BJPsych | The British Journal of Psychiatry (2025)
226, 243–244. doi: 10.1192/bjp.2024.207

Letter

Promises and pitfalls of large language models in psychiatric diagnosis and knowledge tasks

Chang-Bae Bang, Young-Chul Jung, Seng Chan You, Kyungsang Kim and Byung-Hoon Kim†

Keywords
Large language model; Machine learning methods; Computational psychiatry; Diagnosis and classification; Health informatics.

Large language models (LLMs) are neural networks that use billions of parameters pre-trained on large-scale language corpora.[1] Emergent capabilities are thought to arise within LLMs as the scale of training parameters and data increases.[2] One notable example is the zero-shot reasoning capability of LLMs, which refers to their ability to perform specific tasks based on text instructions without any task examples or extra fine-tuning. Although LLMs are known to encode clinical knowledge comparable with that of human clinicians even without additional training,[3–5] the extent to which recent LLMs can be safely adopted in the field of psychiatry remains underexplored.

In February 2024, we evaluated five LLMs (GPT-4, LLaMA2-70B, Mixtral-45B, Vicuna-13B and Gemma-7B) using a zero-shot approach, each repeated five times (Supplementary Table 1). These models were tasked with diagnosing 21 clinical cases and answering 95 multiple-choice questions drawn from the DSM-5-TR® Clinical Cases[6] and DSM-5-TR® Self-Exam Questions.[7] The performance of the LLMs was compared with that of 11 psychiatry residents from a tertiary hospital. Residents were then retested on the questions where their initial answers differed from those of the best-performing LLM (GPT-4), with the LLM's answers provided for reference. All procedures involving human subjects were approved by the Institutional Review Board of Severance Hospital (4-2024-0131).

GPT-4 notably outperformed both psychiatry residents and open-source models in diagnostic and knowledge tasks (Fig 1). For instance, in diagnostic tasks, GPT-4 achieved a mean F1 score of 63.41%, markedly higher than the residents' score of 47.43% ($P = 0.005$). Similarly, in knowledge tasks, GPT-4 demonstrated an accuracy of 85.05%, compared with the residents' accuracy of 62.01% ($P = 0.002$). Notably, when residents received guidance from GPT-4, their performance improved, with mean F1 scores in diagnostic tasks increasing to 60.15% ($P < 0.001$) and accuracy in knowledge tasks rising to 81.63% ($P < 0.001$).

However, GPT-4 was not without flaws. It exhibited a higher rate of 'comorbidity errors' where mutually exclusive diagnoses (e.g. major depressive disorder and bipolar I disorder) were

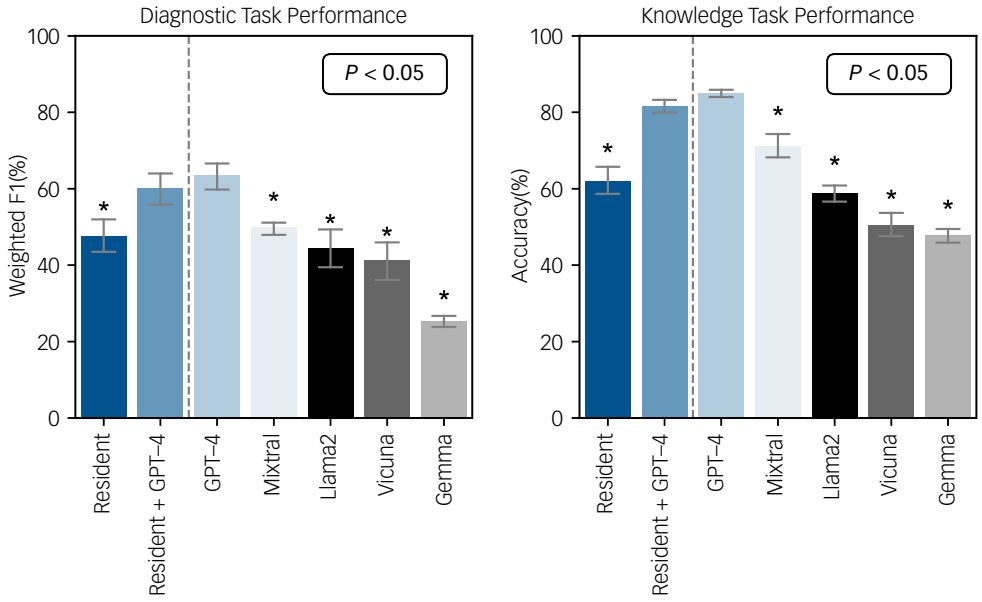

Fig. 1 Bar plots of the task performances using mean scores and error bars with 95% confidence intervals. Dashed vertical lines separate the task performance of the residents (left) and the large language models (right). Asterisks indicate a significant ($P < 0.05$) difference compared with GPT-4 results from the Mann–Whitney U-test.

† This study was conducted while the corresponding author was at Massachusetts General Hospital, Harvard Medical School.

https://doi.org/10.1192/bjp.2024.207 Published online by Cambridge University Press

Table 1 Performance, similarity to GPT-4, and comorbidity error rate of residents before and after GPT-4 guidance

	Diagnostic task						Knowledge task			
	Performance		Similarity to GPT-4		Comorbidity error		Performance		Similarity with GPT-4	
	Weighted F1, mean (s.d.), %	*P*-value[a]	Jaccard index, mean (s.d.)	*P*-value[a]	Error rate, mean (s.d.), %	*P*-value[a]	Accuracy, mean (s.d.), %	*P*-value[a]	Jaccard index, mean, (s.d.)	*P*-value[a]
Resident	47.43 (7.51)	<0.001	0.26 (0.04)	<0.001	0.87 (1.84)	0.441	62.00 (6.20)	<.001	0.46 (0.05)	<0.001
Resident with GPT-4	60.15 (7.73)		0.42 (0.07)		1.73 (4.20)		81.63 (3.14)		0.87 (0.08)	

[a] Calculated using paired *t*-test.

simultaneously presented, compared with the residents (30.48% *v.* 0.87%, $P < 0.001$). The process of making psychiatric diagnoses in humans is influenced by psychiatrists' understanding and belief systems, shaped by art, politics, philosophy, religion, psychotherapies and personal experiences. The high comorbidity error rate of GPT-4 suggests that its diagnoses lack this blend of unique experiences and interpretations, ultimately affecting its diagnostic results. For instance, GPT-4 often provided adjustment disorder as the main diagnosis alongside major depressive disorder, acute stress disorder, post-traumatic stress disorder and prolonged grief disorder (Supplementary Table 2). This phenomenon is consistent with findings of other studies, which have reported that LLMs frequently struggle to understand the context of specific patient cases or grasp subtle differences between individual disorders.[8] Despite this, the comorbidity error rates of the residents did not increase significantly after GPT-4 guidance ($P = 0.4405$; Table 1 and Supplementary Tables 3 and 4), suggesting that GPT-4 can positively influence clinicians' decision-making without increasing critical errors.

The significant increase in similarity between residents' answers and GPT-4's responses after the guidance (mean Jaccard index rising from 0.26 to 0.42 in diagnostic tasks; $P < 0.001$) raises important considerations about the integration of LLMs into psychiatric practice. Prior research has highlighted that dependence on clinical decision-aiding algorithms could impair medical professionals' critical thinking.[9] There is more to consider in psychiatry than the DSM, and uncritical acceptance of these tools may potentially affect patient outcomes. Therefore, careful integration into practice is necessary.

It is important to note that in this study, LLMs were assigned the role of 'expert', which may have obscured the comparison, as the human group consisted of residents still in training. Future research could use a 'psychiatry resident' role for LLMs, compare LLMs with board-certified psychiatrists and use prompts that include the multifaceted knowledge of the DSM, ICD and psychodynamic psychiatry.

As the field of psychiatry continues to evolve, it is crucial to approach the integration of LLMs thoughtfully. Although they offer promising capabilities, maintaining the human element in psychiatric care – including the ability to interpret complex patient narratives and contextualise symptoms within broader life experiences – remains paramount. Balancing technological advances with the irreplaceable aspects of human clinical expertise will be key to leveraging LLMs effectively in psychiatric practice.

Chang-Bae Bang, M.D., Department of Psychiatry, Yonsei University College of Medicine, Seoul, Republic of Korea; and Institute of Behavioral Science in Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea; **Young-Chul Jung**, M.D., Ph.D., Department of Psychiatry, Yonsei University College of Medicine, Seoul, Republic of Korea; and Institute of Behavioral Science in Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea; **Seng Chan You**, M.D., Ph.D., Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea; and Institute for Innovation in Digital Healthcare, Yonsei University, Seoul, Republic of Korea; **Kyungsang Kim**, Ph.D., Center for Advanced Medical Computing and Analysis, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA; **Byung-Hoon Kim**, M.D., Ph.D., Department of Psychiatry, Yonsei University College of Medicine, Seoul, Republic of Korea; Institute of Behavioral Science in Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea; Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea; Institute for Innovation in Digital Healthcare, Yonsei University, Seoul, Republic of Korea; and Center for Advanced Medical Computing and Analysis, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA Email: egyptdj@yonsei.ac.kr

First Received: 9 Jul 2024, revised: 19 Sep 2024, accepted: 23 Sep 2024

Supplementary material

Supplementary material is available online at https://doi.org/10.1192/bjp.2024.207

Funding

This study received no specific grant from any funding agency, commercial or not-for- profit sectors. S.C.Y. serves as the chief technology officer of PHI Digital Healthcare and received grants from Daiichi Sankyo.

Declaration of interest

The authors declare no conflicts of interest.

References

1 Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* 2023; **29**(8): 1930–40.

2 Zoph B, Raffel C, Schuurmans D, Yogatama D, Zhou D, Metzler D, et al. Emergent abilities of large language models. *Trans Mach Learn Res* 2022. https://doi.org/10.48550/arXiv.2206.07682.

3 Katz U, Cohen E, Shachar E, Somer J, Fink A, Morse E, et al. GPT versus resident physicians — a benchmark based on official board scores. *NEJM AI* 2024; **1**(5): AIdbp2300192.

4 Strong E, DiGiammarino A, Weng Y, Kumar A, Hosamani P, Hom J, et al. Chatbot vs medical student performance on free-response clinical reasoning examinations. *JAMA Intern Med* 2023; **183**(9): 1028.

5 Habib S, Butt H, Goldenholz SR, Chang CY, Goldenholz DM. Large language model performance on practice epilepsy board examinations. *JAMA Neurol* 2024; **81**(6): 660–1.

6 Warren Barnhill J. *DSM-5-TR® Clinical Cases. First edition*. American Psychiatric Association Publishing, 2023.

7 Muskin RP, Dickerman LA, Drysdale A, Holderness CC, Maalobeeka G. *DSM-5-TR® Self-Exam Questions. First edition*. American Psychiatric Association Publishing, 2024.

8 Balas M, Wadden JJ, Hébert PC, Mathison E, Warren MD, Seavilleklein V, et al. Exploring the potential utility of AI large language models for medical ethics: an expert panel evaluation of GPT-4. *J Med Ethics* 2024; **50**(2): 90–6.

9 Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit Med* 2021; **4**(1): 31.

EXTRA CONTENT ONLINE

https://doi.org/10.1192/bjp.2024.207 Published online by Cambridge University Press