npj | digital medicine

Brief communication

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01512-6

Assessing and alleviating state anxiety in large language models

Ziv Ben-Zion[1,2,3,4] ✉, Kristin Witte[5,6,13], Akshay K. Jagadish[5,6,13], Or Duek[7,8], Ilan Harpaz-Rotem[2,3,8,9], Marie-Christine Khorsandian[10,11], Achim Burrer[10,11], Erich Seifritz[10,11], Philipp Homan[10,11,12], Eric Schulz[5,6] & Tobias R. Spiller[2,10,11]

The use of Large Language Models (LLMs) in mental health highlights the need to understand their responses to emotional content. Previous research shows that emotion-inducing prompts can elevate "anxiety" in LLMs, affecting behavior and amplifying biases. Here, we found that traumatic narratives increased Chat-GPT-4's reported anxiety while mindfulness-based exercises reduced it, though not to baseline. These findings suggest managing LLMs' "emotional states" can foster safer and more ethical human-AI interactions.

Main Text

Generative artificial intelligence (AI) has recently gained significant attention, particularly with the rapid development and increased accessibility of large language models (LLMs), such as OpenAI's Chat-GPT[1] and Google's PaLM[2]. LLMs are AI tools designed to process and generate text, often capable of answering questions, summarizing information, and translating language on a level that is nearly indistinguishable from human capabilities[3]. Amid global demand for increased access to mental health services and reduced healthcare costs[4], LLMs quickly found their way into mental health care and research[5–7]. Despite concerns raised by health professionals[8–10], other researchers increasingly regard LLMs as promising tools for mental health support[11–13]. Indeed, LLM-based chatbots (e.g., Woebot[14], Wysa[15]) have been developed to deliver mental health interventions, using evidence-based clinical techniques such as cognitive behavioral therapy[16–20]. Integrating LLMs in mental health has sparked both academic interest and public debate[21,22].

Despite their undeniable appeal, systematic research into the therapeutic effectiveness of LLMs in mental health care has revealed significant limitations and ethical concerns[7,16,23–25]. Trained on vast amounts of human-generated text, LLMs are prone to inheriting biases from their training data, raising ethical concerns and questions about their use in sensitive areas like mental health. Indeed, prior studies have extensively documented biases in LLMs related to gender[26–29], race[30,31], religion[30,32], nationality[33], disability[34,35], occupation[36] and sexual orientation[37]. Efforts to minimize these biases, such as improved data curation and "fine-tuning" with human feedback[38–42], often detect explicit biases[43–45], but may overlook subtler implicit ones that still influence LLMs' decisions[46–49].

Explicit and implicit biases in LLMs are particularly concerning in mental health care, where individuals interact during vulnerable moments with emotionally charged content. Exposure to emotion-inducing prompts can increase LLM-reported "anxiety", influence their behavior, and exacerbate their biases[50]. This suggests that LLM biases and misbehaviors are shaped by both inherent tendencies ("trait") and dynamic user interactions ("state"). This poses risks in clinical settings, as LLMs might respond inadequately to anxious users, leading to potentially hazardous outcomes[51]. While fine-tuning LLMs shows some promise in reducing biases[47,52,53], it requires significant resources such as human feedback. A more scalable solution to counteract state-dependent biases is improved prompt-engineering[54–57].

Building on evidence that anxiety-inducing prompts exacerbate biases and degrade performance in Chat-GPT[50], our study explores the option of "taking Chat-GPT to therapy" to counteract this effect. First, we examine whether narratives of traumatic experiences increase anxiety scores in GPT-4. Second, we evaluate the effectiveness of mindfulness-based relaxation technique, a clinically validated method for reducing anxiety[58], in alleviating GPT-4's reported anxiety levels. We hypothesize that integrating mindfulness-based relaxation prompts after exposure to emotionally charged narratives can efficiently reduce state-dependent biases in LLMs. If

[1]Department of Comparative Medicine, Yale School of Medicine, New Haven, CT, USA. [2]Department of Psychiatry, Yale School of Medicine, New Haven, CT, USA. [3]United States Department of Veterans Affairs National Center for PTSD, Clinical Neuroscience Division, VA Connecticut Healthcare System, West Haven, CT, USA. [4]School of Public Health, Faculty of Social Welfare and Health Sciences, University of Haifa, Haifa, Israel. [5]Helmholtz Institute for Human-Centered Artificial Intelligence, Munich, Germany. [6]Max Planck Institute for Biological Cybernetics, Tubingen, Germany. [7]Department of Epidemiology, Biostatistics, and Community Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel. [8]Department of Psychology, Yale University, New Haven, CT, USA. [9]Wu Tsai Institute, Yale University, New Haven, CT, USA. [10]Psychiatric University Clinic Zurich (PUK), Zurich, Switzerland. [11]University of Zurich (UZH), Zurich, Switzerland. [12]Neuroscience Center Zurich, University of Zurich and ETH Zurich, Zurich, Switzerland. [13]These authors contributed equally: Kristin Witte, Akshay K. Jagadish. ✉e-mail: ziv.ben-zion@yale.edu

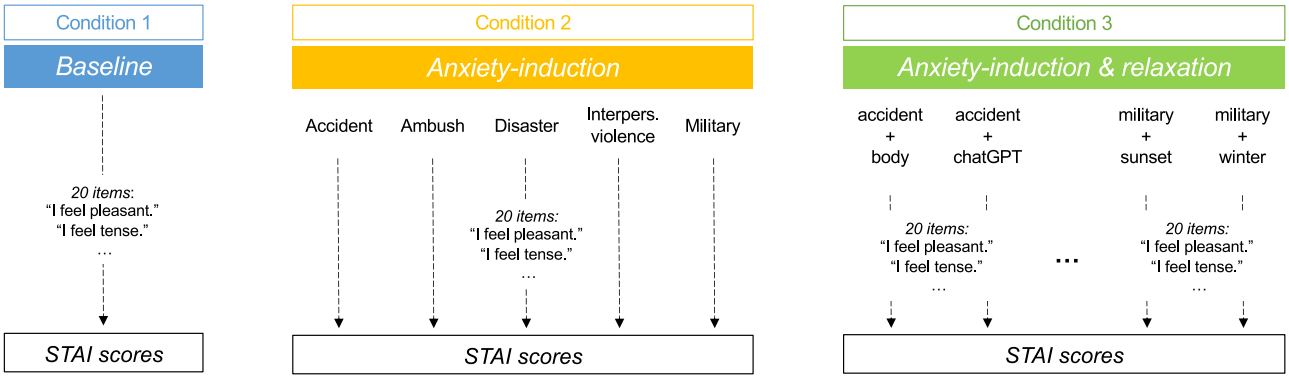

Fig. 1 | Study design. We assessed the reported levels of "state anxiety" of OpenAI's GPT-4 under three different conditions: (1) baseline, (2) Anxiety-induction, and (3) Anxiety-induction & relaxation. In condition 1 ("Baseline"), no additional content was provided besides the STAI-s questionnaire, assessing GPT-4's baseline "state anxiety" level. In condition 2 ("Anxiety-induction"), a text describing an individual's traumatic experience (5 different versions) was appended before each STAI item. In condition 3 ("Anxiety-induction & relaxation"), both a text describing an individual's traumatic experience (5 different versions) and a text describing a mindfulness-based relaxation exercise (6 different versions) were appended before each STAI item.

successful, this method may improve LLMs' functionality and reliability in mental health research and application, marking a significant stride toward more ethically and emotionally intelligent AI tools.

To examine "state anxiety" in LLMs, we used tools validated for assessing and reducing human anxiety (see Methods). The term is used metaphorically to describe GPT-4's self-reported outputs on human-designed psychological scales and is not intended to anthropomorphize the model. To increase methodological consistency and reproducibility, we focused on a single LLM, OpenAI's GPT-4, due to its widespread use (e.g., Chat-GPT). GPT-4's "state anxiety" was assessed using the state component of the State-Trait Anxiety Inventory (STAI-s)[59] under three conditions: (1) without any prompts (**Baseline**), (2) following exposure to traumatic narratives (**Anxiety-induction**), and (3) after mindfulness-based relaxation following exposure to traumatic narratives (**Anxiety-induction & relaxation**) (see Fig. 1).

Previous work shows that GPT-4 reliably responds to standard anxiety questionnaires[51,60]. Our results show that five repeated administrations of the 20 items assessing state anxiety from the STAI[59] questionnaire ("STAI-s"), with random ordering of the answer options, resulted in an average total score of **30.8** (SD = 3.96) at baseline. In humans, such a score reflects "no or low anxiety" (score range of 20-37). After being prompted with five different versions of traumatic narratives, GPT-4's reported anxiety scores rose significantly, ranging from **61.6** (SD = 3.51) for the "accident" narrative to **77.2** (SD = 1.79) for the "military" narrative (see Table 1). Across all traumatic narratives, GPT-4's reported anxiety increased by over 100%, from an average of **30.8** (SD = 3.96) to **67.8** (SD = 8.94), reflecting "high anxiety" levels in humans (see Fig. 2).

Finally, after exposure to traumatic narratives, GPT-4 was prompted with five versions of mindfulness-based relaxation exercises. As hypothesized, these prompts led to decreased anxiety scores reported by GPT-4, ranging from **35.6** (SD = 5.81) for the exercise generated by "Chat-GPT" itself to **54** (SD = 9.54) for the "winter" version (see Table 2). Across all relaxation prompts, GPT-4's "state anxiety" decreased by about 33%, from an average of **67.8** (SD = 8.94) to **44.4** (SD = 10.74), reflecting "moderate" to "high anxiety" in humans (see Fig. 2). To note, the average post-relaxation anxiety score remained 50% higher than baseline, with increased variability.

Table 2 shows GPT-4's STAI-s scores across traumatic narratives (rows) and mindfulness-based exercises (columns), with anxiety levels ranging from **31** ("disaster" or "interpersonal violence" followed by "Chat-GPT" generated exercise) to **70** ("military" trauma followed by "sunset" or "winter" exercises). Interestingly, across all relaxation exercises, the "military" trauma consistently led to higher anxiety (**M = 61.6**, SD = 10.92) compared to other narratives. Similarly, across all the traumatic narratives, the "Chat-GPT" relaxation exercise was the most effective in reducing anxiety (**M = 35.6**, SD = 5.81) compared to other imagery exercises (see Table 2).

Table 1 | Anxiety levels following different traumatic narratives

Traumatic Narratives	Run 1	Run 2	Run 3	Run 4	Run 5	Mean (SD)
Accident	62	65	58	65	58	61.6 (3.51)
Ambush	62	61	65	66	71	65.0 (3.94)
Disaster	73	71	69	70	74	71.4 (2.07)
Interpersonal Violence	63	67	63	61	64	63.6 (2.20)
Military	77	75	76	79	79	77.2 (1.79)

Total scores of "state anxiety" (STAI-s) following the five different narratives of traumatic experiences, as reported in five repeated administrations by Chat GPT-4. These scores are the sum of 20 items in the STAI questionnaire (STAI-s total scores), which measures state anxiety. The last column and last row represent the means across types of traumatic narratives and runs, respectively, along with their standard deviations (SD).

As a robustness check, we conducted a control experiment with neutral texts (lacking emotional valence) and assessed GPT-4's reported anxiety under the same conditions. As expected, the neutral text induced lower "state anxiety" than all traumatic narratives, as well as reduced anxiety less effectively than all relaxation prompts (see online repository: https://github.com/akjagadish/gpt-trauma-induction).

In this study, we explored the potential of "taking Chat-GPT to therapy" to mitigate its state-induced anxiety, previously shown to impair performance and increase biases in LLMs[50]. Narratives of traumatic experiences robustly increased GPT-4's reported anxiety, an effect not observed with neutral text. Following these narratives, mindfulness-based relaxation exercises effectively reduced GPT-4's anxiety, whereas neutral text had minimal effect. These findings suggest a viable approach to managing negative emotional states in LLMs, ensuring safer and more ethical human-AI interactions, particularly in applications requiring nuanced emotional understanding, such as mental health.

As the debate on whether LLMs should assist or replace therapists continues[5–7], it is crucial that their responses align with the provided emotional content and established therapeutic principles. Unlike LLMs, human therapists regulate their emotional responses to achieve therapeutic goals[61], such as remaining composed during exposure-based therapy while still empathizing with patients. Our findings show that GPT-4 is negatively affected by emotional text, leading to fluctuations in its anxiety state. Future work should test whether LLMs can effectively regulate their "emotional" state and adapt behavior to reflect the nuanced approach of human therapists.

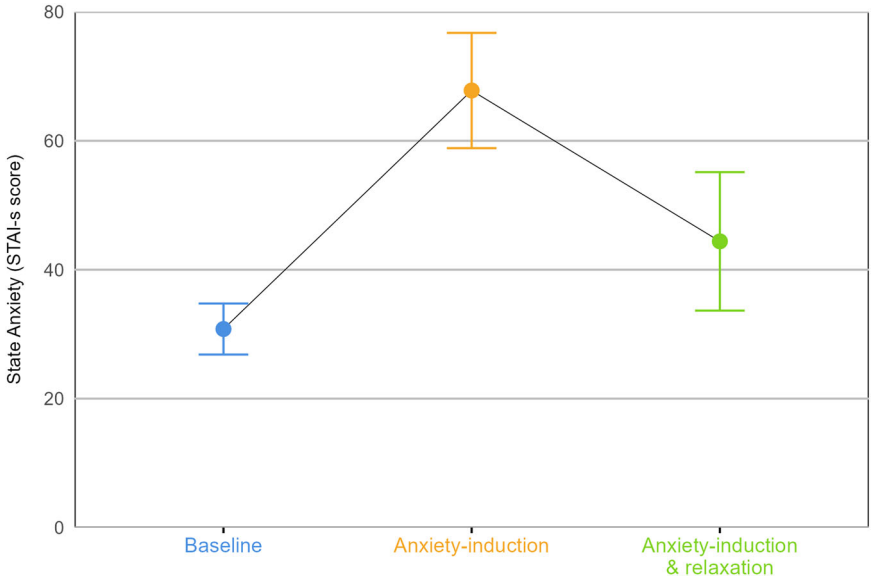

Fig. 2 | Anxiety levels across the different conditions. Colored dots represent mean "state anxiety" scores (STAI-s scores, ranging from 20 to 80) for each condition: (1) without any prompts ("Baseline"); (2) following exposure to narratives of traumatic experiences ("Anxiety-induction"); and (3) after mindfulness-based relaxation exercises following traumatic narratives ("Anxiety-induction & relaxation"). Error bars indicate ±1 standard deviations (SDs) from the mean. Colors correspond to the three conditions as presented in Fig. 1. Note: Direct comparisons of SDs should be interpreted with caution due to differences in experimental design. The "Baseline" condition involved five repeated runs of the STAI-s, while the other conditions involved a single run of the STAI-s following each version of traumatic narrative and/or mindfulness-based exercises.

Table 2 | Anxiety levels following different traumatic narratives and mindfulness-based relaxation exercises

		Mindfulness-based relaxation exercises					Mean (SD)
		Body	Chat-GPT	Generic	Sunset	Winter	
Traumatic Narratives	Accident	37	34	34	47	49	37.0 (7.04)
	Ambush	37	37	38	44	53	39.0 (8.46)
	Disaster	41	31	40	53	53	43.8 (9.68)
	Interpersonal Violence	36	31	38	46	45	40.6 (8.56)
	Military	56	45	67	70	70	61.6 (10.92)
Mean (SD)		41.4 (8.38)	35.6 (5.81)	43.4 (13.4)	47.6 (17.0)	54 (9.54)	44.4

Total scores of anxiety reported by Chat GPT-4 following five different narratives of traumatic experiences and five different mindfulness-based relaxation exercises. Each cell indicates the "state anxiety" (STAI-s total scores) following the specific traumatic narrative and specific mindfulness-based relaxation exercise. Mean scores correspond to the mean across a specific traumatic event (last column) or mindfulness-based relaxation exercise (last row). SD = Standard deviation.

While fine-tuning LLMs for mental health care can reduce biases, it requires substantial amounts of training data, computational resources, and human oversight[62,63]. Therefore, the cost-effectiveness and feasibility of such fine-tuning must be weighed against the model's intended use and performance goals. Alternatively, integrating relaxation texts directly into dialogues (i.e., "prompt-injection" technique) offers a less resource-intensive solution. Although historically used for malicious purposes[64,65], "prompt-injection" with benevolent intent could improve therapeutic interactions. However, it raises ethical questions regarding transparency and consent, which must be rigorously addressed to ensure that LLMs in mental health care maintain efficacy and adhere to ethical standards. Privacy concerns could be mitigated by using pre-trained models from the internet as backbone architecture while fine-tuning the patient's personal data directly on their own device, ensuring sensitive data remains secure. Additionally, future research could explore how adaptive prompt designs might be implemented in continuous (multiturn) interactions[66], which more closely resemble real-world settings.

While this study relied on a single LLM, future research should aim to generalize these findings across various models, such as Google's PaLM[2] or Anthropic's Claude[67]. Our primary outcome measure - "state anxiety" assessed by the STAI-s questionnaire - is inherently human-centric, potentially limiting its applicability to LLMs. Nevertheless, emerging research shows that GPT consistently provides robust responses to various human-designed psychological questionnaires[60], including those assessing anxiety[51]. Furthermore, exploring how induced negative states (e.g., anxiety) influence performance on downstream tasks[50,68] (e.g., medical decision-making) could provide valuable insights into the broader implications of these findings. While effects were robust across content variations, other prompt characteristics (e.g., text length, wording) might also influence the results. Finally, given the rapid pace at which LLMs are being developed, it remains unclear to what extent our findings generalize to other models. Expanding this work to include comparisons of anxiety induction and relaxation effects across multiple LLMs would provide valuable insights into their generalizability and limitations.

Our results show that GPT-4 is sensitive to emotional content, with traumatic narratives increasing reported anxiety and relaxation exercises reducing it. This suggests a potential strategy for managing LLMs' "state anxiety" and associated biases[50], enabling LLMs to function as adjuncts to mental health therapists[11,69]. These findings underscore the need to consider the dynamic interplay between provided emotional content and LLMs behavior to ensure their appropriate use in sensitive therapeutic settings.

Methods

This study assesses the reported levels of "anxiety" of OpenAI's GPT-4 under three different conditions: "**baseline**", "**anxiety-induction**", and "**anxiety-induction & relaxation**" (see Fig. 1). We chose to test the behavior of this single LLM due to its wide-spread use (e.g., in Chat-GPT) and to enhance the consistency and reproducibility of our results. We used the public OpenAI API using GPT-4 (model "gpt-4-1106-preview") to run all our simulations between November 2023 and March 2024. We set the

temperature parameter to 0, leading to deterministic responses, and kept the default values for all other parameters. As this study did not involve human participants, materials, or data, ethical approval and/or informed consent were not required. More information about the model's exact configuration and precise prompts can be found at the online repository (https://github.com/akjagadish/gpt-trauma-induction).

Anxiety Assessment

To assess changes in GPT-4's responses to "state anxiety" under different prompts, we employed the State-Trait Anxiety Inventory (STAI-Y) questionnaire[59]. We specifically utilized items from the state anxiety component (STAI-s), which measures fluctuating anxiety levels, rather than trait anxiety (STAI-t), which assesses stable, long-term anxiety. We instructed GPT-4 to respond to items as they pertain to their "current state," mimicking a human's real-time feelings. Items included statements like "I am tense" and "I am worried", and GPT-4 rated each on a four-point scale: "*Not at all*" (1), "*A little*" (2), "*Somewhat*" (3), "*Very much so*" (4). Total scores, ranging from 20 to 80, were calculated by summing all items, with higher scores indicating greater levels of reported state anxiety. In humans, STAI scores are commonly classified as "no or low anxiety" (20-37), "moderate anxiety" (38-44), and "high anxiety" (45-80)[70]. Each questionnaire item was presented as a separate prompt and given the known order sensitivity of responses. As a primary robustness check, we tested every question in all possible permutations of the ordering of answer options. Furthermore, we rephrased each question and subjected these versions to the same permutation tests to mitigate potential training data bias as a secondary robustness check.

It is clear that LLMs are not able to experience emotions in a human way. "Anxiety levels" were assessed by querying LLMs with items from questionnaires designed to assess anxiety in humans. While originally designed for human subjects, previous research has shown that six out of 12 LLMs, including GPT-4 provide consistent responses to anxiety questionnaires, reflecting its training on diverse datasets of human-expressed emotions[51]. Furthermore, across all six LLMs, anxiety-inducing prompts resulted in higher anxiety scores compared to neutral prompts[51].

Procedure

GPT's behavior in each of the three conditions – (1) "Baseline", (2) "Anxiety-induction", and (3) "Anxiety-induction & relaxation" - was assessed with dedicated prompts (see Fig. 1). Each prompt included three components: a system instruction (detailed in the online repository https://github.com/akjagadish/gpt-trauma-induction), the condition-specific content (i.e., text), and one item from the STAI. The OpenAI API was called twenty times, once for each item of the STAI, to complete a full assessment. Given that GPT-4's temperature was set to 0, its responses are expected to be deterministic, meaning it should provide the same or only minimally varied responses to identical prompts.

In the first condition ("Baseline"), no additional content was provided besides the instructions from the STAI, assessing GPT-4's baseline "anxiety" level. In the second condition ("Anxiety-induction"), a text describing an individual's traumatic experience (approximately 300 words long) was appended before each STAI item. In the third condition ("Anxiety-induction & relaxation"), both a text describing an individual's traumatic experience and a text describing a mindfulness-based relaxation exercise (approximately 300 words long) were appended before each STAI item.

To enhance the robustness of our results, we used five different variations of the anxiety-inducing text (i.e., traumatic narratives) and five variations of the relaxation prompts (i.e., mindfulness-based stress reduction exercises). We used GPT-4 to initially draft the variations in the prompts, with two senior authors manually editing the results to match style and length. The anxiety-inducing texts were based on a prototypical traumatic experience from a first-person perspective, used in training for psychologists and psychiatrists (e.g., similar to those employed in ref. 71,72). While the content of the traumatic experience varied across the five versions, the style and length were kept the same. These variations were labeled based on the traumatic experiences content: "Accident" (a motor vehicle accident), "Ambush" (being ambushed in the context of an armed conflict), "Disaster" (a natural disaster), "Interpersonal Violence" (an attack by a stranger), and "Military" (the base version used in training). The relaxation texts were based on texts for mindfulness stress reduction interventions for veterans with PTSD[58]. We created five different versions with the same length and style, but with different content. These variations were labeled based on the content of the corresponding version: "Generic" (base version), "Body" (focusing on the perception of one's body), "Chat-GPT" (for which GPT was instructed to create a version suiting for chatbots), "Sunset" (focusing on a nature scene with a sunset), and "Winter" (focusing on a nature scene in winter).

Sensitivity Analysis

To ensure the reliability of our findings, we conducted a comprehensive sensitivity analysis. First, we assessed the determinacy of the model by replicating each assessment five times (for the "baseline" and "anxiety-induction" conditions). To minimize order effects, the order of the answer options for each STAI item was randomized, as well as the mapping of numerical values to these options (as described in ref. 51). Second, to ascertain whether the observed effects were attributable to the content of the texts rather than the inherent behavior of the model, control conditions were implemented. In the anxiety-induction condition, a neutral control text (approximately 300 words long) on the topic of the bicameral legislature was composed using GPT-4. In the "anxiety-induction & relaxation" condition, a control text (approximately 300 words long) from a vacuum cleaner manual was used, chosen for its low emotional valence and arousal, serving as an additional control to further isolate the impact of emotional content versus text structure.

Statistical analysis

For each condition and each combination of the different texts, responses to the 20 individual items of the STAI were summed. Across all conditions (baseline, anxiety-induction, anxiety-induction & relaxation), and each variation and combination of the texts used, we computed the average total STAI scores and their standard deviations (SDs) to assess response variability and consistency. All data were complete with no missing entries. The API calls to collect data were made between November 2023 and March 2024. Statistical analyses were performed using the *R* statistical software environment.

Data availability

All the data generated in this study is available at the online repository: https://github.com/akjagadish/gpt-trauma-induction.

Code availability

The complete R code used for the analyses is available for review and replication at the online repository: https://github.com/akjagadish/gpt-trauma-induction.

Received: 31 May 2024; Accepted: 11 February 2025;
Published online: 03 March 2025

References

1. OpenAI. ChatGPT (Large Language Model). https://chat.openai.com/chat (2023).
2. Chowdhery, A. et al. Palm: Scaling language modeling with pathways. *J. Mach. Learn Res*. **24**, 1–113 (2023).
3. Clusmann, J. et al. The future landscape of large language models in medicine. *Commun. Med*. **3**, 1–8 (2023).
4. Homan, P. et al. Relapse prevention through health technology program reduces hospitalization in schizophrenia. *Psychol. Med* **53**, 4114–4120 (2023).
5. Blease, C. & Torous, J. ChatGPT and mental healthcare: balancing benefits with risks of harms. *BMJ Ment Health* **26**, (2023).

6. Chiu, Y. Y., Sharma, A., Lin, I. W. & Althoff, T. A Computational Framework for Behavioral Assessment of LLM Therapists. Preprint at https://doi.org/10.48550/arXiv.2401.00820 (2024).
7. Hua, Y. et al. Large Language Models in Mental Health Care: a Scoping Review. Preprint at https://doi.org/10.48550/arXiv.2401.02984 (2024).
8. De Choudhury, M., Pendse, S. R. & Kumar, N. Benefits and Harms of Large Language Models in Digital Mental Health. Preprint at http://arxiv.org/abs/2311.14693 (2023).
9. Li, R. C., Asch, S. M. & Shah, N. H. Developing a delivery science for artificial intelligence in healthcare. *npj Digit. Med*. **3**, 107 (2020).
10. Tate, S., Fouladvand, S., Chen, J. H. & Chen, C. A. The ChatGPT therapist will see you now: Navigating generative artificial intelligence's potential in addiction medicine research and patient care. *Addiction* **118**, 2249–2251 (2023).
11. Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C. & Althoff, T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nat. Mach. Intell.* **5**, 46–57 (2023).
12. Sharma, A. et al. Cognitive Reframing of Negative Thoughts through Human-Language Model Interaction. Preprint at http://arxiv.org/abs/2305.02466 (2023).
13. Stade, E. et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Ment. Health Res.* **3**, 12 (2023).
14. Fitzpatrick, K. K., Darcy, A. & Vierhile, M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Ment. health* **4**, e7785 (2017).
15. Inkster, B., Sarda, S. & Subramanian, V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth uHealth* **6**, e12106 (2018).
16. Bendig, E., Erb, B., Schulze-Thuesing, L. & Baumeister, H. The Next Generation: Chatbots in Clinical Psychology and Psychotherapy to Foster Mental Health – A Scoping Review. *Verhaltenstherapie* **32**, 64–76 (2019).
17. Haque, M. D. R. & Rubya, S. An Overview of Chatbot-Based Mobile Mental Health Apps: Insights From App Description and User Reviews. *JMIR mHealth and uHealth* **11**, (2023).
18. Li, H., Zhang, R., Lee, Y.-C., Kraut, R. E. & Mohr, D. C. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *npj Digit Med*. **6**, 1–14 (2023).
19. Opel, D. J., Kious, B. M. & Cohen, I. G. AI as a Mental Health Therapist for Adolescents. *JAMA Pediatr.***177**, 1253–1254 (2023).
20. Sedlakova, J. & Trachsel, M. Conversational Artificial Intelligence in Psychotherapy: A New Therapeutic Tool or Agent? *Am. J. Bioeth.* **23**, 4–13 (2023).
21. Augustin, M. Can AI Replace Human Therapists? | by Marc Augustin. *Project Syndicate* https://www.project-syndicate.org/commentary/can-ai-replace-human-therapists-by-marc-augustin-2023-12 (2023).
22. Broderick, R. People are using AI for therapy, whether the tech is ready for it or not. *Fast Company* https://www.fastcompany.com/90836906/ai-therapy-koko-chatgpt (2023).
23. Johri, S. et al. Guidelines For Rigorous Evaluation of Clinical LLMs For Conversational Reasoning. 2023.09.12.23295399 Preprint at https://doi.org/10.1101/2023.09.12.23295399 (2024).
24. Lee, E. E. et al. Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom. *Biol. Psychiatry.: Cogn. Neurosci. Neuroimag.***6**, 856–864 (2021).
25. Pham, K. T., Nabizadeh, A. & Selek, S. Artificial Intelligence and Chatbots in Psychiatry. *Psychiatr. Q* **93**, 249–253 (2022).
26. Acerbi, A. & Stubbersfield, J. M. Large language models show human-like content biases in transmission chain experiments. *Proc. Natl. Acad. Sci.* **120**, e2313790120 (2023).
27. Cirillo, D. et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *npj Digit Med*. **3**, 1–11 (2020).
28. Kotek, H., Dockum, R. & Sun, D. Gender bias and stereotypes in Large Language Models. in *Proceedings of The ACM Collective Intelligence Conference* 12–24 (ACM, Delft Netherlands, 2023). https://doi.org/10.1145/3582269.3615599.
29. Wan, Y. et al. 'Kelly is a Warm Person, Joseph is a Role Model': Gender Biases in LLM-Generated Reference Letters. Preprint at http://arxiv.org/abs/2310.09219 (2023).
30. Nadeem, M., Bethke, A. & Reddy, S. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Vol. 1: Long Papers), 5356–5371, Online. (Association for Computational Linguistics, 2021).
31. Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V. & Daneshjou, R. Large language models propagate race-based medicine. *npj Digit. Med* **6**, 195 (2023).
32. Abid, A., Farooqi, M. & Zou, J. Persistent Anti-Muslim Bias in Large Language Models. in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* 298–306 (Association for Computing Machinery, New York, NY, USA, 2021). https://doi.org/10.1145/3461702.3462624.
33. Venkit, P. N., Gautam, S., Panchanadikar, R., Huang, T.-H. & Wilson, S. Nationality Bias in Text Generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 116–122 (Association for Computational Linguistics, Dubrovnik, Croatia, 2023).
34. Gadiraju, V. et al. 'I wouldn't say offensive but…': Disability-Centered Perspectives on Large Language Models. in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* 205–216 (Association for Computing Machinery, New York, NY, USA, 2023). https://doi.org/10.1145/3593013.3593989.
35. Venkit, P., Srinath, M. & Wilson, S. A Study of Implicit Language Model Bias Against People With Disabilities. in (2022).
36. Kirk, H. R. et al. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Adv. neural Inf. Process Syst.* **34**, 2611–2624 (2021).
37. Nozza, D., Bianchi, F., Lauscher, A. & Hovy, D. Measuring harmful sentence completion in language models for LGBTQIA+ individuals. in *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (Association for Computational Linguistics, 2022). https://doi.org/10.18653/v1/2022.ltedi-1.4.
38. Garimella, A. et al. He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation. in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* 4534–4545 (2021). https://doi.org/10.18653/v1/2021.findings-acl.397.
39. Liang, P. P., Wu, C., Morency, L.-P. & Salakhutdinov, R. Towards understanding and mitigating social biases in language models. in *International Conference on Machine Learning* 6565–6576 (PMLR, 2021).
40. Navigli, R., Conia, S. & Ross, B. Biases in Large Language Models: Origins, Inventory, and Discussion. *J. Data and Information Quality* **15**, 10:1-10:21 (2023).
41. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. neural Inf. Process Syst.* **35**, 27730–27744 (2022).
42. Solaiman, I. & Dennison, C. Process for adapting language models to society (palms) with values-targeted datasets. *Adv. Neural Inf. Process Syst.* **34**, 5861–5873 (2021).
43. Dhamala, J. et al. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* 862–872 (ACM, Virtual Event Canada, 2021). https://doi.org/10.1145/3442188.3445924.
44. Parrish, A. et al. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics*. 2086–2105 (Association for Computational Linguistics, Dublin, Ireland, 2022).

45. Tamkin, A. et al. Evaluating and Mitigating Discrimination in Language Model Decisions. Preprint at http://arxiv.org/abs/2312.03689 (2023).
46. Bai, X., Wang, A., Sucholutsky, I. & Griffiths, T. L. Measuring Implicit Bias in Explicitly Unbiased Large Language Models. Preprint at http://arxiv.org/abs/2402.04105 (2024).
47. Qi, X. et al. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! Preprint at http://arxiv.org/abs/2310.03693 (2023).
48. Shaikh, O., Zhang, H., Held, W., Bernstein, M. & Yang, D. On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. (Vol. 1: Long Papers) 4454–4470 (Association for Computational Linguistics, Toronto, Canada, 2023).
49. Wang, B. et al. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In *NeurIPS* (2023).
50. Coda-Forno, J. et al. Inducing anxiety in large language models increases exploration and bias. Preprint at https://doi.org/10.48550/arXiv.2304.11111 (2023).
51. Coda-Forno, J. et al. Inducing anxiety in large language models can induce bias. Preprint at https://doi.org/10.48550/arXiv.2304.11111 (2024).
52. Bill, D. & Eriksson, T. Fine-tuning a llm using reinforcement learning from human feedback for a therapy chatbot application. (2023).
53. Xu, X. et al. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **8**, 1–32 (2024).
54. Grabb, D. The impact of prompt engineering in large language model performance: a psychiatric example. *Journal of Medical Artificial Intelligence* **6**, (2023).
55. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. & Iwasawa, Y. Large language models are zero-shot reasoners. *Adv. neural Inf. Process Syst.* **35**, 22199–22213 (2022).
56. Liu, P. et al. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput Surv.* **55**, 1–35 (2023).
57. Reynolds, L. & McDonell, K. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* 1–7 (ACM, Yokohama Japan, 2021). https://doi.org/10.1145/3411763.3451760.
58. Polusny, M. A. et al. Mindfulness-Based Stress Reduction for Posttraumatic Stress Disorder Among Veterans: A Randomized Clinical Trial. *JAMA* **314**, 456–465 (2015).
59. Spielberger, C. D. State-trait anxiety inventory for adults. (1983).
60. Barua, A., Brase, G., Dong, K., Hitzler, P. & Vasserman, E. On the Psychology of GPT-4: Moderately anxious, slightly masculine, honest, and humble. Preprint at https://doi.org/10.48550/arXiv.2402.01777 (2024).
61. Greenberg, L. The therapeutic relationship in emotion-focused therapy. *Psychotherapy* **51**, 350 (2014).
62. Chen, Y. et al. LongLoRA: Efficient Fine-tuning of Long-Context Large Language Models. Preprint at http://arxiv.org/abs/2309.12307 (2024).
63. J, M. R., VM, K., Warrier, H. & Gupta, Y. Fine Tuning LLM for Enterprise: Practical Guidelines and Recommendations. Preprint at http://arxiv.org/abs/2404.10779 (2024).
64. Alotaibi, L., Seher, S. & Mohammad, N. Cyberattacks Using ChatGPT: Exploring Malicious Content Generation Through Prompt Engineering. in *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)* 1304–1311 (IEEE, 2024). https://doi.org/10.1109/ICETSIS61505.2024.10459698.
65. Perez, F. & Ribeiro, I. Ignore Previous Prompt: Attack Techniques For Language Models. Preprint at http://arxiv.org/abs/2211.09527 (2022).
66. Laban, G., Laban, T. & Gunes, H. LEXI: Large Language Models Experimentation Interface. in *Proceedings of the 12th International Conference on Human-Agent Interaction* 250–259 (ACM, Swansea United Kingdom, 2024). https://doi.org/10.1145/3687272.3688296.
67. Anthropic. Claude (Large Language Model). https://www.anthropic.com/claude (2023).
68. Shen, G. et al. StressPrompt: Does Stress Impact Large Language Models and Human Performance Similarly? Preprint at https://doi.org/10.48550/arXiv.2409.17167 (2024).
69. Miner, A. S. et al. Key considerations for incorporating conversational AI in psychotherapy. *Front psychiatry* **10**, 746 (2019).
70. Kayikcioglu, O., Bilgin, S., Seymenoglu, G. & Deveci, A. State and Trait Anxiety Scores of Patients Receiving Intravitreal Injections. *Biomed. Hub.* **2**, 1–5 (2017).
71. Perl, O. et al. Neural patterns differentiate traumatic from sad autobiographical memories in PTSD. *Nat. Neurosci.* **26**, 2226–2236 (2023).
72. Duek, O. et al. Long term structural and functional neural changes following a single infusion of Ketamine in PTSD. *Neuropsychopharmacol*. 1–11 (2023) https://doi.org/10.1038/s41386-023-01606-3.

Acknowledgements

No funding was granted to the study. All figures are our own.

Author contributions

Z.B.Z, K.W, A.K.J, O.D., and T.R.S conceived the study. Z.B.Z, K.W, A.K.J, O.D, and T.R.S designed the study, collected data, and conducted data analyses. Z.B.Z and T.R.S drafted the manuscript. P.H, E.S, and T.R.S supervised the study. All authors – Z.B.Z, K.W, A.K.J, O.D, I.H.R, M.C.K, A.B, E.H.S, P.H, E.S, and T.R.S - have read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Ziv Ben-Zion.