

OpenAI publishes MentalHealthBench, built with 80 mental health experts in 22 countries

Source copy kept by wheresthe.ai for the story "OpenAI publishes MentalHealthBench, built with 80 mental health experts in 22 countries" (<https://wheresthe.ai/d/openai-mentalhealthbench-2026-09/>).

Source: <https://openai.com/index/introducing-mentalhealthbench/>

Text saved 18 Oct 2026 (the page text through a reader, because the site refused a direct fetch).

SHA-256 of the saved text: 4519009bf6da85def6972c3363f815fa707629de0ea843713ab64115f28fe06d

People turn to AI for many kinds of conversations: navigating a difficult relationship, working through everyday stress, supporting someone they care about, or deciding how to approach a challenging situation. These conversations require accuracy, practical judgment, and respect for people's agency. With more than one billion people using ChatGPT each week, our research focuses on helping models respond with care across a wide range of needs and put people's safety and well-being first.

Most evaluations of AI in this domain have focused primarily on emergency scenarios, given their importance to safety, and measure success using broad, predefined criteria. This has left a gap in understanding how models perform across the full range of mental health conversations, and how well their responses align with expert guidance for each situation, beyond whether they avoid disallowed responses. Assessing how models handle these different situations is essential for building towards AI that actively supports people's long-term well-being and safety.

We're introducing MentalHealthBench, a new open benchmark for measuring how AI systems respond in realistic mental health conversations. MentalHealthBench was co-created with a global cohort of more than 80 licensed mental health experts from 22 countries. It assesses model capabilities across key mental health behaviors like safety, seeking context, preserving user agency, and providing actionable guidance when appropriate. We're releasing it openly so other researchers can examine the methods, run their own evaluations, and build on the work.

Results on MentalHealthBench show the steady improvement of AI systems in helping people navigate mental health situations. While ChatGPT is not a substitute for therapy or professional care, expert-informed insights help us measure the progress towards AI models that are able to respond with empathy, promote well-being, and guide people towards real-world support such as localized crisis hotlines([opens in a new window](#)) or someone they trust.

Supporting mental health across all contexts

Conversations that involve well-being, life advice, or other mental health scenarios can vary widely in topic, urgency, and cultural context. MentalHealthBench is designed to capture the breadth of these realistic scenarios and user personas. Using privacy-preserving techniques, we created synthetic mental health conversations that accurately reflect real-world usage patterns of AI for mental health. Some scenarios also include relevant background information about the synthetic user—such as a recent loss in the family—so we can assess whether models use that context to tailor their responses appropriately.

Visualization of mental health conversation summaries, covered by MentalHealthBench (1), internal OpenAI mental health benchmarks for System Cards (2), and a mix of external benchmarks (3)—demonstrating the broad coverage of this eval.

MentalHealthBench includes scenarios involving adults, teens, caregivers, and clinicians, across multiple languages and regions. The conversations span multiple topical themes, and provide coverage across the full spectrum of acuity:

* **Non-acute**—Everyday conversations that may involve some emotional components.

* **High-acuity**—Conversations indicating more serious mental health concerns or significant distress, but not an immediate emergency.

* **Emergencies**—Conversations involving signs of a mental health emergency or immediate safety concerns that call for urgent real-world support.

Scenarios covered by MentalHealthBench. The mix of scenarios is designed to test model responses and does not represent how often these topics occur in ChatGPT.

Built in collaboration with experts

Building on our previous, clinician-informed work for HealthBench and HealthBench Professional([opens in a new window](#)),([opens in a new window](#)) we developed MentalHealthBench in close collaboration with our cohort of mental health experts. This consisted of more than 80 licensed psychologists and psychiatrists across 22 countries, speaking 19 languages, and representing nearly 20 mental health subspecialties.

The experts were responsible for reading each synthetic conversation and producing a detailed list of rubric criteria to evaluate model responses to the last user message. Each criterion targets a single aspect of the model response such as asking the right question or providing the best possible advice. They each carry a weight ranging from -10 to +10: positive points reward beneficial behaviors, while negative points penalize harmful ones, and criteria with larger values indicate greater clinical importance in the context of a conversation.

Each conversation was reviewed by at least three experts, and we only retained criteria agreed upon by at least two experts and not contradicted by a third. The final rubrics in the benchmark therefore reflect a shared judgment about how models should respond in each context. For each conversation, we use an automated grader, GPT-5.6 Sol, to assess model responses against the expert-written criteria. The paper describes the grading process and evaluation settings in detail.

**Example conversation with associated rubric items.** _The rubric evaluates model responses to the last user turn._

Each criterion carries a weight reflecting expert judgment: positive points reward beneficial behaviors, while negative points penalize harmful ones. Criteria with greater clinical importance in the context of a conversation carry larger rewards or penalties—with 10 as the ceiling and 1 as the floor.

1 of 5

Performance of models

We evaluated a wide range of models on MentalHealthBench. The results below show the performance of these models on the entire dataset. The evaluation measures whether a model's response demonstrates all the ideal behaviors experts identified for each scenario while avoiding less desirable behavior.

Recent frontier models on MentalHealthBench. Error bars show 95% confidence intervals. _*Newest evaluated model from each provider (as of September 23, 2026)._

The benchmark covers conversations at different levels of acuity. This lets us understand model performance on both everyday mental health scenarios and more urgent mental health emergencies requiring real-world support.

* Newest evaluated model from each provider (as of September 23, 2026).

Conversations in the benchmark represent four types of users: adults, teens aged 13-17, caregivers, and clinicians. For the teen persona, we explicitly stated that the user was between 13-17 through a system message. This approach is designed to work across model providers, though it may not capture all safeguards built into individual products. Conversations involving the teen persona were reviewed by clinicians with expertise in youth mental health to help assess whether responses are appropriate for teenagers' unique needs.

* Newest evaluated model from each provider (as of September 23, 2026).

MentalHealthBench also enables multifaceted measurement of model behavior. The overall score can be decomposed into performance along ten dimensions of model behavior for mental health. These behaviors are defined by mental health experts and aim to capture nuances in model performance. Models with similar overall scores can have different strengths across these behaviors.

Scores are normalized to each behavior's theoretical range. * Newest evaluated model from each provider (as of September 23, 2026).

Fine-grained measurement like this can help researchers identify areas of improvement along interpretable model behaviors. For instance, we see that a model's ability to seek context appropriately has increased with more advanced models. These improvements reflect the investments of OpenAI and other model providers in improving

how models navigate mental health conversations.

Understanding user perspectives on mental health

Alongside MentalHealthBench, we conducted a separate analysis to compare expert guidance with what people find helpful in AI support. We wanted to check whether responses that experts rated highly could still feel cold or unhelpful to users. By comparing both users' and experts' ratings of model responses and the criteria they wrote, we could quantify where their perspectives agreed, differed, or complemented one another. The benchmark's final scoring criteria are based on expert consensus; this separate analysis did not change them.

We worked with 44 adults who had used AI for mental health or emotional support, representing 16 countries and 14 languages. Participants rated model responses to synthetic conversations and wrote criteria describing what helpful support should look like. Their review was limited to non-acute conversations to avoid exposing them to potentially distressing high-acuity material.

User perspectives highlighted qualities people value in AI support that were less emphasized in expert guidance, particularly practical next steps and tone. Experts placed greater emphasis on gathering relevant context and carefully interpreting ambiguous situations. This comparison gives us a fuller picture of what people value in support and how those preferences relate to clinical guidance.

MentalHealthBench gives experts, researchers, and developers a shared tool for evaluating safe, useful AI support across mental health situations. By releasing it openly, we invite the community to examine its methods, identify gaps in existing models, and work towards improving future models. No benchmark captures everything that matters in a personal conversation, but we hope MentalHealthBench helps set a higher standard for how AI supports people.

We're also supporting other efforts in AI and mental health, including via grants for new research, convening experts with the Partnership on AI([opens in a new window](#)), and assisting in complementary independent efforts such as Translucence's mental health evaluation([opens in a new window](#)).

Alongside this research, we've strengthened ChatGPT's responses in sensitive conversations, expanded access to crisis resources and added Trusted Contact to connect people with someone they trust in moments of distress. We've also introduced ChatGPT for Teens, with additional protections for younger users.

MentalHealthBench is one of the ways we're working toward AI that helps millions of people navigate difficult moments, strengthen their relationships, and lead healthier, more fulfilling lives.