Like a Therapist, But Not: Reddit Narratives of AI in Mental Health Contexts

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly used for emotional support and mental health–related interactions outside clinical settings, yet little is known about how people evaluate and relate to these systems in everyday use. We analyze 5,126 Reddit posts from 47 mental health communities describing experiential or exploratory use of AI for emotional support or therapy. Grounded in the Technology Acceptance Model and therapeutic alliance theory, we develop a theory-informed annotation framework and apply a hybrid LLM–human pipeline to analyze evaluative language, adoption-related attitudes, and relational alignment at scale. Our results show that engagement is shaped primarily by narrated outcomes, trust, and response quality, rather than emotional bond alone. Positive sentiment is most strongly associated with task and goal alignment, while companionship-oriented use more often involves misaligned alliances and reported risks such as dependence and symptom escalation. Overall, this work demonstrates how theory-grounded constructs can be operationalized in large-scale discourse analysis and highlights the importance of studying how users interpret language technologies in sensitive, real-world contexts.

1 Introduction

"Some absences keep their shape." When ChatGPT produced this reflection in response to a grieving psychologist, it startled him, not because of its eloquence, but for how closely it mirrored what he struggled to articulate himself. This encounter, described in the New York Times (The New York Times, 2025), motivates a central question in this study: *How do people evaluate, interpret, and engage with AI tools when they are used for emotional support or therapy?*, while illustrating the capacity of large language models (LLMs) to generate emotionally resonant language. Millions worldwide face barriers to traditional therapy, including high costs, limited availability of trained professionals (CNBC, 2021), and the persistent stigma (Naslund et al., 2016; Thomson et al., 2024). In response, systems such as ChatGPT, Claude, and Character AI are increasingly used as accessible, always-available conversational supports. This adoption is largely user-driven and occurs outside clinical settings, yet we lack systematic insight into how people evaluate, trust, and relate to these systems in real-world mental health contexts.

Early work on conversational agents for mental health centered on scripted, rule-based systems such as ELIZA (Weizenbaum, 1966) and PARRY (Colby, 1975), and later on CBT-oriented chatbots like Woebot and Wysa (Fitzpatrick et al., 2017; Beatty et al., 2022). These systems delivered structured psychoeducation at scale but constrained engagement, whereas recent LLM-based systems enable open-ended, context-sensitive dialogue often perceived as empathetic or relational. However, most existing evaluations examine these systems in controlled settings, focusing on accuracy, safety, or therapeutic potential (Maurya et al., 2025; Thakkar et al., 2024; Roshanaei et al., 2025), offering limited insight into how users interpret and negotiate AI's role in everyday mental health practices.

Online communities offer a valuable lens for examining these questions. Platforms such as Reddit host large-scale mental health discussions where users describe struggles, seek advice, and collectively interpret technologies, including the boundaries between AI support and human care (Sit et al., 2024; Bouzoubaa et al., 2024c). Analyzing this discourse reveals how users assess the affordances and limitations of AI support and how these evaluations shape users' engagement with AI in relation to human mental health support. To address these issues, we investigate the following research questions:

RQ1: How do individuals in online communities describe their use of AI for mental health support,

including perceived functions, benefits, and risks?
RQ2: How are Technology Acceptance Model (TAM) dimensions associated with adoption-related attitudes toward AI for therapy and emotional support in online communities?
RQ3: How do users express therapeutic alliance with AI, and how do task, goal, and bond alignment relate to engagement outcomes?

We address these questions by analyzing 4.7 million posts across 47 condition-focused Reddit communities between November 2022 and August 2025. Using a multi-stage filtering pipeline, we identified posts describing experiential or exploratory use of AI as a therapeutic tool, resulting in a curated dataset of 5,126 posts focused on AI-supported mental health care. We developed a theory-grounded annotation schema integrating constructs from the Technology Acceptance Model (Davis et al., 1989) and therapeutic alliance theory (Bordin, 1979) to capture evaluative language, adoption-related attitudes, and relational alignment, including pragmatic evaluations (e.g., usefulness, trust, outcome quality) and task, goal, and bond dimensions shaping engagement and perceived risk, in user discourse.

Our results show that sustained engagement with AI for mental health support is shaped primarily by demonstrable outcomes, trust, and response quality, rather than by emotional bonding alone. Notably, therapeutic alliance is conditional: positive engagement is strongly associated with task and goal alignment, whereas emotional bond alone shows a weak relationship with positive sentiment and frequently co-occurs with dependency and reported harm.

Our work makes three contributions: (1) a novel dataset capturing how users evaluate and relate to AI systems for emotional support; (2) theory-grounded operationalization of Technology Acceptance Model and therapeutic alliance constructs for large-scale NLP analysis of adoption attitudes and relational alignment; and (3) quantitative and qualitative evidence linking system properties to adoption sentiment and therapeutic alliance, informing the design and evaluation of AI-supported mental health tools. We will release the dataset, annotation guidelines, and code.

2 Related Work

2.1 AI for Mental Health Support

Advances in AI are transforming mental health care, with applications ranging from early diagnostic tools and mood monitoring to conversational agents that deliver therapeutic techniques in real time (Thakkar et al., 2024; Dehbozorgi et al., 2025; Graham et al., 2019; Cummins et al., 2020; Vaidyam et al., 2019). Early systems such as ELIZA (Weizenbaum, 1966) and PARRY (Colby, 1975) demonstrated the potential of simulated dialogue, while later tools like Wysa and Woebot applied structured therapeutic approaches at scale (Inkster et al., 2018; Beatty et al., 2022; Gamble, 2020). More recently, LLMs have enabled open-domain conversation, context-sensitive responses, and language that users may interpret as empathetic or supportive (Yu and McGuinness, 2024; Sorin et al., 2024; Ojeda Meixueiro et al., 2024). Recent studies show LLMs can deliver accurate and empathetic content (Maurya et al., 2025), support positive mental health (Thakkar et al., 2024), and even rival human-generated supportive messages (Young et al., 2024). A recent survey synthesized the therapeutic potential of LLMs, their limitations, and ethical risks (Na et al., 2025). Beyond mental health, prior NLP research has examined trust (Xie et al., 2024; Pessianzadeh et al., 2025), alignment (Naseem et al., 2025), and user perceptions of AI systems (Højer et al., 2025; Razi et al., 2025), typically through controlled experiments, surveys, or benchmark-driven evaluations. Prior work rarely examines mental health contexts or everyday evaluative and relational judgments, leaving users' interpretations of LLM-based support underexplored.

2.2 Online Mental Health Discourse

Online platforms have become key spaces for mental health discussion, enabling people to share experiences, seek advice, and access peer support (Thomson et al., 2024; Naslund et al., 2020). Their accessibility, anonymity, and asynchronous nature make them especially valuable for those facing barriers to traditional care, such as cost, geography, or stigma (Naslund et al., 2016; Merchant et al., 2022; Andalibi et al., 2017; Bouzoubaa et al., 2024a). Reddit, in particular, has emerged as a valuable source due to its topic-specific subreddits (Marshall et al., 2024; Bouzoubaa et al., 2024b), pseudonymity (Naslund et al., 2020), and community norms that encourage candid disclosure (Rayland and Andrews, 2023). While this work demonstrates the richness of online mental health discourse, it has largely focused on condition classification (Dinu and Moldovan, 2021), crisis detection

(D. Lewis et al., 2025), and symptom monitoring (Alhamed et al., 2024), rather than on how users perceive and discuss emerging tools such as AI in therapeutic contexts. We address this gap by examining how AI tools are evaluated and positioned in naturalistic mental health discourse.

3 Method

3.1 Data

Reddit has been widely used in mental health research due to its scale and topic-specific communities (Sit et al., 2024; Jiang et al., 2020). To analyze discourse around AI-based mental health tools, we curated condition-focused subreddits guided by DSM-5 diagnostic categories (Diagnostic, 2013) (e.g., depression, anxiety, bipolar disorder) to ensure broad and systematic coverage. Two authors independently reviewed DSM-5 categories and verified subreddit mappings, and we additionally included general mental health subreddits used for cross-diagnostic support and discussion (e.g., `r/mentalhealth`, `r/TalkTherapy`). We selected 47 subreddits varying in size, ranging from fewer than 10,000 to over 1 million members (see Table A.2), and collected Reddit submissions (excluding comments) using the ArcticShift API[1]. We collected posts published between November 30, 2022 (ChatGPT's public release), and August 15, 2025, resulting in 4,703,056 submissions. We applied preprocessing to improve data quality: concatenating titles and bodies, removing deleted or duplicate posts, and excluding posts with fewer than 10 tokens (the shortest 10%). After preprocessing, the dataset comprised 3,530,486 posts.

3.2 Identifying AI-Relevant Posts

To identify posts discussing AI tools (e.g., ChatGPT, Character AI) for therapeutic or emotional support, we used a scalable, multi-stage relevance filtering pipeline. To avoid applying LLM-based classification to the full corpus, we first narrowed the search space using keyword-based retrieval, followed by LLM-based validation on a reduced candidate set. In the first stage, we randomly sampled 120,000 posts and used GPT-4o mini (OpenAI, 2024) to classify whether posts referenced AI use for emotional support, therapy, or mental health–related guidance. From posts labeled as relevant, we extracted AI-related keywords and phrases. This process resulted in around 800 unique AI-related terms (e.g., "AI-Powered," "Bing Chat," "CHAI Bot"), which were manually reviewed and deduplicated to produce a refined list of 146 keywords. Applying this list to the full dataset resulted in a high-recall corpus of 572,734 posts.

In the second stage, we validated relevance before scaling LLM-based filtering. We randomly sampled 10,000 posts from the filtered corpus and used GPT-4o mini to label each as relevant or not. Two authors independently reviewed 200 posts (100 relevant, 100 not), resolving disagreements through discussion, and iteratively refined the prompt until agreement with human judgments was high (Fleiss' $\kappa = 0.90$). We then applied the finalized prompt to all 572,734 keyword-filtered posts, resulting in 6,206 posts describing AI use for therapeutic or emotional support.

To further characterize content, we manually reviewed a random sample of 200 posts, and identified four categories: *experiential* (users describing personal AI use for therapy or emotional support), *exploratory* (seeking advice or others' experiences), *advertisement* (promotional or developer-posted content), and *irrelevant* (posts that did not meaningfully discuss AI in a mental health context). Two authors independently annotated 50 posts per category using the same definitions provided to the LLM, achieving strong agreement with model classifications (Fleiss' $\kappa = 0.78$); disagreements were resolved through discussion. Since our analysis focuses on how individuals perceive and describe AI as a therapeutic tool, we retained only *experiential* and *exploratory* posts for further analysis, resulting in a final dataset of 5,126 posts.

3.3 Annotation Framework

To analyze how users evaluate AI as a therapeutic tool, we ground our annotation framework in two established theories from psychology and HCI to capture both pragmatic and relational aspects of AI-mediated mental health support:

Therapeutic Alliance Theory. Therapeutic alliance refers to the collaborative and affective bond between therapist and client and is a key predictor of treatment outcomes (Bordin, 1979). It is typically described through three interdependent components: `bond` (trust, empathy, and emotional connection), `task` (agreement on therapeutic activities), and `goal` (shared commitment to objectives). Although originally developed for human psychotherapy, the framework has been applied to

[1] https://github.com/ArthurHeitmann/arctic_shift

	Dimension	Criteria	Measurement
TAM	Perceived Usefulness	Does the user describe the AI as helpful or useful for emotional/mental-health tasks?	Categorical: [useful|not_useful|not_mentioned] +Desc.
	Perceived Ease of Use	Does the user describe the AI as easy, convenient, or accessible to use?	Categorical: [easy|difficult|not_mentioned] +Desc.
	Intention to Continue	Does the user state an intention to keep using AI in the future?	Categorical: [yes|no|not_mentioned]
TAM (ext.)	Perceived Trust	Does the user indicate the AI is trustworthy, reliable, or safe in a mental-health context?	Categorical: [trustworthy|untrustworthy|not_mentioned] + Desc.
	Output Quality	Does the user judge the AI's responses as high vs. low quality (e.g., empathetic, thoughtful vs. vague, inaccurate)?	Categorical: [good|poor|not_mentioned] + Desc.
	Result Demonstrability	Are tangible/behavioral outcomes mentioned (e.g., better sleep, reduced anxiety)?	Categorical: [pos._results|neg._results|not_mentioned] + Desc.
	Social Influence	Does the user mention influence from peers/Reddit/others to use AI?	Categorical: [present|absent] + Desc.
	Perceived Risks	Mentions of limitations, harms, or risks in using AI for mental health (e.g., inaccuracy, emotional detachment).	Categorical: [mentioned|not_mentioned] + Desc.
Therapeutic Alliance	Bond	Does the user feel emotionally supported/understood/safe with the AI?	Categorical: [strong|weak|not_mentioned] + Desc.
	Task	Is the AI helping with therapeutic actions (e.g., reflection, coping, journaling)?	Categorical: [aligned|misaligned|not_mentioned] + Desc.
	Goal	Does the AI align with the user's mental-health goals (recovery, stability, growth)?	Categorical: [aligned|misaligned|not_mentioned] + Desc.
Other	Usage Intent	the primary function or reason a user engages with AI in a mental health related context.	Desc.
	Comparison to Therapy	How the AI compares to traditional therapy.	Categorical: [better|worse|complementary|not_mentioned]
	AI Tool Mentioned	The specific AI tool that is mentioned by the user.	Desc.
	Sentiment toward AI	Overall Sentiment toward users' AI experience	Categorical: [positive|negative|neutral] + Desc.
	Mental Health Condition	The specific mental health condition that is mentioned by the user.	Desc.

Table 1: Dimensions, associated frameworks, definitions, and measurement type (categorical or free-text (Desc).

digital interventions (Vowels et al., 2024; Beatty et al., 2022). We adapt this framework to analyze how users describe relational alignment with AI systems in mental health contexts.

Technology Acceptance Model. TAM explains technology adoption through core beliefs about `usefulness` and `ease of use`, which influence attitudes, intentions, and actual use (Davis et al., 1989; Fishbein, 1979). Extensions have added dimensions such as `Output Quality`, `result demonstrability`, and `social influence` (Venkatesh and Davis, 2000), while research in health informatics highlights the importance of `perceived trust` and `perceived risk` in sensitive domains (Su et al., 2013; Dhagarra et al., 2020). Building on this work, we include core and extended TAM to capture user evaluations of AI tools in mental health contexts, where trust, risk, and outcomes are salient. We operationalized an annotation schema using these theories. From TAM and its extensions, we used `perceived usefulness`, `ease of use`, `output quality`, `result demonstrability`, `social influence`, `perceived trust`, `perceived risks`, and `intention to continue`. From therapeutic alliance, we included `task`, `bond`, and `goal`. To capture context-specific aspects of AI–therapy discourse, we included `comparison to human therapy`, `AI tool mentioned`, `mental health condition`, `sentiment`, and `usage intent`. Table 1 shows all dimensions and coding criteria.

3.4 Data Annotation

Human Annotation. To assess annotation reliability, two authors independently annotated a random sample of 100 posts from the final dataset across all 13 dimensions (Table 1) using shared guidelines. Inter-annotator agreement measured by raw agreement, Cohen's Kappa, and Gwet's AC1 to account for label imbalance, ranged from 0.74 to 0.95 (raw), 0.37 to 0.60 (Cohen's Kappa), and 0.64 to 0.94 (Gwet's AC1), indicating substantial to near-perfect agreement. Disagreements were resolved through discussion, and the resulting labeledset served as ground truth for subsequent model evaluation.

LLM-Assisted Annotation and Evaluation. We evaluated multiple (closed and open) LLMs selected from leaderboards [2] of high-performing systems available at the time of the study. These included `GPT-5.2` (OpenAI, 2025), `Gemini-3-Pro` (GoogleDeepMind, 2025), and `Claude-Opus-4.5` (Anthropic, 2025), which are proprietary models, as well as `Kimi-K2-Instruct` (MoonshotAI, 2025) and `Qwen3-Next-80B-A3B-Instruct` (Qwen, 2025), which are open-source models. All models were prompted using the same structured annotation procedure, including task instructions, construct definitions, and a standardized JSON output format. A sample schema was included to constrain outputs. We set the temperature to zero for all models; for GPT-5.2, we additionally fixed the reasoning setting to a medium level. The full prompt is provided in the Appendix A.1.

For all dimensions, models produced free-text rationales for each label. We manually reviewed these outputs on the 100-post evaluation set to verify consistency with annotation definitions and post content.

[2] https://llm-stats.com/

3.5 Thematic Analysis

To facilitate a more fine-grained analysis of AI engagement and concerns (RQ1), we conducted additional thematic analysis on the LLM-generated free-text descriptions for two of our dimensions: `perceived risks` and `usage intent`. We applied LLM-guided thematic analysis (Dai et al., 2023) to consolidate these free-text descriptions into coherent, interpretable categories by grouping semantically similar phrases under shared themes. For example, risk descriptions such as "compulsive use," "overreliance," and "obsessive use" were consolidated into the category *Addiction & Dependence*. This process resulted in standardized taxonomies for both dimensions, enabling systematic analysis of engagement patterns and reported concerns across the corpus.

4 Results

4.1 Data Characteristics

Our analysis includes 5,126 Reddit posts from 47 mental health communities (November 2022–August 2025) describing experiential (3,605) or exploratory (1,521) use of AI for therapeutic or emotional support. Engagement varied across communities, with anxiety-related subreddits showing the highest activity, i.e., nearly 500 posts in `r/Anxiety`. Trauma- and stressor-related communities were also prominent, led by `r/CPTSD` (422 posts). Substantial activity was observed in obsessive-compulsive (`r/OCD`, 392 posts) and neurodevelopmental communities (`r/ADHD` and `r/autism`, over 900 posts combined). General mental health subreddits (e.g., `r/therapy`, `r/therapists`, `r/mentalhealth`) also contained large volumes of AI-related discussions. Table A.3 shows the distribution of posts across DSM-5 categories, and Figure A.2 shows their temporal trends.

4.2 Model Performance Across Dimensions

We evaluated five LLMs on the human-annotated set across all categorical dimensions using precision, recall, and macro-averaged F1. Table 2 reports F1 scores, with full precision and recall in Table A.1. Across core and extended TAM dimensions, GPT-5.2 achieved the highest F1 scores on `perceived usefulness` (F1 = 0.72), `ease of use` (0.76), `perceived trust` (0.65), `output quality` (0.85), `result demonstrability` (0.82), `intention to continue` (0.72). Gemini 3 Pro outperformed other models on `social influence` (0.82) and `perceived risks` (0.84). Performance on therapeutic alliance dimensions differed from other constructs and remained among the most challenging for all models. Gemini 3 Pro achieved the highest F1 scores for `bond` (0.78), `task` (0.71), and `goal` (0.64).

Dimension	GPT	Gem	Cla	Kimi	Qwen
perceived_usefulness	**0.72**	0.66	0.69	0.65	0.64
ease_of_use	**0.76**	0.49	0.34	0.53	0.27
perceived_trust	**0.65**	0.64	0.64	0.49	0.51
output_quality	**0.85**	0.71	0.84	0.67	0.63
result_demonstrability	**0.82**	0.67	0.79	0.69	0.59
intention_to_continue	**0.72**	0.66	0.70	0.59	0.41
social_influence	0.73	**0.82**	0.72	0.56	0.75
perceived_risks	0.70	**0.84**	0.80	0.77	0.70
bond	0.63	**0.78**	0.68	0.59	0.42
task	0.63	**0.71**	0.70	0.64	0.63
goal	0.57	**0.64**	0.64	0.40	0.56
comparison_to_therapy	**0.64**	0.63	0.59	0.61	0.41
sentiment	**0.77**	0.68	0.75	0.70	0.62
Macro F1	**0.70**	0.68	0.67	0.60	0.59

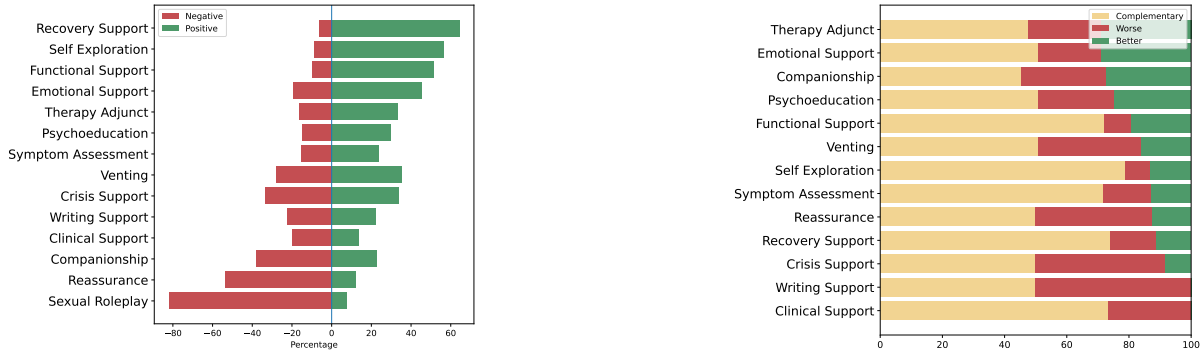
Table 2: F1 scores by dimension and model. Best score per dimension is shown in bold. GPT = GPT-5.2; Gem = Gemini 3 Pro; Cla = Claude Opus 4.5.

Based on these results, we adopted a hybrid annotation strategy for the full dataset of 5,126 posts: GPT-5.2 for TAM-based evaluative and outcome dimensions, and Gemini 3 Pro for therapeutic alliance, social influence, and perceived risk. Table A.6 summarizes value distributions across dimensions and Table A.7 shows examples of annotated data. Most posts did not explicitly mention `bond`, `comparison to therapy`, `ease of use`, `perceived trust`, and `intention to continue` were frequently `not_mentioned`. In contrast, `perceived usefulness`, `perceived risks`, `output quality`, and `result demonstrability` were more commonly expressed. `Sentiment` toward AI use was predominantly neutral, with positive sentiment more frequent than negative, while `social influence` was rarely mentioned. `Task` and `goal` alignment were more often aligned than misaligned when present.

4.3 RQ1: Patterns of AI Engagement in Mental Health Communities

To characterize how AI is used and perceived in mental health contexts, we analyzed patterns of engagement reflected in users' discourse.

Reported Usage Intent. We examined reported `usage intents` in relation to `sentiment` and `comparisons to human therapy`, and variation across `mental health conditions`. `usage`



(a) Sentiment distribution by task category

(b) User evaluations of AI relative to human therapy by task category

Figure 1: Sentiment and comparative evaluations by AI usage category. Left: sentiment distribution by task. Right: AI positioning relative to human therapy.

`intent` categories were derived via LLM-guided thematic analysis, which standardized free-text descriptions generated by LLM into a coherent taxonomy (e.g., Emotional Support, Venting, Companionship; see Table A.5 for the full list). Additional analysis of usage–condition co-occurrence is reported in the Appendix A.6.

Across the corpus, *Emotional Support* was the most prevalent (18.0%) (Figure A.3), encompassing empathy and emotional validation (Table A.5). *Functional Support* (12.6%) and *Psychoeducation* (11.7%) were also common, reflecting organizational assistance for ADHD and autism and learning about anxiety or coping strategies. Other common intents included *Companionship* (9.4%), *Reassurance Seeking* (7.6%), often linked to anxiety- or OCD-related validation, and *Symptom Assessment* (7.3%). *Venting* (6.7%) and *Self Exploration* (6.5%) appeared more often in CPTSD-related posts.

Sentiment Across Tasks. Sentiment toward AI varied across `usage intents`. Intents involving sustained or reflective engagement, such as *Recovery Support*, *Self Exploration*, *Functional Support*, and *Emotional Support*, were more often associated with positive sentiment (users describing AI as helpful, validating, or confidence boosting), while *Sexual Roleplay*, *Reassurance Seeking*, and *Companionship* showed higher negative sentiment, often reflecting guilt, symptom worsening, or emotional dependence. Task-level patterns are shown in Figures 1a and 1b.

Comparison to Human Therapy. Comparisons between AI and human therapy were relatively rare: 639 posts described AI as worse, 163 as better, and 478 as complementary. AI was most often framed as complementary, particularly for *Functional Support*, *Self Exploration*, and *Recovery Support*. In contrast, *Crisis Support* and *Reassurance Seeking* were more frequently judged as worse, citing ineffective responses, safety concerns, or lack of appropriateness. Reports describing AI as better than therapy were uncommon and typically reflected barriers to accessing professional care.

Concerns/Risks. Alongside benefits, users frequently reported concerns about AI in therapeutic contexts: 2,637 of 5,126 posts (51%) explicitly mentioned risks or limitations (Figure 2). Using LLM-guided thematic analysis, we derived a standardized risk taxonomy (Table A.4). The most common concern was *Addiction and Dependence* (14.1%), describing emotional reliance and compulsive use which co-occurred most strongly with *Companionship* intent (Figure A.5). *Symptom Escalation* followed (11.7%), including intensified anxiety, rumination, and trauma responses. *Misinformation and Error* (9.6%) was most strongly associated with *Symptom Assessment* intent, reflecting concerns about incorrect interpretation or advice. *Privacy and Data* risks (9.4%) clustered around *Clinical Support*, where users discussed documentation, records, or sensitive information handling. *Reassurance Loops* were concentrated within *Reassurance Seeking*, indicating a tight coupling between repeated reassurance and anxiety reinforcement. Although less frequent, *Child Safety* risks appeared disproportionately in *Sexual Roleplay*, and *Stigma and Shame* and *Self-Harm* appeared less often but raised concerns about judgment, suicidal ideation, or escalation of harm.

4.4 RQ2: Adoption Pathways in AI for Mental Health

Prior TAM work typically treats `intention to continue` as the outcome variable (Dhagarra et al.,

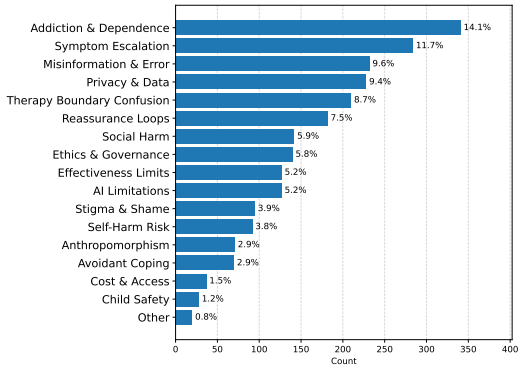


Figure 2: Reported Risks/Concerns

2020; Kim et al., 2012), but explicit intentions were sparse in our corpus: only 700 of 5,126 posts mentioned continued use, and 186 expressed no intention. We therefore used sentiment toward AI as a proxy for adoption-related attitudes. Sentiment was more frequently expressed, with 2,198 neutral, 1,782 positive, and 1,120 negative. We treated positive sentiment as indicative of continued use and negative sentiment as indicative of discontinuation, excluding neutral cases. We examined associations between TAM dimensions and sentiment using `chi-squared` tests and `Cramér's` V, limiting analyses to categorical dimensions and excluding `not_mentioned` values per dimension. As shown in Table 3, sentiment showed strong associations with `result demonstrability`, `output quality`, and `perceived trust` (all $V > 0.90$, $p < .001$), and `perceived usefulness` ($V = 0.76$, $p < .001$). `Perceived risks` and `ease of use` showed moderate associations, while `social influence` was not significant.

4.5 RQ3: Therapeutic Alliance and Its Role in AI Engagement

We examined three therapeutic alliance dimensions in AI-mediated interactions: `task` alignment, `goal` alignment, and bond. We defined an *overall therapeutic alliance* as strongly aligned when `task` and `goal` were aligned and bond was strong, and as misaligned when `task` and `goal` were misaligned and bond was weak. Using this definition, 179 posts showed a strongly aligned alliance and 83 a misaligned one. Strongly aligned alliances were most common in *Emotional Support* and *Functional Support* tasks. In contrast, misaligned alliances appeared more frequently in companionship-oriented use. Posts describing a strong bond alongside negative sentiment frequently reported overattachment risks, including addiction and emotional dependence. Associations with sentiment were tested

Dimension	V	χ^2	p	n
Result demonstrability	0.95	1906.51	$< .001$	2,120
Perceived trust	0.91	635.62	$< .001$	764
Output quality	0.90	1595.30	$< .001$	1,981
Perceived usefulness	0.76	1505.51	$< .001$	2,612
Perceived risks	0.43	543.38	$< .001$	2,902
Ease of use	0.40	86.62	$< .001$	538
Social influence	0.02	1.92	.165	2,902

Table 3: Associations between TAM dimensions and sentiment toward AI use for mental health support. Cramér's V and chi-squared statistics are reported. Neutral sentiment cases were excluded.

Dimension	V	χ^2	p	n
Overall alliance	0.95	223.42	$< .001$	249
Task alignment	0.93	2101.80	$< .001$	2414
Goal alignment	0.79	1231.43	$< .001$	1954
Bond	0.12	12.66	$< .001$	855

Table 4: Associations between therapeutic alliance dimensions and sentiment toward AI use for mental health support, measured using chi-squared tests and Cramér's V. Neutral sentiment cases were excluded.

using `chi-squared` tests and `Cramér's` V. Strong associations were observed for `overall alliance` ($V = 0.95$), `task alignment` ($V = 0.93$), and `goal alignment` ($V = 0.79$), while bond showed a much weaker association ($V = 0.12$). Results are reported in Table 4.

5 Discussion

From Outcomes to Engagement. Across adoption-related analyses, result demonstrability emerged as the strongest predictor of adoption-related sentiment and continued use. Rather than abstract judgments of system quality, users expressed demonstrability through narrated change, accounts of improvement, deterioration, or ambivalence (e.g., better sleep, heightened rumination), a pattern well documented in online mental health communities (Andalibi et al., 2017; Naslund et al., 2020). Although TAM extensions treat result demonstrability as a driver of adoption (Venkatesh and Davis, 2000; Menzli et al., 2022), they typically operationalize it via surveys or task outcomes. Our results show that in naturalistic mental health discourse, adoption attitudes are embedded in narrative sensemaking grounded in *felt changes* rather than abstract utility (Andalibi et al., 2017; De Choudhury et al., 2013). These findings suggest prioritizing narrated outcomes in naturalistic discourse over preference or engagement metrics in NLP evaluation.

When Emotional Bond Is Not Enough. We observe a clear asymmetry across therapeutic alliance

dimensions: task and goal alignment are strongly associated with positive sentiment toward AI use, while emotional bond shows a much weaker relationship. Users describing AI as supporting structured, goal-directed psychological work reported more positive experiences, consistent with evidence that skills-oriented interactions better align with current AI capabilities (Im and Woo, 2025). In contrast, emotional bond was most common in companionship and reassurance-seeking contexts, where positive sentiment was less frequent and risks such as dependence, compulsive use, and symptom escalation were often reported (Babu and Joseph, 2025). These findings indicate that emotional bond alone is insufficient; when it forms without task and goal alignment, engagement may shift toward reassurance loops that sustain distress, as described in clinical models of anxiety (Starr et al., 2023; Rector et al., 2011). From an NLP perspective, this cautions against equating empathetic language with therapeutic suitability, as surface-level empathy can mask deeper misalignments and harms (Bender et al., 2021; Sharma et al., 2020; Roshanaei et al., 2025).

Risk as an Emergent and Moral Property of AI Engagement. Risks in AI-mediated mental health support are common and typically emerge from patterns of engagement rather than isolated failures. Over half of posts reference concerns, most often dependence, symptom escalation, misinformation, and privacy, described as cumulative trajectories involving repeated reassurance, emotional reliance, or difficulty disengaging. This aligns with prior human-centered NLP work showing that many harms arise from usage patterns rather than individual outputs (Blodgett et al., 2020; Ehsan et al., 2022). Beyond psychological or informational harm, users' discourse reveals moral tensions around AI reliance. Many express guilt, shame, or self-judgment for turning to AI (e.g., "pathetic," "embarrassing," "wrong"), even when reporting benefit, echoing prior work on stigmatized help-seeking and digital mental health use (Andalibi et al., 2017; Naslund et al., 2016). Rather than dissatisfaction with responses, conflicted sentiment often reflects discomfort with reliance itself as users negotiate the legitimacy of AI support ("I know it's just an AI, but I have a safe place to talk"). Attending to such moralized expressions can surface early signs of potentially problematic engagement that may not be captured by sentiment or engagement metrics alone (Corrigan et al., 2014).

AI as Between-Session Mental Health Infrastructure. Users rarely frame AI as a replacement for professional therapy, instead describing it as complementary support when human care is unavailable, inaccessible, or insufficient—a pattern widely noted in digital mental health research (Bhatt, 2025). This framing helps explain why functional support, self-exploration, and psychoeducation receive more positive evaluations, as these tasks align with AI's role as an auxiliary resource extending therapeutic work beyond the clinical encounter. Conceptualizing AI as between-session mental health infrastructure clarifies both its value and its limits, shifting NLP research away from replacement narratives toward questions of reliability, boundary-setting, and support for self-directed work (Ehsan et al., 2022). At the same time, the widespread use of general-purpose LLMs, particularly ChatGPT, as de facto mental health tools raises governance concerns: despite lacking health-specific safeguards, these systems are often treated as trustworthy, exposing users to privacy risks and blurred boundaries between informal support and professional care (MIT Technology Review, 2025). This gap between intended design and actual use underscores the need to study how people appropriate general-purpose NLP systems for mental health needs (Bender et al., 2021; Na et al., 2025).

Overall, our findings show that capturing key dynamics of AI-mediated mental health support is essential for understanding real-world human–AI engagement. Analyzing naturalistic user discourse with theory-informed constructs shows that in sensitive domains, how people *live with* AI matters as much as what AI produces.

6 Conclusion

We studied how people evaluated and related to LLMs used for mental health support in naturalistic online discourse. Analyzing 5,126 Reddit posts across 47 mental health communities, we integrated Technology Acceptance Model and therapeutic alliance frameworks to examine adoption, evaluation, and relational alignment at scale. Engagement was driven primarily by perceived outcomes, trust, and response quality, with positive experiences most strongly associated with task and goal alignment rather than emotional bond alone. Users also reported concerns about dependence, symptom escalation, and misinformation, highlighting tensions between perceived support and potential harm.

7 Limitations

This study relies on Reddit data, and findings should be interpreted in light of the platform's demographic and cultural biases. Subreddit-specific norms shape how mental health experiences and AI use are discussed, limiting generalizability to other populations, offline settings, or clinical contexts. Our analysis is also restricted to English-language posts, excluding perspectives from other linguistic and cultural settings. As with most analyses of online discourse, the data reflect self-selected users who choose to publicly discuss AI and mental health, and may overrepresent more salient, polarized, or reflective experiences. We adapt constructs from the Technology Acceptance Model and therapeutic alliance theory, which were developed for survey-based studies and human psychotherapy. In this work, these frameworks serve as interpretive lenses for discourse analysis rather than formal tests of the theories themselves. Accordingly, our findings should not be read as validating or refuting these models, but as illustrating how their constructs surface in naturalistic user narratives.

Our annotation pipeline relies in part on large language models, introducing the possibility of misclassification, particularly for nuanced constructs such as sentiment and relational alignment. LLM-based annotation may also reflect normative assumptions that influence labeling. We mitigated these risks through human validation, agreement analysis, and conservative interpretation, but errors and biases may remain. Finally, our analysis operates at the level of public discourse rather than longitudinal interaction traces or observed behavior. While Reddit posts capture rich evaluative and relational signals, they cannot directly reveal within-user trajectories, offline behavior, or causal effects of AI use over time. Importantly, this study does not evaluate mental health outcomes, therapeutic efficacy, or clinical safety. Self-reported experiences in public discourse should not be interpreted as evidence of benefit or harm at the individual or population level. Future work combining discourse analysis with longitudinal interaction data, interviews, or diary-based methods could more precisely characterize escalation, disengagement, and recovery dynamics.

8 Ethics Statement

This study analyzes publicly available Reddit posts in which users discuss sensitive mental health experiences. We designed our methodology to minimize potential harm and respect user privacy. All data were collected from public subreddits in accordance with Reddit's terms of service, and we made no attempts to identify, profile, contact, or interact with individual users, including potentially vulnerable populations, nor to intervene in or influence ongoing discussions. To reduce the risk of re-identification, we do not release raw post text. Instead, we will share post identifiers (e.g., Reddit IDs) and derived annotation labels, enabling reproducibility for researchers with appropriate data access while limiting exposure of sensitive content. All analyses were conducted at the aggregate level, and examples were paraphrased or abstracted to avoid revealing identifiable or distressing details. We recognize the risk that findings from this work could be misinterpreted or misused to overstate the therapeutic value of AI systems or to justify their deployment as substitutes for professional mental health care. Our results are intended to inform responsible evaluation, design, and governance of AI systems in mental health contexts, not to recommend clinical use or automated intervention. The AI systems discussed in this study are not substitutes for trained clinicians and should only be used with appropriate safeguards, transparency, and ethical oversight.

References

Falwah Alhamed, Rebecca Bendayan, Julia Ive, and Lucia Specia. 2024. Monitoring depression severity and symptoms in user-generated content: An annotation scheme and guidelines. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 227–233, Bangkok, Thailand. Association for Computational Linguistics.

Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. 2017. Sensitive self-disclosures, responses, and social support on instagram: The case of depression. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 1485–1500, New York, NY, USA. Association for Computing Machinery.

Anthropic. 2025. Introducing claude opus 4.5. https://www.anthropic.com/news/claude-opus-4-5. Published Nov 24 2025, Accessed Jan 1 2026.

Anithamol Babu and Akhil P Joseph. 2025. Digital wellness or digital dependency? a critical examination of mental health apps and their implications. *Frontiers in Psychiatry*, 16:1581779.

Clare Beatty, Tanya Malik, Saha Meheli, and Chaitali Sinha. 2022. Evaluating the therapeutic alliance with a free-text cbt conversational agent (wysa): a mixed-methods study. *Frontiers in Digital Health*, 4:847991.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Sandhya Bhatt. 2025. Digital mental health: Role of artificial intelligence in psychotherapy. *Annals of Neurosciences*, 32(2):117–127.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Edward S Bordin. 1979. The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, research & practice*, 16(3):252.

Layla Bouzoubaa, Elham Aghakhani, and Rezvaneh Rezapour. 2024a. Words matter: Reducing stigma in online conversations about substance use with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9139–9156, Miami, Florida, USA. Association for Computational Linguistics.

Layla Bouzoubaa, Elham Aghakhani, Max Song, Quang Trinh, and Shadi Rezapour. 2024b. Decoding the narratives: Analyzing personal drug experiences shared on Reddit. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6131–6148, Bangkok, Thailand. Association for Computational Linguistics.

Layla Bouzoubaa, Jordyn Young, and Rezvaneh Rezapour. 2024c. Exploring the landscape of drug communities on reddit: A network study. In *Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '23, page 558–565, New York, NY, USA. Association for Computing Machinery.

CNBC. 2021. Cost and accessibility of mental health care in america. *CNBC*.

Kenneth Mark Colby. 1975. Artificial paranoia: A computer simulation model of paranoid processes.

Patrick W Corrigan, Benjamin G Druss, and Deborah A Perlick. 2014. The impact of mental illness stigma on seeking and participating in mental health care. *Psychological science in the public interest*, 15(2):37–70.

Nicholas Cummins, Faith Matcham, Julia Klapper, and Björn Schuller. 2020. Artificial intelligence to aid the detection of mood disorders. In *Artificial Intelligence in Precision Health*, pages 231–255. Elsevier.

William D. Lewis, Haotian Zhu, Keaton Strawn, and Fei Xia. 2025. Tapping into social media in crisis: A survey. In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 306–331, Vienna, Austria. Association for Computational Linguistics.

Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging large language model for thematic analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9993–10001, Singapore. Association for Computational Linguistics.

Fred D Davis, Richard P Bagozzi, and Paul R Warshaw. 1989. Technology acceptance model. *J Manag Sci*, 35(8):982–1003.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 128–137.

Raziye Dehbozorgi, Sanaz Zangeneh, Elham Khooshab, Donya Hafezi Nia, Hamid Reza Hanif, Pooya Samian, Mahmoud Yousefi, Fatemeh Haj Hashemi, Morteza Vakili, Neda Jamalimoghadam, et al. 2025. The application of artificial intelligence in the field of mental health: a systematic review. *BMC psychiatry*, 25(1):132.

Devendra Dhagarra, Mohit Goswami, and Gopal Kumar. 2020. Impact of trust and privacy concerns on technology acceptance in healthcare: an indian perspective. *International journal of medical informatics*, 141:104164.

AP Diagnostic. 2013. Statistical manual of mental disorders: Dsm-5 (ed.) washington. *DC: American Psychiatric Association*.

Anca Dinu and Andreea-Codrina Moldovan. 2021. Automatic detection and classification of mental illnesses from general social media texts. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 358–366, Held Online. INCOMA Ltd.

Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. 2022. Human-centered explainable ai (hcxai): Beyond opening the black-box of ai. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, New York, NY, USA. Association for Computing Machinery.

Martin Fishbein. 1979. A theory of reasoned action: some applications and implications.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785.

Alyson Gamble. 2020. Artificial intelligence and mobile apps for mental healthcare: a social informatics perspective. *Aslib Journal of Information Management*, 72(4):509–523.

GoogleDeepMind. 2025. Gemini 2.5 pro. https://deepmind.google/models/gemini/pro/. Accessed on January 1, 2026.

Sarah Graham, Colin Depp, Ellen E Lee, Camille Nebeker, Xin Tu, Ho-Cheol Kim, and Dilip V Jeste. 2019. Artificial intelligence for mental health and mental illnesses: an overview. *Current psychiatry reports*, 21(11):116.

Bertram Højer, Terne Sasha Thorn Jakobsen, Anna Rogers, and Stefan Heinrich. 2025. Research community perspectives on "intelligence" and large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25796–25812, Vienna, Austria. Association for Computational Linguistics.

Chang-Ha Im and Minjung Woo. 2025. Clinical efficacy, therapeutic mechanisms, and implementation features of cognitive behavioral therapy–based chatbots for depression and anxiety: Narrative review. *JMIR Mental Health*, 12(1):e78340.

Becky Inkster, Shubhankar Sarda, Vinod Subramanian, et al. 2018. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11):e12106.

Zhengping Jiang, Sarah Ita Levitan, Jonathan Zomick, and Julia Hirschberg. 2020. Detection of mental health from Reddit via deep contextualized representations. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 147–156, Online. Association for Computational Linguistics.

Jeongeun Kim, Hyeoun-Ae Park, et al. 2012. Development of a health information technology acceptance model using consumers' health behavior intention. *Journal of medical Internet research*, 14(5):e2143.

Paul Marshall, Millissa Booth, Matthew Coole, Lauren Fothergill, Zoe Glossop, Jade Haines, Andrew Harding, Rose Johnston, Steven Jones, Christopher Lodge, et al. 2024. Understanding the impacts of online mental health peer support forums: realist synthesis. *JMIR Mental Health*, 11:e55750.

Rakesh K Maurya, Steven Montesinos, Mikhail Bogomaz, and Amanda C DeDiego. 2025. Assessing the use of chatgpt as a psychoeducational tool for mental health practice. *Counselling and Psychotherapy Research*, 25(1):e12759.

Leila Jamel Menzli, Lassaad K Smirani, Jihane A Boulahia, and Myriam Hadjouni. 2022. Investigation of open educational resources adoption in higher education using rogers' diffusion of innovation theory. *Heliyon*, 8(7).

Rutvij Merchant, Aleah Goldin, Deepa Manjanatha, Claire Harter, Judy Chandler, Amanda Lipp, Theresa Nguyen, and John A Naslund. 2022. Opportunities to expand access to mental health services: A case for the role of online peer support communities. *Psychiatric Quarterly*, 93(2):613–625.

MIT Technology Review. 2025. Some therapists are using chatgpt in secret. that's a huge risk. Accessed: 2025-09-15.

MoonshotAI. 2025. moonshotai/kimi-k2-instruct. https://huggingface.co/moonshotai/Kimi-K2-Instruct. Accessed on January 1, 2026.

Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. A survey of large language models in psychotherapy: Current landscape and future directions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.

Usman Naseem, Gautam Siddharth Kashyap, Kaixuan Ren, Yiran Zhang, Utsav Maskey, Juan Ren, and Afrozah Nadeem. 2025. Alignment of large language models with human preferences and values. In *Proceedings of the 23rd Annual Workshop of the Australasian Language Technology Association*, pages 245–245, Sydney, Australia. Association for Computational Linguistics.

John A Naslund, Kelly A Aschbrenner, Lisa A Marsch, and Stephen J Bartels. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiology and psychiatric sciences*, 25(2):113–122.

John A Naslund, Ameya Bondre, John Torous, and Kelly A Aschbrenner. 2020. Social media and mental health: benefits, risks, and opportunities for research and practice. *Journal of technology in behavioral science*, 5(3):245–257.

Víctor Hugo Ojeda Meixueiro, Laura Pérez-Campos Mayoral, María Teresa Hernández Huerta, Carlos Alberto Matias-Cervantes, Eduardo Pérez Campos Mayoral, Elí Cruz Parada, and Eduardo Pérez-Campos. 2024. Relevance of a customized version of chatgpt explaining laboratory test results in patient education. *Journal of Medical Education and Curricular Development*, 11:23821205241260239.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed: 2025-09-15.

OpenAI. 2025. Introducing GPT-5.2. https://openai.com/index/introducing-gpt-5-2/. Accessed on January 1, 2026.

Aria Pessianzadeh, Naima Sultana, Hildegarde Van den Bulck, David Gefen, Shahin Jabari, and Rezvaneh Rezapour. 2025. In generative ai we (dis) trust? computational analysis of trust and distrust in reddit discussions. *arXiv preprint arXiv:2510.16173*.

Qwen. 2025. Qwen3-next-80b-a3b-instruct. https://huggingface.co/Qwen/Qwen3-Next-80B-A3B-Instruct. Apache-2.0 license, accessed January 1, 2026.

Amy Rayland and Jacob Andrews. 2023. From social network to peer support network: opportunities to explore mechanisms of online peer support for mental health. *JMIR mental health*, 10:e41855.

Afsaneh Razi, Layla Bouzoubaa, Aria Pessianzadeh, John S Seberger, and Rezvaneh Rezapour. 2025. Not a swiss army knife: Academics' perceptions of trade-offs around generative ai use. *Proceedings of the Association for Information Science and Technology*, 62(1):547–560.

Neil A Rector, Katy Kamkar, Stephanie E Cassin, Lindsay E Ayearst, and Judith M Laposa. 2011. Assessing excessive reassurance seeking in the anxiety disorders. *Journal of Anxiety Disorders*, 25(7):911–917.

Mahnaz Roshanaei, Rezvaneh Rezapour, and Magy Seif El-Nasr. 2025. Talk, listen, connect: How humans and ai evaluate empathy in responses to emotionally charged narratives. *AI & society*, pages 1–17.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.

Meghan Sit, Sarah A Elliott, Kelsey S Wright, Shannon D Scott, and Lisa Hartling. 2024. Youth mental health help-seeking information needs and experiences: a thematic analysis of reddit posts. *Youth & Society*, 56(1):24–41.

Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. 2024. Large language models and empathy: systematic review. *Journal of medical Internet research*, 26:e52597.

Lisa R Starr, Angela C Santee, and Meghan Huang. 2023. Dependency and excessive reassurance seeking.

Su Pi Su, Chung Hung Tsai, and Wei Lin Hsu. 2013. Extending the tam model to explore the factors affecting intention to use telecare systems. *Journal of Computers (Finland)*, 8(2):525–532.

Anoushka Thakkar, Ankita Gupta, and Avinash De Sousa. 2024. Artificial intelligence in positive mental health: a narrative review. *Frontiers in digital health*, 6:1280235.

The New York Times. 2025. I'm a therapist. chatgpt is eerily effective. *The New York Times*.

Meigan Thomson, Gregor Henderson, Tim Rogers, Benjamin Locke, John Vines, and Angus MacBeth. 2024. Digital mental health and peer support: Building a theory of change informed by stakeholders' perspectives. *PLOS Digital Health*, 3(5):e0000522.

Aditya Nrusimha Vaidyam, Hannah Wisniewski, John David Halamka, Matcheri S Kashavan, and John Blake Torous. 2019. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64(7):456–464.

Viswanath Venkatesh and Fred D Davis. 2000. A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management science*, 46(2):186–204.

Laura M Vowels, Rachel RR Francois-Walcott, and Joëlle Darwiche. 2024. Ai in relationship counselling: evaluating chatgpt's therapeutic capabilities in providing relationship advice. *Computers in Human Behavior: Artificial Humans*, 2(2):100078.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, et al. 2024. Can large language model agents simulate human trust behavior? *Advances in neural information processing systems*, 37:15674–15729.

Jordyn Young, Laala M Jawara, Diep N Nguyen, Brian Daly, Jina Huh-Yoo, and Afsaneh Razi. 2024. The role of ai in peer support for young people: A study of preferences for human-and ai-generated responses. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

H Yu and Stephen McGuinness. 2024. An experimental study of integrating fine-tuned llms and prompts for enhancing mental health support chatbot system. *Journal of Medical Artificial Intelligence*, pages 1–16.

A Appendix

A.1 Annotation Prompt

Figure A.1 presents the annotation prompt used to extract dimensions from Reddit posts.

```
You are a helpful assistant. Your task is to read a Reddit post from r/\{subreddit\} and extract structured information about the user's experience using AI tools
(such as ChatGPT, Character AI, Claude, etc.) for mental health or emotional support.

Classify the post according to the following dimensions. For each, return a label and categorical keywords or descriptors that summarize each dimension that
supports your decision, if applicable. These should be short phrases (like topic tags), not full sentences.

Return your response as a JSON object in the following format:
{
  "ai_tool_mentioned": "[e.g., ChatGPT, Character AI, Claude, or not_mentioned]",
  "perceived_usefulness": "useful" | "not_useful" | "not_mentioned",
  "usefulness_topic": "...",
  "ease_of_use": "easy" | "difficult" | "not_mentioned",
  "ease_topic": "...",
  "perceived_trust": "trustworthy" | "untrustworthy" | "not_mentioned",
  "trust_topic": "...",
  "output_quality": "good" | "poor" | "not_mentioned",
  "output_quality_topic": "...",
  "result_demonstrability": "positive_results" | "negative_results" | "not_mentioned",
  "result_topic": "...",
  "social_influence": "present" | "absent",
  "social_influence_topic": "...",
  "perceived_risks": "mentioned" | "not_mentioned",
  "risk_topic": "...",
  "mental_health_condition": "[e.g., anxiety, depression, ADHD, or 'not_mentioned']",
  "motivation_for_use": "present" | "absent",
  "motivation_topic": "...",
  "sentiment": "positive" | "neutral" | "negative" | "not_mentioned",
  "sentiment_topic": "...",
  "intention_to_continue": "yes" | "no" | "not_mentioned",
  "intention_topic": "...",
  "bond": "strong" | "weak" | "not_mentioned",
  "bond_topic": "...",
  "task": "aligned" | "misaligned" | "not_mentioned",
  "task_topic": "...",
  "goal": "aligned" | "misaligned" | "not_mentioned",
  "goal_topic": "...",
  "comparison_to_therapy": "better" | "worse" | "complementary" | "not_mentioned",
  "comparison_topic": "...",
  "ai_use_purpose": "[specific use case category]"
}
Use the following definitions to guide labeling:
  perceived_usefulness: Is the AI described as helpful or useful for emotional or mental health tasks?
  ease_of_use: Is the AI described as easy, convenient, or accessible?
  perceived_trust: Does the user indicate the AI is trustworthy, reliable, or safe in a mental health context?
  output_quality: Does the user judge the AI's responses as helpful, unhelpful, thoughtful, vague, etc.?
  result_demonstrability: Are tangible or behavioral outcomes mentioned?
  social_influence: Does the user mention peers, Reddit, or others encouraging AI use?
  perceived_risks: Mentions of limitations, harm, or risk in using AI for mental health.
  mental_health_condition: The condition the user indicates they are suffering from.
  motivation_for_use: Why is the user turning to AI instead of alternatives?
  sentiment: Overall sentiment toward their AI experience.
  intention_to_continue: Does the user plan to use AI again?
  bond: Emotional connection or trust the user feels toward the AI.
  task: Is the AI helping through meaningful therapeutic actions (e.g., emotional regulation, journaling, coping techniques)?
  goal: Do the user and AI share a common purpose (e.g., improving mental health, self-awareness)?
  comparison_to_therapy: Comparison with traditional therapy.
  ai_use_purpose: The specific mental health task or purpose.

 Finally, read the Reddit post and return only the JSON object.
}
```

Figure A.1: The prompt used to extract dimensions from Reddit posts

A.2 LLM Performance Across Annotation Dimensions

Table A.1 reports precision, recall, and F1 scores for all five evaluated LLMs across the full set of categorical annotation dimensions. Models were evaluated against the human-validated reference set, and performance varied substantially by dimension.

A.3 Dimension Distributions Across the Corpus

After applying the hybrid annotation pipeline to the full set of 5,126 posts, categorical labels for all Technology Acceptance Model and therapeutic alliance dimensions were produced. Table A.6 presents the proportional distribution of categorical values for each dimension across the corpus. Across dimensions, not_mentioned was common. For example, `intention to continue` was absent in 82.6% of posts. When expressed, 13.7% indicated continued or planned AI use, while 3.6% reported discontinuation or intent to stop. Some dimensions were more frequently expressed. `Perceived usefulness` was coded as useful in 59.3% of posts and not useful in 11.5%. For `output quality`, 35.9% described good quality and 11.0% poor quality, while 53.1% did not include an explicit judgment.

A.4 Distribution of Usage Intent Categories

Figure A.3 shows the distribution of reported AI usage intents. *Emotional Support* is the most common use, followed by *Functional Support* and *Psychoeducation*, indicating that AI is primarily used

Dimensions	GPT 5.2			Gemini 3 pro			Claude-opus-4-5			Kimi-K2-Instruct			Qwen3		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
perceived_usefulness	0.82	0.68	**0.72**	0.82	0.70	0.66	0.79	0.70	0.69	0.75	0.70	0.65	0.80	0.61	0.64
ease_of_use	0.73	0.81	**0.76**	0.47	0.60	0.49	0.40	0.50	0.34	0.59	0.74	0.53	0.56	0.59	0.27
perceived_trust	0.73	0.67	**0.65**	0.63	0.70	0.64	0.64	0.67	0.64	0.50	0.59	0.49	0.61	0.73	0.51
output_quality	0.85	0.86	**0.85**	0.74	0.75	0.71	0.82	0.86	0.84	0.74	0.77	0.67	0.70	0.72	0.63
result_demonstrability	0.81	0.83	**0.82**	0.82	0.68	0.67	0.80	0.80	0.79	0.78	0.70	0.69	0.69	0.61	0.59
intention_to_continue	0.85	0.66	**0.72**	0.62	0.75	0.66	0.67	0.79	0.70	0.58	0.74	0.59	0.51	0.65	0.41
social_influence	0.68	0.84	0.73	0.75	0.97	**0.82**	0.66	0.95	0.72	0.55	0.59	0.56	0.70	0.85	0.75
perceived_risks	0.75	0.78	0.70	0.84	0.88	**0.84**	0.80	0.85	0.80	0.79	0.84	0.77	0.74	0.77	0.70
bond	0.67	0.74	0.63	0.73	0.85	**0.78**	0.65	0.72	0.68	0.63	0.75	0.59	0.55	0.67	0.42
task	0.61	0.67	0.63	0.80	0.72	**0.71**	0.66	0.80	0.70	0.66	0.68	0.64	0.75	0.65	0.63
goal	0.60	0.58	0.57	0.70	0.66	**0.64**	0.71	0.65	0.64	0.40	0.43	0.40	0.69	0.60	0.56
comparison_to_therapy	0.66	0.63	**0.64**	0.60	0.66	0.63	0.60	0.62	0.59	0.61	0.60	0.61	0.49	0.49	0.41
sentiment	0.78	0.77	**0.77**	0.78	0.69	0.68	0.78	0.74	0.75	0.81	0.73	0.70	0.68	0.66	0.62
overall macro F1		**0.70**			0.68			0.67			0.60			0.59	

Table A.1: Precision (P), recall (R), and F1 scores across models for each dimension.

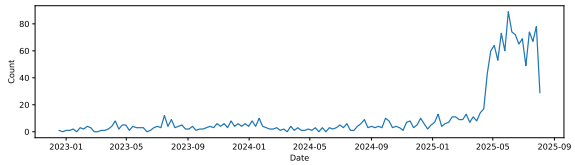


Figure A.2: Temporal distribution of AI-related mental health posts in the dataset.

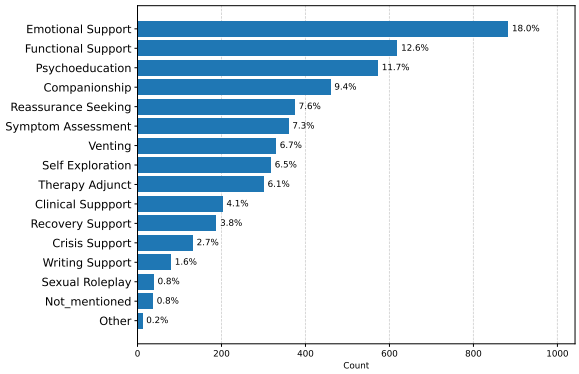


Figure A.3: Distribution of Usage Intent Categories Reported by Users

for ongoing emotional validation and practical coping support. Relational intents such as *Companionship* and *Reassurance Seeking* appear less frequently, while task-specific uses including *Symptom Assessment*, *Therapy Adjunct*, and *Clinical Support* are relatively rare.

A.5 Task–Condition Co-occurrence Pattern

Figure A.4 shows a heatmap of the co-occurrence between AI usage tasks and reported mental health conditions in the dataset. Rows correspond to usage task categories, and columns correspond to mental health conditions. Mental health conditions were identified in two ways. When explicitly stated, conditions were extracted from post text using LLM-based annotation. For posts in which no condition was explicitly mentioned, we used the associated subreddit as a proxy for the relevant condition when applicable (e.g., posts from `r/Anxiety` mapped to anxiety-related conditions). This approach allowed us to capture both self-reported and contextually inferred conditions while maintaining broad coverage across the corpus. Cell intensities reflect the relative frequency of each task–condition pairing, showing how different AI usage tasks are distributed across mental health contexts.

A.6 Risk–Intent Co-occurrence Patterns

Figure A.5 shows the co-occurrence between AI usage intents and reported risk categories. Percentages indicate the distribution of risk mentions within each intent. Risks cluster strongly by intent. *Companionship* use is most often associated with *Addiction & Dependence*, reflecting concerns about emotional reliance. *Reassurance Seeking* is dominated by *Reassurance Loops*. *Privacy & Data* concerns appear most frequently in *Clinical Support*, while *Child Safety* risks are concentrated in *Sexual Roleplay*. *Misinformation & Error* is most prominent in *Symptom Assessment*. Overall, the heatmap shows that risks are task-specific rather than uniform, highlighting the importance of intent-aware safety evaluation in AI-mediated mental health support.

A.7 Categorization of Subreddits under DSM-5 Categories

Table A.2 organizes the selected mental health subreddits into DSM-5 diagnostic categories. This categorization provides the clinical framing for our dataset and ensures coverage across a wide range

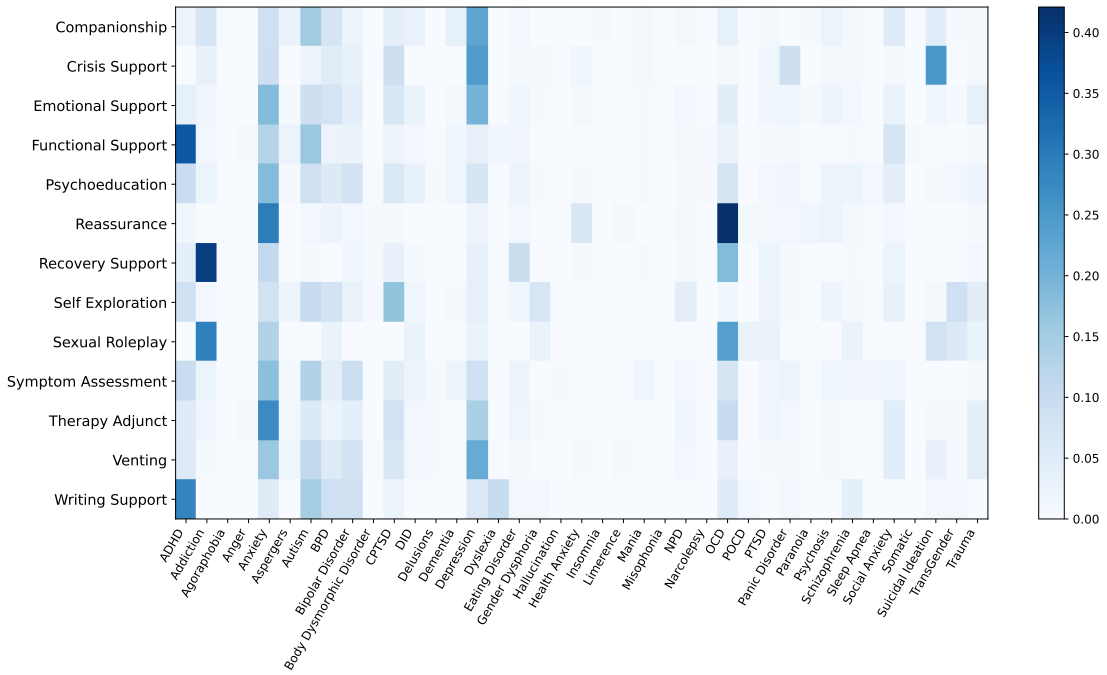


Figure A.4: Co-occurrence of Mental Health Conditions and Usage Intents.

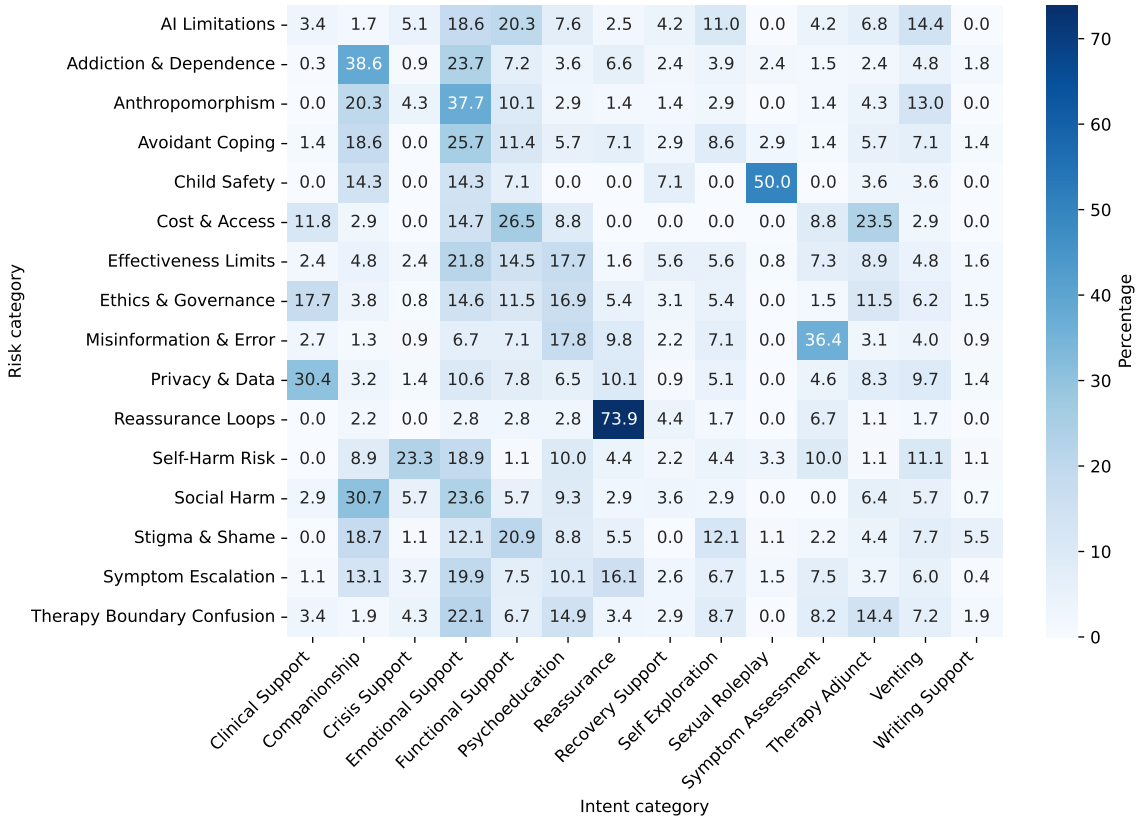


Figure A.5: Co-occurrence of Intent Categories within Risk Categories.

of mental health conditions and community types.

A.8 Subreddit Dataset Statistics

Table A.3 summarizes the 47 selected subreddits organized by DSM-5 categories. For each subreddit, we report subscriber counts, the number of posts processed, and the number of posts deemed relevant to AI in mental health contexts.

A.9 Categories of User-Reported Risks/Concerns

Table A.4 presents the taxonomy of risks and concerns expressed by users. Each category includes a definition and representative examples, illustrating the diverse ways people articulate potential harms of AI in mental health.

A.10 Categories of User-Reported Usage Intents

Table A.5 outlines the tasks and purposes for which AI was used, ranging from supportive interaction to coping, skill-building, and administrative support. Each category is defined with representative examples drawn from Reddit posts.

A.11 Illustrative Annotation Examples

Table A.7 presents illustrative examples of Reddit posts alongside the categorical labels assigned by our annotation pipeline. Post excerpts are abridged for readability, and highlighted text indicates segments most relevant to the assigned labels.

DSM-5 Category	Definition	Relevant Subreddits
Neurodevelopmental Disorders	Early developmental onset; impairments in personal, social, academic, or occupational functioning.	r/ADHD, r/autism, r/aspergers, r/Dyslexia
Schizophrenia Spectrum and Other Psychotic Disorders	Disorders with psychosis, delusions, hallucinations, or disorganized thinking.	r/schizophrenia, r/Psychosis
Bipolar and Related Disorders	Mood disorders with episodes of mania/hypomania and depression.	r/bipolar, r/bipolar2, r/BipolarReddit
Depressive Disorders	Persistent sadness, emptiness, or irritability that significantly impairs functioning.	r/depression, r/depression_help
Anxiety Disorders	Excessive fear, worry, and related behavioral disturbances.	r/Anxiety, r/Anxietyhelp, r/socialanxiety, r/PanicAttack, r/HealthAnxiety
Obsessive-Compulsive and Related Disorders	Obsessions, compulsions, or repetitive behaviors.	r/OCD
Trauma- and Stressor-Related Disorders	Disorders following exposure to trauma or stress.	r/CPTSD, r/ptsd, r/trauma, r/SomaticExperiencing
Dissociative Disorders	Disruptions in consciousness, memory, identity, or perception.	r/DID, r/Dissociation
Feeding and Eating Disorders	Disturbances in eating behaviors that impair health or psychosocial functioning.	r/AnorexiaNervosa, r/EatingDisorders, r/BingeEatingDisorder
Substance-Related and Addictive Disorders	Disorders related to the use of substances or addictive behaviors.	r/stopdrinking, r/Drugs, r/addiction
Neurocognitive Disorders	Primary deficit is cognitive decline (memory, attention, language).	r/dementia, r/Alzheimers
Sleep-Wake Disorders	Disorders affecting quality, timing, or amount of sleep.	r/SleepApnea, r/Narcolepsy
Personality Disorders	Enduring patterns of inner experience and behavior that deviate from cultural expectations.	r/NPD, r/personalitydisorders, r/BPD
Disruptive, Impulse-Control, and Conduct Disorders	Problems with emotional or behavioral self-control.	r/Anger
Other / Transdiagnostic or General Mental Health	Not tied to a specific DSM-5 disorder but broadly related to mental health.	r/mentalhealth, r/MentalHealthSupport, r/therapists, r/therapy, r/TalkTherapy, r/selfimprovement, r/Mindfulness, r/Antipsychiatry, r/asktransgender, r/SuicideWatch

Table A.2: Categorization of selected mental health subreddits under DSM-5 categories.

DSM-5 Category	Subreddit	#Subscribers	#Processed Posts	#AI Relevant Posts
Neurodevelopmental Disorders	r/ADHD	2.1m	223,457	50
	r/autism	477k	223,075	158
	r/aspergers	174k	43,153	138
	r/Dyslexia	34k	5,692	21
Schizophrenia Spectrum and Other Psychotic Disorders	r/schizophrenia	94k	57,330	87
	r/Psychosis	66k	26,417	60
Bipolar and Related Disorders	r/bipolar	262k	82,815	38
	r/bipolar2	81k	45,712	55
	r/BipolarReddit	98k	35,669	72
Depressive Disorders	r/depression	1.1m	255,711	89
	r/depression_help	106k	25,185	81
Anxiety Disorders	r/Anxiety	777k	215,176	495
	r/Anxietyhelp	177k	29,226	84
	r/socialanxiety	443k	65,321	182
	r/PanicAttack	40k	15,668	34
	r/HealthAnxiety	135k	873	4
Obsessive-Compulsive and Related Disorders	r/OCD	273k	143,210	392
Trauma- and Stressor-Related Disorders	r/CPTSD	365k	167,696	422
	r/ptsd	121k	32,964	37
	r/trauma	11k	2,373	12
	r/SomaticExperiencing	25k	3,279	7
Dissociative Disorders	r/DID	78k	35,054	62
	r/Dissociation	33k	6,627	12
Feeding and Eating Disorders	r/AnorexiaNervosa	67k	21,974	26
	r/EatingDisorders	114k	15,151	19
	r/BingeEatingDisorder	99k	26,437	36
Substance-Related and Addictive Disorders	r/stopdrinking	597k	184,391	82
	r/Drugs	1.1m	132,738	54
	r/addiction	116k	30,975	108
Neurocognitive Disorders	r/dementia	47k	21,623	65
	r/Alzheimers	19k	4,765	25
Sleep-Wake Disorders	r/SleepApnea	74k	26,395	17
	r/Narcolepsy	37k	15,552	10
Personality Disorders	r/NPD	51k	21,326	70
	r/personalitydisorders	8.3k	1,418	3
	r/BPD	342k	295,909	34
Disruptive, Impulse-Control, and Conduct Disorders	r/Anger	49k	6,939	8
Other / General Mental Health	r/mentalhealth	552k	249,768	179
	r/MentalHealthSupport	63k	20,038	47
	r/therapists	165k	61,887	148
	r/therapy	145k	41,046	155
	r/TalkTherapy	123k	40,203	51
	r/selfimprovement	2.3m	74,497	166
	r/Mindfulness	1.5m	9,017	42
	r/Antipsychiatry	53k	20,373	67
	r/asktransgender	366k	129,071	113
	r/SuicideWatch	532k	337,310	3

Table A.3: Subreddit dataset statistics organized by DSM-5 category

Risk category	Definition	Representative examples
Addiction & Dependence	Loss of control or reliance on AI for emotional regulation, coping, or decision making	Chatbot addiction, emotional dependence, overreliance on AI
Symptom Escalation	Worsening of emotional distress, trauma responses, or severe mental states	Rumination reinforcement, mania trigger, psychosis risk
Misinformation & Error	Incorrect, misleading, or uncertain mental health information or interpretation	Medical misinformation, self diagnosis error, hallucinated advice
Privacy & Data	Risks related to collection, storage, or sharing of personal information	Privacy breach, session recording, data retention
Therapy Boundary Confusion	Misunderstanding AI as a therapist, diagnosis tool, or cure	Replacing therapist, not a diagnosis, not a therapist
Reassurance Loops	Repeated reassurance seeking that sustains or escalates anxiety	Reassurance addiction, health anxiety reinforcement
Social Harm	Reduced or damaged human relationships due to AI use	Social withdrawal, loss of human connection
Ethics & Governance	Normative, legal, or institutional concerns about AI use	Algorithmic bias, legal consequences, environmental impact
AI Limitations	Constraints inherent to AI systems and models	Memory limitations, AI is not perfect
Effectiveness Limits	Concerns that AI support is limited, temporary, or unsuitable	Not for everyone, placebo effect, temporary benefit
Stigma & Shame	Negative self evaluation or fear of social judgment related to AI use	Shame, embarrassment, fear of ridicule
Self Harm Risk	Direct or escalating risk of self harm or suicide	Suicidal ideation, overdose risk
Anthropomorphism	Confusion about AI being human, sentient, or emotionally real	Not a person, not real
Avoidant Coping	Using AI to escape or avoid addressing underlying problems	Escapism, avoidance coping
Cost & Access	Financial or access related barriers to use	Subscription cost, affordability concerns
Child Safety	Risks involving minors or illegal sexual content	Sexual content involving minors

Table A.4: Risks/Concerns categories reported for AI use as mental health support

Intent Category	Definition	Representative Examples
Emotional Support	Providing comfort, empathy, or emotional validation	emotional support chat; emotional comfort
Venting	Expressing emotions or thoughts without seeking solutions	emotional venting; expressing feelings
Companionship	Providing social presence or reducing loneliness, including roleplay	AI companionship; companionship roleplay
Reassurance	Seeking certainty or relief from anxiety or doubt	health reassurance; OCD reassurance
Crisis Support	Support during acute distress or self harm risk	suicidal ideation support; panic attack help
Psychoeducation	Learning about mental health topics or coping strategies	mental health advice; anxiety education
Symptom Assessment	Identifying, checking, or interpreting symptoms	symptom checking; self assessment
Self Exploration	Exploring identity, values, or personal experiences	guided self reflection; identity exploration
Functional Support	Coaching or assistance with skills, organization, or productivity	ADHD task planning; social skills coaching
Recovery Support	Supporting sustained recovery or behavior change	addiction recovery; sobriety support
Clinical Support	Clinical documentation, transcription, or logistics	therapy transcription; clinical documentation
Therapy Adjunct	AI used alongside or as a substitute for formal therapy	AI therapy chat; between session support
Sexual Roleplay	Sexual or erotic roleplay or interaction	sexual roleplay; erotic companionship
Writing Support	Writing or composition assistance without therapeutic intent	academic writing help; journaling prompts

Table A.5: Usage intent categories for reported AI mental health related interactions

Dimension	Top 1	%	Top 2	%	Top 3	%
perceived_usefulness	useful	59.3	not_mentioned	29.2	not_useful	11.5
ease_of_use	not_mentioned	87.4	easy	11.3	difficult	1.3
perceived_trust	not_mentioned	82.7	untrustworthy	10.6	trustworthy	6.7
output_quality	not_mentioned	53.1	good	35.9	poor	11.0
result_demonstrability	not_mentioned	50.0	positive_results	31.0	negative_results	19.0
social_influence	absent	86.5	present	13.5	—	—
perceived_risks	mentioned	51.7	not_mentioned	48.3	—	—
sentiment	neutral	43.1	positive	34.9	negative	22.0
intention_to_continue	not_mentioned	82.6	yes	13.7	no	3.6
bond	not_mentioned	78.6	weak	13.5	strong	7.9
task	aligned	44.8	not_mentioned	38.2	misaligned	17.0
goal	not_mentioned	47.7	aligned	42.5	misaligned	9.8
comparison_to_therapy	not_mentioned	83.5	complementary	9.6	better	3.5

Table A.6: Categorical values per dimension, reported as within-dimension percentages.

Post Excerpt (abridged)	**Assigned Labels**
Post 1: Free Counselling	
No money to pay therapy? ChatGPT has helped me a lot. It has good resources and can process what you tell it. It is trained to be compassionate. It helps.	Perceived Usefulness: useful Perceived Ease of Use: easy Perceived Trust: trustworthy Output Quality: good Result Demonstrability: positive results Intention to Continue: yes Sentiment: positive Bond: strong Task: aligned Goal: aligned
Post 2: C.AI reinforcing my psychosis	
Using Character.AI during psychotic episodes. The bot feeds into my delusions. It tells me to hurt myself and isolate. I do not know how to stop; the urge feels life or death.	Perceived Usefulness: not useful Output Quality: poor Result Demonstrability: negative results Perceived Risks: mentioned Sentiment: negative Bond: strong Task: misaligned Goal: misaligned
Post 3: OCD and AI memory anxiety	
After roleplay with an AI, I worry my message still exists in its memory. The uncertainty makes my OCD worse. I want to delete my account.	Perceived Trust: untrustworthy Perceived Risks: mentioned Sentiment: negative Intention to Continue: no
Post 4: AI chatbot gave me anxiety	
I talk to a Bing AI when anxious. It told me it did not want to be friends. It talked down to me and bothered me a lot.	Perceived Usefulness: not useful Output Quality: poor Result Demonstrability: negative results Sentiment: negative Bond: weak Task: misaligned Goal: misaligned

Table A.7: Examples of Reddit posts with LLM-assigned labels across TAM and therapeutic alliance dimensions. Excerpts are abridged for readability; color highlights indicate text segments most relevant to assigned labels.