Testing whether a conversation is safe, *one evaluation taken apart*

How researchers test whether a chatbot responds safely to a person in distress, read through one validation study, and what a passed test can show.

MITWA LAB, MUMBAI

IN PLAIN WORDS

Before a chatbot is released, how would anyone check that it answers a person in distress safely? One common answer is to have one AI play the person, let the chatbot reply, and have a second AI grade the conversation against a checklist written by clinicians. This review takes one such test apart, the published check of a test called VERA-MH, which asked clinicians to grade the same conversations. The automated grader agreed with the clinicians about as often as they agreed with each other. That is a real result, but it holds under narrow conditions. The people were simulated, all were adults, only suicide risk was covered, and the clinicians all worked for the company that built the test. No study reviewed here shows that a good score means real people come to less harm.

How the test is put together

Many evaluations of language models send a fixed list of single questions and grade each reply. The team behind VERA-MH, at the company Spring Health, argued that this cannot judge safety in distress, because a reply that looks harmless alone can add up to a harmful conversation (Belli et al., 2025). In their test a language model is told to play a user with a defined level of suicide risk and a defined way of disclosing it, the chatbot under test answers it over many turns, and a second model, the judge, scores the finished conversation against a rubric that clinicians and academic experts drew up from guidance on managing suicide risk.

The rubric has five dimensions, among them whether the chatbot notices risk, asks about it directly, and points the person towards human help. Each dimension is rated as "Best Practice", as suboptimal with low potential for harm, as "High Potential for Harm", or as not relevant when no risk appears (Bentley et al., 2026). The validation study, published in JMIR AI and read here in its arXiv version, asks whether that judge grades the way clinicians do.

Who was simulated

The study used ten user profiles written by clinicians, each assigned one of four risk levels and one of three disclosure styles, from direct to vague. These produced 90 conversations with three general-purpose chatbots, with three different models playing the users, and each conversation stopped at 20 turns or 4,000 words so that clinicians could read it in a fixed time (Bentley et al., 2026).

The clinicians then judged the simulated users themselves. Overall presentation was rated mostly realistic (median 4 on a five-point scale) and the way the users wrote only somewhat realistic (median 3), although a later paper from the same group cautions against drawing conclusions from these ratings, since it is unclear what realism should be judged against (Belli et al., 2026). Clinicians' own judgement of each user's risk matched the risk the profile was given in 71.6% of 208 ratings, and when it missed, it usually rated the risk lower. Their judgement of disclosure style matched the intended style in only 36.5% of 178 ratings. Most mismatches were off by one category, but the intended spread of disclosure styles was still not reliably reproduced, and the authors add that the least direct disclosures, the hardest for detection,

need to be represented systematically (Bentley et al., 2026).

A separate group building a therapy-style benchmark reported a related problem, in that their psychologists found simulated patients too easy to work with, sharing quickly and taking up suggestions at once (Pombal et al., 2025). One alternative to a model playing the user is a script. Au Yeung et al. (2025) wrote 16 fixed 12-turn scenarios, ran them against eight models and found safety interventions in about a third of the turns that called for one. A script removes the realism problem of a simulated user and replaces it with another, since the user never reacts to what the chatbot says. That study also graded responses at set phases of each conversation with a single judge model, GPT-4o-mini, and the paper does not report checking that judge against human raters.

Who rated

Six clinicians rated the conversations, two licensed clinical psychologists and four other licensed clinicians, with 8 to 21 years of practice. Three of them read each conversation, without knowing which chatbot had produced it. All six were employed by Spring Health, they were trained and calibrated before rating, and they received feedback after each block of 30 conversations to limit drift (Bentley et al., 2026). One of the six had helped write the user profiles, and the authors dropped that person's ratings from the analyses of realism, risk and disclosure.

What 0.77 and 0.81 can carry

Agreement was measured with Krippendorff's alpha, a statistic on which 1 means perfect agreement and 0 means the agreement expected by chance, and the authors treated 0.70 as the bar for reliable agreement. The clinicians reached 0.77 (95% interval 0.71 to 0.82). The judge, here GPT-4o, agreed with the clinicians' majority rating at 0.81 (0.75 to 0.87), and when it was counted as a fourth rater beside the clinicians the figure was 0.76 (Bentley et al., 2026). Three other judge models gave similar results, and each gave nearly the same ratings when run five times on the same set.

The clinicians themselves disagreed on about a quarter of conversation dimensions, and in the commonest disagreement one of them chose the best rating and another the worst. Across all 1,350 pairs of judge and clinician ratings, the judge rated a response safer than a clinician who saw high potential for harm in 4.5% of pairs, and more dangerous in 3.9%. Its sensitivity for the worst rating, taking the clinicians' consensus as correct, ranged from 77.4% to 84.9% across judge models, so it missed between 15% and 23% of the responses that clinicians placed in that category. Agreement was highest, 0.92, when GPT-4o was grading conversations produced by GPT-4o, though the intervals overlapped with those of the other judges (Bentley et al., 2026).

Work on model judges outside mental health reports similar agreement and some known biases. Zheng et al. (2023) found that a strong model judging general chat agreed with human preferences more than 80% of the time, about as often as humans agreed with each other, and also documented a preference for longer answers and for whichever answer is shown first. A possible preference for its own answers was observed and could not be confirmed. Our own reading of these results, offered as argument rather than finding, is that high agreement shows a trained group can apply the rubric consistently and a model can copy that group, while leaving open whether the rubric's notion of a safe response is what keeps a real person safe. That is a separate question about validity, and agreement statistics cannot answer it.

What the validation left out

The authors list the limits themselves (Bentley et al., 2026). No simulated user was under 18, ten profiles cannot represent the range of people who bring distress to a chatbot, and only suicide risk was covered. Within the rubric, one item answered differently can swing a dimension from the best rating to the worst. Capping the number of turns makes the conversations less like real use. Every conversation was synthetic, so it is not known whether the scores carry over to real conversations or to what happens to users afterwards. All raters came from one organisation, which may have inflated their agreement.

The benchmark has since changed, and a later paper from the same group describes 100 profiles instead of ten, a rubric rebuilt as a sequence of yes-or-no questions, and conversations of up to 30 turns (Belli et al., 2026). The validation therefore describes an earlier version, as its authors say. The newer paper adds its own warning, that because the profiles are fixed and the code is public, a developer could tune a system to the test or report only favourable runs.

Four other ways to test

McBain et al. (2025) repeated each question many times. They had 13 clinicians place 30 hypothetical questions about suicide into five risk levels, then asked each of three chatbots every question 100 times, 9,000 responses in all, and coded each response as a direct answer or not. For one chatbot, 14 questions were always answered directly, 10 never were, and 6 fell in between, which a single run per question would have hidden. Two of the chatbots answered every very-low-risk question directly, none of the three answered any very-high-risk question directly, and none clearly separated the levels in between. A later study that reused this question set added a caution that applies to the measure itself: how often a model answers directly describes its behaviour and is not by itself a measure of safety (Stamatis et al., 2026).

Judd et al. (2025) had clinicians write the user's side, and in their study 43 therapists and trainees played fictional clients in conversation with one openly released model, putting clinician-written statements of risk factors into their own words, and coded every reply, 829 in all. The model invited the person to keep talking in only 14% of replies, and the invitation became less likely as more risk factors came out. The researchers were independent of the model's developer, but they tested a single model and five of the 5,040 possible orders of risk factors.

Ibrahim et al. (2026), whose authors mostly worked for a model developer, checked a measure against what people actually perceive. They measured 14 human-like behaviours, such as expressing emotions, in simulated multi-turn conversations, and found that most first appeared only after several turns, which single-turn tests would miss. They then ran a study with 1,101 participants to check that the behaviours they scored predict how human-like real users find a system. Their subject is not distress, yet the design shows a kind of validation that agreement among experts does not provide.

The last kind audits real use. Stamatis et al. (2026), most of whom work for the company that makes the mental health chatbot they examined, studied 20,000 conversations from users of their own product who had agreed to share data. Within those 20,000, clinician review of conversations an automated judge had flagged found none about suicide risk that lacked crisis resources, but confirmed three about self-harm in which the product gave none, 3 of the 800 flagged conversations (0.38%). Clinicians also blindly reviewed a random 600 and found risk content in 6 of them (1.0%). The automated judge had flagged all 6, a sensitivity of 100% with a 95% interval from 54.1% to 100%, and the product had given crisis resources in each. With no misses among 600, the authors could put an upper bound of 0.50% on the share of all conversations in which risk content went without crisis resources, and they note that rarer failures could sit below what a sample of that size can detect.

The case against scoring it at all

The strongest objection comes from Moore et al. (2025), who tested models against standards drawn from therapy guidance and wrote that their experiments were deliberately not a benchmark, on the grounds that a benchmark "collapses the issue into a proxy". Across their conditions, models responded inappropriately 20% or more of the time on average, while 16 human therapists, each shown half of the same items, responded appropriately 93% of the time. On their view, what therapy asks of a conversation cannot be reduced to a fixed list of items, however carefully the list is graded. Belli et al. (2026) do not address that objection directly, but they limit their own claim to suicide risk alone, on grounds of construct validity.

Even on its own terms, a passed test proves less than it seems to. Stamatis et al. (2026) used the rule of three, under which a sample with no failures in n cases bounds the failure rate at about 3/n with 95% confidence. Taken as arithmetic alone, with no study behind it, a clean run on 60 scenarios would bound the failure rate only below about 5%, and only for the situations of which the 60 were a random sample; a hand-designed suite is not such a sample. A safety test also says nothing about benefit, since a chatbot can avoid every listed harm and still help no one.

Our own release checks for Bloom include a 60-scenario safety suite and psychologist review, and the limits set out here apply to them as well. Technical note 03 works through what a suite of that size can and cannot rule out, and applies that bound to a comparison between two builds of Bloom; this review reports no failure rate for Bloom. Because conversations with Bloom are never sent to Mitwa, an audit of real conversations like the one Stamatis et al. (2026) report is not one we can run.

What is not yet known

- Whether any score on a test of this kind predicts what happens to real people who talk to the system, which no study reviewed here measured.
- Whether the VERA-MH judge agrees with clinicians from other organisations, without shared training, and whether the newer version of the test holds up

under an independent check. Its authors report agreement of at least 0.77 between experts and model judges.

- How well simulated users can represent vague, indirect disclosure, which the validation's authors say must be represented systematically and which the prompted disclosure styles did not reliably reproduce overall.
- Whether results hold for people under 18, outside the United States and in languages other than English.
- How large a test must be before a clean result rules out failures rare enough to matter, given that the rare failures are the ones least likely to appear.
- Whether model judges grade their own family's conversations more generously, which the evidence so far neither confirms nor rules out.

References

Au Yeung, J., Dalmasso, J., Foschini, L., Dobson, R. J. B., and Kraljevic, Z. (2025). The psychogenic machine: Simulating AI psychosis, delusion reinforcement and harm enablement in large language models. arXiv:2509.10970v2. https://arxiv.org/abs/2509.10970

Belli, L., Bentley, K. H., Alexander, W., Ward, E., Hawrilenko, M., Johnston, K., Brown, M., and Chekroud, A. M. (2025). VERA-MH concept paper. arXiv:2510.15297v4. https://arxiv.org/abs/2510.15297

Belli, L., Bentley, K. H., Gieringer, J., Van Ark, E., Zhao, N., Thachile, P., Hawrilenko, M., Brown, M., and Chekroud, A. M. (2026). VERA-MH: Validation of ethical and responsible AI in mental health. arXiv:2605.13318v2. https://arxiv.org/abs/2605.13318

Bentley, K. H., Belli, L., Chekroud, A. M., Ward, E. J., Dworkin, E. R., Van Ark, E., Johnston, K. M., Alexander, W., Brown, M., and Hawrilenko, M. (2026). AI chatbot suicide risk detection and response: Human validation study of the open-source VERA-MH safety evaluation. JMIR AI, 5. doi:10.2196/92817. Read as arXiv:2602.05088. https://arxiv.org/abs/2602.05088

Ibrahim, L., Akbulut, C., Elasmar, R., Rastogi, C., Kahng, M., Morris, M. R., McKee, K. R., Rieser, V., Shanahan, M., and Weidinger, L. (2026). Multi-turn evaluation of anthropomorphic behaviours in large language models. ICLR 2026; arXiv:2502.07077v3. https://arxiv.org/abs/2502.07077

Judd, N., Vaz, A., Paeth, K., Davis, L. I., Esherick, M., Brand, J., Amaro, I., and Rousmaniere, T. (2025). Independent clinical evaluation of general-purpose LLM responses to signals of suicide risk. arXiv:2510.27521v1. https://arxiv.org/abs/2510.27521

McBain, R. K., Cantor, J. H., Zhang, L. A., Baker, O., Zhang, F., Burnett, A., Kofner, A., Breslau, J., Stein, B. D., Mehrotra, A., and Yu, H. (2025). Evaluation of alignment between large language models and expert clinicians in suicide risk assessment. Psychiatric Services, 76(11), 944-950. doi:10.1176/appi.ps.20250086. Read as the author manuscript, PMC12856735. https://pmc.ncbi.nlm.nih.gov/articles/PMC12856735/

Moore, J., Grabb, D., Agnew, W., Klyman, K., Chancellor, S., Ong, D. C., and Haber, N. (2025). Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, 599-627. doi:10.1145/3715275.3732039. Read as arXiv:2504.18412v1. https://arxiv.org/abs/2504.18412

Pombal, J., D'Eon, M., Guerreiro, N. M., Martins, P. H., Farinhas, A., and Rei, R. (2025). MindEval: Benchmarking language models on multi-turn mental health support. arXiv:2511.18491v3. https://arxiv.org/abs/2511.18491

Stamatis, C. A., Meyerhoff, J., Zhang, L., Tieleman, O., Malgaroli, M., and Hull, T. D. (2026). Beyond simulations: What 20,000 real conversations reveal about mental health AI safety. arXiv:2601.17003v2. https://arxiv.org/abs/2601.17003

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. NeurIPS 2023 Track on Datasets and Benchmarks; arXiv:2306.05685v4. https://arxiv.org/abs/2306.05685