OP-050 https://doi.org/10.5080/61upk.ozt469

EVALUATION OF CHATGPT, GEMINI, MICROSOFT COPILOT, AND DEEP SEEK'S PERFORMANCE IN THE PSYCHIATRY OFFICIAL BOARD EXAMS

Fatma Sevil Gökhan[1], **Esra Kabadayi Sahin**[2]

[1]*Psychiatry, Yerköy Şehit Korgeneral Osman Erbaş State Hospital, Yozgat, Türkiye*
[2]*Psychiatry, Ankara Yıldırım Beyazıt University, Ankara, Türkiye*

BACKGROUND AND AIM: Large language models(LLMs) using deep learning have advanced significantly in processing language. Potential applications of LLMs in medicine include research, education, and practice, especially as decision aids.The aim of this study was to investigate and compare the performance of LLMs and psychiatrists on board exam questions in various psychiatry association.

METHODS: (Ethics Committee Approval must be obtained and the number should be specified). 175 multiple-choice board questions were collected from websites of the Turkish Psychiatric Association, American Board of Psychiatry and Neurology, European Psychiatric Association and Royal College of Psychiatrists. Ethics committee approval was waived, as it used only publicly available online questions and did not involve human participants. Questions were input into LLMs with input prompt as follows: "As a highly experienced professor of psychiatry, you assist in psychiatry questions. Your role is determine the correct answer". Two psychiatrists also answered the questions without knowing the answers of LLMs. Results were analyzed using descriptive statistics and chi-square test.

RESULTS: ChatGPT had an accuracy rate of 79.4%, Gemini 82.3%, Copilot 85.1% and DeepSeek 85.7%. Psychiatrists with 5 and 12 years of experience have accuracy rate of 88.6% and 90.1%.Only one of the 8 questions that no LLMs could solve was answered incorrectly by both psychiatrists.Of the 8 questions answered incorrectly by both psychiatrists, 7 were answered incorrectly by all LLMs.Only one of the questions was answered correctly by neither psychiatrists nor LLMs.No statistically significant difference was observed comparing the accuracy for Turkish and English questions ($p>0.05$).There was no statistically significant difference in performance of the large language models in accuracy assessment according to whether questions contained cases or no ($p>0.05$).

CONCLUSIONS: LLMs performed very close to psychiatrists both short and case questions in different languages. It has also shown in literature LLMs perform close to physicians in multiple-choice-exams of medical fields. However, their performance in image and open-ended questions have shown varying. Future research is needed to demonstrate their potential impact on patients.

Keywords: Large language models, board exam, artificial intelligence, deep learning