The Spanish Journal of Psychology (2025), **28**, e3, 1–5
doi:10.1017/SJP.2024.32

Review Article

Building Trust with AI: How Essential is Validating AI Models in the Therapeutic Triad of Therapist, Patient, and Artificial Third? Comment on What is the Current and Future Status of Digital Mental Health Interventions?

Alejandro Garcia-Rudolph[1,2], David Sánchez-Pinsach[1,2], Anna Gilabert[1,2], Joan Saurí[1,2], Maria Dolors Soler[1,2] and Eloy Opisso[1,2]

[1]Universitat Autònoma de Barcelona, Spain and [2]Fundació Institut d'Investigació en Ciències de la Salut Germans Trias i Pujol, Spain

Abstract

Since the publication of "What is the Current and Future Status of Digital Mental Health Interventions?" the exponential growth and widespread adoption of ChatGPT have underscored the importance of reassessing its utility in digital mental health interventions. This review critically examined the potential of ChatGPT, particularly focusing on its application within clinical psychology settings as the technology has continued evolving through 2023 and 2024. Alongside this, our literature review spanned US Medical Licensing Examination (USMLE) validations, assessments of the capacity to interpret human emotions, analyses concerning the identification of depression and its determinants at treatment initiation, and reported our findings. Our review evaluated the capabilities of GPT-3.5 and GPT-4.0 separately in clinical psychology settings, highlighting the potential of conversational AI to overcome traditional barriers such as stigma and accessibility in mental health treatment. Each model displayed different levels of proficiency, indicating a promising yet cautious pathway for integrating AI into mental health practices.

Keywords: artificial intelligence; Chatbot; ChatGPT

(Received: 24 May 2024; revised: 12 November 2024; accepted: 20 December 2024)

"What is the Current and Future Status of Digital Mental Health Interventions?" (Baños et al., 2022) was published online in February 2022. Later, that same year, in November 2022, OpenAI released ChatGPT (OpenAI, 2024). Remarkably, within just two months, by January 2023, ChatGPT had already attracted over 100 million users, setting a record as the fastest-growing consumer application in history. For comparison, platforms like Instagram took about 2.5 years to reach the same number of users (Lungren et al., 2023).

This unprecedented adoption continued, and as of May 2024, its user base had expanded to more than 180.5 million, marking an 80% growth since January 2023 (Exploding Topics, 2024). The explosive growth and widespread acceptance of ChatGPT not only underscore the relevance but also amplify the importance of examining how such technologies can be harnessed to enhance and possibly transform digital mental health interventions.

Baños et al. (2022) have already discussed chatbots and conversational agents, particularly those based on cognitive behavioral therapy (CBT) (Bendig et al., 2019). However, considering the rapid and widespread adoption of ChatGPT during 2023 and 2024, it becomes essential to specifically reevaluate and validate its use in the field of mental health. This reassessment will help determine whether ChatGPT is prepared for application in clinical psychology settings.

Building on this premise, our review not only synthesizes recent advancements in validation from the existing literature but also introduces findings from our original validation study. which contributes by analyzing 100 multiple-choice questions derived from vignettes in official Spanish government examinations for public administration roles in psychology. Notably, our findings are new insights as they have yet to be published or considered in any academic journal.

Given that ChatGPT is the most widely adopted chatbot in use today (Exploding Topics, 2024) and there are two versions of ChatGPT—GPT-3.5, which is freely available, and GPT-4, accessible via subscription—it is crucial to analyze both models to provide comprehensive insights. This dual approach ensures inclusivity, accommodating not only those who can afford the subscription model but also psychology students and practitioners who might only have access to the free version. By examining the performances of both GPT-3.5 and GPT-4, our analysis can cater to a broader audience, making it relevant and accessible to all, regardless of

Corresponding author: Alejandro Garcia-Rudolph, Departmento de Investigación e Innovación, Institut Guttmann, Institut Universitari de Neurorehabilitació adscrit a la UAB, Badalona, Barcelona, Spain; Emails: alejandropablogarcia@gmail.com; agarciar@guttmann.com

Cite this article: Garcia-Rudolph, A., Sánchez-Pinsach, D., Gilabert, A., Saurí, J., Soler, M. D., & Opisso, E. (2025). Building Trust with AI: How Essential is Validating AI Models in the Therapeutic Triad of Therapist, Patient, and Artificial Third? Comment on What is the Current and Future Status of Digital Mental Health Interventions?. *The Spanish Journal of Psychology* **28**, e3, 1–5. https://doi.org/10.1017/SJP.2024.32

financial constraints. This inclusive approach is especially important in educational settings, where access to advanced tools may be limited. Thus, our review aimed to offer a balanced perspective on the capabilities and limitations of these AI models, ensuring that all potential users in the field of psychology, can make informed decisions about their applicability and utility.

The rapid adoption of ChatGPT has raised significant concerns in the medical community regarding biases, misinformation, ethics in publishing, and potential plagiarism, prompting calls for cautious implementation in practice (Eysenbach et al., 2023).

Recent reports envision a vast potential for future GPT applications in mental health including psychotherapy in clinical settings (Cheng et al., 2023). Nevertheless, the use of ChatGPT in mental health is also associated with several potential drawbacks and limitations (Wong et al., 2024). It offers an output text that seems knowledgeable and coherent, but the AI does not truly "understand" its output. This could lead to "hallucinations"—instances where the AI produces plausible sounding but incorrect or unrelated information (Alkaissi & McFarlane, 2023). ChatGPT's absence of clinical reasoning and accumulated experience may result in omission of important clinical information from patient summaries and medical records. Therefore, it is highly recommended to require professionals to verify and revise ChatGPT-generated content. (Cheng et al., 2023).

The US Medical Licensing Examination (USMLE)

Given the significance of the USMLE as a critical assessment in medical licensing, ChatGPT has recently been evaluated against this rigorous examination. The USMLE does not specifically disclose the percentage of the examination dedicated to psychology or behavioral sciences. However, for USMLE step 1, the behavioral sciences section, which includes psychology, is one of the seven broad categories tested. The weighting for each category can vary slightly based on the individual examination form, but typically, behavioral sciences might constitute around 10% to 15% of the questions (Parry et al., 2019).

Psychological aspects are often embedded also in step 3, in clinical case simulations and the multiple-choice questions where a physician's knowledge of psychiatric conditions, their management, and understanding of the psychological dimensions of patient care are tested. This includes recognizing and planning management for psychiatric disorders, dealing with behavioral and social factors that affect health, and understanding the interactions between psychological and physical health (Schwartz et al., 2018).

Kung et al. (2023) conducted an assessment of ChatGPT's capabilities on the USMLE, encompassing steps 1, 2CK, and 3. The study found that GPT-3.5 performed close to or above the passing thresholds (60%) across all three examinations (350 USMLE items in total) without specific training or reinforcement.

Gilson et al. (2023) assessed USMLE steps 1 and 2 using two datasets, one derived from AMBOSS, a commonly used question bank for medical students, and the second set was the National Board of Medical Examiners (NBME). GPT-3.5 achieved accuracies of 44% (44/100), 42% (42/100), 64.4% (56/87), and 57.8% (59/102).

Mihalache et al. (2024) reported that GPT-4 responded to 319 text-based multiple-choice questions from USMLE practice test material. ChatGPT-4 answered 82 of 93 (88%) questions correctly on USMLE step 1, 91 of 106 (86%) on step 2CK, and 108 of 120 (90%) on step 3 (Mihalache et al., 2024).

In our discussion of AI performance on the USMLE, it is also beneficial to consider similar assessments globally. The study by Guillen-Grima et al. (2023) evaluates the performance of GPT-3.5 and GPT-4 on the Spanish Medical Intern (MIR) examination, essential for medical specialist training in Spain. Their results indicate variability in performance depending on the type of question (theoretical vs. practical) and the examination's language (Spanish vs. English). For instance, GPT-4 achieved an overall performance of 86.81%, while GPT-3.5 scored 63.18% when the language was Spanish.

Clinical Vignettes

Franco D'Souza et al. (2023) evaluated GPT-3.5 "in enhancing mental health and well-being" using 100 clinical case vignettes. from a reference book (Wright et al., 2017). After each case, the book includes one or more open questions about it and GPT-3.5 replies to such questions that were assessed by expert faculties from the Department of Psychiatry. First, each vignette was categorized into one of ten themes (such as i) diagnosis, ii) differential diagnosis, iii) assessment, iv) investigation, and so on). The evaluation of GPT-3.5 was completed by taking the mean value of the scores in all themes provided by the experts. Results suggests that ChatGPT 3.5 achieved "Grade A" in 61% of cases, "Grade B" in 31%, and "Grade C" in 8% (Franco D'Souza et al., 2023). Limitations to this study include that the evaluations are reliant on experts' faculties whose biases might influence grading, compounded by the lack of a standardized scoring rubric, which can impact the consistency and reliability of the results. Besides, using the mean of expert scores to evaluate performance might obscure individual variance in assessments. Importantly, the study focuses solely on GPT-3.5 without examining newer models like GPT-4.0; a comparative analysis between versions could offer insights into advancements or improvements in AI capabilities over time.

Levkovich and Elyoseph (2023a) assigned ChatGPT to analyze vignettes of a hypothetical patient with varying levels of perceived burdensomeness and thwarted belongingness, comparing ChatGPT's evaluations to those by mental health professionals, using GPT-3.5 and GPT-4. The study aimed to assess GPT-4's accuracy in evaluating suicide risk aspects, finding its risk assessments comparable to professionals, especially in identifying suicidal ideation, though it overestimated psychache. Despite its potential in clinical decision-making, further extensive studies are necessary, particularly as GPT-3.5 often underestimated severe suicide risk, potentially downplaying actual risks (Levkovich & Elyoseph, 2023a).

Capacity to Interpret Human Emotions Using Specific Tests

Elyoseph et al. (2024) conducted a pilot evaluation study to assesses the ability of GPT-4 and Google Bard models to understand human emotions through both visual and textual inputs The Reading the Mind in the Eyes Test, created by Baron-Cohen and colleagues, was employed to evaluate the models' ability to interpret visual emotional cues. Concurrently, the Levels of Emotional Awareness Scale assessed the large language models' skill in understanding emotions from text. Together, these tests offered a comprehensive assessment of the mentalizing abilities of ChatGPT-4 and Bard. ChatGPT-4 demonstrated effectiveness in visual mentalizing, closely matching human performance levels. While both models showed proficiency

https://doi.org/10.1017/SJP.2024.32 Published online by Cambridge University Press

in interpreting textual emotions, Bard's performance in visual emotion recognition requires additional examination and improvement. (Elyoseph et al., 2024).

ChatGPT and Depression

Researchers evaluated the management recommendations for depressive episodes by GPT-3.5 and GPT-4 compared to those from primary care physicians, using specifically designed vignettes of hypothetical patients. The study found that for mild depression, both AI models significantly favored psychotherapy far more than physicians, who suggested it only 4.3% of the time. For severe cases, while both preferred combined treatments, the AIs recommended antidepressants more exclusively, highlighting their consistent, unbiased approach compared to the more varied human recommendations (Levkovich & Elyoseph, 2023b).

Our discussion can be enhanced by considering the recent publication from Obradovich et al. (2024), which underscores the emerging role of large language models (LLMs) in mental healthcare and psychiatric research. These models show significant promise in improving diagnostic accuracy and patient outcomes through early detection, treatment, and evaluation of mental health conditions. However, challenges such as ethical concerns, unpredictability of outputs, opacity in operational mechanisms, and the potential introduction of biases highlight the need for a multidisciplinary approach to ensure responsible deployment. Transparency and the establishment of ethical guidelines are crucial to build trust and effectively integrate LLMs into mental health practices (Obradovich et al., 2024).

Performance of GPT-3.5 and GPT-4.0 on 100 Multiple-Choice Questions from Official Spanish Government Clinical Psychology Examinations across Four Clinical Cases

Objectives

We aimed to evaluate the accuracy of AI models GPT-3.5 and GPT-4.0 by examining their responses to a series of 100 multiple-choice questions derived from official clinical psychology examinations administered by the Spanish government.

Methodology

Experimental Setup

The study was conducted in May 2024 at the Psychology Department of Institut Guttmann Hospital. The responses were generated by the AI models GPT-3.5 and GPT-4.0, each response was produced once and in situ, ensuring real-time processing by the models without repeated attempts.

Selection of Examinations and Questions

We selected two official public examinations from 2021 and 2024, utilized in the public selection processes for positions within the Senior Technical Corps, specifically the psychology option. These examinations were chosen for their relevance in assessing specific clinical psychology skills required in local government positions in Spain. Each examination encompassed two clinical cases, with each case initially comprising 25 multiple-choice questions. Notably, the 2021 examination for Case 1 was adjusted by the official evaluation committee, reducing two questions and resulting in 23 questions for that case. The full text of both examinations, including questions and answers, is available on the respective city council's website (Murcia, 2024).

Testing Procedure

Each clinical case and associated questions were individually prompted to both AI models by authors AG-R and DS-P. Each model's response to the questions was captured one at a time and registered in an Excel sheet to ensure accurate data collection and analysis. Importantly, no prompts were provided as responses to the chatbots for any of the questions, maintaining a standard question-response format throughout the study.

Results and Discussion

Overall Performance

In total, 98 questions were analyzed -48 from the 2021 examination and 50 from the 2024 examination.

GPT-3.5 achieved an overall accuracy of 67.3% across both examinations, while GPT-4.0 showed a higher overall accuracy of 76.5%.

Performance Across Examination Periods and Cases

Notable differences were observed across the two distinct examination periods and four clinical cases. The results from the 2021 examination for Case 1 showed both GPT-3.5 and GPT-4.0 scoring equally, with a correct response rate of 60.8%. However, for Case 2 of the same year, GPT-4.0 slightly outperformed GPT-3.5, achieving a 68.0% success rate compared to GPT-3.5's 64.0%. The 2024 examination further highlighted the advancements in GPT-4.0's capabilities. In Case 1, GPT-4.0 answered 80.0% of the questions correctly, while GPT-3.5 remained consistent with its earlier performance at 64.0%. Case 2 of the 2024 examination saw GPT-4.0 reaching a remarkable 96.0% correctness, significantly surpassing GPT-3.5's 80.0%.

This reflects an evolution in the model's performance from 2021 to 2024, marking an improvement in its ability to handle complex clinical psychology questions, suggesting a promising potential for AI applications in clinical settings. The progression in accuracy from GPT-3.5 to GPT-4.0 across these examinations illustrates not only advancements in AI technology but also its increasing utility in educational and professional domains.

Main Weaknesses Identified Across Studies

The analysis of GPT-3.5 and GPT-4.0 across multiple studies reveals several weaknesses that merit attention, especially as these models are considered for broader application in clinical psychology settings.

One of the main weak points observed for GPT-3.5 is its inconsistency in handling complex clinical scenarios. This model often underestimates the severity of conditions, particularly in critical cases, which could lead to suboptimal recommendations in a real-world clinical context. This trend was evident in the simulation exercises where GPT-3.5 frequently scored lower than GPT-4.0, particularly in more challenging or severe case scenarios.

GPT-4.0, while showing marked improvements over its predecessor, also exhibited some deficiencies. Notably, it tends to overestimate certain psychological conditions, such as psychache, which might lead to overly cautious or aggressive treatment strategies if relied upon without human oversight. This tendency to overestimate could be attributed to the model's deep learning algorithms,

https://doi.org/10.1017/SJP.2024.32 Published online by Cambridge University Press

which might be overly sensitive to certain input cues or patterns that do not necessarily correlate with higher risk in a consistent manner.

Both models also demonstrate a lack of adaptability to nuanced human emotions and social contexts that can be critical in clinical psychology. While they can process and respond to standardized test scenarios effectively, their ability to interpret subtleties in human behavior and emotional expressions remains limited. This limitation underscores the need for further refinement in their training datasets and algorithms to better understand and react to the complexities of human psychological conditions.

In conclusion, while GPT-3.5 and GPT-4.0 offer significant potential for supporting clinical psychology through automated assessments and interventions, their current limitations in accuracy, sensitivity to nuance, and the potential for biased or unbalanced responses highlight the necessity for continuous development and careful integration into clinical practices. These weaknesses must be addressed through more sophisticated training, better model tuning, and integration of comprehensive human oversight to ensure their efficacy and safety in clinical applications.

The Artificial Third

The concept of "thirdness" traditionally enriched the two-party dynamic of therapist and patient by introducing a third element into their interaction, as discussed by Tal et al. (2023). This element, historically seen as the therapist's ability to balance personal perspective with an understanding of the patient's viewpoint, has taken a new form with the advent of generative AI (GenAI) in therapy. Coined as the "artificial third" by researcher Yuval Haber (Tal et al., 2023), this role of GenAI adds a distinctive dimension to the therapy setting, creating a triadic relationship comprising the patient, the therapist, and the AI. Tal and colleagues note that this arrangement not only can reshape clinical interactions but also provide a fresh lens through which both therapist and patient can reassess themselves, facilitated by the interpretive capacities of this nonhuman participant (Tal et al., 2023).

ChatGPT exemplifies the concept of the "artificial third" in therapeutic settings, particularly enhancing CBT through its capability to simulate therapeutic dialogues. ChatGPT can simulate dialogue and provide responses based on conditioned learning from vast datasets, which might align with the mechanisms of conditioning discussed in verbal behavior therapy. The concept of effective verbal interaction in therapeutic settings was recently supported by Pardo-Cebrián et al. (2022), analyzing the impact of different verbalizations by therapists using the Socratic method, highlighting the importance of specific verbal strategies in influencing client responses.

This AI integration may support in identifying and correcting patients' cognitive distortions that might influence their emotional and behavioral responses. Additionally, ChatGPT may assist in devising exposure therapy plans and support patients in completing therapeutic "homework," ensuring continuity and consistency in treatment approaches outside of conventional sessions. Its round-the-clock availability could provide crucial, timely access to support and information, pivotal for CBT enhancement and immediate patient support. This AI tool does not aim to replace the therapist, but act as a significant adjunct by offering analytical perspectives that may not be immediately evident in traditional CBT sessions.

Considering the introduction of GenAI as an "artificial third" in therapeutic settings (Tal et al., 2023), how critical is the rigorous validation of AI models in building trust between the therapist, patient, and the AI as an "artificial third" in therapy? What specific validation strategies could effectively establish this trust within the therapeutic triad?

Conclusions

Our comprehensive review and empirical analysis contribute to the evolving dialogue on digital mental health interventions, an area that has gained unprecedented relevance in the wake of the COVID-19 pandemic. By comparing the capabilities of GPT-3.5 and GPT-4.0 in clinical psychology settings, our study highlights the potential of conversational AI to mitigate some of the barriers traditionally associated with mental health treatment, such as stigma and accessibility. Both models demonstrated varying degrees of proficiency, revealing a promising yet cautious path forward for integrating AI into mental health practices. GPT-4.0, in particular, showed marked improvements in understanding and responding to complex clinical scenarios, which suggests that newer generations of AI could enhance the personalization and cultural adaptation of digital therapies. However, challenges such as their ability to interpret subtleties in human behavior and emotional expressions remain limited, as highlighted in our review.

Despite these advancements, the implementation of such technologies in real-world settings is fraught with complexities, including ethical considerations, the need for hybrid care models, and the ongoing development of predictive methodologies. Our findings underscore the importance of continuous validation and adaptation of AI tools to meet diverse patient needs and align with the evolving standards of precision medicine.

Data availability statement. The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgements. Special thanks to Olga Araujo from Institut Guttmann Neurorehabilitation Hospital, Documentation Department, for her continued support in accessing related literature.

Author contribution. A. G.-R.: Conceptualization, investigation, data curation, formal analysis, investigation, methodology, writing – original draft, writing review & editing. D. S.-P.: Conceptualization, data curation, formal analysis, investigation, writing review & editing. A. G.: Conceptualization, investigation, writing review & editing. J. S.: Conceptualization, Investigation, formal analysis, methodology, writing-review & editing. M.D. S.: Conceptualization, formal analysis, investigation, methodology, supervision, project administration, writing-review & editing. E. O.: Conceptualization, formal analysis, methodology, supervision, project administration, writing-review & editing.

Competing interest. The authors declare no competing interests exist.

References

Alkaissi, H., & **McFarlane, S. I**. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, **15**(2). http://doi.org/10.7759/cureus.35179.

Baños, R. M., **Herrero, R.**, & **Vara, M. D**. (2022). What is the current and future status of digital mental health interventions?. *The Spanish Journal of Psychology*, **25**, e5. http://doi.org/10.1017/SJP.2022.2.

Bendig, E., **Erb, B.**, & **Schulze-Thuesing, L.**, & **Baumeister, H**. (2019). The next generation: Chatbots in clinical psychology and psychotherapy to foster mental health–A scoping review. *Verhaltenstherapie*, **32**(1), 1–13. https://doi.org/10.1159/000501812

Cheng, S. W., **Chang, C. W.**, **Chang, W. J.**, **Wang, H. W.**, **Liang, C. S.**, **Kishimoto, T.**, **Chang, J. P.**, **Kuo, J. S.**, & **Su, K. P**. (2023). The now and future of ChatGPT and GPT in psychiatry. *Psychiatry Clinical Neuroscience*, **77**(11), 592–596. https://doi.org/10.1111/pcn.13588.

Elyoseph, Z., **Refoua, E.**, **Asraf, K.**, **Lvovsky, M.**, **Shimoni, Y.**, & **Hadar-Shoval, D**. (2024). Capacity of generative AI to interpret human emotions from visual and textual data: Pilot evaluation study. *JMIR Mental Health*, **6**(11), 54369. https://doi.org/10.2196/54369.

Exploding Topics. (2024). *Number of ChatGPT Users.* Retrieved May 2024, from https://explodingtopics.com/blog/chatgpt-users.

Eysenbach, G. (2023). The role of ChatGPT, generative language models, and artificial intelligence in medical education: A conversation with ChatGPT and a Call for papers. *JMIR Medical Education*, **6**(9), 46885. https://doi.org/10.2196/46885.

D'Souza, R. F., **Amanullah, S.**, **Mathew, M.**, & **Surapaneni, K. M**. (2023). Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. *Asian Journal of Psychiatry*, **89**, 103770. https://doi.org/10.1016/j.ajp.2023.103770.

Guillen-Grima, F., **Guillen-Aguinaga, S.**, **Guillen-Aguinaga, L.**, **Alas-Brun, R.**, **Onambele, L.**, **Ortega, W.**, **Montejo, R.**, **Aguinaga-Ontoso, E.**, **Barach, P.**, & **Aguinaga-Ontoso, I**. (2023). Evaluating the efficacy of ChatGPT in navigating the Spanish medical residency entrance examination (MIR): Promising horizons for AI in clinical medicine. *Clinics and Practice*, **13**(6),1460–1487. https://doi.org/10.3390/clinpract13060130.

Gilson, A., **Safranek, C.W.**, **Huang, T.**, **Socrates, V.**, **Chi, L.**, **Taylor, R.A.**, & **Chartash, D**. (2023). 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 Medical Education*, **8**(9), 45312. https://doi.org/10.2196/45312.

Kung, T. H., **Cheatham, M.**, **Medenilla, A.**, **Sillos, C.**, **De Leon, L.**, **Elepaño, C.**, **Madriaga, M.**, **Aggabao, R.**, **Diaz-Candido, G.**, **Maningo, J.**, & **Tseng, V**. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*, **2**(2), 0000198. https://doi.org/10.1371/journal.pdig.0000198

Levkovich, I., & **Elyoseph, Z**. (2023a). Suicide risk assessments through the eyes of ChatGPT-3.5 Versus ChatGPT-4: Vignette study. *JMIR Mental Health*, **20**(10), 51232. https://doi.org/10.2196/51232.

Levkovich, I., & **Elyoseph, Z**. (2023b). Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. *Family Medicine and Community Health*, **11**(4), 002391. https://doi.org/10.1136/fmch-2023-002391.

Lungren, M. P., **Fishman, E. K.**, **Chu, L. C.**, **Rizk, R. C.**, & **Rowe, S. P.** (2023). More is different: Large language models in health care. *Journal of the American College of Radiology*, **1**, 1546–1440. https://doi.org/10.1016/j.jacr.2023.11.021.

Mihalache, A., **Huang, R. S.**, **Popovic, M. M.**, & **Muni, R. H**. (2024). ChatGPT-4: An assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination. *Medical Teacher*, **46**(3), 366–372. https://doi.org/10.1080/0142159X.2023.2249588.

Murcia. (2024). *Convocatorias de procesos selectivos—Convocatoria al cuerpo superior facultativo, opción Psicología [Announcements of selective processes—Announcement for the higher technical body, Psychology option].* Retrieved May 2024, from https://empleopublico.carm.es/web/pagina?IDCONTENIDO=2340&IDTIPO=200&CODIGO_CUERPO=AFX17&CODIGO_CONVOCATORIA=AFX17C22&TIPO_ACCESO=L&RASTRO=c$m61986,61991,62006.

Obradovich, N., **Khalsa, S. S.**, **Khan, W. U.**, **Suh, J.**, **Perlis, R. H.**, **Ajilore, O.**, & **Paulus, M. P**. (2024). Opportunities and risks of large language models in psychiatry. *NPP-Digital Psychiatry and Neuroscience*, **2**(1), 8. https://doi.org/10.1038/s44277-024-00010-z

OpenAI. (2024). *Creating safe AGI that benefits all of humanity.* Retrieved May 2024, from https://openai.com/chatgpt/

Pardo-Cebrián, R., **Calero-Elvira, A.**, & **Guerrero-Escagedo, M. C**. (2022). Verbal behavior analysis of expert and inexperienced therapists applying the socratic method. *The Spanish Journal of Psychology*, **25**, e19. https://doi.org/10.1017/SJP.2022

Parry, S., **Pachunka, J.**, & **Beck Dallaghan, G. L**. (2019). Factors predictive of performance on USMLE Step 1: Do Commercial Study aids improve scores? *Medical Science Educator*, **29**(3), 667–672. https://doi.org/10.1007/s40670-019-00722-4.

Schwartz, L. F., **Lineberry, M.**, **Park, Y. S.**, **Kamin, C. S.**, & **Hyderi, A. A.** (2018). Development and evaluation of a student-initiated test preparation program for the USMLE Step 1 Examination. *Teaching and Learning in Medicine*, **30**(2), 193–201. https://doi.org/10.1080/10401334.2017.1386106.

Tal, A., **Elyoseph, Z.**, **Haber, Y.**, **Angert, T.**, **Gur, T.**, **Simon, T.**, & **Asman, O**. (2023). The artificial third: Utilizing ChatGPT in mental health. *The American Journal of Bioethics*, **23**(10), 74–77. https://doi.org/10.1080/15265161.2023.2250297.

Wong, R. S. Y. (2024). ChatGPT in psychiatry: Promises and pitfalls. *The Egyptian Journal of Neurology, Psychiatry and Neurosurgery*, **60**, 14. https://doi.org/10.1186/s41983-024-00791-2

Wright, B., **Dave, S.**, & **Dogra, N**. (2017). *100 cases in psychiatry*. Taylor & Francis Group.

https://doi.org/10.1017/SJP.2024.32 Published online by Cambridge University Press