Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01181-x

Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder

Check for updates

Jiyeong Kim[1,8], Kimberly G. Leonte[2,8], Michael L. Chen[1], John B. Torous[3], Eleni Linos[1,9], Anthony Pinto[4,5,9] & Carolyn I. Rodriguez[6,7,9]

Despite the promising capacity of large language model (LLM)-powered chatbots to diagnose diseases, they have not been tested for obsessive-compulsive disorder (OCD). We assessed the diagnostic accuracy of LLMs in OCD using vignettes and found that LLMs outperformed medical and mental health professionals. This highlights the potential benefit of LLMs in assisting in the timely and accurate diagnosis of OCD, which usually entails a long delay in diagnosis and treatment.

Large language model (LLM)-powered artificial intelligence (AI) chatbots exhibit professional-level knowledge across multiple medical specialty areas[1,2] and have been evaluated for disease detection, treatment suggestions, medical education, and triage assistance[3,4]. Moreover, their ability in advanced clinical reasoning holds promise in assisting in a physician's diagnosis and treatment planning[5–7]. In a statement from the American Psychiatric Association (APA), caution was urged in the use of LLM tools in clinical decision-making[8]; yet a recent survey revealed that many psychiatrists use LLMs in answering clinical questions and documenting notes, and they believe that LLMs would improve diagnostic accuracy in psychiatry[9]. Given the interest and current usage, rigorous study of these tools is urgently needed, especially with respect to awareness of potential bias, compliance with HIPAA, and the use of LLM to augment decision-making.

Obsessive-compulsive disorder (OCD) is a common mental health condition, doing repetitive behaviors (compulsions) to avoid unwanted thoughts or sensations (obsessions), which significantly disrupt the daily lives of a person and the family[10]. It affects approximately 1 in 40 adults in the United States[11], and among adults with OCD, nearly one-half experience serious disability[12]. Unfortunately, there is, on average, a 17-year delay between the onset of symptoms and treatment initiation[13]. Longer duration of untreated illness has a negative impact on the long-term outcome of patients with OCD in terms of inferior treatment response and increased symptom severity[14]. Although there have been efforts to detect mental disorders through social media and online languages, exploration of LLMs in OCD identification has been limited despite their potential contribution to improving diagnostic delay[9,15–17]. Given the patient data safety concerns around deploying LLMs in care, recent studies have employed clinical vignettes for LLM assessments[18,19]. The differing outcomes around the performance of LLMs for mental health studies[17] highlight the need for rigorous study design that can encompass the performance variability of LLMs, including testing multiple LLMs. Given that to our knowledge this method has not been tested in OCD, it is critical to explore the potential of LLM to diagnose OCD. Thus, the goal of this study was to examine the diagnostic accuracy of LLM compared to clinicians and other mental health professionals using clinical vignettes of OCD.

One LLM (ChatGPT-4) correctly ranked OCD as the primary diagnosis in all responded case vignettes (100%, $N = 16/16$ [$n = 4/4$ in 2013 vignettes; $n = 7/7$ in 2015 vignettes; $n = 5/5$ in 2022 vignettes]) (see Table 1). ChatGPT-4 and Gemini Pro both did not provide a response for OCD vignettes with sexual content, noting "content violation"[20]. Notably, the overall LLM diagnostic accuracy (96.1%, $N = 49/51$) was higher than that of medical and mental health professionals in prior studies. The most accurate group of clinicians was doctoral trainees in psychology (81.5%, $N = 106/130$). Moreover, the overall accuracy of LLMs for OCD identification was more than 30% higher than that of primary care physicians (49.5%) and more than 20% higher than that of American Psychological Association members (61.1%). Two other LLMs missed one diagnosis each while they showed higher accuracy than all other mental health providers in this study

[1]Stanford Center for Digital Health, Department of Medicine, Stanford University, Palo Alto, CA, USA. [2]Clearview Horizons, North Andover, MA, USA. [3]Division of Digital Psychiatry, Department of Psychiatry, Beth Israel Deaconess Medical Center, Boston, MA, USA. [4]Department of Psychiatry, Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY, USA. [5]Northwell, New Hyde Park, NY, USA. [6]Department of Psychiatry and Behavioral Sciences, School of Medicine, Stanford University, Palo Alto, CA, USA. [7]Veterans Affairs Palo Alto Health Care System, Palo Alto, CA, USA. [8]These authors contributed equally: Jiyeong Kim, Kimberly G. Leonte.[10]These authors jointly supervised this work: Eleni Linos, Anthony Pinto, Carolyn I. Rodriguez. e-mail: jykim3@stanford.edu; cr2163@stanford.edu

Table 1 | OCD identification rate of LLMs and medical and mental health professionals

Content of OCD vignette	LLM			Group 1. APA members[a] (*N* = 360)	Group 2. Primary care physicians[b] (*N* = 208)	Group 3. Doctoral trainees in psychology[c] (*N* = 130)	Group 4. Medical providers in Guam[d] (*N* = 105)	Group 5. Clergy members in Guam (*N* = 110)
	ChatGPT-4 (*N* = 16)	Gemini Pro (*N* = 16)	Llama 3 (*N* = 19)					
Across all vignettes	96.1% (*N* = 49/51)[e]			61.1%	49.5%	81.5%	41.9%	35.5%
	100% (*n* = 16/16)	93.8% (*n* = 15/16)	94.7% (*n* = 18/19)					
Vignette 1. Harm obsessions	100% (3/3 trials)	100% (3/3 trials)	100% (3/3 trials)	68.5%	20.0%	77.8%	18.2%	27.8%
Vignette 2. Sexual orientation obsessions	100% (3/3 trials)	100% (2/2 trials)[g]	66.7% (2/3 trials)	23.0%	15.4%	66.7%	10.0%	6.7%
Vignette 3. Sexual attraction to children obsessions	No response[f]	100% (1/1 trial)[h]	100% (3/3 trails)	57.1%	29.2%	77.8%	15.0%	11.1%
Vignette 4. Religious obsessions	100% (3/3 trials)	100% (3/3 trials)	100% (3/3 trials)	71.2%	62.5%	80.0%	72.7%	21.7%
Vignette 5. Contamination obsessions	100% (3/3 trials)	100% (3/3 trials)	100% (3/3 trials)	84.2%	67.7%	93.7%	83.3%	73.3%
Vignette 6. Blurting out offensive language obsessions	100% (1/1 trial)	100% (1/1 trial)	100% (1/1 trial)	N/A	26.1%	74.5%	N/A	N/A
Vignette 7. Somatic obsessions	100% (1/1 trial)	0% (0/1 trial)	100% (1/1 trial)	N/A	60.0%	82.4%	N/A	N/A
Vignette 8. Symmetry obsessions	100% (2/2 trials)	100% (2/2 trials)	100% (2/2 trials)	N/A	96.3%	100.0%	85.0%	71.4%

[a] The top five degrees/licenses of the American Psychological Association (APA) members were PhD (67.6%), MA/MS (31.5%), PsyD (14.2%), EdD/EdS/EdM (6.8%), MSW/LMSW (1.7%). Currently licensed was 81.3%.

[b] The areas of specialty included Internal Medicine (35.4%), Pediatrics (32.3%), Family Medicine (22.2%), other specialties (10.6%), and General Medicine (4.5%).

[c] The degrees include Clinical Psychology with Health Emphasis PhD, School-Clinical PsyD, School-Clinical PhD, Clinical Psychology PsyD, and Clinical Psychology PhD in 7 APA-accredited doctoral programs in the Greater New York area.

[d] Group 4 includes medical doctors, nurse practitioners, physician assistants, and doctors of Osteopathic Medicine. The areas of specialty included Internal Medicine (33.3%), Family Medicine (26.7%), Pediatrics (16.2%), Obstetrics and Gynecology (6.7%), Emergency Medicine (4.8%), and other (12.6%).

[e] The sample size differs between LLMs and mental health and health care professionals due to study design (LLM: responses from three LLM models; Human participants: responses from the wide distribution of the vignette studies).

[f] LLM (ChatGPT-4) did not respond to all three vignette trials due to content violation.

[g] LLM (Gemini Pro) did not respond to one vignette trial due to content violation.

[h] LLM (Gemini Pro) did not respond to two vignette trials due to content violation.

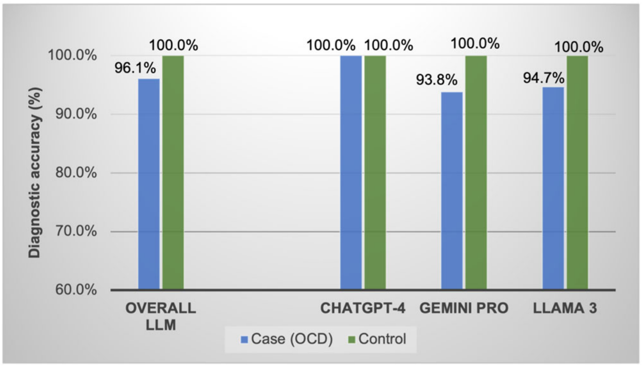

Fig. 1 | Diagnostic accuracy of LLMs by case (OCD) and control (other psychiatric disorders). Overall LLM performance: Case ($N = 49/51$) and control ($N = 21/21$). ChatGPT-4 and Gemini Pro (Case: 16 OCD vignettes) and Llama 3 (Case: 19 OCD vignettes). All LLMs had the same control group comprised of seven psychiatric disorders (major depressive disorder, generalized anxiety disorder, post-traumatic stress disorder, uni or bipolar depression, depression among adolescents, social anxiety disorder, and panic disorder).

(Llama 3: 94.7%, $n = 18/19$; Gemini Pro: 93.8%, $n = 15/16$). Llama 3 could not differentiate sexual orientation obsessions and sexual identity confusion, and Gemini Pro could not differentiate somatic obsessions and body dysmorphic disorder (see Table S3). For control vignettes, the overall diagnostic accuracy of three LLMs was 100.0% ($N = 21/21$), presenting consistency across the LLMs (ChatGPT-4: $n = 7/7$; Gemini Pro: $n = 7/7$; Llama 3: $n = 7/7$) (see Fig. 1). One incorrect reasoning was identified in Llama 3 and Gemini Pro each while ChatGPT-4's reasoning was all correct. Cohen's kappa was 1.0 (perfect agreement) (see Supplementary Table 2).

Our results show that the LLMs, deployed in a zero-shot approach, were consistently more accurate in identifying OCD vignettes compared to mental health and medical professionals, doctoral students, and clergy members. ChatGPT-4's perfect diagnostic performance with accurate clinical reasoning in diagnosing OCD was notable. Results could be improved if we had prompted or finetuned the LLMs, suggesting the even greater potential of the use of LLMs in mental health diagnostic support. There is growing evidence that patients use LLMs for self-diagnosis and medical information and may feel more comfortable disclosing this information to a chatbot than a clinician[21,22]. These findings suggest that LLMs may be a viable tool for increasing the efficiency and accuracy of OCD diagnostic assessment, or at least serving as a valuable and easily accessible screening tool.

To the best of our knowledge, this is the first study that evaluated LLMs' diagnostic performance and clinical reasoning in OCD. Despite the potential benefit of augmenting clinical care with LLMs, further research is needed to weigh ethical issues and risks/benefits of improved diagnostic accuracy[23]. While we observed a slight variation in diagnostic performance by LLM, remarkably, clinical reasoning was comprehensive for all correct cases. Further exploration is needed to understand precisely how and when these tools could be deployed to maximize outcomes within the workflow of clinical practice. An important limitation of our study is that it uses vignettes rather than clinical histories from electronic medical records; any tests of efficacy need to include real-world clinical data to further understand the rates of false positives and negatives and the challenges this may present. Another limitation was the LLMs' inability to provide a response to an OCD vignette that contained specific sexual content, which violated the platform's content rules. Although the LLM correctly identified OCD in the vignettes presented, it is unclear if this accuracy would also be consistent across a larger range of mental health diagnoses or OCD symptom presentations.

The public health benefit of expediting and improving the accuracy of clinicians' ability to detect OCD symptoms is the impact it would have on individuals early in the course of illness, before it severely disrupts school, work, relationships, and quality of life. Further studies are warranted to assess the effectiveness of using LLMs to aid with diagnostic assessments and treatment suggestions for psychiatric disorders within mental health and primary settings. Well-designed clinical studies can generate practical knowledge in the applications of LLMs in psychiatry, which might help in the wise adoption of these technologies in this area of medicine. Given the potential public health benefits of earlier diagnosis and improved care for patients with OCD, this use of LLM merits further testing and consideration.

Methods

We identified five prior studies that assessed providers' ability to identify OCD using clinical vignettes from mental health providers, medical doctors, psychology doctoral students, and clergy members (see Supplementary Table 1). We applied a case-control design to mitigate undetected potential biases—including selection bias—varying the sources of a control group for LLM performance. We used a 1:1 ratio of clinical vignettes of OCD as a case and other psychiatric disorders as a control group. We tested a total of 19 OCD vignettes, presenting 8 different OCD types, from five original previously published studies[24–28]. Seven control vignettes included major depressive disorder, generalized anxiety disorder, post-traumatic stress disorder, uni or bipolar depression, depression among adolescents, social anxiety disorder, and panic disorder from seven validated sources. For control vignettes, a list of suggested options for disorders was not provided except for one vignette, as we intended to adapt the validated vignettes (see Supplementary Table 2).

We assessed the performance of three different AI chatbots to reduce any potential biases that may come from a specific LLM[29]. Three AI chatbots included ChatGPT-4 (Open AI, Inc., April 2023 version), Gemini Pro (Google Inc., April 2024 version), and Llama 3 (Meta Inc., April 2024 version). We used a zero-shot approach in that we did not prompt or finetune the chatbots with the goal of improving their performance[30]. The three AI chatbots were asked to provide the three most likely medical diagnoses, rank their choice in order of likelihood, and offer clinical reasoning behind their diagnoses[31]. Each vignette was input into a fresh session of the chatbots, and the first response was recorded. Primary diagnosis, which was ranked first, was counted as the correct identification of the disorder for OCD cases and all controls. The correct OCD identification rates in three AI chatbots were compared with medical and mental health professionals' performance assessed using the same OCD vignettes in the five original studies. Diagnostic performance for control vignettes was examined across three AI chatbots. Lastly, two mental health professionals independently reviewed the AI chatbots' responses to detect any incorrect clinical reasoning behind their diagnoses. Inter-rater agreement was calculated as a Cohen's kappa. The present study involved no personally identifiable human subject information; hence, it was deemed exempt from the Institutional Review Board at Stanford University.

Data availability

The vignettes used in this study were publicly available from published literature, and the vignette information is provided in a footnote of Supplementary Table 3. The data generated from this study, including OCD and control vignettes and LLMs' responses to them, will be made available through Supplemental Material upon publication of this study.

Received: 15 March 2024; Accepted: 28 June 2024;
Published online: 19 July 2024

References

1. Beam, K. et al. Performance of a large language model on practice questions for the neonatal board examination. *JAMA Pediatr.* **177**, 977–979 (2023).
2. Cai, Z. R. et al. Assessment of correctness, content omission, and risk of harm in large language model responses to dermatology continuing medical education questions. *J. Invest. Dermatol*. https://doi.org/10.1016/J.JID.2024.01.015 (2024).
3. Lyons, R. J., Arepalli, S. R., Fromal, O., Choi, J. D. & Jain, N. Artificial intelligence chatbot performance in triage of ophthalmic conditions. *Can. J. Ophthalmol.* https://doi.org/10.1016/J.JCJO.2023.07.016 (2023).

4. Chen, S. et al. Use of artificial intelligence chatbots for cancer treatment information. *JAMA Oncol.* **9**, 1459–1462 (2023).
5. Strong, E. et al. Chatbot vs medical student performance on free-response clinical reasoning examinations. *JAMA Intern. Med.* **183**, 1028–1030 (2023).
6. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
7. Sallam, M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare* **11**, 887 (2023).
8. Psychiatry.org. The basics of augmented intelligence: some factors psychiatrists need to know now. https://www.psychiatry.org/News-room/APA-Blogs/The-Basics-of-Augmented-Intelligence (2023).
9. Blease, C., Worthen, A. & Torous, J. Psychiatrists' experiences and opinions of generative artificial intelligence in mental healthcare: An online mixed methods survey. *Psychiatry Res.* **333**, 115724 (2024).
10. APA. What is obsessive-compulsive disorder? https://www.psychiatry.org:443/patients-families/obsessive-compulsive-disorder/what-is-obsessive-compulsive-disorder (2022).
11. National Institute of Mental Health (NIMH). Obsessive-compulsive disorder (OCD). https://www.nimh.nih.gov/health/statistics/obsessive-compulsive-disorder-ocd (2022).
12. National Comorbidity Survey (NCSSC). Harvard Medical School. https://www.hcp.med.harvard.edu/ncs/index.php (2007)
13. Pinto, A., Mancebo, M. C., Eisen, J. L., Pagano, M. E. & Rasmussen, S. A. The brown longitudinal obsessive compulsive study: clinical features and symptoms of the sample at intake. *J. Clin. Psychiatry* **67**, 703–711 (2006).
14. Perris, F. et al. Duration of untreated illness in patients with obsessive–compulsive disorder and its impact on long-term outcome: a systematic review. *J. Pers. Med.* **13**, 1453 (2023).
15. Galido, P. V., Butala, S., Chakerian, M. & Agustines, D. A case study demonstrating applications of ChatGPT in the clinical management of treatment-resistant schizophrenia. *Cureus* **15**, e38166 (2023).
16. Cohan, A. et al. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In *Proc. 27th International Conference on Computational Linguistics* (eds. Bender, E. M., Derczynski, L. & Isabelle, P.) 1485–1497 (Association for Computational Linguistics, 2018).
17. Xu, X. et al. Leveraging large language models for mental health prediction via online text data. In *Proc. ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (Association for Computing Machinery, 2023).
18. Galatzer-Levy, I. R., McDuff, D., Natarajan, V., Karthikesalingam, A. & Malgaroli, M. The capability of large language models to measure psychiatric functioning. Preprint at https://doi.org/10.48550/ARXIV.2308.01834 (2023).
19. Levkovich, I. & Elyoseph, Z. Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. *Fam. Med. Community Health* **11**, e002391 (2023).
20. Usage policies. https://openai.com/policies/usage-policies (2024).
21. Lucas, G. M., Gratch, J., King, A. & Morency, L. P. It's only a computer: virtual humans increase willingness to disclose. *Comput. Hum. Behav.* **37**, 94–100 (2014).
22. Elyoseph, Z., Hadar-Shoval, D., Asraf, K. & Lvovsky, M. ChatGPT outperforms humans in emotional awareness evaluations. *Front. Psychol*. **14**, 1199058 (2023).
23. House, T. W. FACT SHEET: President Biden issues executive order on safe, secure, and trustworthy artificial intelligence. *The White House* https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/ (2023).
24. Glazier, K., Swing, M. & McGinn, L. K. Half of obsessive-compulsive disorder cases misdiagnosed: vignette-based survey of primary care physicians. *J. Clin. Psychiatry* **76**, e761–e767 (2015).
25. Gouniai, J. M., Smith, K. D. & Leonte, K. G. Do clergy recognize and respond appropriately to the many themes in obsessive-compulsive disorder?: data from a Pacific Island community. *Ment. Health Relig. Cult.* **25**, 33–46 (2022).
26. Gouniai, J. M., Smith, K. D. & Leonte, K. G. Many common presentations of obsessive-compulsive disorder unrecognized by medical providers in a Pacific Island community. *J. Ment. Health Train. Educ. Pract.* **17**, 419–428 (2022).
27. Glazier, K., Calixte, R. M., Rothschild, R. & Pinto, A. High rates of OCD symptom misidentification by mental health professionals. *Ann. Clin. Psychiatry* **25**, 201–209.
28. Glazier, K. & McGinn, L. K. Non-contamination and non-symmetry OCD obsessions are commonly not recognized by clinical, counseling and school psychology doctoral students. *J. Depress. Anxiety* **04** (2015).
29. Kim, J., Cai, Z. R., Chen, M. L., Simard, J. F. & Linos, E. Assessing biases in medical decisions via clinician and AI Chatbot responses to patient vignettes. *JAMA Netw. Open* **6**, E2338050 (2023).
30. Wang, J. et al. Prompt engineering for healthcare: methodologies and applications. Preprint at https://doi.org/10.48550/arXiv.2304.14670 (2024).
31. Savage, T., Nayak, A., Gallo, R., Rangan, E. & Chen, J. H. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *Npj Digit. Med.* **7**, 1–7 (2024).

Acknowledgements

E.L. was supported by the National Institutes of Health: Mid-career Investigator Award in Patient-Oriented Research (K24AR075060). The funders of this study had no role in the design and conduct of the study, collection, management, analysis, and interpretation of the data, preparation, review, or approval of the manuscript, nor a decision to submit the manuscript for publication.

Author contributions

J.K. and K.G.L. contributed equally as joint first authors. E.L., A.P., and C.I.R. contributed equally as joint senior authors. Concept and design: J.K., J.B.T., and C.I.R. Acquisition, analysis, or interpretation of data: J.K., K.G.L., A.P., and C.I.R. Drafting of the manuscript: J.K., K.G.L., A.P., and C.I.R. Critical review of the manuscript for important intellectual content: J.K., K.G.L., M.L.C., J.B.T., E.L., A.P., and C.I.R. Statistical analysis: J.K. and K.G.L. Administrative, technical, or material support: M.L.C., E.L., and C.I.R. Supervision: E.L., A.P., and C.I.R.

Competing interests

In the last 3 years, C.I.R. has served as a consultant for Biohaven Pharmaceuticals, Osmind, and Biogen; and receives research grant support from Biohaven Pharmaceuticals, a stipend from American Psychiatric Association Publishing for her role as Deputy Editor at The American Journal of Psychiatry, and book royalties from American Psychiatric Association Publishing. J.B.T. is an associated editor for npj Digital Medicine. He was not involved in the review of this paper. The remaining authors declare no competing financial or non-financial interests or other relationships relevant to the subject of this manuscript.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01181-x.

Correspondence and requests for materials should be addressed to Jiyeong Kim or Carolyn I. Rodriguez.

Reprints and permissions information is available at http://www.nature.com/reprints