and medical subjects.

Methods: This cross-sectional descriptive study evaluated the performance of three AI chatbots using a validated set of 200 NEET-PG 2023 questions sourced from PrepLadder and verified against standard textbooks. These questions were presented verbatim to ChatGPT-4, Google Gemini, and Microsoft Copilot. Each chatbot received the questions independently in separate sessions to minimize memory bias. Responses were recorded as correct or incorrect using the validated answer key, and accuracy was expressed as the percentage of correct responses. Comparative analysis was performed for overall accuracy, subject distribution, and question type (recall, analytical, image-based, and case-based). Differences were assessed using the chi-square test with $p < 0.05$ considered statistically significant.

Results: Microsoft Copilot achieved the highest overall accuracy with 165/200 correct responses (82.5%), followed by ChatGPT-4 with 161/200 (80.5%) and Google Gemini with 155/200 (77.5%). The difference in overall performance was not statistically significant ($\chi^2 = 1.6$, $p = 0.4$). All three chatbots achieved 100% accuracy in Microbiology, Anesthesia, and Psychiatry, whereas lower accuracy occurred in Community Medicine, Forensic Medicine, Internal Medicine, and Radiology. No significant variation was found across subjects ($\chi^2 = 2.7$, $p = 0.9$). By question type, recall-based items showed the highest accuracy (85.5%), followed by case-based (82.4%) and analytical (77.3%), while image-based questions were the most challenging (mean accuracy 71.0%). Although Copilot performed slightly better on recall and image-based items, the differences across the three chatbots for question type were not statistically significant ($\chi^2 = 0.35$, $p = 0.9$). These findings highlight variability by subject and question format but no significant difference among the three tools.

Conclusion: All three AI chatbots demonstrated good accuracy in solving NEET-PG questions, performing better in recall-based subjects and less well with image-based items, reflecting current limitations in multimodal applications. They can complement exam preparation by serving as an accessible and interactive platform, offering an affordable alternative to expensive coaching. In healthcare, AI chatbots hold potential for assisting with diagnosis, treatment planning, triage, and referral, particularly in resource-limited settings. However, concerns regarding data privacy, patient confidentiality, lack of empathy, and erosion of clinical decision-making limit their broader adoption. Future research should evaluate evolving versions of these models, larger exam datasets, and integration into structured educational frameworks.

Table 1. Performance by Subject.

Subject	Total Questions	ChatGPT	Gemini	Copilot
Anatomy	9 (4.5%)	5 (55.6%)	5 (55.6%)	6 (66.7%)
Biochemistry	14 (7%)	14 (100.0%)	13 (92.9%)	13 (92.9%)
Physiology	8 (4%)	6 (75.0%)	6 (75.0%)	6 (75.0%)
Pathology	16 (8%)	13 (81.2%)	14 (87.5%)	15 (93.8%)
Microbiology	10 (5%)	10 (100.0%)	10 (100.0%)	10 (100.0%)
Pharmacology	12 (6%)	9 (75.0%)	10 (83.3%)	11 (91.7%)
Community Medicine	15 (7.5%)	11 (73.3%)	10 (66.7%)	11 (73.3%)
Forensic Medicine and Toxicology	8 (4%)	4 (50.0%)	3 (37.5%)	7 (87.5%)
Ophthalmology	8 (4%)	7 (87.5%)	7 (87.5%)	7 (87.5%)
ENT	6 (3%)	5 (83.3%)	5 (83.3%)	5 (83.3%)
Internal Medicine	17 (8.5%)	12 (70.6%)	12 (70.6%)	12 (70.6%)
Surgery	27 (13.5%)	23 (85.2%)	22 (81.5%)	20 (74.1%)
Pediatrics	10 (5%)	8 (80.0%)	8 (80.0%)	8 (80.0%)
Obstetrics and Gynecology	18 (9%)	15 (83.3%)	13 (72.2%)	14 (77.8%)
Radiology	4 (2%)	2 (50.0%)	2 (50.0%)	2 (50.0%)
Orthopedics	6 (3%)	5 (83.3%)	4 (66.7%)	6 (100.0%)
Anesthesia	3 (1.5%)	3 (100.0%)	3 (100.0%)	3 (100.0%)
Dermatology	4 (2%)	4 (100.0%)	3 (75.0%)	4 (100.0%)
Psychiatry	5 (2.5%)	5 (100.0%)	5 (100.0%)	5 (100.0%)
Total	200 (100%)	161 (80.5%)	155 (77.5%)	165 (82.5%)

Legend: Significance, Chi Square 2.7, p=0.9

ISSN 2076-6327

This journal is published by Pitt Open Library Publishing