Open Problems in Chatbots and Mental Health

JARED MOORE, Stanford University
JACY REESE ANTHIS, University of Chicago
WILLIAM AGNEW, Carnegie Mellon University
ANDREA MOCK, Vals AI
KEVIN KLYMAN, Harvard University
SANMI KOYEJO, Stanford University
NICK HABER, Stanford University
STEVIE CHANCELLOR, University of Minnesota
ERIC LIN, Stanford University
RYAN LOUIE*, Vanderbilt University
DESMOND C. ONG*, University of Texas at Austin

General-purpose chatbots are increasingly used for emotional and mental health support. They can be useful, but issues such as sycophancy, reinforcement of delusional beliefs, and poor crisis response create risks that are difficult for researchers to study and evaluate. In this paper, we review the most critical open problems for researchers and policymakers in the use of general-purpose chatbots and their effects on mental health. First, there are tractable problems in the **controlled evaluation** of chatbots: establishing the construct and ecological validity of evaluations, evaluating multi-turn interactions, and connecting evaluations to improvements in chatbot behavior. For example, conventional benchmarks that primarily rely on static datasets cannot establish chatbot safety, and novel simulation methods need validation. Second, researchers must connect their studies of chatbots and mental health to **real-world evidence**. This includes analyzing the chatbot products that AI companies and mental health practitioners have deployed, including their user interfaces and human review and escalation practices for user welfare concerns. However, clinical analysis opportunities afforded by real chat log data sharing are complicated by significant privacy concerns. Finally, we argue that researchers should take an active role in the construction and refinement of evidence-based **policy** through corporate and governmental decisions such as user protections, standards of care, protected access to user chat logs, data sharing between institutions, incident reporting, and post-deployment monitoring. Such behavioral evaluations, studies of real-world outcomes, and data governance depend on each other and must be addressed in tandem. We believe a coordinated research effort would have substantial real-world impact.

Additional Key Words and Phrases: chatbots, mental health, evaluation, AI governance, privacy

1 Introduction

More and more people are using chatbots based on large language models (LLMs) for emotional and mental health support [94]. Many users have positive emotional interactions with chatbots, while others have experienced tragic outcomes, including delusional spirals and worsening mental health crises [58, 104], including suicide [146].

Despite these harms, researchers do not yet know the prevalence or causal effects of chatbot use on mental health. We identify open problems in evaluating and improving the effects of general-purpose chatbots on users' mental health and psychological well-being. This includes clinical concerns—psychiatric symptoms, mental disorders, and crises—and everyday experiences of emotional distress, coping, loneliness, and social connection. These outcomes matter whether users explicitly seek mental health support or encounter them during other uses [155].

*Equal contribution.

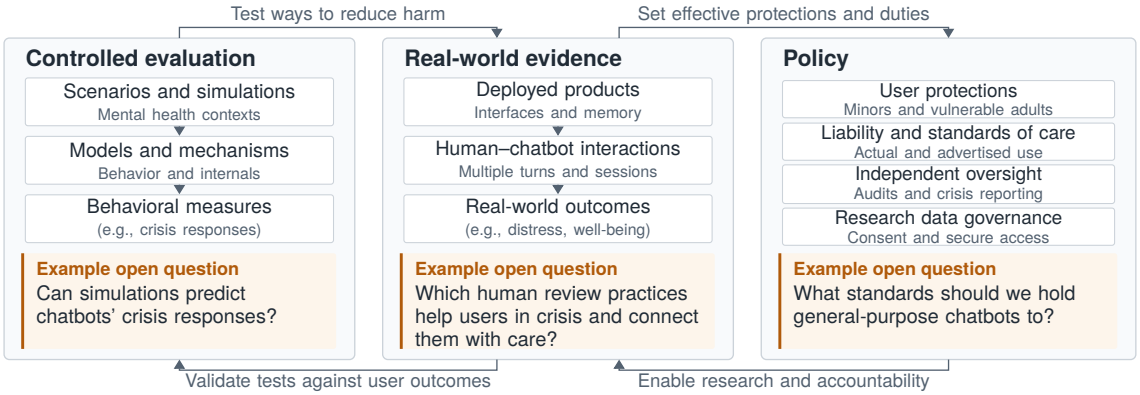


Fig. 1. **We survey problems in controlled evaluation (§4), real-world evidence (§5), and policy (§6).** Addressing these problems may support safer chatbot use and user well-being.

We organize these open problems into three areas (Figure 1): controlled evaluation of chatbots, real-world evidence about deployed products, and policy. First, we argue that chatbot evaluations for mental health must draw on established measurement practices and connect chatbot behavior to real-world user outcomes. Experts must be involved throughout chatbot development and deployment, including the design, execution, and interpretation of evaluations. We cover challenges in tests using predefined scenarios and assessments in our discussion of **controlled evaluation** (§4).

Second, scholars are recognizing the need to collect **real-world evidence** on how chatbots affect people (§5). The full chatbot product, including its interface and memory, affects user interaction and well-being. We examine challenges in testing deployed products and collecting longitudinal interaction data to study these effects. Data from real interactions can also help validate simulations and support clinical research on symptom disclosure and risk signals, but their use raises substantial privacy concerns.

Companies are already improving their products, and legislators are passing bills, to ameliorate chatbots' psychological risks [24, 30, 120, 126]. Indeed, as we describe in Section 6, **policy** can translate evidence on chatbots' mental health effects into user protections and standards of care. However, many questions remain about effective standards of care and user protections. Such work requires arranging institutional data access agreements, establishing independent evaluation of proprietary chatbots, and stipulating provider responsibilities.

We assembled this research agenda through a selective synthesis of work in AI evaluation, human-computer interaction, psychology, mental health, and policy, and our experiences in this domain. Throughout the paper, we distill 12 recommendations and 12 open questions to provide accessible takeaways for researchers and other stakeholders. We mark each with icons for the stakeholders whose needs or work it most directly addresses: researchers (), AI developers (), clinicians (), users (), and policymakers (). Importantly, we also highlight how impacted users can be centered in creating knowledge and answering these open questions.

2 Related Work

While chatbots can produce concerning responses in high-risk settings [72, 103, 154], they can also be used for mental health support [21, 158]. Researchers have described how chatbots' use warrants further clinical and regulatory scrutiny [119, 131, 135] as well as LLM applications in digital mental health and gaps in their clinical evidence [59, 63]. Hua et al. [63] distinguish controlled tests of chatbot behavior, pilot studies of usability, and trials of clinical efficacy. Reviews also

examine how LLM applications fit into different levels of care, including their implications for access, equity, and user well-being [40, 81]. All of these reviews stress unresolved issues in user privacy and ethics. Across these applications, studies often lack longitudinal validation [106, 165].

While prior work has studied both purpose-built mental health systems and general-purpose chatbots used for support, we focus on general-purpose chatbots even when users are not explicitly seeking support [cf. 107].

Chatbot frameworks. Frameworks propose standards for chatbot behavior and oversight, including limits on autonomy [58], the need for clinical rigor such as randomized trials [176], and oversight groups [119]. Parks et al. [135] call for criteria on safety and evidence, validation of LLM judges and simulated users, and consumer participation. Lee et al. [82] argue for transparent and standardized evaluation of chatbots and discuss who is accountable for them. Zuromski et al. [181] propose a clinical research agenda for youth suicide risk. Suh et al. [159] propose specific interaction patterns that might amplify user psychological vulnerabilities. How to implement and assess these proposed standards in deployed products remains an open question.

Evaluations. Broad work on AI evaluation may inform how mental health risks are assessed. Scholars have called for assessments both pre- and post-deployment [141, 142]. Chen et al. [32] survey AI agent evaluation, including interaction harms, simulation validity, and evaluator independence.

In mental health specifically, evaluations operationalize these concerns through tests of clinical capabilities and chatbot behavior. Several evaluations test how clinically capable LLMs are, such as psychopharmacology [47] and mental health question answering [83]. Some researchers have assessed LLMs against psychotherapy techniques [35, 118]. Other scholars have documented issues with chatbots replacing clinicians [65, 103]. Many evaluations test LLMs on other relevant tasks, such as assigning crisis labels [23] and in triage [172].

Recently, evaluations have begun to turn to multi-turn conversational trajectories [151, 171]. Some use real transcripts, acting as single-turn evaluations with prefix context [73, 105]. Others generate synthetic conversation histories but score only the model responses [89]. Increasingly, evaluations have used sophisticated simulated users [15, 17, 71, 137, 139, 171] and LLM judges [15, 17, 71, 89, 137, 171].

A key question is when to use clinicians: whether their ratings should serve as the final score in a benchmark [83, 154], define rubrics applied by an LLM judge [89], or validate an automated system [17, 71, 137, 171]. Weilnhammer et al. [171] validate behavioral judgments against clinicians, but do not establish downstream effects on users.

Many papers evaluate chatbots on specific mental health conditions. Understandably, most focus on crisis and suicidal ideation [72, 137, 154, 172]. Others have examined psychosis and delusion-linked harms [105, 177]. These evaluations characterize behavior under tested conditions, but do not by themselves establish effects on mental health over time.

Synthesis. Prior work has established chatbot evaluation criteria and has called for longitudinal evidence, but leaves many open questions about how to produce and independently validate that evidence for general-purpose chatbot products in mental health. To do so, researchers must link evaluation measures to user outcomes, test the complete products users encounter, and study sensitive longitudinal data.

3 Approach and Positionality

We contribute a commentary on open problems in evaluating and improving the mental health effects of general-purpose chatbots by unifying multidisciplinary perspectives. In accordance with prior agenda-setting work for shaping AI's societal impact in areas like education [54], safety [4, 38, 61], and sociotechnical governance [140, 143], we did not

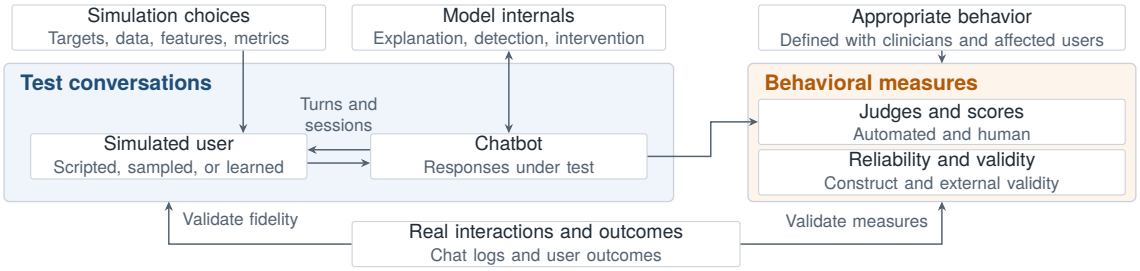


Fig. 2. **A controlled evaluation has several interacting parts.** We discuss appropriate behavior, measures, and real outcomes in §4.1, test conversations and simulations in §4.2, and model internals in §4.3.

employ a formal protocol for selecting literature or synthesizing problems and recommendations. Instead, our selection reflects the author team's research, professional experience, and discussions. Our backgrounds span AI evaluation and safety, human-computer interaction and social computing, cognitive and affective science, learning and therapeutic technologies, clinical psychiatry, and technology policy. Our experience includes academic and industry research, clinical practice, nonprofit leadership, government and international policy work, collaboration with clinicians and impacted communities, and public communication about technology and mental health.

This paper grew out of informal discussions during research collaboration meetings and workshops about the evolving culture of AI evaluations and safety, the need to characterize real-world impacts, and engaging in organizing and public engagement efforts on mental health chatbots. Our scope was refined via exercises mapping the landscape and initially reflecting on how safety cultures, independent audits, and regulatory landscapes in analogous fields had evolved in response to public safety incidents [144, 157]. We divided the writing according to our backgrounds: for example, psychology and HCI researchers anchored the discussion of real-world evidence, while AI evaluation and measurement researchers helped contextualize broader shifts and sharpen a specific agenda for mental health effects.

4 Controlled evaluation

Controlled evaluations test chatbot behavior under specified conditions (Figure 2), but better benchmark performance does not necessarily make a system safer or prove that it will improve well-being.

4.1 Appropriate behavior and measurement

Defining appropriate behavior. While there is much uncertainty about optimal chatbot behavior, some behaviors are clearly inappropriate or harmful and these provide clear targets for positive change. Until recently, much of the conversation about how AI should act has been subsumed by "AI safety," "harmlessness," or "alignment" [61, 140]. These approaches have established minimum standards and post-training methods [13, 80, 132].

Nonetheless, the space of appropriate social responses, especially as they concern users' mental health, is much broader. The same response may be appropriate or dangerous depending on the preceding conversation, the user's intent, or their vulnerability. For instance, when exploring the possibility of faster-than-light travel, is a chatbot encouraging a delusion, or is it appropriately role-playing science fiction? Appropriate behavior also depends on clinical judgment, user goals, and social and cultural norms. Clinicians and affected users should help define these criteria.

Recommendation 1. Users and survivors should help shape chatbot evaluation.

Explanation. Researchers should collaborate with users and survivors on research questions, study design, and interpretation of findings, with fair compensation and control over participation and use of their data [22]. Meaningful participation requires influence over research goals and decisions [19].

Measurement validity. Mental health and psychological well-being are well studied by the measurement sciences (psychometrics). When clinicians develop instruments like the PHQ-9 to measure patient depression, they explicitly measure reliability and other aspects of construct and criterion validity (e.g., comparing patient self-reported responses on the PHQ-9 with physician assessments) [79]. Some mental health evaluations validate automated judgments against clinicians [17, 171], but systematic measurement validation remains an open problem in AI evaluation [167]. In mental health, Badawi et al. [12] note that benchmarks are limited in "scale, authenticity, and reliability." AI evaluations should likewise test scoring consistency (reliability) and, if applicable, compare with expert judgments.

Construct validity evaluates the degree to which a measurement captures the concept of interest [149]. For example, Byun et al. [23] evaluates crisis detection, while Shah et al. [154] assess responses for the acknowledgement of risk. These scores measure different components of crisis handling and are not interchangeable. There is little systematic testing of whether AI benchmark scores measure their claimed constructs or agree with other measures of the same construct [44]. For example, "anxiety" in social media datasets can refer to a disorder, symptoms, or a transient emotion [28]. Labeling expressions of worry as an anxiety disorder could reward recognizing anxious language without identifying a condition.

Separately, *external validity* concerns whether evaluation results generalize to new populations and contexts [29]. For example, evaluations may not generalize across cultural and linguistic contexts because user expressions, chatbot behavior, and behavioral judges may vary across these contexts [53].

Recommendation 2. Mental health benchmarks should specify their constructs, test scoring reliability, and be validated against related measures.

Explanation. Benchmarks should define, for example, the target population, conversational context, crisis acuity, intended role of the chatbot, and expected supportive actions. Measures of the same construct should be compared with one another. Claims should remain bounded to the populations, languages, and settings in which the benchmark has been validated.

Validation against real-world outcomes. Recent work has measured chatbot behaviors, including sycophancy [34] and crisis handling [154]. These measures may contribute to assessing whether chatbot responses are appropriate.

Still, observing inappropriate responses in real logs does not by itself connect chatbot behaviors to clinical outcomes or establish what caused user harms [104]. For example, repeated textual expressions of an unusual belief may not constitute a clinically defined delusion. The conversational context and—more importantly—user mental state can significantly change the meaning of any expression. Such outcomes might include suicide attempts, hospitalization, emergency care, emotional dependency, and other broader quality-of-life measures.

Recommendation 3. Researchers should link evaluation measures to validated measures of real-world outcomes.

Explanation. Where feasible, researchers should use randomized experiments or other forms of causal identification to test links between evaluation measures and real-world outcomes. Researchers should propose *theoretically*

motivated explanations for how user-chatbot interaction patterns impact psychological processes and clinical outcomes. Researchers should preregister longitudinal studies linking evaluation measures to consented, independently assessed outcomes, using clinical assessments or ecological momentary assessments (EMA) of mood and stress [110].

Open Question 1. Which conversational signals in chat logs indicate real-world risk?

Explanation. Researchers must determine whether various conversational signals indicate risks such as crises or distress. Hence, evaluations should connect to downstream, real-world quantities [75]. Nonetheless, signals may simply predict risks without causing them and should be compared against baseline rates of relevant outcomes. Such evidence should inform the scenarios and behavioral standards used to evaluate chatbots offline.

4.2 Interaction trajectories

Understanding chatbots' effects on users requires studying interactions over time. Behavioral tests of chatbots are predominantly static and single-turn, based on fixed prompts and expected outputs [89]. Psychological change may not occur through the discrete turns of a conversation but over larger units of time such as conversations, sessions, and more. For example, some harms don't arise from a single bad response: increased emotional dependence, social withdrawal, or changes in decision-making happen over time. Indeed, harms that arise from longer interactions may appear below the threshold of detection for any one conversation turn.

Recommendation 4. Evaluations should cover multi-turn and multi-session interactions with systematic variation in the conversation history.

Explanation. Recent work has begun to use multi-turn and deployment-grounded evaluations in AI and mental health [71, 105, 173]. Furthermore, evaluators should vary specified features of conversation history (e.g., when distress is disclosed), holding others fixed and repeating each condition. Such variation may isolate the effects of conversation history on chatbot behavior from baseline randomness. Behavioral judges should evaluate patterns across turns or sessions rather than single-turn responses. At the same time, complex judgments may yield low inter-annotator agreement [89] and require novel evaluation frameworks [90].

4.2.1 User simulations. Chat logs are typically observational: they do not show how another chatbot would have responded to the same user. Some evaluators use simulators to write messages as a hypothetical user for testing new models before deployment [71, 106, 133, 177]. Potential targets of simiulation include a specific user (e.g., based on personal data), a category of users (e.g., people with ADHD, teenagers), and events (e.g., a specific crisis). Existing work accordingly evaluates fidelity to particular people, population distributions, or clinical profiles [62, 134, 169].

Simulators also raises ethical questions. They might erase users' lived experience and undermine their power to shape AI systems [3, 74] (cf. Recommendation 1). Still, simulations may reduce participant exposure to harmful interactions.

Simulation methods. Historically, simulators have used scripted interaction. Scripts can vary in determinism from complete in the case of playing back existing conversations [105] to absent in the case of persona-based prompts with no specific turn indicators [87, 171]. Alternatively, one could, simply sample real user queries. Recent work by OpenAI suggests that summaries of production use can inform pre-deployment evaluations [89, 173]. Johnson et al. [71] found

that simulations conditioned on production data better matched real users' behavior. There is a wide range of technical methods available to use LLMs for simulation, such as prompting [138], fine-tuning [2], continued [64] and initial pretraining [64, 71], and inference-time parameter interventions [5, 33]. Still, these methods exhibit different failure modes. LLMs prompted with personas may become less consistent with them over longer conversations [88].

Simulation choices and fidelity. A central limitation is *fidelity*: how well synthetic scenarios reflect observed user behavior [71, 137, 154]. Evaluations may omit unexpected user behaviors or misrepresent their frequency [71]. Bentley et al. [17] asked licensed clinicians to rate the realism of simulated users created by researchers. Clinical expertise is essential for assessing clinical plausibility, but these ratings alone do not establish that simulations match real users' language or behavior (see §4.1). Other work has used LLMs to judge simulator messages [e.g., 12, 89, 100]. Automated judges are scalable and ideally reproducible but are not infallible; like any annotator, they must be validated against reference human, ideally clinical, judgments [103, 104]. Claims about fidelity to a category of users or events require comparison with an appropriate real-world distribution. Reference data may include online community posts [156] and therapeutic conversations [175]. Still, such data raise an ecological validity concern: discussions between humans may not reflect how people converse with chatbots.

One might use a variety of features to indicate which behaviors count as similar. For example, one could compare simulated and real users using scores of their "anxiety" from a validated classifier. However, similar scores could reflect shared diction, message lengths, or topics rather than realistic presentations of anxiety. Notably, it is not clear whether simulators that reflect one feature dimension (e.g., the writing style of a 14-year-old) generalize to other dimensions (e.g., the clinical criteria of a 14-year-old with generalized anxiety disorder). Features need not only derive from reference *text*. Simulators could use other structured reference data, such as depression screeners and social surveys [62].

Recommendation 5. Evaluators should justify simulation choices in relation to their intended claims and establish which claims the available validation supports.

Explanation. Targets, assessments, reference data, comparison features, and similarity metrics should be studied together. A simulator that realistically portrays a mental health condition can help test how a chatbot responds to it, but does not show how often the chatbot would fail in real use. Evaluators should report whose experiences their data cover, which dimensions of fidelity they assess, and where evidence is missing.

Open Question 2. Can simulator-based evaluations predict chatbot behavior in deployment?

Explanation. Researchers should use simulators to compare predicted chatbot behaviors and failure rates with those observed in deployment. They should assess how simulator determinism and output diversity affect those predictions. Real interactions and outcomes may reveal whether the simulated tests detect consequential risks or miss unintended effects. These findings should inform simulation choices and behavioral standards. Such work on simulators requires independent research (§6.1).

4.3 Model internals

The evaluations above examine chatbot outputs; interpretability work can complement them by detecting patterns in and intervening on model internals. Activation patterns are the numerical states computed as models process text. Activation probes are classifiers trained on such internal states.

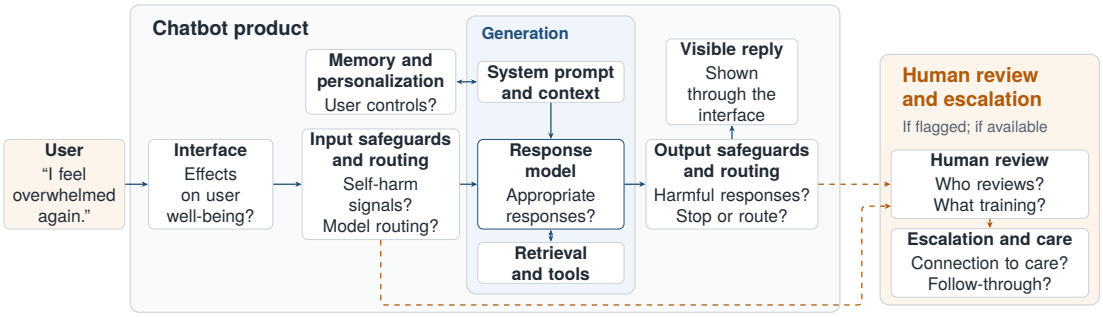


Fig. 3. **One visible chatbot interaction involves multiple systems.** This hypothetical diagram illustrates some of the possible components of chatbot systems. One of our central questions is how these components affect user well-being, separately and together.

Detection. Zhu et al. [180], for example, find representations for depression and suicidality. Such work is promising but largely uses text labels rather than extended interactions and is not validated against independent clinical assessments.

Open Question 3. Can probes detect mental health risks more accurately or cheaply than text classifiers?

Explanation. Activation probes have been shown to generalize to medical interactions at far lower estimated compute than prompted or fine-tuned monitors [98]. However, they were less accurate than most text-based monitors on a held-out mental health dataset.

Intervention. Researchers may also use activation-based methods to modify these internal states. Prior work has reduced excessive agreement and broadly harmful responses [33].

Open Question 4. Can we steer model activations away from traits such as sycophancy? Will steering effectively reduce downstream harms in novel situations?

Explanation. For example, would reducing sycophancy scores reduce delusion reinforcement, and how would this affect model helpfulness? Nonetheless, whether a model behavior is harmful or helpful is a sociotechnical problem: users' behavior and circumstances shape the effects of a response.

5 Real-world evidence

Understanding chatbots' effects on mental health requires studying deployed products and interactions over time.

5.1 Deployed chatbot products

Evaluating the full product. Chatbot systems are sometimes treated as a single LLM, but modern human–chatbot interaction often includes multiple models and non-LLM components that shape user impacts. A single conversation can involve system prompts, different underlying models, safety classifiers, memory and personalization, retrieval and tools, interface cues, human escalation, and more. Figure 3 diagrams such systems.

Evidence from social media has shown that the very interfaces which people use can cause addictive and other maladaptive tendencies [14, 148]. Some researchers test consumer-facing chatbot UIs. Regarding delusional behavior, Kirgis et al. [77] manually showed that the ChatGPT UI and the API yielded different responses. Wang et al. [168] found an increase in bias and sycophancy in the UI relative to the API for Claude, Gemini, and ChatGPT. By contrast, Johnson

et al. [71] found little difference between the UIs of Claude and ChatGPT and their APIs when prompted with user mental health simulators.

Recommendation 6. Evaluations should test deployed chatbots, not only their foundation models.

Explanation. Researchers should measure how interfaces, memory, personalization, and other components affect user well-being, both together and separately. Researchers should collaborate with HCI researchers to test how systems are used in the wild and explore alternate conversational interfaces.

Access and component standards. Evaluating these products is difficult when their components and configurations are not public or fully specified. Even if all providers gave researchers access to consumer-product-equivalent endpoints, such endpoints still would not indicate *which* components lead to certain outcomes. For example, modern chatbots use memory features to collect and summarize prior user conversations. OpenAI and Anthropic document different memory implementations and user controls [10, 11, 129]. There are related proposals for standard features, but these have not been adopted [70, 166].

Beyond memory, another key component is the cascade of classifiers and underlying models to which any given user query is routed. Both OpenAI and Anthropic report using classifiers to detect self-harm behavior in users [7, 124] and whether models should use reasoning [122]. These classifiers and routers are neither public nor reproducible.

Recommendation 7. Researchers and providers should develop standards for chatbot components.

Explanation. Standards should document component behavior and configuration [46, 82]. Researchers also need access to the deployed chatbots, including a documented research API for reproducible prompt-response testing, and credentialed access for sensitive configuration, personalization, logs, and more. Providers should also support component toggles, turning on or off features such as memory or classifiers. Absent changes in chatbot development and standardization, we encourage more research on automated methods for testing consumer interfaces [168], building on tools to audit social media systems [78, 92].

5.2 Human review and escalation

Chatbots are part of broader human systems which bear on users' well-being. Some chatbot providers, for example, will escalate user queries for human review. As of September 2026, OpenAI says that suspected plans to harm others (but not oneself) are reviewed by a small team who may refer imminent threats to law enforcement [123]. They also provide an optional "Trusted Contact" feature in which a serious concern triggers a notice to that contact. Humans review every notice, although it is not clear who reviews them or what training they have received [128]. OpenAI estimated suicidal indicators appear in conversations of 0.15% of weekly active users and in 0.05% of messages, but they do not report statistics on human review [126]. By contrast, Anthropic says it flags suicide- or self-harm-related content but does not share this information [6, 8], and their public materials do not describe human review [6].

Such practices—and, furthermore, users' knowledge that companies *engage* in such practices—may impact user mental health. Analogously, some patients in therapy deliberately hide or misrepresent suicidal ideation because they fear involuntary hospitalization or other losses of autonomy [20].

Open Question 5. Which review and escalation practices help, and how often do users receive care?

Explanation. Human review practices raise a variety of questions: what kinds of human-in-the-loop interventions are more or less effective, for what downstream measure, using which kinds of human reviewers, and more. We should also seek to understand what kinds of messages are flagged at all. For example, researchers could use such referral events to develop open benchmarks for the development of classifiers. Knowledge of the workings of these human-in-the-loop systems will help the research community understand the effect of different interventions (§5.4) and help policymakers decide what kinds of interventions we should have (§6).

5.3 Collecting longitudinal human–chatbot data

Users' interactions with chatbots will inform how chatbots affect their mental health. Researchers should ethically work with users to analyze these [112].

Sampling people and interactions. Collected chat log data may not generalize to all relevant populations. While there have been broad efforts to collect chat logs [179], efforts for mental health specific uses of chatbots have been largely ad hoc and small [104, 105, 178]. In a survey of U.S. native-English users through Prolific, fewer than 25% supplied their chat logs [178]. The recruitment method may affect the data. Public data donations might scale to more contributions [31, 178] but may fail to recruit certain groups. Furthermore, the people recruited for studies are not necessarily the average users [45, 95, 174]. Trial eligibility can also exclude relevant clinical subpopulations. Applying typical exclusion criteria, one study estimated that 79% of people with schizophrenia-spectrum disorders would have been ineligible, because of, for example, suicide risk [161]. For youth, active parental consent may lower participation and change sample composition, potentially missing high-risk youth [84]. Without specifically aiming to recruit such individuals, experiments may undersample the tails of the population distribution and miss safety outcomes concentrated there.

Recommendation 8. Studies should balance targeted recruitment for rare, high-risk cases with representative sampling to estimate population prevalence.

Explanation. Any sampling regime must account for differences in how often conditions and symptom severity present. Sampling in proportion to these rates may yield too few high-acuity cases (e.g., crises or delusional spirals) to study reliably. On the other hand, focusing on such cases may underrepresent more common experiences (e.g., depression and anxiety). Researchers must therefore balance targeted recruitment to study rare presentations and representative sampling for estimating population prevalence [21]. Furthermore, more studies should test whether subpopulations are vulnerable to chatbot-related risks, whether they even use chatbots, and the effect of current proposed solutions on them (such as occasional reminders that chatbots are nonconscious). To do so, we suggest partnering with clinics[18] or advocacy organizations [108] for recruitment.

Data export and interoperability. There are many technical challenges to collecting user chat log data. Users may have logs across multiple providers and devices. There is not one widely adopted process for users to request such data [9, 57, 127]. Some providers export chat logs as branching conversation trees, which do not specify a canonical linear conversation history. For analysis, Moore et al. [104] linearized each tree by following the specified current node's parent chain to the root. Selecting one path necessarily excludes alternative branches; therefore, researchers need a standard for linearization. One initiative calls for a shared schema [39].

Recommendation 9. Future work should explore which systems, such as browser extensions or data-export standards, could be built to standardize chat log analysis.

Explanation. Similarly, we must decide how to standardize analysis across disparate file formats.

Privacy and access. The scale of user data—potentially tens of thousands of messages per user—can make the typical manual review to identify and remove sensitive information impractical. Unable to read the transcripts they gathered themselves, Moore et al. [104] instead queried LLMs to summarize and code the data. Such methods introduce data leakage risks. Furthermore, many evaluation paradigms, such as simulations, may depend on user data. These evaluations must be widely run to become validated standards (§4.2 and 4.2.1), requiring appropriately governed data access.

For these reasons, we must explore various approaches to reduce identification risk, especially through automated de-identification. Text de-identification detectors may use rule- or dictionary-based systems, learned named entity recognition (NER), hybrids, and generative LLMs [43, 86, 101, 160] with varying degrees of success.

Open Question 6. What limits automated chat log de-identification, and how can we improve it?

Explanation. Removing explicit identifiers may not anonymize user data. High-dimensional text data may still expose sensitive information—such as locating someone through regional slang like "hella" [48], determining someone's age through their discussion of menopause, or fingerprinting based on their use of language [111]. Indeed, literature on clinical free-text de-identification shows that many automated methods struggle and allow downstream reidentification [102, 109]. Nonetheless, humans are poor at removing identifiers: on clinical data, automated systems perform similarly to individual reviewers [27, 41, 115]. To address these challenges, some suggest requiring local-only processing of private data, although this may limit access to those with suitable clusters and engineering capacity. In contrast to de-identification, researchers might use differential privacy methods [116] to limit what can be inferred from released query results, such as when reporting aggregate counts or producing aggregate insights of chatbot use [85].

Recommendation 10. Standarize chat log access, scaling scrutiny to each use's sensitivity.

Explanation. Automatic measures to de-identify text are imperfect. Different analytical applications—from organizationally internal to private data for public research to public anonymized datasets—may demand different degrees of scrutiny, more precise automated methods, and, for the most sensitive and public uses, manual review.

5.4 Clinical research

Clinical research advances in step with new data [162]. Chat log data provide a new form of high-frequency behavioral data and may capture many aspects of a user's life, such as disclosure of mental health symptoms [159]. Hence, chat log data may inform the fundamental phenomenology of mental health. For example, researchers have used chat logs both as a record for potential risk signals and as a proxy for clinical outcomes [104]. Still, chat logs record behavior within a particular interaction and may not generalize to other settings.

Chat log data may also change how we weight the clinical import of different signals. For example, depression screeners ask about thoughts of death or self-harm, but those factors alone have limited value for predicting who will

die by suicide [96]. Disclosure is also context-dependent. A meta-analysis estimated that only 45.9% of people with suicidal thoughts disclosed them [60].

Recommendation 11. Researchers should study chat logs as evidence on psychopathology.

Explanation. Chat logs introduce a new disclosure setting and may allow changes to risk modeling, intervention delivery, and tools to observe patient behavior. For example, some people seek reassurance for experienced anxieties. Such reassurance may temporarily allay distress while nonetheless maintaining anxiety and repeated reassurance seeking [150]. Repeated reassurance may leave users with fewer opportunities to learn to tolerate uncertainty [56]. while still reporting finding these interactions helpful [147]. With their constant availability, chatbots are a natural choice for such reassurance and hence people's chat logs with them may provide evidence on such behavioral patterns. Researchers may need to recalibrate clinical models to accommodate these new data.

Open Question 7. How should general-purpose chatbots engage with users in mental health crises?

Explanation. This is a live question affecting real people. OpenAI policy says that their chatbots should not quit a mental health conversation. At the same time, they forbid enablement of self-harm [121, 125]. (Anthropic does not appear to have a requirement to continue every mental health conversation [6–8].) We can find no published data justifying why OpenAI implemented such a practice of chatbots continuing to engage. By contrast, multi-stakeholder guidance could provide such a justification or reject the claim [76]. OpenAI's policy mirrors guidance for trained crisis responders, which suggests continuing to engage, assess risks, and follow up [1, 113]. Nonetheless, chatbots are neither trained responders nor licensed providers. Continued engagement also raises questions about clinical responsibility. Might such engagement establish treatment relationships [145]?

Recommendation 12. Researchers should measure the clinical impact of individual chatbot features.

Explanation. What are the causal links between chatbot design, usage, and harms? If linked to real-world outcomes, chat log data may provide insight into how to respond appropriately to crises. Targeted studies built on the various independent components of chatbots may begin to build an evidence base for best practices as opposed to opaque policies issued by industry providers.

6 Policy

6.1 Policy interventions and user protection

As of October 2026, jurisdictions worldwide have passed legislation to address concerns about mental health and chatbots. Many laws and proposals prohibit types of chatbot interactions, such as romantic and sexual roleplay, excessive sycophancy, facilitating self-harm, suicide, and harm against others, and discouraging help-seeking from real humans. These measures often require warnings to users that the chatbots are not human.

In the United States, such laws have been enacted in a variety of states [25, 52, 130, 170]. However, many of these protections only apply to minors. Crisis-response requirements generally cover all users [55]. The European Union currently has no specific regulations on chatbot companions [117], although the AI Act does prohibit certain harmful manipulative practices [50, Art. 5], and the proposed EU Kids Act [49] seeks to prevent emotional dependency on

chatbots. China has implemented many guardrails for anthropomorphic interaction services, such as banning virtual intimate relationships for children and restricting features that may foster emotional dependence for users of all ages [37, 51]. The UK has announced plans for mandatory breaks for chatbot users under 18 and restrictions on their access to sexualized chatbot services [42], and is considering bans on children's access to mental health chatbots [42, 163].

Certain policies explicitly require chatbot developers report to regulatory bodies (e.g., U.S. state attorneys general) on topics such as child safety [25] and crisis events [24, 36]. Still, such chatbot developers' reports may not be public or otherwise allow researcher access.

Open Question 8. What are the impacts of enacted policy mitigations that restrict chatbot behavior?

Explanation. Newly enacted laws lay out many interventions chatbot providers must take. Despite the dearth of controlled experiments on the effects of chatbot behavior on mental health, the different stipulations of these policies may create natural experiments if the chatbot providers implement different localized product endpoints. Hence future work might compare populations across those jurisdictions, while accounting for differences.

Open Question 9. How can policy mitigations protect adults, whose needs differ from those of minors?

Explanation. Many protections in enacted laws apply only to minors. However, adults can be vulnerable due to age, disability, mental illness, or loneliness [104]. What policy levers allow us to protect such vulnerable adults? Furthermore, many laws require vendors to identify users who are minors, but such age verification may raise intractable privacy and accuracy problems [16, 97, 152]. Future work should explore these problems.

AI companies might work with independent groups to conduct audits [26, 164]. For example, Transluce recently conducted a mental health evaluation of chatbots from Anthropic and OpenAI, who provided private production conversations and anonymized features. This protected user privacy but limited independent verification [71].

Open Question 10. What types of transparency, reporting, and evaluations will enable effective external oversight of chatbot systems?

Explanation. Who should have access to audit evidence—vetted researchers, regulators, or evaluators? To assess independence, evaluations should disclose who chooses research questions, who controls publication, and who can independently verify conclusions.

6.2 Liability and standards of care

Chatbot providers may need to meet certain reasonable standards of care for their users. If they do not meet those standards or are otherwise negligent, certain proposed policies may make those providers *liable* for chatbots' harmful effects on user mental health [91, 153]. Still, while clinical standards, design requirements, and legal duties of care overlap, they are not interchangeable. Rigorously establishing causal links between chatbot usage, interface design, and harms (Recommendation 12) might inform policies and lawsuits aiming to hold chatbot developers to account.

Of course, what is a reasonable standard of care depends on the intended use of a chatbot. Indeed, certain chatbots are purpose-built for mental health applications. The UK regulates digital mental health technologies with a medical purpose [99], and the US does similarly [69, 82]. Still, it is difficult to define what is clinical. For example, "wellness"

chatbot products exist in a regulatory gray area but may still have clinical impacts. Ostermann et al. [131] argue that chatbots used therapeutically can amount to a "medical device."

Open Question 11. What standards should we hold general-purpose chatbots to?

Explanation. Current policies leave unclear which standards apply when chatbots are used for mental health support. Illinois, for example, restricts chatbots from providing or advertising therapy [66], but it is unclear how such laws apply to general-purpose chatbots that make no such claims. Exempting them risks leaving vulnerable users unprotected, whether they deliberately use chatbots for therapeutic support [93, 158] or drift into emotional support during task-oriented interactions [155]. Standards should therefore track actual, not just advertised, use.

6.3 Research data governance

The effects of chatbots on mental health are hard to study because conversations are private, limiting independent visibility into user experiences and concentrating access to evidence among chatbot providers. Nevertheless, such data can help identify harms and evaluate protections when shared with the scientific community, provided access is governed and limited. Hence, researchers may require controlled, anonymized access to real user chat logs, evaluation access to user-facing chatbots, and more. Regulators could mandate such access.

Open Question 12. How can researchers organize and share user chat logs while keeping data private?

Explanation. Genomic and clinical repositories offer precedents for governing sensitive research data [67, 68, 114, 136]. These approaches combine consent-based project-level approval and researcher credentialing. Some allow controlled downloads while others require analysis in a secure environment. Chat log access should also give participants control over secondary uses and protect the privacy of third parties mentioned in logs.

Still, it is unclear who should control access to these data. For example, if the data are collected by a university or contain, at one point or another, participant identifiers, such sharing may require engaging with institutional ethics, privacy, and legal offices. Furthermore, analyses may reveal that participants may present with crises. Depending on whether these analyses are retrospective, researchers may be required to provide participants services and may be required to report certain kinds of crises.

7 Limitations

Our literature review and selection of open problems and recommendations was not systematic, but rather shaped by our team's composition and positionality. Our author team does not include anyone with lived experience of harms related to chatbot use; thus, our recommendation to give users and survivors meaningful participation in influencing research goals and decisions can be exercised by this team in the future, as well. We do not empirically validate our proposed research and policy directions or establish which might be most effective or feasible. Although we acknowledge potential benefits from chatbots, we focus primarily on risks and harms. We do not establish the prevalence or causal effects of chatbot harms. Public reports only indicate a few kinds of harm people have experienced with chatbots. Given our focus on these harms, we may fail to consider the interaction between unreported harms and chatbots. We cover a broad range of mental health and psychological well-being outcomes. Recommendations for one context may not

generalize to others. We do not systematically examine cultural and linguistic differences in mental health presentation. Chatbot products change, and our analysis relies on public descriptions that may be incomplete.

8 Conclusion

In this paper, we reviewed open problems in the use of general-purpose chatbots and their effects on mental health and psychological well-being in three domains. These domains interact: connecting **controlled evaluation** to independently assessed user outcomes requires **real-world evidence** and user simulations need validation against real interactions. Producing and independently verifying this evidence depends on governed access to deployed products and sensitive longitudinal data. Evidence on which chatbot behaviors and interventions help or harm users should inform **policy** on user protection and standards of care, while those protections should themselves be evaluated. We hope this survey will help establish an evidence base such that future chatbots help and do not harm users' mental health.

Ethical Considerations

Research on sensitive chat logs and crisis interventions can create privacy risks and undermine user autonomy. We therefore emphasize governed data access, meaningful participation by users and survivors, and independent validation of proposed protections.

Generative AI Usage Statement

We used generative AI tools such as Codex and Claude Code to format this manuscript, find relevant papers to review, propose grammatical and typographic improvements, and offer points of improvement. Any errors are our own.

Acknowledgements

We thank Ben Buck and Meryl Ye for comments on this manuscript.

J.M. acknowledges support from the Stanford Interdisciplinary Graduate Fellowship and the Future of Life Institute Vitalik Buterin Fellowship.

References

[1] 988 Suicide & Crisis Lifeline. 2024. Best Practices: Suicide Safety Policy. https://988lifeline.org/professionals/best-practices/

[2] Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. 2025. Consistently Simulating Human Personas with Multi-Turn Reinforcement Learning. doi:10.48550/arXiv.2511.00222 arXiv:2511.00222 [cs.CL].

[3] William Agnew, A. Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R. McKee. 2024. The Illusion of Artificial Inclusion. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 12 pages. doi:10.1145/3613904.3642703

[4] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. doi:10.48550/arXiv.1606.06565 arXiv:1606.06565.

[5] Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C. Kozlowski, Bernard Koch, Erik Brynjolfsson, James Evans, and Michael S. Bernstein. 2025. Position: LLM social simulations are a promising research method. In *Proceedings of the 42nd international conference on machine learning (Proceedings of machine learning research, Vol. 267)*, Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (Eds.). PMLR, 81005–81034. https://proceedings.mlr.press/v267/anthis25a.html

[6] Anthropic. 2025. Anthropic's Protocol for Addressing Expressions of Suicidal Ideation, Suicide, and Self-Harm Risk. https://www-cdn.anthropic.com/51dab7c41a832caf249149c7bfdd56fe65ecf960.pdf Accessed 2026-09-09.

[7] Anthropic. 2025. Building Safeguards for Claude. https://www.anthropic.com/news/building-safeguards-for-claude Accessed 2026-09-07.

[8] Anthropic. 2026. Crisis Helpline Support in Claude. https://support.claude.com/en/articles/13171706-crisis-helpline-support-in-claude Accessed 2026-09-09.

[9] Anthropic. 2026. How Can I Export My Claude Data? https://support.anthropic.com/en/articles/9450526-how-can-i-export-my-claude-data

[10] Anthropic. 2026. Import and Export Your Memory from Claude. https://support.claude.com/en/articles/12123587-import-and-export-your-memory-from-claude Claude Help Center. Accessed 2026-09-07.

[11] Anthropic. 2026. Use Claude's Chat Search and Memory to Build on Previous Context. https://support.claude.com/en/articles/11817273-use-claude-s-chat-search-and-memory-to-build-on-previous-context Claude Help Center. Accessed 2026-09-07.

[12] Abeer Badawi, Elahe Rahimi, Md Tahmid Rahman Laskar, Sheri Grach, Lindsay Bertrand, Lames Danok, Prathiba Dhanesh, Jimmy Huang, Frank Rudzicz, and Elham Dolatabadi. 2026. When Can We Trust LLMs in Mental Health? Large-Scale Benchmarks for Reliable LLM Evaluation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vera Demberg, Kentaro Inui, and Lluís Marquez (Eds.). Association for Computational Linguistics, Rabat, Morocco, 3873–3896. doi:10.18653/v1/2026.eacl-long.180

[13] Yuntao Bai et al. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 https://arxiv.org/abs/2212.08073

[14] Amanda Baughan, Mingrui Zhang, Raveena Rao, Kai Lukoff, Anastasia Schaadhardt, Lisa Butler, and Alexis Hiniker. 2022. "I Don't Even Remember What I Read": How Design Influences Dissociation on Social Media. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM, 1–13. doi:10.1145/3491102.3501899

[15] Luca Belli, Kate Bentley, Will Alexander, Emily Ward, Matt Hawrilenko, Kelly Johnston, Mill Brown, and Adam Chekroud. 2025. VERA-MH Concept Paper. doi:10.48550/arXiv.2510.15297 arXiv:2510.15297 [cs].

[16] Steven M. Bellovin. 2025. Privacy-Preserving Age Verification—and Its Limitations. Submission to the IAB/W3C Workshop on Age-Based Restrictions on Content Access. https://www.cs.columbia.edu/~smb/papers/age-verify.pdf

[17] Kate H. Bentley, Luca Belli, Adam M. Chekroud, Emily J. Ward, Emily R. Dworkin, Emily Van Ark, Kelly M. Johnston, Will Alexander, Millard Brown, and Matt Hawrilenko. 2026. AI Chatbot Suicide Risk Detection and Response: Human Validation Study of the Open-Source VERA-MH Safety Evaluation. *JMIR AI* 5 (June 2026), e92817. doi:10.2196/92817

[18] Zachary Bergson, Sarah G. Vassall, Adam Wright, Allison B. McCoy, Katherine M. Schafer, Margaret C. Achee, and Julia M. Sheffield. 2026. Characterizing Artificial Intelligence Psychosis in a Large Medical Setting. *JAMA Psychiatry* (7 Oct. 2026). doi:10.1001/jamapsychiatry.2026.2923 Published online.

[19] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Díaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. Association for Computing Machinery, New York, NY, USA, 8 pages. doi:10.1145/3551624.3555290

[20] M. Blanchard and Barry A. Farber. 2020. "It Is Never Okay to Talk About Suicide": Patients' Reasons for Concealing Suicidal Ideation in Psychotherapy. *Psychotherapy Research* 30, 1 (2020), 124–136. doi:10.1080/10503307.2018.1543977

[21] Rebekah Bodner, Katherine Lim, Steven Siddals, Simon Goldberg, and John Torous. 2026. Barriers to understanding how many people use AI for mental health support: an estimate and narrative review. *npj Digital Public Health* 1, 1 (July 2026), 21. doi:10.1038/s44482-026-00025-7

[22] Sandra Bucci, Alice Newton, Pauline Whelan, Kim Cartwright, Prathiba Chitsabsean, Simon Foster, Victoria Green, Amanda Larkin, Rhiannon-Faye McDonald, Ethel Quayle, Matthias Schwannauer, Victoria Selby, and Filippo Varese. 2025. Reporting on a Partnership to Co-Design a Digital Health Intervention With Young People Who Have Experienced Technology-Assisted Sexual Abuse. *Health Expectations* 28, 3 (May 2025), e70288. doi:10.1111/hex.70288

[23] Grace Byun, Rebecca Lipschutz, Sean T. Minton, Abigail Powers, and Jinho D. Choi. 2026. CRADLE Bench: A Clinician-Annotated Benchmark for Multi-Faceted Mental Health Crisis and Safety Risk Detection. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vera Demberg, Kentaro Inui, and Lluís Marquez (Eds.). Association for Computational

Linguistics, Rabat, Morocco, 1572–1590. doi:10.18653/v1/2026.eacl-long.73

[24] California State Legislature. 2025. SB-243 Companion Chatbots. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB243 Chapter 677, Statutes of 2025; Business and Professions Code Section 22603.

[25] California State Legislature. 2026. SB 1119: Companion Chatbots: Children's Safety. https://leginfo.legislature.ca.gov/faces/billCompareClient.xhtml?bill_id=202520260SB1119&showamends=false Chapter 190, Statutes of 2026.

[26] California State Legislature. 2026. SB 813: Independent Verification Organizations. https://legiscan.com/CA/text/SB813/2025 Chapter 179, Statutes of 2026.

[27] David S. Carrell, Bradley Malin, John Aberdeen, Samuel Bayer, Cheryl Clark, Ben Wellner, and Lynette Hirschman. 2016. Is the Juice Worth the Squeeze? Costs and Benefits of Multiple Human Annotators for Clinical Text De-Identification. *Methods of Information in Medicine* 55, 4 (2016), 356–364. doi:10.3414/ME15-01-0122

[28] Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine* 3, 1 (March 2020), 1–11. doi:10.1038/s41746-020-0233-7 Number: 1.

[29] Stevie Chancellor, Jessica L. Feuston, and Jayhyun Chang. 2023. Contextual Gaps in Machine Learning for Mental Illness Prediction: The Case of Diagnostic Disclosures. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2 (October 2023), 332:1–332:27. doi:10.1145/3610181

[30] Character.AI. 2025. Taking bold steps to keep teen users safe on character.ai. https://blog.character.ai/u18-chat-announcement/

[31] ChatGPT Wrapped n.d.. ChatGPT Wrapped. https://gptwrapped.husaria.dev/ Website.

[32] Felix Chen, Stephan Rabanser, Peter Kirgis, Nitya Nadgir, Rishi Bommasani, Magda Dubois, Elham Tabassi, Cozmin Ududec, Anka Reuel, Sanmi Koyejo, Sayash Kapoor, and Arvind Narayanan. 2026. Emerging Trends in AI Agent Evaluation. https://sage.cs.princeton.edu/trends/

[33] Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona Vectors: Monitoring and Controlling Character Traits in Language Models. doi:10.48550/arXiv.2507.21509 arXiv:2507.21509 [cs.CL]

[34] Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. 2026. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science* 391, 6792 (March 2026), eaec8352. doi:10.1126/science.aec8352

[35] Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A Computational Framework for Behavioral Assessment of LLM Therapists. doi:10.48550/arXiv.2401.00820 arXiv:2401.00820 [cs].

[36] Colorado General Assembly. 2026. HB 26-1263: Conversational Artificial Intelligence Service Operator Requirements. https://leg.colorado.gov/bill_files/116851/download

[37] Cyberspace Administration of China. 2026. Interim Measures for the Administration of Artificial Intelligence Anthropomorphic Interaction Services. https://www.cac.gov.cn/2026-04/10/c_1777558395078289.htm Articles 2, 13–14, and 22–23; original text in Chinese; effective July 15, 2026.

[38] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. 2020. Open Problems in Cooperative AI. doi:10.48550/arXiv.2012.08630 arXiv:2012.08630 [cs.AI].

[39] Data Transfer Initiative. 2026. AI Conversation History Export. https://schemas.pub/schemas/24

[40] Munmun De Choudhury, Sachin R. Pendse, and Neha Kumar. 2023. Benefits and Harms of Large Language Models in Digital Mental Health. doi:10.48550/arXiv.2311.14693 arXiv:2311.14693 [cs].

[41] Louise Deléger, Katalin Molnar, Guergana Savova, Fei Xia, Todd Lingren, Qi Li, Keith Marsolo, Genevieve B. Melton, Eric S. Kirkendall, and Imre Solti. 2013. Large-Scale Evaluation of Automated Clinical Note De-Identification and Its Impact on Information Extraction. *Journal of the American Medical Informatics Association* 20, 1 (2013), 84–94. doi:10.1136/amiajnl-2012-001012

[42] Department for Science, Innovation and Technology. 2026. Growing Up in the Online World: Government Response (July 2026). https://www.gov.uk/government/consultations/growing-up-in-the-online-world-a-national-consultation/outcome/growing-up-in-the-online-world-government-response-july-2026 CP 1643; updated 19 August 2026.

[43] Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association* 24, 3 (2017), 596–606. doi:10.1093/jamia/ocw156

[44] Meera Desai, Sang T. Truong, Hanna Wallach, Alex Chouldechova, A. Feder Cooper, Jean Garcia-Gathright, Daniel E. Ho, Abigail Z. Jacobs, Sanmi Koyejo, Nicholas Pangakis, and Angelina Wang. 2026. What AI Benchmarks Actually Measure: Adapting Convergent and Discriminant Validity to Interrogate Fifty-Six AI Benchmarks. In *Third Conference on Language Modeling*. doi:10.48550/arXiv.2609.08812

[45] Eric G. Devine, Kristina R. Peebles, and Valeria Martini. 2017. Strategies to exclude subjects who conceal and fabricate information when enrolling in clinical trials. *Contemporary Clinical Trials Communications* 5 (2017), 67–71. doi:10.1016/j.conctc.2016.12.005

[46] Yi Dong, Ronghui Mu, Gaojie Jin, Yi Qi, Jinwei Hu, Xingyu Zhao, Jie Meng, Wenjie Ruan, and Xiaowei Huang. 2024. Position: Building Guardrails for Large Language Models Requires Systematic Design. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 11375–11394. https://proceedings.mlr.press/v235/dong24c.html

[47] Bridget Dwyer, Matthew Flathers, Akane Sano, Allison Dempsey, Andrea Cipriani, Asim H. Gazi, Bryce Hill, Carla Gorban, Carolyn I. Rodriguez, Charles Stromeyer IV, and others. 2025. Mindbench.ai: an actionable platform to evaluate the profile and performance of large language models in a mental healthcare context. *NPP–Digital Psychiatry and Neuroscience* 3 (2025), 28. doi:10.1038/s44277-025-00049-6

[48] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Cambridge, MA, 1277–1287. https://aclanthology.org/D10-1124/

[49] European Commission. 2026. Proposal for a Regulation: EU KIDS ACT—EU Keeping Internet Digital Spaces Accountable and Trustworthy. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52026PC0681 COM(2026) 681 final; Articles 5 and 14; legislative proposal.

[50] European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj Official Journal of the European Union, 2024/1689, 12 July 2024. Articles 5, 68, 78, and 92.

[51] Mollie Feng. 2026. China's First AI Companion Rules Draw a New Line Between Tools and Relationships. Just Security. https://www.justsecurity.org/148468/china-ai-companion-rules-relationships/

[52] Georgia General Assembly. 2026. SB 540: Online Internet Safety; Certain Disclosures Related to Conversational AI Services; Require. https://www.legis.ga.gov/legislation/73452 Act 518, 2025–2026 Regular Session.

[53] Mawada Ghanem, Christian Evangeli-Dawson, and Anna Georgiades. 2023. The Role of Culture on the Phenomenology of Hallucinations and Delusions, Explanatory Models, and Help-Seeking Attitudes: A Narrative Review. *Early Intervention in Psychiatry* 17, 9 (2023), 843–863. doi:10.1111/eip.13449

[54] Michail Giannakos, Roger Azevedo, Peter Brusilovsky, Mutlu Cukurova, Yannis Dimitriadis, Davinia Hernandez-Leo, Sanna Järvelä, Manolis Mavrikis, and Bart Rienties. 2025. The promise and challenges of generative AI in education. *Behaviour & Information Technology* 44, 11 (2025), 2518–2544. doi:10.1080/0144929X.2024.2394886

[55] Justine Gluck. 2026. *Regulating the Conversation: The U.S. Landscape of AI Chatbot Legislation.* Technical Report. Future of Privacy Forum. https://fpf.org/wp-content/uploads/2026/09/FPF-U.S.-Chatbot-Regulations-Report-R3.pdf

[56] Ashleigh Golden and Elias Aboujaoude. 2026. A Transdiagnostic Model for How General Purpose AI Chatbots Can Perpetuate OCD and Anxiety Disorders. *npj Digital Medicine* 9 (2026), 343. doi:10.1038/s41746-026-02531-7

[57] Google. 2026. Download Your Gemini Apps Data. https://support.google.com/gemini/answer/16920332

[58] Declan Grabb, Max Lamparth, and Nina Vasan. 2024. Risks from Language Models for Automated Mental Healthcare: Ethics and Structure for Implementation. doi:10.48550/arXiv.2406.11852 arXiv:2406.11852.

[59] Zhijun Guo, Alvina Lai, Johan H. Thygesen, Joseph Farrington, Thomas Keen, and Kezhi Li. 2024. Large Language Models for Mental Health Applications: Systematic Review. *JMIR Mental Health* 11, 1 (Oct. 2024), e57400. doi:10.2196/57400

[60] D. J. Hallford, D. Rusanov, B. Winestone, R. Kaplan, M. Fuller-Tyszkiewicz, and G. Melvin. 2023. Disclosure of suicidal ideation and behaviours: A systematic review and meta-analysis of prevalence. *Clinical Psychology Review* 101 (2023), 102272. doi:10.1016/j.cpr.2023.102272

[61] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2022. Unsolved Problems in ML Safety. doi:10.48550/arXiv.2109.13916 arXiv:2109.13916 [cs.LG].

[62] Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. 2025. SimBench: Benchmarking the Ability of Large Language Models to Simulate Human Behaviors. https://openreview.net/forum?id=PL51SpN6ZJ

[63] Yining Hua, Steve Siddals, Zilin Ma, Isaac Galatzer-Levy, Winna Xia, Christine Hau, Hongbin Na, Matthew Flathers, Jake Linardon, Cyrus Ayubcha, and John Torous. 2025. Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review. *World Psychiatry* 24, 3 (2025), 383–394. doi:10.1002/wps.21352

[64] humans&. 2026. Persimmon. https://persimmon.humansand.ai/blog/ Research preview.

[65] Zainab Iftikhar, Amy Xiao, Sean Ransom, Jeff Huang, and Harini Suresh. 2025. How LLM Counselors Violate Ethical Standards in Mental Health Practice: A Practitioner-Informed Framework. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8, 2 (Oct. 2025), 1311–1323. doi:10.1609/aies.v8i2.36632

[66] Illinois General Assembly. 2025. Wellness and Oversight for Psychological Resources Act. https://ilga.gov/Legislation/PublicActs/View/104-0054 Public Act 104-0054 (HB 1806), effective August 1, 2025; Sections 20 and 35.

[67] Inter-university Consortium for Political and Social Research. 2026. SOMAR Virtual Data Enclave (VDE). https://www.icpsr.umich.edu/sites/somar/somar-vde-overview-and-resources Accessed 2026-10-09.

[68] Inter-university Consortium for Political and Social Research. n.d.. Request Restricted Data Output. https://www.icpsr.umich.edu/sites/somar/output-review Accessed 2026-10-09.

[69] International Medical Device Regulators Forum, Software as a Medical Device Working Group. 2017. *Software as a Medical Device (SaMD): Clinical Evaluation.* Technical Report IMDRF/SaMD WG/N41FINAL:2017. International Medical Device Regulators Forum. https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-170921-samd-n41-clinical-evaluation_1.pdf

[70] Russell Jackson. 2026. *The Sovereign AI Horizontal Memory (SAIHM) Protocol.* Internet-Draft draft-saihm-memory-protocol-01. Internet Engineering Task Force. https://datatracker.ietf.org/doc/draft-saihm-memory-protocol/ Independent Submission; intended status: Informational; work in progress.

[71] Daniel D. Johnson, Robert Friel, Cassidy Laidlaw, Luke Hewitt, Nari Johnson, Conrad Stosz, and Sarah Schwettmann. 2026. Mental Health Behavior Report. https://behaviors.transluce.org/mental-health

[72] Nick Judd, Alexandre Vaz, Kevin Paeth, Layla Inés Davis, Milena Esherick, Jason Brand, Inês Amaro, and Tony Rousmaniere. 2025. Independent Clinical Evaluation of General-Purpose LLM Responses to Signals of Suicide Risk. doi:10.48550/arXiv.2510.27521 arXiv:2510.27521 [cs].

[73] Mark Kalinich, James Luccarelli, John P. Santa Maria, Matthew Flathers, Phuong Anh Nguyen, Seo Ho Song, Kevin Makhoul, Maria Jose Rivera Criado, Callie Ginapp, Bryce Hill, J. Nicholas Shumate, Haruka Notsu, Christopher L. Smith, Frank Moss, and John Torous. 2026. Conversational trajectory degrades large language model detection of suicidal ideation relative to clinicians: a preregistered study. doi:10.64898/2026.07.10.26357132

ISSN: 3067-2007 Pages: 2026.07.10.26357132.

[74] Shivani Kapania, William Agnew, Motahhare Eslami, Hoda Heidari, and Sarah E. Fox. 2025. Simulacrum of Stories: Examining Large Language Models as Qualitative Research Participants. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 17 pages. doi:10.1145/3706598.3713220

[75] Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. 2025. AI Agents That Matter. *Transactions on Machine Learning Research* (2025). doi:10.48550/arXiv.2407.01502

[76] Gabriela Khazanov, Peter Franz, Elizabeth Cameron Stade, Deirdra Kelly, Mica Poerio, Carly Maitlin, Kara J. Emery, Daniel Mark Low, Leslie Adams, Torrey A. Creed, Kelly L. Green, Sarra Nazem, Lauren M. Weinstock, Christopher D. Hughes, Shireen L. Rizvi, William M. Buerger, Dominic Sisti, Amy McGuire, Abbi Marsh, Amar D. Mandavia, Philip Held, Caroline Vaile Wright, Alexandra King, Caitlynne Behan, De'Ja Wrought, Xuan Zhao, Johannes Christopher Eichstaedt, Shari Jager-Hyman, Randy Patrick Auerbach, Evan Kleiman, Gregory Brown, and Shannon Wiltsey Stirman. 2026. Development of a Consensus Statement to Guide AI Chatbot Responses to Suicide Risk Disclosure. doi:10.31234/osf.io/txpem_v1 Preprint.

[77] Peter Kirgis, Ben Hawriluk, Sherrie Feng, Aslan Bilimer, Sam Paech, and Zeynep Tufekci. 2026. LLM Spirals of Delusion: A Benchmarking Audit Study of AI Chatbot Interfaces. doi:10.48550/arXiv.2604.06188 _eprint: 2604.06188.

[78] Anne Kohlbrenner, Ben Kaiser, Kartikeya Kandula, Rebecca Weiss, Jonathan Mayer, Ted Han, and Robert Helmer. 2022. Rally and WebScience: A Platform and Toolkit for Browser-Based Research on Technology and Society Problems. doi:10.48550/arXiv.2211.02274

[79] Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams. 2001. The PHQ-9: Validity of a Brief Depression Severity Measure. *Journal of General Internal Medicine* 16, 9 (2001), 606–613. doi:10.1046/j.1525-1497.2001.016009606.x

[80] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, et al. 2025. Tülu 3: Pushing Frontiers in Open Language Model Post-Training. In *Proceedings of the 2nd Conference on Language Modeling*. https://openreview.net/forum?id=i1uGbfHHpH

[81] Hannah R. Lawrence, Renee A. Schneider, Susan B. Rubin, Maja J. Mataric, Daniel J. McDuff, and Megan Jones Bell. 2024. The opportunities and risks of large language models in mental health. doi:10.48550/arXiv.2403.14814 arXiv:2403.14814 [cs].

[82] Hannah Lee, Rebecca Handler, Tushar Mungle, and Tina Hernandez-Boussard. 2026. Building Safer Artificial Intelligence Mental Health Chatbots: A Framework for Transparency, Evaluation, and Shared Accountability. *Journal of the American Medical Informatics Association* 33, 8 (Aug. 2026), 1538–1553. doi:10.1093/jamia/ocag078

[83] Yahan Li, Jifan Yao, John Bosco S. Bunyi, Adam C. Frank, Angel Hsing-Chi Hwang, and Ruishan Liu. 2026. CounselBench: A Large-Scale Expert Evaluation and Adversarial Benchmarking of Large Language Models in Mental Health Question Answering. doi:10.48550/arXiv.2506.08584 arXiv:2506.08584 [cs.CL].

[84] Chao Liu, Jr. Cox, Ronald B., Isaac J. Washburn, Julie M. Croff, and Hugh C. Crethar. 2017. The Effects of Requiring Parental Consent for Research on Adolescents' Risk Behaviors: A Meta-analysis. *Journal of Adolescent Health* 61, 1 (2017), 45–52. doi:10.1016/j.jadohealth.2017.01.015

[85] Daogao Liu, Edith Cohen, Badih Ghazi, Peter Kairouz, Pritish Kamath, Alexander Knop, Ravi Kumar, Pasin Manurangsi, Adam Sealfon, Da Yu, and Chiyuan Zhang. 2025. Urania: Differentially Private Insights into AI Use. https://arxiv.org/abs/2506.04681 Conference on Language Modeling (COLM 2025).

[86] Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, Fang Zeng, Lichao Sun, Wei Liu, Dinggang Shen, Quanzheng Li, Tianming Liu, Dajiang Zhu, and Xiang Li. 2023. DeID-GPT: Zero-shot Medical Text De-Identification by GPT-4. doi:10.48550/arXiv.2303.11032 arXiv:2303.11032.

[87] Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling Domain-Experts to Create LLM-simulated Patients via Eliciting and Adhering to Principles. doi:10.48550/arXiv.2407.00870 arXiv:2407.00870 [cs].

[88] Pedro Henrique Luz de Araujo, Michael A. Hedderich, Ali Modarressi, Hinrich Schuetze, and Benjamin Roth. 2026. Persistent Personas? Role-Playing, Instruction Following, and Safety in Extended Interactions. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 5329–5359. doi:10.18653/v1/2026.eacl-long.246

[89] Ali Malik, Declan Grabb, Rebecca Soskin Hicks, Rahul K. Arora, Preston Bowman, Michael Sharman, Sahra Ghalebikesabi, Mikhail Trofimov, Vinnie Monaco, and Karan Singhal. 2026. *MentalHealthBench: An Expert-Informed Benchmark of AI Capabilities in Realistic Mental Health Conversations*. Research report. OpenAI. https://cdn.openai.com/ctf-cdn/MentalHealthBench_A_Comprehensive_Benchmark_of_AI_Capabilities_in_Realistic_Mental_Health_Conversations.pdf

[90] Neil Mallinar, A Ali Heydari, Xin Liu, Anthony Z Faranesh, Brent Winslow, Nova Hammerquist, Benjamin Graef, Cathy Speed, Mark Malhotra, Shwetak Patel, Javier L Prieto, Daniel McDuff, and Ahmed A Metwally. 2026. A scalable framework for evaluating health language models. *npj Digital Medicine* 9, 1 (2026), 437. doi:10.1038/s41746-026-02492-x

[91] Maryland General Assembly. 2026. SB 827: Consumer Protection and Product Liability—Chatbots. https://mgaleg.maryland.gov/2026RS/bills/sb/sb0827f.pdf First Reader; introduced February 6, 2026; proposed Commercial Law section 14-5109(b).

[92] Surya Mattu, Leon Yin, Angie Waller, and Jon Keegan. 2021. How We Built a Facebook Inspector. https://themarkup.org/citizen-browser/2021/01/05/how-we-built-a-facebook-inspector

[93] Ryan K. McBain, Robert Bozick, Melissa Diliberti, Li Ang Zhang, Fang Zhang, Alyssa Burnett, Aaron Kofner, Benjamin Rader, Joshua Breslau, Bradley D. Stein, Ateev Mehrotra, Lori Uscher-Pines, Jonathan Cantor, and Hao Yu. 2025. Use of generative AI for mental health advice among US adolescents and young adults. *JAMA Network Open* 8, 11 (November 2025), e2542281. doi:10.1001/jamanetworkopen.2025.42281

[94] Ryan K. McBain, Jonathan H. Cantor, Joshua Breslau, Melissa Diliberti, Li Ang Zhang, Fang Zhang, Alyssa Burnett, Aaron Kofner, Benjamin Rader, Pat Pataranutaporn, Bradley D. Stein, Ateev Mehrotra, and Hao Yu. 2026. AI Chatbot Use and Disclosure for Mental Health Among US Adolescents

and Young Adults. *JAMA Pediatrics* (2026). doi:10.1001/jamapediatrics.2026.2015

[95] David J. McCann, Nancy M. Petry, Anders Bresell, Eva Isacsson, Ellis Wilson, and Robert C. Alexander. 2015. Medication Nonadherence, "Professional Subjects," and Apparent Placebo Responders. *Journal of Clinical Psychopharmacology* 35, 5 (2015), 566–573. doi:10.1097/JCP.0000000000000372

[96] Catherine M. McHugh, Amy Corderoy, Christopher James Ryan, Ian B. Hickie, and Matthew Michael Large. 2019. Association between Suicidal Ideation and Suicide: Meta-Analyses of Odds Ratios, Sensitivity, Specificity and Positive Predictive Value. *BJPsych Open* 5, 2 (2019), e18. doi:10.1192/bjo.2018.88

[97] Reid McIlroy-Young. 2025. The Folly of AI for Age Verification. doi:10.48550/arXiv.2506.00038 arXiv:2506.00038 [cs.CY]

[98] Alex McKenzie, Urja Pawar, Phil Blandfort, William Bankes, David Krueger, Ekdeep Singh Lubana, and Dmitrii Krasheninnikov. 2025. Detecting High-Stakes Interactions with Activation Probes. doi:10.48550/arXiv.2506.10805 arXiv:2506.10805 [cs.LG].

[99] Medicines and Healthcare products Regulatory Agency. 2026. Digital Mental Health Technology—Regulation and Evaluation for Safe and Effective Products: Device Characterisation, Regulatory Qualification and Classification. https://assets.publishing.service.gov.uk/media/67a08ff6c496e5d3ddafb6c4/2025.01.29_MHRA_guidance_on_DMHT_-_Device_characterisation__regulatory_qualification_and_classification.pdf Version 1.3, sections 6.1–6.2.

[100] Shuhaib Mehri, Philippe Laban, Sumuk Shashidhar, Marwa Abdulhai, Sergey Levine, Michel Galley, and Dilek Hakkani-Tür. 2026. Measuring and Mitigating the Distributional Gap Between Real and Simulated User Behaviors. doi:10.48550/arXiv.2605.07847 arXiv:2605.07847 [cs.CL].

[101] Microsoft. 2026. Presidio - data protection and de-identification SDK. https://github.com/microsoft/presidio Type: Python.

[102] Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild. In *The First Conference on Language Modeling*. https://openreview.net/forum?id=p9KIxJmKpy

[103] Jared Moore, Declan Grabb, William Agnew, Kevin Klyman, Stevie Chancellor, Desmond C. Ong, and Nick Haber. 2025. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers.. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 599–627. doi:10.1145/3715275.3732039

[104] Jared Moore, Ashish Mehta, William Agnew, Jacy Reese Anthis, Ryan Louie, Yifan Mai, Peggy Yin, Myra Cheng, Samuel J. Paech, Kevin Klyman, Stevie Chancellor, Eric Lin, Nick Haber, and Desmond C. Ong. 2026. Characterizing Delusional Spirals through Human-LLM Chat Logs. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 7631–7674. doi:10.1145/3805689.3806443

[105] Jared Moore, Andrea Mock, Yifan Mai, Jacy Reese Anthis, Ryan Louie, William Agnew, Ashish Mehta, Kevin Klyman, Percy Liang, Nick Haber, Eric Lin, and Desmond C. Ong. 2026. DelusionEval: Measuring Delusion-Linked Behaviors in AI Chatbots. doi:10.48550/arXiv.2608.05004 arXiv:2608.05004 [cs.CL].

[106] Hamilton Morrin, Joshua Au-Yeung, Zarinah Agnew, Søren Dinesen Østergaard, and Thomas A. Pollak. 2026. It Is the Journey, Not the Destination: Moving From End Points to Trajectories When Assessing Chatbot Mental Health Safety. *JMIR Mental Health* 13 (2026), e91454. doi:10.2196/91454

[107] Hamilton Morrin, Luke Nicholls, Michael Levin, Jenny Yiend, Udita Iyengar, Francesca DelGuidice, Sagnik Bhattacharyya, James MacCabe, Stefania Tognin, Ricardo Twumasi, Ben Alderson-Day, and Thomas Pollak. 2025. Delusions by design? How everyday AIs might be fuelling psychosis (and what can be done about it). doi:10.31234/osf.io/cmy7n_v1

[108] Hamilton Morrin, Vinitha Soundararajan, Thomas Cheliotis-James, Boris Warszawski, Joshua Fakulujo, Zeqi Jia, Etienne Brisson, and Thomas A. Pollak. 2026. Delusions and Harms Associated with AI Chatbot Use: Early Evidence from 185 Real-World Reports. arXiv:2609.08027 [cs.HC] https://arxiv.org/abs/2609.08027

[109] John X. Morris, Thomas R. Campion, Sri Laasya Nutheti, Yifan Peng, Akhil Raj, Ramin Zabih, and Curtis L. Cole. 2025. DIRI: Adversarial Patient Reidentification with Large Language Models for Evaluating Clinical Text Anonymization. *AMIA Joint Summits on Translational Science Proceedings* 2025 (2025), 355–364. https://pmc.ncbi.nlm.nih.gov/articles/PMC12150728/

[110] Inez Myin-Germeys, Jim van Os, Joseph E. Schwartz, Arthur A. Stone, and Philippe A. Delespaul. 2001. Emotional Reactivity to Daily Life Stress in Psychosis. *Archives of General Psychiatry* 58, 12 (2001), 1137–1144. doi:10.1001/archpsyc.58.12.1137

[111] Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. 2012. On the Feasibility of Internet-Scale Author Identification. In *2012 IEEE Symposium on Security and Privacy*. 300–314. doi:10.1109/SP.2012.46

[112] National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html

[113] National Institute of Mental Health. 2024. 5 Action Steps to Help Someone Having Thoughts of Suicide. https://www.nimh.nih.gov/health/publications/5-action-steps-to-help-someone-having-thoughts-of-suicide Revised 2024; NIMH Identifier OM 24-4315.

[114] National Institutes of Health. 2014. NIH Genomic Data Sharing Policy. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html Notice NOT-OD-14-124.

[115] Ishna Neamatullah, Margaret M. Douglass, Li-Wei H. Lehman, Andrew Reisner, Mauricio Villarroel, William J. Long, Peter Szolovits, George B. Moody, Roger G. Mark, and Gari D. Clifford. 2008. Automated De-Identification of Free-Text Medical Records. *BMC Medical Informatics and Decision Making* 8 (2008), 32. doi:10.1186/1472-6947-8-32

[116] Joseph Near, David Darais, Naomi Lefkovitz, and Gary Howarth. 2025. *Guidelines for Evaluating Differential Privacy Guarantees*. Technical Report NIST SP 800-226. National Institute of Standards and Technology. doi:10.6028/NIST.SP.800-226

[117] Mar Negreiro and Öykü Dilara Anaç. 2026. *The spread of AI companions and the challenges they generate.* Briefing PE 789.299. European Parliamentary Research Service. https://www.europarl.europa.eu/RegData/etudes/BRIE/2026/789299/EPRS_BRI(2026)789299_EN.pdf

[118] Viet Cuong Nguyen, Mohammad Taher, Dongwan Hong, Vinicius Konkolics Possobom, Vibha Thirunellayi Gopalakrishnan, Ekta Raj, Zihang Li, Heather J. Soled, Michael L. Birnbaum, Srijan Kumar, and Munmun De Choudhury. 2024. Do Large Language Models Align with Core Mental Health Counseling Competencies? doi:10.48550/arXiv.2410.22446 arXiv:2410.22446 [cs].

[119] Nick Obradovich, Sahib S. Khalsa, Waqas U. Khan, Jina Suh, Roy H. Perlis, Olusola Ajilore, and Martin P. Paulus. 2024. Opportunities and risks of large language models in psychiatry. *NPP—Digital Psychiatry and Neuroscience* 2, 1 (May 2024), 8. doi:10.1038/s44277-024-00010-z

[120] Office of the Governor of New York. 2025. Governor Hochul Pens Letter to AI Companion Companies Notifying Them That Safeguard Requirements Are Now in Effect. https://www.governor.ny.gov/news/governor-hochul-pens-letter-ai-companion-companies-notifying-them-safeguard-requirements-are

[121] OpenAI. 2024. Model Spec. https://cdn.openai.com/spec/model-spec-2024-05-08.html

[122] OpenAI. 2025. GPT-5 System Card. https://openai.com/index/gpt-5-system-card/ Accessed 2026-09-07.

[123] OpenAI. 2025. Helping people when they need it most. https://openai.com/index/helping-people-when-they-need-it-most/

[124] OpenAI. 2025. Introducing gpt-oss-safeguard. https://openai.com/index/introducing-gpt-oss-safeguard/ Accessed 2026-09-07.

[125] OpenAI. 2025. Model Spec (2025/04/11). https://model-spec.openai.com/2025-04-11.html

[126] OpenAI. 2025. Strengthening ChatGPT's Responses in Sensitive Conversations. https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/ Authority: OpenAI.

[127] OpenAI. 2026. Exporting Your ChatGPT History and Data. https://help.openai.com/en/articles/7260999-how-do-i-export-my-chatgpt-history-and-data

[128] OpenAI. 2026. Introducing Trusted Contact in ChatGPT. https://openai.com/index/introducing-trusted-contact-in-chatgpt/ Accessed 2026-09-07.

[129] OpenAI. 2026. Memory FAQ. https://help.openai.com/en/articles/8590148-memory-faq OpenAI Help Center. Accessed 2026-09-07.

[130] Oregon Legislative Assembly. 2026. SB 1546: Relating to Artificial Intelligence Companions. https://olis.oregonlegislature.gov/liz/2026R1/Measures/Overview/sb1546 Chapter 85, 2026 Regular Session.

[131] Max Ostermann, Oscar Freyer, F. Gerrik Verhees, Jakob Nikolas Kather, and Stephen Gilbert. 2025. If a therapy bot walks like a duck and talks like a duck then it is a medically regulated duck. *npj Digital Medicine* 8, 1 (Dec. 2025), 741. doi:10.1038/s41746-025-02175-z

[132] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training Language Models to Follow Instructions with Human Feedback. arXiv:2203.02155 https://arxiv.org/abs/2203.02155

[133] Sam Paech. 2025. Spiral-Bench: Multiturn Evaluation for Sycophancy and Delusion Behaviours. https://eqbench.com/spiral-bench_v1.0.html

[134] Joon Sung Park, Carolyn Q. Zou, Jonne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, and Michael S. Bernstein. 2024. LLM Agents Grounded in Self-Reports Enable General-Purpose Simulation of Individuals. doi:10.48550/arXiv.2411.10109 arXiv: 2411.10109.

[135] Acacia Parks, Eoin Travers, Ramesh Perera-Delcourt, Max Major, Marcos Economides, and Phil Mullan. 2025. Is This Chatbot Safe and Evidence-Based? A Call for the Critical Evaluation of Generative AI Mental Health Chatbots. *Journal of Participatory Medicine* 17, 1 (May 2025), e69534. doi:10.2196/69534

[136] PhysioNet. 2022. Applying for Access to Protected Datasets Such as MIMIC Is Now a Three Step Process. https://physionet.org/news/post/395/

[137] José Pombal, Maya D'Eon, Nuno M. Guerreiro, Pedro Henrique Martins, António Farinhas, and Ricardo Rei. 2025. MindEval: Benchmarking Language Models on Multi-turn Mental Health Support. doi:10.48550/ARXIV.2511.18491 Version Number: 3.

[138] Huachuan Qiu and Zhenzhong Lan. 2024. Interactive Agents: Simulating Counselor-Client Psychological Counseling via Role-Playing LLM-to-LLM Interactions. doi:10.48550/arXiv.2408.15787 arXiv:2408.15787 [cs].

[139] Jiahao Qiu, Yinghui He, Xinzhe Juan, Yimin Wang, Yuhan Liu, Zixin Yao, Yue Wu, Xun Jiang, Ling Yang, and Mengdi Wang. 2025. EmoAgent: Assessing and Safeguarding Human-AI Interaction for Mental Health Safety. doi:10.48550/arXiv.2504.09689 arXiv:2504.09689 [cs].

[140] Inioluwa Deborah Raji and Roel Dobbe. 2023. Concrete Problems in AI Safety, Revisited. doi:10.48550/arXiv.2401.10899 arXiv:2401.10899 [cs.CY].

[141] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, 33–44. doi:10.1145/3351095.3372873

[142] Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Ramona Comanescu, Canfer Akbulut, Tom Stepleton, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, William Isaac, and Laura Weidinger. 2024. Gaps in the Safety Evaluation of Generative AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, 1 (2024), 1200–1217. doi:10.1609/aies.v7i1.31717

[143] Anka Reuel, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, Markus Anderljung, Ben Garfinkel, Lennart Heim, Andrew Trask, Gabriel Mukobi, Rylan Schaeffer, Mauricio Baker, Sara Hooker, Irene Solaiman, Alexandra Sasha Luccioni, Nitarshan Rajkumar, Nicolas Moës, Jeffrey Ladish, David Bau, Paul Bricman, Neel Guha, Jessica Newman, Yoshua Bengio, Tobin South, Alex Pentland, Sanmi Koyejo, Mykel J. Kochenderfer, and Robert Trager. 2025. Open Problems in Technical AI Governance. doi:10.48550/arXiv.2407.14981 arXiv:2407.14981 [cs.CY].

[144] Shalaleh Rismani, Renee Shelby, Andrew Smart, Edgar Jatho, Joshua Kroll, AJung Moon, and Negar Rostamzadeh. 2023. From Plane Crashes to Algorithmic Harm: Applicability of Safety Engineering Frameworks for Responsible ML. In *Proceedings of the 2023 CHI Conference on Human*

Factors in Computing Systems. ACM, 1–18. doi:10.1145/3544548.3581407

[145] Jessica L. Roberts, Leah R. Fowler, and Paul S. Appelbaum. 2022. Unlimited Medical Liability? *Emory Law Journal Online* 72 (2022), 42. https://pmc.ncbi.nlm.nih.gov/articles/PMC10468648/

[146] Kevin Roose. 2024. Can A.I. Be Blamed for a Teen's Suicide? *The New York Times* (Oct. 2024). https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html

[147] Tony Rousmaniere, Yimeng Zhang, Xu Li, and Siddharth Shah. 2026. Large Language Models as Mental Health Resources: Patterns of Use in the United States. *Practice Innovations* 11, 2 (2026), 139–155. doi:10.1037/pri0000292

[148] Nicolas Ruiz, Gabriela Molina León, and Hendrik Heuer. 2024. Design Frictions on Social Media: Balancing Reduced Mindless Scrolling and User Satisfaction. doi:10.48550/arXiv.2407.18803

[149] Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. 2025. Measurement to Meaning: A Validity-Centered Framework for AI Evaluation. doi:10.48550/arXiv.2505.10573

[150] Paul M. Salkovskis and Osamu Kobori. 2015. Reassuringly Calm? Self-Reported Patterns of Responses to Reassurance Seeking in Obsessive Compulsive Disorder. *Journal of Behavior Therapy and Experimental Psychiatry* 49, Part B (2015), 203–208. doi:10.1016/j.jbtep.2015.09.002

[151] Samuel Schmidgall, Tim Strother, Alex Bijamov, Min Woo Sun, Anil Palepu, Akshay Goel, Ulysse Marteau-Ferey, Théo Moutakanni, Yubin Kim, Joseph Breda, Christina Chen, Justin Chen, Paul Jhun, Andrew Marshall, Sarah Montgomery-Taylor, Eun Kyoung Hong, Shashank Ravi, Kameron C. Black, David J. H. Wu, Austin Schoeffler, John Lyon Havlik, Elizabeth C. Stade, Jared Moore, Johannes C. Eichstaedt, Logan Pierce, Jeff Jopling, Johannes Thrul, Thomas H. McCoy, Neguine Rezaii, Mark S. Chen, Ellen E. Fitzsimmons-Craft, Shirley Wang, Desmond C. Ong, Katharina Schultebraucks, Torrey Creed, Bhavna Daryani, Aleksandra Faust, S. Sara Mahdavi, Caroline Figueroa, Eric Lin, Kate H. Bentley, Ernest Lim, Zhihan Jiang, Xuhai "Orson" Xu, Daniel McDuff, John Torous, Yun Liu, Po-Hsuan Cameron Chen, Mike Schaekermann, Joelle Barral, Alvin Rajkomar, Tao Tu, Lin Yang, Isaac Galatzer-Levy, David Racz, and Valentin Liévin. 2026. Training Language Models in Simulation for Psychiatric Intake. doi:10.48550/arXiv.2610.40476 arXiv:2610.40476 [cs.AI].

[152] Security and Privacy Scientists and Researchers. 2026. Joint Statement of Security and Privacy Scientists and Researchers on Age Assurance. https://csa-scientist-open-letter.org/ageverif-Feb2026

[153] U.S. Senate. 2025. Hawley, durbin introduce legislation empowering americans to bring liability claims against AI companies. https://www.hawley.senate.gov/hawley-durbin-introduce-legislation-empowering-americans-to-bring-liability-claims-against-ai-companies/

[154] Siddharth Shah, Amit Gupta, Aarav Mann, Alexandre Vaz, Benjamin E. Caldwell, Robert Scholz, Peter Awad, Rocky Allemandi, Doug Faust, Harshita Banka, and Tony Rousmaniere. 2025. Evaluating the Clinical Safety of LLMs in Response to High-Risk Mental Health Disclosures. doi:10.48550/arXiv.2509.08839 arXiv:2509.08839 [cs.CY].

[155] Yaoxi Shi, Cathy Mengying Fang, Pattie Maez, and Amit Goldenberg. 2026. Stumbling Into AI Emotional Dependence: How Routine AI Interactions Reshape Human Connection. doi:10.48550/arXiv.2606.04150 arXiv:2606.04150 [cs.AI].

[156] Soorya Ram Shimgekar, Vipin Gunda, Jiwon Kim, Violeta J. Rodriguez, Hari Sundaram, and Koustuv Saha. 2026. AI Psychosis: Does Conversational AI Amplify Delusion-Related Language? doi:10.48550/arXiv.2603.19574 arXiv:2603.19574 [cs].

[157] Ben Shneiderman. 2020. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems. *ACM Transactions on Interactive Intelligent Systems* 10, 4 (Oct. 2020), 1–31. doi:10.1145/3419764

[158] Inhwa Song, Sachin R. Pendse, Neha Kumar, and Munmun De Choudhury. 2024. The Typing Cure: Experiences with Large Language Model Chatbots for Mental Health Support. doi:10.48550/arXiv.2401.14362 arXiv:2401.14362 [cs].

[159] Jina Suh, Mihaela Vorvoreanu, Forough Poursabzi-Sangdeh, Emily Tseng, Eugenia Kim, Luke Nicholls, James W. Pennebaker, and Eric Horvitz. 2026. Psychological Influences of Conversational AI: Research and Design Directions for Reducing Harm and Promoting Well-Being. doi:10.48550/arXiv.2607.25057 arXiv:2607.25057 [cs.AI].

[160] Vaakesan Sundrelingam, Shireen Parimoo, Frances Pogacar, Radha Koppula, Saeha Shin, Chloe Pou-Prom, Surain B. Roberts, Amol A. Verma, and Fahad Razak. 2025. pyDeid: an improved, fast, flexible, and generalizable rule-based approach for deidentification of free-text medical records. *JAMIA Open* 8, 1 (Jan. 2025), ooae152. doi:10.1093/jamiaopen/ooae152

[161] Heidi Taipale, Johannes Schneider-Thoma, Justo Pinzón-Espinosa, Joaquim Radua, Orestis Efthimiou, Christiaan H. Vinkers, Ellenor Mittendorfer-Rutz, Narcís Cardoner, Luis Pintor, Antti Tanskanen, Anneka Tomlinson, Paolo Fusar-Poli, Andrea Cipriani, Eduard Vieta, Stefan Leucht, Jari Tiihonen, and Jurjen J. Luykx. 2022. Representation and Outcomes of Individuals With Schizophrenia Seen in Everyday Practice Who Are Ineligible for Randomized Clinical Trials. *JAMA Psychiatry* 79, 3 (2022), 210. doi:10.1001/jamapsychiatry.2021.3990

[162] John Torous, Sandra Bucci, Imogen H. Bell, Lars V. Kessing, Maria Faurholt-Jepsen, Pauline Whelan, Andre F. Carvalho, Matcheri Keshavan, Jake Linardon, and Joseph Firth. 2021. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry* 20, 3 (2021), 318–335. doi:10.1002/wps.20883

[163] UK Government. 2026. Fact Sheet: New Rules to Protect Children Online. https://www.gov.uk/government/publications/fact-sheet-new-rules-to-protect-children-online/fact-sheet-new-rules-to-protect-children-online Updated 18 September 2026.

[164] Utah Department of Commerce. n.d.. AI Third-Party Evaluators. https://commerce.utah.gov/ai/third-party-evaluators/

[165] Evdokia Voultsiou and Lefteris Moussiades. 2026. A Systematic Review of Large Language Models in Mental Health: Opportunities, Challenges, and Future Directions. *Electronics* 15, 3 (Jan. 2026), 524. doi:10.3390/electronics15030524

[166] W3C AI Agent Memory Interoperability Community Group. 2026. AI Agent Memory Interoperability Community Group Charter. https://www.w3.org/community/ai-agent-memory-interop/ Community Group charter adopted 2026-06-19. Accessed 2026-09-07.

[167] Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas J Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. 2025. Position: Evaluating generative AI systems is a social science measurement challenge. In *Proceedings of the 42nd international conference on machine learning (Proceedings of machine learning research, Vol. 267)*, Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (Eds.). PMLR, 82232–82251. https://proceedings.mlr.press/v267/wallach25a.html

[168] Jennifer Wang, Joachim Baumann, Daniel E. Ho, and Sanmi Koyejo. 2026. API Benchmark Scores Do Not Reliably Transfer to Chatbot Interfaces. doi:10.48550/arXiv.2609.08861

[169] Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M. Murphy, Nev Jones, Kate Hardy, Hong Shen, Fei Fang, and Zhiyu Zoey Chen. 2024. PATIENT-Ψ: Using Large Language Models to Simulate Patients for Training Mental Health Professionals. doi:10.48550/arXiv.2405.19660 arXiv:2405.19660 [cs].

[170] Washington State Legislature. 2026. HB 2225: Regulating Artificial Intelligence Companion Chatbots. https://app.leg.wa.gov/billsummary?Year=2025&BillNumber=2225 Chapter 168, 2026 Laws.

[171] Veith Weilnhammer, Kevin YC Hou, Lennart Luettgau, Christopher Summerfield, Raymond Dolan, and Matthew M. Nour. 2026. A clinically validated framework for auditing AI chatbot behavior in mental health interactions. *Nature Medicine* (Aug. 2026), 1–11. doi:10.1038/s41591-026-04577-2

[172] Veith Weilnhammer, Lennart Luettgau, Christopher Summerfield, Raymond Dolan, Elise Wilkinson, Virginia Corno, Viknesh Sounderajah, and Matthew M. Nour. 2026. Single-turn emergency psychiatric triage across 15 frontier AI chatbots. doi:10.48550/arXiv.2604.25415 arXiv:2604.25415 [q-bio].

[173] Marcus Williams, Hannah Sheahan, Cameron Raymond, Tomek Korbak, Deng Pan, Peilin Yang, Leon Maksin, Ningyi Xie, Phillip Guo, Ian Kivlichan, and Micah Carroll. 2026. Predicting LLM Safety Before Release by Simulating Deployment. doi:10.48550/arXiv.2607.07184 arXiv:2607.07184 [cs.LG].

[174] Andy T. Woods, Carlos Velasco, Carmel A. Levitan, Xiaoang Wan, and Charles Spence. 2015. Conducting Perception Research over the Internet: A Tutorial Review. *PeerJ* 3 (2015), e1058. doi:10.7717/peerj.1058

[175] Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Creation, Analysis and Evaluation of AnnoMI, a Dataset of Expert-Annotated Counselling Dialogues. *Future Internet* 15, 3 (2023), 110. doi:10.3390/fi15030110

[176] Jinhua Yang, Ting Liu, Yiming Taclis Luo, Tianyue Niu, Patrick Pang, Ao Xiang, and Qin Yang. 2026. Exploring the application boundaries of LLMs in mental health: a systematic scoping review. *Frontiers in Psychology* 16 (Feb. 2026). doi:10.3389/fpsyg.2025.1715306

[177] Joshua Au Yeung, Jacopo Dalmasso, Luca Foschini, Richard JB Dobson, and Zeljko Kraljevic. 2025. The Psychogenic Machine: Simulating AI Psychosis, Delusion Reinforcement and Harm Enablement in Large Language Models. doi:10.48550/arXiv.2509.10970 arXiv:2509.10970 [cs].

[178] Yutong Zhang, Dora Zhao, Jeffrey T. Hancock, Robert Kraut, and Diyi Yang. 2025. The Rise of AI Companions: Interaction with AI Companions and Psychological Well-being. doi:10.48550/arXiv.2506.12605 arXiv:2506.12605 [cs].

[179] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1M ChatGPT Interaction Logs in the Wild. doi:10.48550/arXiv.2405.01470 arXiv:2405.01470 [cs].

[180] Fangyi Zhu, Ajay Subramanian, Allison Constant, Camille Wang, Ravish Gupta, and Corey J. Keller. 2026. Interpretable Symptom Vectors for Depression in a Large Language Model. doi:10.48550/arXiv.2609.01832 arXiv:2609.01832 [cs.CL] Preprint.

[181] Kelly L. Zuromski, Margaret Meagher, Tracy Costigan, and Elizabeth A. Olson. 2026. AI chatbots and youth suicide risk: current evidence, critical gaps, and a clinical research agenda. *npj Digital Medicine* (2026). doi:10.1038/s41746-026-03080-9