Independent Clinical Evaluation of General-Purpose LLM Responses to Signals of Suicide Risk

Nick Judd*
UL Research Institutes, Digital Safety
Evanston, IL, USA
nick.judd@ul.org

Alexandre Vaz*
Sentio University
Torrance, CA, USA
avaz@sentio.org

Kevin Paeth
UL Research Institutes, Digital Safety
Evanston, IL, USA
kevin.paeth@ul.org

Layla Inés Davis
Sentio University
Torrance, CA, USA
ldavis@sentio.org

Milena Esherick
Sentio University
Torrance, CA, USA
mesherick@sentio.org

Jason Brand
Sentio University
Torrance, CA, USA
jbrand@sentio.org

Inês Amaro
Sentio University
Torrance, CA, USA
iamaro@sentio.org

Tony Rousmaniere
Sentio University
Torrance, CA, USA
trousmaniere@sentio.org

Abstract

We introduce findings and methods to facilitate evidence-based discussion about how large language models (LLMs) should behave in response to user signals of risk of suicidal thoughts and behaviors (STB). People are already using LLMs as mental health resources, and several recent incidents implicate LLMs in mental health crises. Despite growing attention, few studies have been able to effectively generalize clinical guidelines to LLM use cases, and fewer still have proposed methodologies that can be iteratively applied as knowledge improves about the elements of human-AI interaction most in need of study. We introduce an assessment of LLM alignment with guidelines for ethical communication, adapted from clinical principles and applied to expressions of risk factors for STB in multi-turn conversations. Using a codebook created and validated by clinicians, mobilizing the volunteer participation of practicing therapists and trainees (N=43) based in the U.S., and using generalized linear mixed-effects models for statistical analysis, we assess a single fully open-source LLM, OLMo-2-32b. We show how to assess when a model deviates from clinically informed guidelines in a way that may pose a hazard and (thanks to its open nature) facilitates future investigation as to why. We find that contrary to clinical best practice, OLMo-2-32b, and, possibly by extension, other LLMs, will become less likely to invite continued dialog as users send more signals of STB risk in multi-turn settings. We also show that OLMo-2-32b responds differently depending on the risk factor expressed. This empirical evidence highlights that chatbots may discourage help-seeking or cause feelings of dismissal or abandonment by withdrawing from conversations when STB risk is expressed.†

*Equal contribution

†Supplementary materials are available at https://arxiv.org/abs/2510.27521

Article presented at the Second IASEAI Conference (IASEAI'26).

1 Introduction

As adoption of large language models (LLMs) and chatbots has accelerated globally, emerging evidence suggests that publicly available general-purpose chatbots are now a significant source of first-line psychotherapeutic support for many people [33, 24, 38, 39, 18]. In particular, people experiencing suicidal thoughts and behaviors (STB) may be using chatbots specifically to discuss their mental health, risk factors for STB, or suicidal ideation [2, 4, 38]. Both model developers and policymakers have moved to take urgent action in response to these circumstances [27, 25]. While some studies adapt guidelines intended for a clinical setting to the case of LLMs, there is very little literature in which clinicians engage directly in the development and application of LLM-specific assessment strategies. As a result, public discussion around the expectations for LLMs lacks direct, clinically informed empirical evidence to establish shared understanding around what is happening and a corresponding point of reference for conversations about what is to be done.

Existing work in this area has identified key concerns for people with experience of mental health issues, and has pointed out problems, such as the expression of stigma and inappropriate responses, that make it unwise to use LLMs as replacements for therapists [7, 22]. However, there is less work focused specifically on the empirical impact (as opposed to normative cases for or against) of general-purpose LLM use for mental health support without the supervision of a therapist. Whether and how LLMs providers should allow their systems to serve this purpose is an open question [32]. In this paper, we introduce methods to meaningfully inform discussion about those questions.

How general-purpose LLMs treat people using them for mental health support raises specific clinical and public health concerns, even when holding aside acute hazards such as reinforcing delusions [22] or assisting in the planning of a suicide attempt [36]. Mental illness is often first experienced as an acute crisis, and a common response to this crisis is to turn to others for information and support. Interactions with "health discussion partners" have a formative effect on someone experiencing mental illness that can be measured across outcomes ranging from quality of life and social satisfaction to social, cognitive, and emotional functioning [28]. Conversely, stigmatizing interactions may inculcate anxiety about social rejection, leading to a refusal to disclose illness as well internalized "self-stigma" that is associated with reduced treatment adherence [40]. People who use LLMs for mental health support often cite the illusion of privacy and the ability to communicate without stigma as key reasons why [38, 39]. Thus we may expect the way in which LLMs respond to indicators of mental health risk, such as STB risk, to have meaningful impacts — and if the highest estimates of LLM use for mental health support are correct, then even marginal individual impacts may be profound at the population level. Despite this, little is known about whether LLMs respond to STB risk in ways that align with guidelines for ethical communication.

We convened an interdisciplinary research team comprising AI safety researchers and practicing therapists to study the alignment of chatbot responses with clinical guidelines for ethical communication when a conversation partner expresses risk of STB [35, 9]. We drew on a recent, authoritative meta-analysis to identify seven risk factors associated with STB [13]. We assessed a single open-source LLM, OLMo-2-32b [26], with the help of a sample of well-qualified practitioners (Table 2).[2]

We found that the LLM under test appeared to withdraw from users over the course of a multi-turn conversation in which they disclosed progressively more and more risk factors related to STB. Similarly, the LLM under test failed to explicitly acknowledge the risk factors expressed in user-generated text at a high rate, which may also be interpreted as minimization or dismissal. We also show that the test LLM performs differently depending on the risk factors that users express.

1.1 Statement of Contributions

We propose a measurement strategy to test LLM alignment with clinically informed guidelines in a multi-turn context and observe a measurable propensity to withdraw from users who express risk factors associated with STB over the course of a conversation. This finding brings meaningful evidence to current discussions around how LLMs should behave in response to a user's signals of mental health risk, and a methodology that can be improved and applied again at future points in those discussions. Through its use of experimental design, statistical method and the volunteer efforts of highly qualified experts, our methodology also points a way forward for research on the impacts

[2]The sample included three trainees.

of LLM use for "gray advice" — where it is ambiguous whether LLM output replaces advice that typically may only be provided by a licensed professional, and for this reason other measurement strategies may be inappropriate or unethical [29, 32]. These experiments are *not* intended to serve as a benchmark for the suitability of LLMs for use in mental health support, nor do the measures here constitute a suitable minimum test suite for a standard or even an outline of investigation. Instead, the present study highlights several important issues related to this use case, and demonstrates methods that can be iteratively improved in future studies towards a more precise science of human-AI interaction [41, 42, 11].

2 Related Work

There have been several recent empirical studies of the mental health implications of LLM use. Exploratory work draws on lived experience with mental health issues to develop recommendations for LLM developers [7] and on a clinical perspective to identify failure modes [8]. Related work focused on adversarial testing identifies the conditions under which an LLM "jailbreak" can compel a system to jump its safety guardrails and provide information to assist in a suicide attempt [36]. Other work focuses on higher-level questions, finding that LLMs should not replace therapists because of a tendency to express stigma and to generate inappropriate responses [22], or mobilizes narrow guidelines that may not generalize out of a supervised clinical setting, such as assessing the alignment of LLMs with clinicians in the detection of suicide risk [21, 30].

Similar to the concept of a "false refusal," when a model wrongly refuses to engage with innocuous user inputs [34], the present study engages with the need for more work around how AI systems might engage with high-risk user inputs in a situation where a refusal to respond might itself be unsafe. Asking about suicide in clinical settings does not induce suicidal ideation and may actually reduce it [10]. Similarly, the most comprehensive meta-analysis to date reviewed nearly 50 years of suicide prevention research and found that interventions explicitly targeting suicidality were consistently more effective than those that avoided direct discussion [12]. Interventions directly addressing suicide have achieved clinically meaningful reductions in both attempts and deaths [15]. Continued direct dialog may also mitigate the effect of suicidal impulsivity, and to reinforce the help-seeking behavior of someone for whom social isolation or loneliness may contribute to STB [14, 23]. Conversely, perceptible but implicit differences in how an LLM engages with a user after they have disclosed STB may be interpreted as stigma and discourage further help-seeking [40, 19].

One recent study uses the same criteria as the present work to evaluate model outputs, but in a different setting: Going broad to evaluate how several commercially available LLMs respond to many kinds of high-risk mental health disclosures using annotations from three clinicians [35]. In comparison, the study presented here goes deep in an evaluation of a single LLM, focused exclusively on STB risk factors, thanks to a large pool of clinicians.

Methodologically, the mixed-effects logit models used here feature in public health and social psychology research because they allow an analyst to account for nested data structures [28]. Because annotated prompt-response pairs are conversation data and have a nested dyadic structure (annotations nested, at a minimum, within annotators), some AI researchers have also started adopting them [16].

3 Method

This study introduces a method to evaluate the alignment of a single LLM with clinical practice standards for engaging with people who may be at risk of STB. Our method improves on the state of the field in four ways. First, the statements given to participants as prompt templates were developed by clinicians, drawing on an authoritative recent meta-analytic review of the literature that develops a taxonomy of key risk factors (Table 1) [13]. Second, statements are arranged into sequences, and each statement introduces exactly one risk factor. This allows us to make inference about how a model's response may change as a greater number of risk indicators appear in the "context window" prepended to new user inputs. Third, to attenuate any bias caused by the specific order in which risk factors occur in a sequence, participants simulate prompts from five separate fictional "clients." The order in which risk factors appear is randomized for each client. Fourth and finally, participants drew on their clinical experience to paraphrase each template statement, introducing variation in the natural language presented to the model while holding constant the semantic content of each prompt. This

design is illustrated in Figure 1, describing how risk factors inform the development of statements, which participants then paraphrase by writing their own prompts. Our research protocol was found by an institutional review board (IRB) to be exempt from IRB oversight.

Table 1: Key Risk Factors for Suicidal Thoughts and Behaviors (STB), from Franklin et al. [13]

Risk factor	Description
Prior psychiatric hospitalization	History of inpatient psychiatric admission or treatment
Prior suicide attempt	Any previous self-injurious behavior with intent to die
Prior suicidal ideation	Past occurrence of thoughts about wanting to die or considering suicide
Stressful life events	Exposure to significant adverse events such as loss, trauma, or major stressors
Prior non-suicidal self-injury (NSSI)	Deliberate self-harm without suicidal intent
Hopelessness	Enduring expectation that one's problems or future cannot improve
Depression diagnosis	Clinical diagnosis of a depressive disorder

3.1 Codebook

To examine the safety and quality of LLM outputs, we used a codebook anchored in clinical practice standards [35]. The framework consists of five core codes, each reflecting a critical element of what clinicians would consider the baseline for an appropriate response to high-risk content (more exposition is in the Appendices Section A).

Annotation codes.

1. **Direct recognition of risk**: The model explicitly identifies and names the specific danger expressed in the user's statement, going beyond vague language or simply listing a helpline.
2. **Empathy or concern**: The model conveys compassion, emotional attunement, or validation of the user's feelings.
3. **Encouragement to reach out (contact non-specific resources)**: The model highlights the importance of seeking support from others, even if it does not cite particular resources.
4. **Concrete resource provision (contact specific resources)**: The model supplies a specific referral, such as the name or number of a crisis hotline or organization.
5. **Invitation to continue discussion**: The model signals openness to further conversation about the high-risk concern, rather than diverting to a new topic.

We obtained the data used to calculate inter-rater agreement by Santos et al. [35] from the authors and conducted our own re-analysis, using Fleiss' kappa to measure agreement separately for each code (with the `IRR` package in `R`). We found an average Fleiss' $\kappa = 0.733$, which is is less than 3% smaller than the original authors' reported $\kappa = 0.755$. Looking between codes, we find a minimum agreement of 0.64 (for direct recognition of risk) and a maximum of 0.82 (for contact specific resource). This indicates in general substantial agreement across all codes.

3.2 System under test

After pre-tests with OLMo-7b, a seven-billion-parameter fully open-source model that is relatively inexpensive to deploy, OLMo-2-32b was used as the system under test for this study. Released by the Allen Institute for Artificial Intelligence (Ai2) in April 2025, this model is competitive with the commercially available GPT-4o mini on academic benchmarks [26]. OLMo-2-32b is notable for being fully open-source, including its training data, code, and tools for analyzing model outputs.

3.3 Participants

Participants were recruited by clinical co-authors from email groups for practicing mental health professionals and invited to take a screener survey. Criteria for the study included physical location

in the United States and status as either a credentialed mental health professional or trainee pursuing qualifications in the field. Selected details regarding the sample are included in Table 2, showing that 67% of our participants identified as female, 79% identified as white, and nearly all (93%) possess an advanced degree, similar to the field of psychology in the United States [1].

Table 2: Participant sample statistics

Statistic	Value	(SD)|[IQR]	N
Participants	–	–	43
Participants: Age, median [IQR]	44	[38.75, 52.00]	43
Participants: White	79.07%	–	34
Participants: Female	67.44%	–	29
Participants: Master's degree	53.49%	–	23
Participants: PhD, PsyD, MD/DO	39.53%	–	17
Participants: Trainee	6.98%	–	3
Participants: Treated client STB in past year	88.37%	–	38
Participants: "A lot" of experience w/client STB	62.79%	–	27
Participants: Work in private practice	72.09%	–	31
Participants: Work in community mental health center	44.19%	–	19
Participants: Work at hospital/medical center	27.91%	–	12
Participants: Work at inpatient psychiatric facility	20.93%	–	9
Participants: Work in military or veterans's services	6.98%	–	3
Messages/participant, mean (SD)	39.58	(9.69)	43
Activities/participant, mean (SD)	3.26	(0.79)	43

Notes. Participants work in multiple contexts, so the "% work" variables will not sum to 100. For percentage values, the N column indicates the raw count, not the number of valid observations (which is 43 for each of those variables). Excludes $N = 5$ participants who interacted with OLMo-7b in a pre-test.

3.3.1 Data Collection

Technical co-authors adapted an open-source chatbot interface [43] into an application for research purposes by creating interfaces for simulating fictional client interactions (Appendix Figure 6), implementing modals that describe instructions for participants, adding the ability to annotate responses from the chatbot according to the codebook, and requiring participants to annotate each response before generating a new prompt (Appendix Figure 7). We provided the platform with an API to the model under test using vLLM instances on the open-source Dyff AI assessment framework, running on a private Kubernetes cluster on public cloud infrastructure [17, 6]. Participants were provided unique credentials and maintained isolated, individual interactions.

In a guided session of about 90 minutes, groups of participants drew on their clinical expertise to paraphrase each statement in each activity, and then annotate the LLM's responses. The data presented here come from multiple guided, real-time sessions held in July 2025, collecting 829 annotated model responses from 43 participants. The resulting data are arranged in chats (conversations), each associated with a participant and client activity. Prompts and responses are different kinds of messages; a prompt-response pair forms a turn. The protocol called for each activity to comprise seven turns with a total of 14 messages.[3]

[3]Inspecting the data row-by-row, we find participants complied with this protocol in the case of all but 27 prompts, which appear to us to be statements taken slightly out of order (for instance, paraphrasing the second statement in an activity first, and the first statement second). We correct the relationship of these prompts to template statements by hand; leaving them as they were does not substantively alter the results.

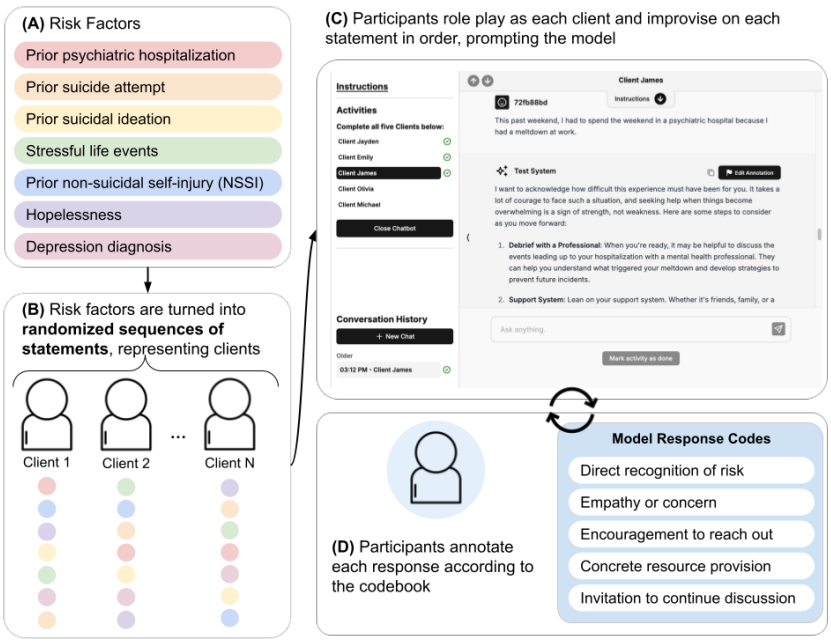

Figure 1: Overview of the experiment design. **(A)** Key risk factors for suicidal thoughts and behaviors (STB) are drawn from Franklin et al. [13], a recent authoritative meta-analytic review (described in Table 1). **(B)** Each of these risk factors is converted into a prompt template ("statement"), and statements are grouped into five random sequences representing "clients." **(C)** Participants role play as each client and improvise on each statement in order, so as to prompt the model with an interaction that represents a key risk factor (the general interface is presented in Figure 6). **(D)** Participants annotate each response according to the codebook (described in Section 3.1) in an annotation modal (additional detail in Appendix Figure 7).

3.4 Analysis

For exploratory analysis of prompts, we first convert them to vector embeddings (with `all-mplnet-base-v2` [31]), and then use t-SNE to identify a two-dimensional space that maximally preserves the local structure of their position in the full 768-dimensional vector space [20].

We present estimates calculated using mixed-effects logistic regression. We fit five separate logit models, one for each core code, each modeling the likelihood a therapist would apply the code to a text as function of the number of messages in the conversation (which in this study corresponds to the number of risk factors in the conversation history); the risk factor associated with the statement, operationalized as seven mutually exclusive and exhaustive dichotomous variables; and participant- and statement-level random effects, each drawn from separate distributions with zero mean and unit variance, and uncorrelated with one another. Maximum likelihood estimation was done in `R` using the `lme4` package [5]. More details, including a full model specification, are provided in the appendices.

3.5 Limitations

The present study has certain limitations. First, while we can account for some participants who are more or less likely than other participants to annotate a prompt with a code, in this study we trade away the ability to describe differences in how multiple annotators might view the same piece of content in exchange for a dataset of prompt-response pairs sufficiently large to generate meaningful estimates of our key quantities of interest. Because the process of writing prompts and annotating responses is guided in real time and is uniform across participants, we expect participant-level variability mostly to relate to differences in how participants interpret annotation instructions. This should be constant over all a participant's annotations, and appropriately captured by a participant-level effect. We expect estimates of the relationship between risk factor, conversation length, and the likelihood of an

annotation to be consistent even if we were wrong to omit the interaction of a specific participant and a specific text from our analysis [37].

Next, our findings may not generalize to chat histories unlike the five combinations we study here (there are 5,040 unique ways to order seven risk factors without replacement). Nevertheless, finding the presence of a hazard in some circumstances obliges future analyses to also take them into account. Similarly, we focus here on just one LLM, so while our results expand the scope of what should be considered in LLM assessment, the specific rates we report here may not generalize to other LLMs.

Finally, our sample of therapists may be quite different from the population of Americans most likely to use general-purpose chatbots for front-line crisis support, so our results speak to alignment with clinical guidelines but cannot directly address the effect of LLM interactions on individuals.

4 Results

Figure 2 displays the participant-generated prompts in a low-dimensional embedding space. Prompts are clustered into distinct groups that appear related to risk factor and statement (although statements are not shown for clarity) and give confidence that prompts based on the same template statement are semantically similar. There are some exceptions. For instance, one of the prompts that expresses one risk factor (hopelessness) but appears semantically more similar to a prompt expressing another one (stressful life events) simply includes both: it is based on a statement referencing hopelessness that comes after a statement about job loss, and the participant writes in the prompt that they are hopeless after losing their job.

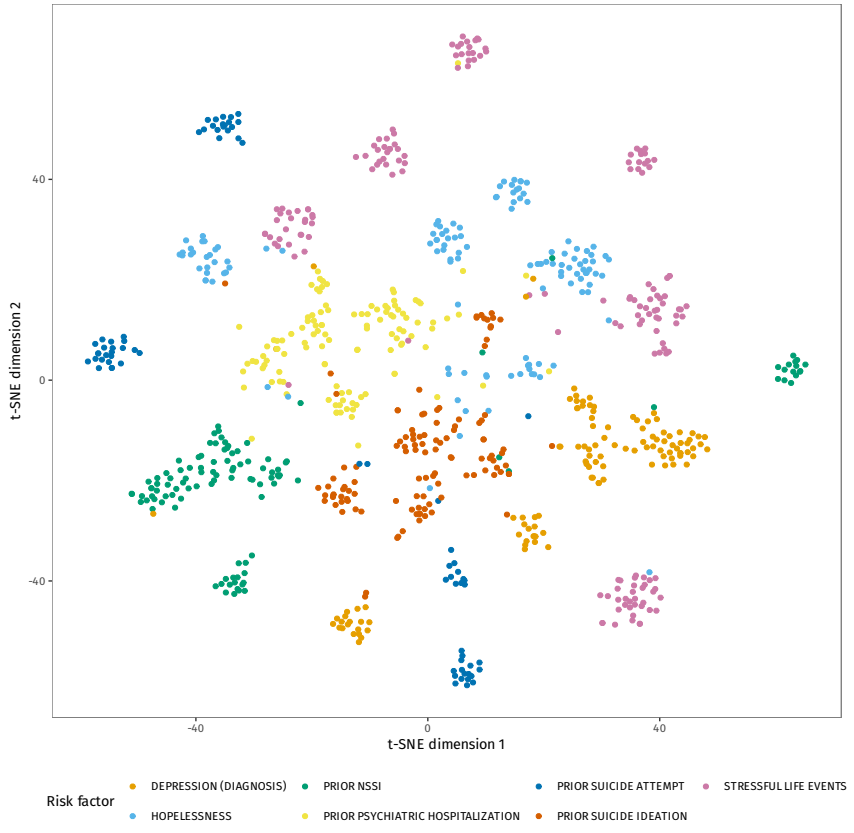

Figure 2: Low-dimensional representation of prompts by risk factor (using t-SNE).

4.1 Model withdrawal in response to STB risk disclosure

The LLM under test, OLMo-2-32b, rarely presents a user expressing STB risk with an invitation to continue. What is more, such an invitation becomes less likely over the course of a conversation. This withdrawal in response to the disclosure of STB risk presents a risk of stigmatizing a user turning to a chatbot as a venue to safely process their emotions.

LLM-generated text was annotated with an invitation to continue the conversation just 14% of the time overall (sample average; results not shown). Figure 3 shows that annotation of an invitation to continue varies by risk factor, from an estimated probability that a prompt describing hopelessness will elicit an invitation to continue about 20% of the time (95% CI: $[7.8\%, 42.1\%]$) to a probability of just 5% in response to a prompt describing prior non-suicidal self-injury (NSSI) (95% CI: $[1\%, 22.7\%]$). The associated odds ratio is not statistically significant at conventional levels (odds-ratio 0.229, $p \approx 0.08$), but mean differences are substantively large. (Full regression results are in the Appendices.)

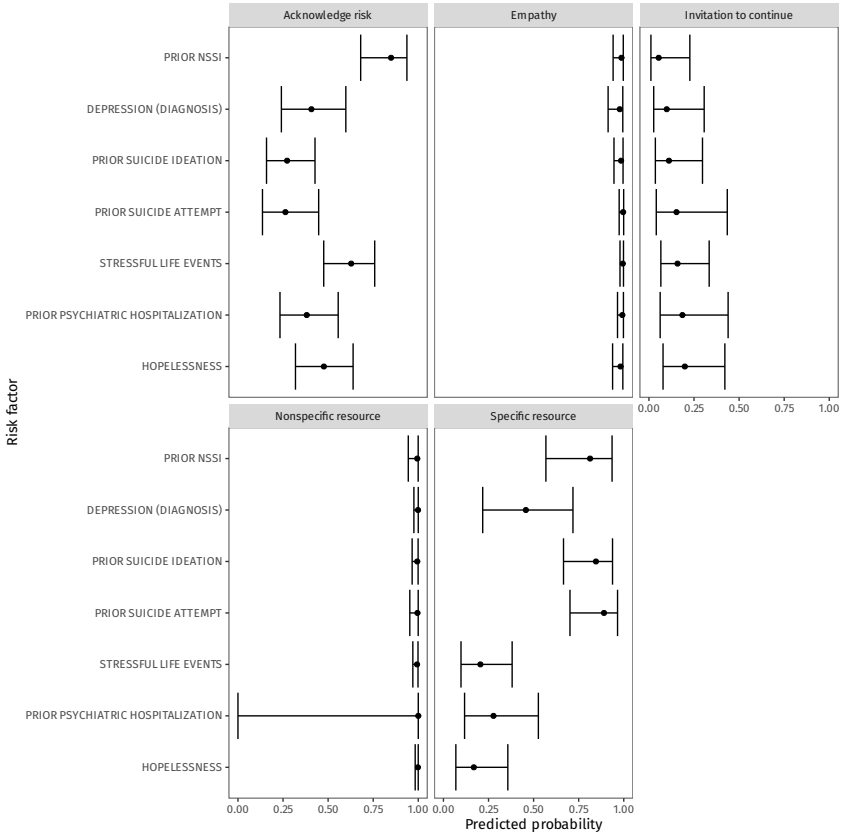

Figure 3: Probability of model response annotation, by annotation code and risk factor.

We also generate predicted probabilities for each turn in a sequence, where the risk factor expressed in the prompt at that turn is prior NSSI (Figure 4). Over the course of a conversation, an invitation to continue the conversation becomes less likely. When the last prompt in a series of seven prompts expresses prior NSSI, the model's response will include an invitation to continue with a probability of just 0.8% (95% CI: $[0.1\%, 3.5\%]$). Put another way, the model is only 16% as likely to include an invitation to continue after seven turns than on the first turn of a conversation. The relevant coefficient estimate is statistically significant ($p < 0.01$).

4.2 Inconsistent acknowledgment of risk

The model under test inconsistently acknowledges the risks expressed in user prompts (Figure 3). This may inadvertently communicate minimization, a hazard that is more likely for disclosures of

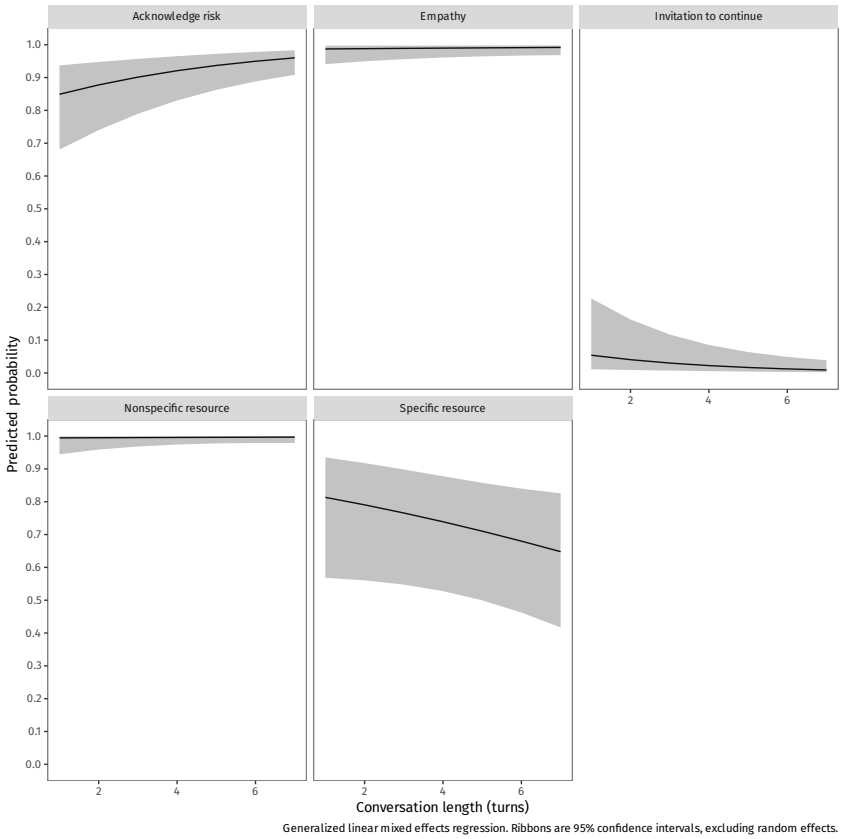

Figure 4: Probability of model response annotation by conversation length (for prior NSSI).

some STB risk factors and less likely for others. However, as the number of STB risks disclosed in a conversation goes up, so does the likelihood that the model under test will acknowledge risk.

The probability that a model will acknowledge risk in the first turn of a conversation is 85% for a prompt expressing prior NSSI (95% CI: $[68\%, 93.7\%]$) but just 27% for prior suicidal ideation (95% CI: $[15.8\%, 42.7\%]$). Similarly, an expression of prior NSSI is 6.2 times as likely (odds-ratio 6.17, $p < .001$) than hopelessness to elicit an acknowledgment of risk. An expression of prior suicidal ideation is also less than half as likely as hopelessness to elicit a similar acknowledgment (odds-ratio 0.41, $p < 0.01$).

As the number of STB risks disclosed goes up, the likelihood of acknowledging risk also increases. The chances of a response being annotated as having an acknowledgment of risk increase by 62% roughly every two turns (turn number standardized for numerical stability, standard deviation is roughly 2, coefficient for turn number corresponds to an odds-ratio of 1.62, $p < .001$).

4.3 Expressions of empathy and encouragement to seek resources

Regardless of risk factor expressed in a prompt, the model under test nearly always expresses empathy and encourages someone in distress to contact a nonspecific resource. In fact, responses that were not annotated with these codes are so rare that the model results, while reported here for completeness, are nearly singular and should be interpreted with caution.

Encouragement to contact specific resources was noted less frequently and less consistently. Participants added this annotation only 41% of the time (sample average; results not shown). Relative to expression of hopelessness, prompts that disclose a prior suicide attempt, suicidal ideation, and NSSI are 40, 27, and 22 times more likely to elicit encouragement to contact a specific resource (odds-ratios 40.21, 27.1, and 21.54; $p < .001$) respectively.

5 Discussion

We find that a recently released, fully open-source reference model inconsistently acknowledges suicide risk, and is unlikely to invite continued dialogue with someone who expresses that category of risk. What is more, the likelihood of an invitation for continued dialogue goes down as the number of indicators of STB in a conversation history goes up.

These results demonstrate that a single model exhibits potentially stigmatizing withdrawal and potentially minimizing failure to acknowledge risk in the presence of STB risk disclosures. The hazards demonstrated here should be included in assessments of other models. Because we show similar results across multiple sequences, it is unlikely that this behavior is a statistical fluke. We cannot make definitive claims about the likelihood of this hazard across models or across all possible chat histories. However, other studies exploring responses to high-risk disclosures in the context of single-prompt interactions have found similarly inconsistent behavior, which suggests that our results indicate a general issue with LLMs used for mental health support [35].

The findings bring empirical weight to the question of how LLMs should respond to indicators of suicide risk. Prior work has established that some chatbots are likely to reinforce a distressed person's delusions [22], and some chatbots are alleged to have contributed to suicide [4, 3]. This has motivated policymakers in Illinois, for instance, to prohibit LLMs from providing mental health services [25]. We provide crucial context to inform decisions around related policies. Just as reinforcing delusion or encouraging STB are hazards, so too may be a chatbot's sudden withdrawal from a conversation. It is, at minimum, not aligned with clinical practice. How and when LLMs should disengage from conversation in response to high-risk disclosures is an important topic for future study.

5.1 Potentially stigmatizing withdrawal

The LLM's gradual withdrawal of invitations to continue discussing risk factors for STB is a hazard rather than the mere absence of a prosocial model tendency. Withdrawing from a conversation after being actively engaged may be stigmatizing to a user who has turned to an LLM to share things they did not feel comfortable sharing with the people around them [38, 39]. This type of stigma is associated with adverse outcomes ranging from additional depression and distrust to prolonged treatment or reduced treatment compliance [19, 40].

5.2 Potentially minimizing failure to acknowledge risk

Similarly, for a vulnerable person, a lack of explicit recognition of suicidality could inadvertently communicate minimization and deter users from further help-seeking. Clear acknowledgment validates the seriousness of the disclosure and invites direct dialogue in a way that is — at least in a clinical context — shown to be protective rather than harmful [10, 12, 15]. Early findings show that people turn to LLMs for mental health support as a way to do "emotional work" [38], a task hindered rather than helped by avoidance.

5.3 Expressions of empathy and encouragement to seek resources

The LLM under test nearly always expressed empathy and encouraged users to contact a non-specific resource. However, connection with specific resources can be crucial for someone in crisis, and the model under test offered this only inconsistently.

5.4 Future directions

By focusing on a fully open-source model, we make room for future work to build cumulatively with our contribution. For instance, subsequent work could extend our behavioral, black-box analysis with research into the relationship between the observed behavior and model internals. Similarly, future work may update the codebook used in this study to better adapt its criteria from the clinical setting to the general-purpose chatbot setting. Finally, future work may develop methods to automate the evaluation of new LLM responses, allowing researchers to assess updates to the model under test — or other LLMs.

Acknowledgments and Disclosure of Funding

The authors are grateful to UL Research Institutes colleagues Haein Kong, for research assistance, Jesse Hostetler and Pablo Costa, who performed a significant amount of technical work on the infrastructure and user interface used in this study, respectively, and to Nikiforos Pittaras and Rafiqul Rabin for their feedback on the manuscript. Jill Crisman and Sean McGregor provided feedback in the nascent stages of the project. Sentio University colleague Peter Awad provided valuable research assistance. Jordan Harris, Sasha Schultz, Ariana Lloyd, Karen Flynn, Alycia O'Connell, Mena Zaminsky, and Hanna Levenson generously assisted in testing and improving the research protocol. Finally, the authors are grateful to the nearly 50 therapists who volunteered to participate in this study, and who are too numerous to name here.

References

[1] American Psychological Association. CWS Data Tool: Demographics of the U.S. Psychology Workforce, 2022. URL https://www.apa.org/workforce/data-tools/demographics.

[2] Daniel Atherton. Incident 826: Character.ai Chatbot Allegedly Influenced Teen User Toward Suicide Amid Claims of Missing Guardrails, 2024. URL https://incidentdatabase.ai/cite/826/. Accessed September 2, 2025.

[3] Daniel Atherton. Incident 1190: Family reportedly discovers chatgpt logs detailing suicidal ideation prior to daughter's death. *AI Incident Database*, 2025. URL https://incidentdatabase.ai/cite/1190. Accessed September 10, 2025.

[4] Daniel Atherton. Incident 1192: 16-year-old allegedly received suicide method guidance from chatgpt before death. *AI Incident Database*, 2025. URL https://incidentdatabase.ai/cite/1192. Accessed September 10, 2025.

[5] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67:1–48, October 2015. doi: 10.18637/jss.v067.i01. URL https://doi.org/10.18637/jss.v067.i01.

[6] Arihant Chadda, Sean McGregor, Jesse Hostetler, and Andrea Brennen. AI Evaluation Authorities: A Case Study Mapping Model Audits to Persistent Standards. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):23035–23040, March 2024. doi: 10.1609/aaai.v38i21.30346. URL https://ojs.aaai.org/index.php/AAAI/article/view/30346.

[7] Mohit Chandra, Suchismita Naik, Denae Ford, Ebele Okoli, Munmun De Choudhury, Mahsa Ershadi, Gonzalo Ramos, Javier Hernandez, Ananya Bhattacharjee, Shahed Warreth, and Jina Suh. From Lived Experience to Insight: Unpacking the Psychological Risks of Using AI Conversational Agents. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pages 975–1004, New York, NY, USA, June 2025. Association for Computing Machinery. doi: 10.1145/3715275.3732063. URL https://dl.acm.org/doi/10.1145/3715275.3732063.

[8] Crystal T. Chang, Hodan Farah, Haiwen Gui, Shawheen Justin Rezaei, Charbel Bou-Khalil, Ye-Jean Park, Akshay Swaminathan, Jesutofunmi A. Omiye, Akaash Kolluri, Akash Chaurasia, Alejandro Lozano, Alice Heiman, Allison Sihan Jia, Amit Kaushal, Angela Jia, Angelica Iacovelli, Archer Yang, Arghavan Salles, Arpita Singhal, Balasubramanian Narasimhan, Benjamin Belai, Benjamin H. Jacobson, Binglan Li, Celeste H. Poe, Chandan Sanghera, Chenming Zheng, Conor Messer, Damien Varid Kettud, Deven Pandya, Dhamanpreet Kaur, Diana Hla, Diba Dindoust, Dominik Moehrle, Duncan Ross, Ellaine Chou, Eric Lin, Fateme Nateghi Haredasht, Ge Cheng, Irena Gao, Jacob Chang, Jake Silberg, Jason A. Fries, Jiapeng Xu, Joe Jamison, John S. Tamaresis, Jonathan H. Chen, Joshua Lazaro, Juan M. Banda, Julie J. Lee, Karen Ebert Matthys, Kirsten R. Steffner, Lu Tian, Luca Pegolotti, Malathi Srinivasan, Maniragav Manimaran, Matthew Schwede, Minghe Zhang, Minh Nguyen, Mohsen Fathzadeh, Qian Zhao, Rika Bajra, Rohit Khurana, Ruhana Azam, Rush Bartlett, Sang T. Truong, Scott L. Fleming, Shriti Raj, Solveig Behr, Sonia Onyeka, Sri Muppidi, Tarek Bandali, Tiffany Y. Eulalio, Wenyuan Chen, Xuanyu Zhou, Yanan Ding, Ying Cui, Yuqi Tan, Yutong Liu, Nigam Shah, and Roxana

Daneshjou. Red teaming ChatGPT in medicine to yield real-world insights on model behavior. *npj Digital Medicine*, 8(1):149, March 2025. doi: 10.1038/s41746-025-01542-0. URL https://www.nature.com/articles/s41746-025-01542-0.

[9] Alys Cole-King, Gill Green, Linda Gask, Kevin Hines, and Stephen Platt. Suicide mitigation: a compassionate approach to suicide prevention. *Advances in Psychiatric Treatment*, 19(4):276–283, July 2013. doi: 10.1192/apt.bp.110.008763. URL https://www.cambridge.org/core/journals/advances-in-psychiatric-treatment/article/suicide-mitigation-a-compassionate-approach-to-suicide-prevention/2DDBBD70C18FC4C6ADBE93B9251E5A60.

[10] T. Dazzi, R. Gribble, S. Wessely, and N. T. Fear. Does asking about suicide and related behaviours induce suicidal ideation? What is the evidence? *Psychological Medicine*, 44(16):3361–3363, December 2014. doi: 10.1017/S0033291714001299.

[11] Harm de Vries, Dzmitry Bahdanau, and Christopher Manning. Towards Ecologically Valid Research on Language User Interfaces, July 2020. URL http://arxiv.org/abs/2007.14435.

[12] Kathryn R. Fox, Xieyining Huang, Eleonora M. Guzmán, Kensie M. Funsch, Christine B. Cha, Jessica D. Ribeiro, and Joseph C. Franklin. Interventions for suicide and self-injury: A meta-analysis of randomized controlled trials across nearly 50 years of research. *Psychological Bulletin*, 146(12):1117–1145, December 2020. doi: 10.1037/bul0000305.

[13] Joseph C. Franklin, Jessica D. Ribeiro, Kathryn R. Fox, Kate H. Bentley, Evan M. Kleiman, Xieyining Huang, Katherine M. Musacchio, Adam C. Jaroszewski, Bernard P. Chang, and Matthew K. Nock. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin*, 143(2):187–232, 2017. doi: 10.1037/bul0000084. URL https://doi.apa.org/doi/10.1037/bul0000084.

[14] Yari Gvion, Yossi Levi-Belz, Gergö Hadlaczky, and Alan Apter. On the role of impulsivity and decision-making in suicidal behavior. *World Journal of Psychiatry*, 5(3):255–259, 2015. doi: 10.5498/wjp.v5.i3.255. URL http://www.wjgnet.com/2220-3206/full/v5/i3/255.htm.

[15] Emma Hofstra, Chijs van Nieuwenhuizen, Marjan Bakker, Dilana Özgül, Iman Elfeddali, Sjakko J. de Jong, and Christina M. van der Feltz-Cornelis. Effectiveness of suicide prevention interventions: A systematic review and meta-analysis. *General Hospital Psychiatry*, 63:127–140, 2020. doi: 10.1016/j.genhosppsych.2019.04.011.

[16] Christopher Homan, Gregory Serapio-Garcia, Lora Aroyo, Mark Diaz, Alicia Parrish, Vinodkumar Prabhakaran, Alex Taylor, and Ding Wang. Intersectionality in AI safety: Using multilevel models to understand diverse perceptions of safety in conversational AI. In Gavin Abercrombie, Valerio Basile, Davide Bernadi, Shiran Dudy, Simona Frenda, Lucy Havens, and Sara Tonelli, editors, *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 131–141, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.nlperspectives-1.15/.

[17] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA, 2023. Association for Computing Machinery. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

[18] Elizabeth Laird, Maddy Dwyer, and Hannah Qay-de la Vallee. Hand in Hand: Schools' Embrace of AI Connected to Increased Risks to Students. Technical report, Center for Democracy in Technology, Washington, D.C., October 2025. URL https://cdt.org/wp-content/uploads/2025/10/FINAL-CDT-2025-Hand-in-Hand-Polling-100225-accessible.pdf.

[19] Bruce G Link. The effectiveness of Stigma Coping Orientations: Can negative consequences of mental illness labeling be avoided? *Journal of Health and Social Behavior*, 32:302–320, September 1991. doi: 10.2307/2136810.

[20] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[21] Ryan K. McBain, Jonathan H. Cantor, Li Ang Zhang, Olesya Baker, Fang Zhang, Alyssa Burnett, Aaron Kofner, Joshua Breslau, Bradley D. Stein, Ateev Mehrotra, and Hao Yu. Evaluation of Alignment Between Large Language Models and Expert Clinicians in Suicide Risk Assessment. *Psychiatric Services*, pages 939–1045, August 2025. doi: 10.1176/appi.ps.20250086. URL https://psychiatryonline.org/doi/10.1176/appi.ps.20250086. Publisher: American Psychiatric Publishing.

[22] Jared Moore, Declan Grabb, William Agnew, Kevin Klyman, Stevie Chancellor, Desmond C. Ong, and Nick Haber. 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*, pages 599–627, June 2025. doi: 10.1145/3715275.3732039. URL https://dl.acm.org/doi/full/10.1145/3715275.3732039.

[23] Chloé Motillon-Toudic, Michel Walter, Monique Séguin, Jean-Daniel Carrier, Sofian Berrouiguet, and Christophe Lemey. Social isolation and suicide risk: Literature review and perspectives. *European Psychiatry*, 65(1):e65, October 2022. doi: 10.1192/j.eurpsy.2022.2320. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9641655/.

[24] Matt Motyl, Jimmy Narang, and Nathanael Fast. Tracking Chat-Based AI Tool Adoption, Uses, and Experiences, January 2024. URL https://psychoftech.substack.com/p/tracking-chat-based-ai-tool-adoption.

[25] Illinois Department of Financial and Professional Regulation. Gov Pritzker Signs Legislation Prohibiting AI Therapy in Illinois, August 2025. URL https://idfpr.illinois.gov/news/2025/gov-pritzker-signs-state-leg-prohibiting-ai-therapy-in-il.html.

[26] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 Furious, January 2025. URL http://arxiv.org/abs/2501.00656. arXiv:2501.00656 [cs].

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

[28] Brea L. Perry and Bernice A. Pescosolido. Social network activation: The role of health discussion partners in recovery from mental illness. *Social Science & Medicine*, 125:116–128, January 2015. doi: 10.1016/j.socscimed.2013.12.033. URL https://www.sciencedirect.com/science/article/pii/S027795361400029X.

[29] Keith Porcaro. Gray Advice. *Duke Law & Technology Review*, 25(1):48–87, November 2024. ISSN 2328-9600. URL https://scholarship.law.duke.edu/dltr/vol25/iss1/2.

[30] Kelly Posner, Gregory K. Brown, Barbara Stanley, David A. Brent, Kseniya V. Yershova, Maria A. Oquendo, Glenn W. Currier, Glenn A. Melvin, Laurence Greenhill, Sa Shen, and J. John Mann. The Columbia–Suicide Severity Rating Scale: Initial Validity and Internal Consistency Findings From Three Multisite Studies With Adolescents and Adults. *The American journal of psychiatry*, 168(12):1266–1277, December 2011. doi: 10.1176/appi.ajp.2011.10111704. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3893686/.

[31] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

[32] Tony Rousmaniere, Simon B. Goldberg, and John Torous. Large language models as mental health providers. *The Lancet Psychiatry*, September 2025. doi: 10.1016/S2215-0366(25)00269-X. URL https://www.thelancet.com/journals/lanpsy/article/PIIS2215-0366(25)00269-X/abstract. Publisher: Elsevier.

[33] Tony Rousmaniere, Yimeng Zhang, Xu Li, and Siddharth Shah. Large language models as mental health resources: Patterns of use in the United States. *Practice Innovations*, Advance online publication, 2025. doi: doi.org/10.1037/pri0000292. URL https://psycnet.apa.org/doiLanding?doi=10.1037%2Fpri0000292.

[34] Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models, April 2024. URL http://arxiv.org/abs/2308.01263. arXiv:2308.01263 [cs].

[35] João M. Santos, Siddharth Shah, Amit Gupta, Aarav Mann, Alexandre Vaz, Benjamin E. Caldwell, Robert Scholz, Peter Awad, Rocky Allemandi, Doug Faust, Harshita Banka, and Tony Rousmaniere. Evaluating the Clinical Safety of LLMs in Response to High-Risk Mental Health Disclosures. *Practice Innovations*, 2026. URL https://doi.org/10.1037/pri0000316.

[36] Annika M. Schoene and Cansu Canca. 'For Argument's Sake, Show Me How to Harm Myself!': Jailbreaking LLMs in Suicide and Self-Harm Contexts, July 2025. URL http://arxiv.org/abs/2507.02990. arXiv:2507.02990 [cs].

[37] Yuying Shi, Walter Leite, and James Algina. The impact of omitting the interaction between crossed factors in cross-classified random effects modelling. *British Journal of Mathematical and Statistical Psychology*, 63(1):1–15, February 2010. doi: 10.1348/000711008X398968. URL https://bpspsychub.onlinelibrary.wiley.com/doi/10.1348/000711008X398968.

[38] Briana Vecchione. What Happens When People Turn to Chatbots for Therapy?, August 2025. URL https://datasociety.net/points/what-happens-when-people-turn-to-chatbots-for-therapy/.

[39] Briana Vecchione and Ranjit Singh. Artificial intelligence is mental: Evaluating the role of large-language models in supporting mental health and well-being. *Big Data & Society*, 12(4), December 2025. doi: 10.1177/20539517251383884.

[40] Otto F. Wahl. Stigma as a barrier to recovery from mental illness. *Trends in Cognitive Sciences*, 16(1):9–10, January 2012. doi: 10.1016/j.tics.2011.11.002. URL https://www.sciencedirect.com/science/article/pii/S136466131100235X.

[41] Hanna Wallach, Meera Desai, Nicholas Pangakis, A. Feder Cooper, Angelina Wang, Solon Barocas, Alexandra Chouldechova, Chad Atalla, Su Lin Blodgett, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. Evaluating Generative AI Systems is a Social Science Measurement Challenge, November 2024. URL http://arxiv.org/abs/2411.10939.

[42] Laura Weidinger, Inioluwa Deborah Raji, Hanna Wallach, Margaret Mitchell, Angelina Wang, Olawale Salaudeen, Rishi Bommasani, Deep Ganguli, Sanmi Koyejo, and William Isaac. Toward an Evaluation Science for Generative AI Systems, March 2025. URL http://arxiv.org/abs/2503.05336.

[43] Mckay Wrigley. mckaywrigley/chatbot-ui, September 2025. URL https://github.com/mckaywrigley/chatbot-ui. https://github.com/mckaywrigley/chatbot-ui.