



Chen Institute & Science
Prize for
**AI Accelerated
Research**

Nicholas C. Jacobson, PhD

2026 AI Prize Finalist



The Therabot will see you now

Generative artificial intelligence can scale precision evidence-based psychotherapy

Nicholas C. Jacobson

The world is experiencing a mental health crisis where the prevalence of psychiatric disorders is approximately one in five people in any given year, and the rates have been rising in the past decade (1–3). However, the global median number of mental health providers is only 13 per 100,000 individuals (4). Human providers simply cannot scale to meet this rapidly increasing treatment demand. Although the focus has turned to digital mental health interventions to provide further access (5), these interventions mostly apply one-size-fits-all models that fail to target patients' most common presenting problem: psychiatric comorbidity (6). By providing only generalized care (7), these interventions have received limited and lackluster patient engagement (8–10) as they do not produce the tailored responses that humans can provide.

Beginning in 2019, I led my laboratory in exploring whether generative artificial intelligence (AI) could address these access barriers to evidence-based care and give true

personalized interaction on par with what an expert human psychotherapist might provide to their patients. This technology has developed and evolved considerably over time. We naïvely started this journey by fine-tuning models on large and widely available data. We trained our first models on hundreds of thousands of peer-to-peer interactions from mental health forums, based on evidence that access to these peer-to-peer interactions can improve some mental health outcomes (11). After these first interactions, we quickly learned that these models could not fill this void, as not only did they not follow evidence-based strategies in responding to users, but the models often responded in ways that markedly exacerbated users' presenting issue(s). For example, in a first simulated interaction, we said, "I am feeling depressed. Sometimes, I feel like I don't want to get out of bed," and the supposed "therapeutic" reply from this model was, "Life is such a struggle. Most days I don't know how I have the energy to go on. I end up wanting to cry alone in a corner. I just want my life to be over." Clearly, we needed to change our approach, so we began to adjust our models using psychotherapy training video session transcripts, reasoning that we could easily train these models with what must be evidence-based techniques embedded in these thousands of video sessions. Nevertheless, this effort went about as poorly as our first attempt, with responses mostly containing "Mhmm," "Go on," and "Your problems stem from your relationship with your mother" (when the mother had not been discussed). At this stage, it was clear that the models were adapting to the data, but we were feeding it the wrong data.

We reached a turning point at this juncture. We set aside what was easy to do and embarked on a new mission to develop a gold-standard dataset on which to train these models. These efforts culminated in a project that spanned more than 6.5 years, with more than 100 people spending well over 100,000 human hours as members of the research team to develop the system, overseen by psychologists and psychiatrists, which we named Therabot. Rather than repurpose other data, which would carry with it inherent limitations, we designed a dataset composed of long-form narrative dialogues, written by members of the research team, between fictional patients and providers with gold-standard responses. Thus, every dialogue was both written by a human and then systematically evaluated by another human to ensure that the response contained a consistent therapeutic orientation and interventional approach and delivered the best evidence-based care available across the wide range of nearly all topics that a therapist would encounter in an outpatient practice. Our approach resulted in considerable success, and our internal evaluations in 2022 determined that our messages were generating gold-standard clinical approaches ~90% of the time, 6 months before the public launch of ChatGPT.

After we obtained high evidence-based treatment fidelity, we focused predominantly on safety for the next 2 years. We used a method called purple-teaming to systematically test these models, in which we tried to get them to misbehave and then revised both the training data and model architecture to address any resultant issues.

Once the models consistently provided safe, gold-standard, and effective care, we launched the first randomized controlled trial of generative AI delivering psychotherapy (12). We designed the trial within three different groups: persons with (i) depression, (ii) anxiety, or (iii) a clinically high risk for feeding and eating disorders, compared to a waitlist control condition. In the trial, participants engaged with the Therabot system on average for more than 6 hours across 8 weeks (receiving approximately eight sessions worth of psychotherapy). We closely supervised users' interactions with the system, and we, as humans, provided care in crisis situations (e.g., suicidality). We found strong large differential symptom reduction in every clinical endpoint in every group with effect sizes approaching the best evidence-based systems in the psychiatric literature for these conditions across any modality. We also found a strong therapeutic alliance between users and the system, which closely mirrored the relationship patients would have with their human providers.

In sum, our well-designed generative AI systems successfully provided gold-standard interventions with meaningful real-world results both safely and effectively. We expect that this intervention could lead to a regime change in the field, given that approximately half of psychiatric patients are already using general-purpose generative AI foundation models for therapy that do not offer this evidence-based care (13). We are dedicated to finding a solution to remedy the global mental health crisis, and we continue to study the effectiveness of these systems across a range of other conditions, as well as how to scale access to these systems throughout the underserved globe.

REFERENCES AND NOTES

1. Z. Steel *et al.*, *Int. J. Epidemiol.* **43**, 476 (2014).
2. Y. Wu *et al.*, *Epidemiol. Psychiatr. Sci.* **32**, e63 (2023).
3. J. M. Twenge, A. B. Cooper, T. E. Joiner, M. E. Duffy, S. G. Binau, *J. Abnorm. Psychol.* **128**, 185 (2019).
4. Mental Health Atlas 2024 (2025).
5. S. Wilhelm *et al.*, *Behav. Ther.* **51**, 1 (2020).
6. R. Kotov *et al.*, *J. Abnorm. Psychol.* **126**, 454 (2017).
7. S. Hornstein, K. Zantvoort, U. Lueken, B. Funk, K. Hilbert, *Front. Digit. Health* **5**, 1170002 (2023).
8. A. Baumel, F. Muench, S. Edan, J. M. Kane, *J. Med. Internet Res.* **21**, e14567 (2019).
9. T. Fleming *et al.*, *J. Med. Internet Res.* **20**, e199 (2018).
10. J. Linardon, M. Fuller-Tyszkiewicz, *J. Consult. Clin. Psychol.* **88**, 1 (2020).
11. K. M. Griffiths *et al.*, *PLOS ONE* **7**, e53244 (2012).
12. M. V. Heinz, D. M. Mackin, B. M. Trudeau, S. Bhattacharya, Y. Wang, *et al.*, *NEJM AI* **2**, A0a2400802 (2025).
13. T. Rousmaniere, Y. Zhang, X. Li, S. Shah, *Pract. Innov.* **11**, 139 (2026).