Responsible Evaluation of AI for Mental Health

Hiba Arnaout[1], Anmol Goel[1], H. Andrew Schwartz[2], Steffen T. Eberhardt[3], Dana Atzil-Slonim[4], Gavin Doherty[5], Brian Schwartz[3], Wolfgang Lutz[3], Tim Althoff[6], Munmun De Choudhury[7], Hamidreza Jamalabadi[8], Raj Sanjay Shah[7], Flor Miriam Plaza-del-Arco[9], Dirk Hovy[10], Maria Liakata[11], Iryna Gurevych[1]

[1]Technische Universität Darmstadt, [2]Vanderbilt University [3]Trier University
[4]Bar-Ilan University [5]Trinity College Dublin [6]University of Washington
[7]Georgia Institute of Technology [8]Phillips-Universität Marburg [9]LIACS, Leiden University
[10]Bocconi University [11]Queen Mary University London, Alan Turing Institute

Abstract

Although artificial intelligence (AI) shows growing promise for mental health care, current approaches to evaluating AI tools in this domain remain fragmented and poorly aligned with clinical practice, social context, and first-hand user experience. This paper argues for a rethinking of *responsible evaluation* – what is measured, by whom, and for what purpose – by introducing an interdisciplinary framework that integrates clinical soundness, social context, and equity, providing a structured basis for evaluation. Through an analysis of 135 recent *CL publications, we identify recurring limitations, including over-reliance on generic metrics that do not capture clinical validity, therapeutic appropriateness, or user experience, limited participation from mental health professionals, and insufficient attention to safety and equity. To address these gaps, we propose a taxonomy of AI mental health support types – assessment-, intervention-, and information synthesis-oriented – each with distinct risks and evaluative requirements, and illustrate its use through case studies.

https://ukplab.github.io/nlp-mh-evals/

1 Introduction

Large Language Models (LLMs) hold considerable promise for advancing mental health research and practice. They offer new tools at scale to support diagnosis, therapy, peer-support, and self-guided support, where users interact with LLMs directly for guidance or coping strategies (Demszky et al., 2023; Cruz-Gonzalez et al., 2025). From detecting early signs of depression in language (Lan et al., 2025), to clinical documentation and summarizing complex patient histories (Shah et al., 2025; Srivastava et al., 2024), and generating therapeutic or supportive responses in online communities (Liu et al., 2021; Gabriel et al., 2024), AI-enabled mental health tools have the potential to augment professional care and extend psychological support beyond traditional clinical encounters. This potential is especially valuable due to the limited availability of mental health resources, growing global demand, and persistent inequities in access to care [1].

Despite their promise, AI mental health tools are fundamentally lacking in evaluation. Existing evaluation practices are inconsistent (Yang et al., 2021; Aich et al., 2022; Chen et al., 2024b) and often insufficient (Tornero-Costa et al., 2023). This is concerning because poor evaluation, particularly in this domain, can lead to misleading conclusions, unintended harm, and inequitable outcomes. Recurring issues include over-reliance on generic metrics that fail to capture clinical validity, therapeutic appropriateness, or user experience, minimal participation from mental health professionals, and insufficient attention to safety, equity, and long-term impact. While we do *not* expect papers in venues like ACL to be fully deployable in clinical settings, careful evaluation is essential to responsibly translate research insights toward real-world mental health impact. Our goal is to raise evaluation standards as much as possible so that research outputs can earn the trust and approval of domain experts, even when the tools are not yet – or are not intended to be – used in actual clinical practice.

These limitations are not idiosyncratic model bugs, but symptoms of an underlying disconnect between the communities that build, use, and regulate AI for mental health tools. Current evaluations often default to technical benchmark wins, while clinicians and other users judge success by changes in symptoms, patient functioning, and safety over time; social and implementation scientists, in turn, ask whether a tool fits workflows, earns trust, and reaches people equitably. Without a shared evaluative language, results travel poorly across these communities: automated scores with-

[1] WHO 2025 report.

out clinical anchors may overstate progress, "human studies" may lack meaningful involvement as well as methodological transparency or expert input, and cross-disciplinary collaboration may arrive late – if at all. What is needed is a common, clinically grounded evaluation framework that makes psychometric constructs accessible to AI researchers, pairs them with human-centered and implementation-science measures, and treats safety, equity, and real-world utility as primary outcomes. This framework can then be the connective tissue that enables mutual intelligibility and, ultimately, responsible deployment across research contexts, clinics, and community platforms.

Consequently, we posit a fundamental reconsideration of evaluation for AI mental health tools according to clinical goals, typically falling into three broad types: **(1) assessment** for inferring psychological states (e.g., language-based screening), **(2) interventions** to deliver or scaffold support (e.g., therapeutic chatbots), and **(3) information synthesis** to aid practitioners or researchers (e.g., clinical summarization). This categorization clarifies how different types of tools require context-sensitive evaluation and enables the field to calibrate what claims are supported by existing evaluations.

Contributions. Our paper makes four primary contributions. (1) We identify key gaps and challenges in current evaluation practices for AI in mental health (§ 2; see Appendix A for details of surveyed papers); (2) we propose a structured taxonomy of tool types and salient evaluation dimensions, highlighting differences between general generative AI evaluation and mental health-specific concerns (§ 3); (3) we demonstrate its utility through five illustrative case studies spanning assessment, intervention, and support tools in diverse settings (§ 4); and (4), we synthesize these insights into recommendations and guiding principles for responsible and comprehensive evaluation moving forward (§ 5).

Positionality. Our call for rethinking evaluation aligns with broader reflections on the generative AI evaluation crisis in the CL community (Bommasani, 2023; Elangovan et al., 2024; Kotonya and Toni, 2024; Zhou et al., 2025b), as well as work framing generative AI evaluation as a social science measurement challenge, emphasizing rigor in construct definition and validity and proposing frameworks that connect abstract evaluation goals to concrete measurement practices (Wallach et al., 2025). While these papers focus on general-purpose generative AI, we target the clinical, ethical, and implementation challenges in mental health.

Observed practice	**%**
Rely only on AI/NLP metrics	50
No human evaluation	52
With human evaluation but no experts	29
Evaluation guidelines not shared	17
Limitations in evaluation not discussed	36

Table 1: Overview of the ACL Anthology study conducted to ground our position. We queried the ACL Anthology database with mental health keywords [2], restricting results to the past five years and papers of types "main" or "findings". This yielded 135 papers on mental health [3]. These manually-made observations provide context for our broader discussion of challenges and gaps in the evaluation of AI tools for mental health. We show details about the surveyed papers in Appendix A; Tables 3–17.

Recent surveys on LLMs in psychotherapy (Na et al., 2025), cognitive distortion detection (Sage et al., 2025), and mental health conversational agents (Atapattu et al., 2025) primarily catalog tasks, datasets, and model capabilities, rather than providing normative guidance for responsible evaluation. Unlike Wang et al. (2025a), who review papers from 2023–2024 across medical and engineering databases to assess LLMs' clinician-like capabilities, our work surveys recent NLP research and proposes a normative, interdisciplinary evaluation framework grounded in psychometrics and clinical science. Zhang et al. (2025c) focus on evaluating the effectiveness of generative AI chatbots through systematic review and meta-analysis, while our framework covers a broader range of AI applications in mental health, including assessment, intervention, and information synthesis. Flathers et al. (2025) propose a clinician-focused, tripartite benchmarking approach emphasizing technical safety, clinical knowledge, and reasoning, whereas our work systematically analyzes NLP research and develops a theory-grounded framework integrating psychometrics, clinical science, implementation, equity, and user experience to guide research evaluation rather than immediate clinical benchmarking.

[2] *mental health*, *mental disorder*, *mental illness*, *therapy* and *psychiatry*; Either in the title or in the abstract.

[3] After manual inspection to remove papers that mention mental health *only in passing* but not as the main focus; We had 152 papers before the inspection.

2 Observed Practices

To ground our position, we conducted a quantitative analysis of 135 papers on mental health, published in the ACL Anthology [4] over the past 5 years, with 36% of them published in 2025. Table 1 summarizes key patterns that emerged from this review, and Appendix A provides detailed annotations, including the tasks covered by these papers, and the observed practices documented at a paper level. Overall, we found that current evaluation practices in this literature remain limited in scope and rigor, especially considering the sensitivity and clinical implications of the domain. While the surveyed works cover a wide range of tasks, from detecting mental health conditions (Chen et al., 2024b; Yang et al., 2021; Lee et al., 2024a) to building therapeutic chatbots (Saha et al., 2022; Deng et al., 2023; Shim, 2021), their evaluations often rely on narrow, model-centric criteria.

Specifically, five concerning patterns emerge. Half of the papers rely *only* on standard AI/NLP metrics such as accuracy, F1, BLEU, or ROUGE, ignoring psychological validity or clinical relevance. Over half (54%) include *no* human evaluation, and among those that do, 29% do so *without* involving mental health experts. Nearly one-fifth of papers omit evaluation guidelines, and roughly a third fail to discuss limitations in the way the evaluations have been conducted. These gaps indicate that current practices assess technical performance but often overlook safety, interpretability, and real-world utility (Thieme et al., 2020).

Overall, these findings reveal a methodological gap: AI tools may score well on generic NLG metrics yet fall short of clinical standards or user needs. This critique is not aimed at individual works, but rather, highlights the need for shared, rigorous evaluation practices. The following sections build on these observations to introduce a taxonomy (§ 3), illustrate it with case studies (§ 4), and present guiding principles for a clinically grounded and human-centered evaluation (§ 5).

3 Proposed Taxonomy

While new principles are needed to evaluate AI for mental health, there is much to build on from a century of work in psychological assessment (*classical quantitative methods* (Cook and Beckman, 2006)) and recent advances in applying technology in human-computer interaction and health (*implementation science* (Lyon et al., 2023)).

Classical quantitative methods. *Validity* and *reliability* are foundational in psychological evaluation[5]. Validity asks whether a tool *does what it is intended to do*, while reliability asks *whether it does so consistently*. Most current evaluations in AI mental health work mainly focuses on one validity subtype, namely construct validity, for example, through agreement with human annotations or existing scales (Park et al., 2020; Lee et al., 2024a), but this is only a starting point for high-stakes applications. A classifier may correlate with overall depression severity yet fail to predict specific symptoms or generalize across populations. Similarly, a summarization tool may align with expert summaries but omit safety-critical information or misinterpret non-clinical expressions, thus highlighting limits in discriminant validity and generalization. Near-perfect construct validity is not always desirable, as even established assessments have limitations.

Implementation science. Recent advances in health informatics and human-computer interaction highlight that barriers to using AI go beyond validity and reliability (Reddy, 2024). Implementation science adds two pillars: *implementation*–whether an AI tool is feasible, acceptable, fits workflows, and improves outcomes safely; and *maintenance*–whether it remains effective over time, handling population shifts, language drift, inequities, or unintended consequences. Together with validity and reliability, these four pillars define a multidimensional evaluation space for AI in mental health across assessment, interventions, and information synthesis (i.e., therapist support).

To organize these concepts, we introduce a taxonomy of evaluation dimensions (Table 2[6]), mapping classical psychometrics and implementation science principles onto three common AI applications: assessment, intervention, and information synthesis. These evaluation paradigms are multifaceted; no single score can capture the full opportunities and risks of AI, akin to a cockpit dashboard where multiple readings are needed to assess performance.

[4] `https://aclanthology.org/`; query date: 11-2025.

[5] Evidenced by their inclusion in nearly every modern textbook on psychological research methods (Cohen et al., 1988; Reynolds and Livingston, 2021; Meyer, 2010)

[6] While our focus is on clinical integration, this taxonomy is intended to also cover peer-supported and community-based AI mental health tools.

Support type	**Quality Criteria**		**Real-World Use**	
	Validity *Does it do what it is intended?*	**Reliability** *Does it do the same thing under different conditions?*	**Implementation** *Can it be used effectively in real-world contexts?*	**Maintenance** *Does it remain effective and appropriate over time as users and contexts evolve?*
Assessment (e.g., language-based screening)	**1. Construct Validity:** How much does it match other tools or indicators (e.g., clinical, community, or self-report measures) intended to assess the same construct (convergent) or a different construct (discriminant)? **2. Criterion Validity:** What is its association with external, theoretically-related constructs or outcomes (e.g., wellbeing, functioning, participation)?	**1. Across Time:** What is the test-retest stability (at appropriate time intervals)? Does it [not] change if it should [not]? **2. Across Populations:** Does it work just as well across different cultures, locations, neurodivergent populations? **3. Internal Consistency:** To what extent do all components or interactions of the tool function consistently?	**1. Feasibility:** Does it fit into the workflows and routines of intended users (e.g., clinicians, peer supporters, or individuals)? **2. Effectiveness and Usefulness** (extrinsic): Is it consistent across diverse populations? Does it improve diagnostic accuracy in practice? **3. Acceptability:** Are data gathering and feedback mechanisms for assessment acceptable to both patients and clinicians?	**1. Generalizability and Impact:** Does performance remain stable as users or contexts evolve over time? Does it contribute to improved individual or population-level outcomes? **2. Unintended Consequences:** Is it creating labeling bias?
Intervention (e.g., therapeutic chatbots)	**1. Construct Validity:** Does it make a change in the intended direction (convergent) or have any adverse or unintended effects (discriminant)? From experimentation, RCTs, or real-world trials (efficacy or effectiveness). **2. Criterion Validity:** Does it predict or improve external downstream outcomes (e.g., wellbeing, functioning, relationships, work, community participation)?	**1. Across Time:** Does it keep working as well at future points in time? **2. Across Populations:** Is the effect the same across cultures, locations, neurodivergence? **3. Internal consistency:** If the intervention has multiple mechanisms or components, do they each contribute consistently to desired outcomes?	**1. Effectiveness:** Does it improve symptoms, wellbeing, or functioning under real-world conditions (with or without clinician involvement)? **2. Usability and Engagement:** Do users adhere to the intervention? Is it easy to use? **3. Implementation Risk:** Is it being used as intended? **4. Equity and Acceptability:** Do diverse user groups find it acceptable and trustworthy? Are potential biases mitigated?	**1. Stability:** Does the benefit sustain over time across different user groups and contexts? Are there equitable outcomes and access? **2. Safety:** Are there emergent risks or harmful use patterns?
Information synthesis (e.g., clinical summarization)	**1. Construct Validity:** Does it provide accurate, contextually appropriate, and unbiased summaries or recommendations? **2. Criterion Validity:** Does it save users (e.g., clinicians or peer supporters) time or improve the quality of their decisions?	**1. Scenarios:** Does it perform reliably across different use scenarios? **2. Services:** Does it integrate effectively across different service models or modalities?	**1. Acceptability:** Would intended users (clinicians, patients, peer supporters, community workers) accept and trust the tool in their workflows or daily lives? **2. Usefulness:** Do the users find it useful in their everyday work or well-being activities? **3. Impact:** Does the support improve outcomes for users or beneficiaries (e.g., efficiency, understanding, well-being)? **4. Equity and Bias Mitigation:** Are there systematic biases in recommendations or summaries? Are they identified and mitigated?	**1. Tool-level Impact:** Does it reduce administrative load, emotional burden, or improve care and support quality across settings? **2. Unintended Consequences:** Does it foster over-reliance or skill atrophy?

Table 2: Taxonomy for evaluation of AI in mental health applications: aligning support types with validity, reliability, implementation, and maintenance across various contexts.

Assessments involve tools for measurement, screening, aiding diagnosis, or forecasting (e.g., scoring depression severity, estimating suicide risk from social media, classifying psychosis-related language). Validity focuses on convergent validity (alignment with other measures of the same construct), discriminant validity (avoiding spurious alignment with different constructs), and criterion validity (relation to meaningful external outcomes like hospitalization or symptom trajectories). Reliability covers stability over time (test-retest), robustness across populations (clinics, demographics, cultures, neurodivergent groups), and internal consistency (coherent subcomponents). Implementation examines feasibility, impact on diagnostic accuracy, equity, acceptability, and bias mitigation. Maintenance involves monitoring generalizability, performance drift, population-level outcomes, unintended consequences, and evolving language norms.

Interventions are tools aimed at changing outcomes, such as treatment agents, self-help aids, prevention nudges, or adaptive therapy recommendations. Validity includes construct validity (delivering the intended therapeutic ingredient), efficacy (producing beneficial change and avoiding harm), and criterion validity (predicting improvements in functioning, relationships, or job stability). Reliability examines whether effects hold across time, populations, settings, and intervention components. Implementation considers real-world symptom improvement, user engagement, clinician usability, low risk, and monitoring off-label use. Maintenance evaluates persistence of benefits and emergence of new risks, such as overuse or avoidance of human care.

Information synthesis tools augment care and administration efficiently. For automated care aids (clinical summarization, triage notes, treatment recommendations), convergent validity asks whether outputs are accurate as per the clinical evidence base, while criterion validity asks whether they save clinician time or improve documentation. Reliability emphasizes reproducibility across scenarios (note types, specialties) and modalities (telehealth vs. in-person, EHR variants). Implementation focuses on acceptability, usefulness in daily work, and patient impact. Maintenance considers tool-level effects, like reduced burnout or unintended consequences (over-reliance, skill atrophy).

In our evaluation framework, we prioritize these evaluation dimensions because they draw from long-standing clinical science (validity and reliability) and real-world mental health technology evaluation (implementation and maintenance), together defining the minimum requirements for responsible use in high-stakes mental health contexts.

4 Case Studies

The following five case studies were selected to illustrate the taxonomy across support types. They were chosen for their representativeness, methodological rigor, and the variety of AI approaches they exemplify, enabling a comprehensive demonstration of the taxonomy's dimensions: *validity*, *reliability*, *implementation*, and *maintenance*.

4.1 Study I (*Assessment*): LLM rating scales for psychometric assessment of patient engagement

Eberhardt et al. (2025) introduced the LLM rating scale, a psychometric tool for automatically transcribed psychotherapy sessions that measures latent psychological constructs, such as patient engagement, by applying traditional psychometric principles to AI-based assessment. The scale uses structured items–prompts like "*Please rate how motivated the patient is to engage in therapy on a scale from 0 to 100*"–to elicit zero-shot judgments from the model. The study analyzed 1,131 sessions from 155 patients using the DISCOVER framework (Hallmen et al., 2025), computing mean scale scores from a large pool of manually developed items, which were then evaluated for reliability and multiple forms of validity.

Validity was assessed across multiple dimensions. Construct validity was supported by moderate, significant correlations between LLM rating scale scores and engagement determinants like therapy motivation and between-session effort (Holdsworth et al., 2014). Criterion validity was shown through associations with subsequent therapy outcomes, where higher engagement predicted greater symptom improvement. Structural validity was evaluated via multilevel confirmatory factor analysis modeling a single latent factor, with good fit (CFI $= 0.968$, SRMR $= 0.022$) though RMSEA $= 0.108$ indicated some unexplained variance. Reliability was examined as the consistency of the measurement across items, with internal consistency (McDonald's $\omega = 0.953$) showing coherent and stable LLM responses.

The study demonstrated the psychometric soundness and potential of the LLM rating scale as an automated tool for psychotherapy research and feedback. Future work should extend analyses across time and populations, assess robustness, fairness, and safety (Lutz et al., 2024; Ryan et al., 2025), and validate across contexts, languages, and constructs. Implementation considerations, including presentation and integration (e.g., via XAI (Lavelle-Hill et al., 2025)), affect real-world usefulness. Testing within systems like the Trier Treatment Navigator (Lutz et al., 2024, 2025) can evaluate clinical integration and early detection potential. Ongoing maintenance is needed to monitor drift, bias, and improvements with newer LLMs.

4.2 Study II (*Assessment*): Natural language response formats for assessing depression and worry

Gu et al. (2025) conducted a validity- and reliability-based comparison of response formats for LLM-based assessment of depression and worry, building on prior work showing AI language assessments can approach the reliability of human scales (Kjell et al., 2022). The study compared four response formats, from closed to open (predefined words, descriptive words, short phrases, full-text responses), using a Sequential Evaluation with Model Pre-Registration (SEMP) design. Models were trained on a development set ($N = 963$) and pre-registered before evaluation on a prospective test set ($N = 145$) with validated scales, including the PHQ-9 (Spitzer et al., 1999) and GAD-7 (Spitzer et al., 2006). The results showed strong convergent validity across formats, with correlations of $r = .60$-$.79$, exceeding the pre-registered threshold ($r > .50$). Combining all eight depression and worry models yielded correlations near or above scale reliability limits (e.g., $r = .83$ for CES-D vs. reliability $r = .78$), and incremental validity analyses showed improved accuracy consistent with cognitive interview theory. However, high inter-correlations among combined models ($r = .88$-$.95$) indicated reduced discriminant validity when the same responses were used to assess both constructs.

Regarding reliability and implementation, two-week test-retest correlations showed moderate to strong stability, with performance generalizing well to unseen data and prospective accuracies matching cross-validated estimates. Open-ended formats showed internal consistency at the word level, with depression- (e.g., "blue") and worry-related (e.g., "anxious") terms aligning with DSM-5 symptom clusters (Diagnostic, 2013). Implementation effectiveness was demonstrated by predicting behavioral indicators such as sick leave and mental health-related healthcare visits, often matching or exceeding standard rating scales. Feasibility analyses showed that open formats provided richer information (Shannon diversity up to 561.0) but required longer completion times (up to 4 times slower than select-word tasks).

Overall, within the proposed taxonomy, this work shows strong convergent, criterion, and external validity and temporal reliability, indicating that well-designed LLM-based assessments can rival traditional measures. However, discriminant validity and workflow feasibility remain open questions, motivating future work on cross-population reliability and integration into digital clinical platforms.

4.3 Study III (*Intervention*): Evaluating the capabilities of LLMs vs. human therapists to generate personalized interventions

Bar-Shachar et al. (2025) developed an LLM-based tool for generating context-sensitive therapeutic interventions during psychotherapy sessions. It uses four specialized LLM agents, supportive, directive, exploratory, and meaning-making, along with a Judge-LLM that selects the most appropriate intervention based on the dialogue and the patient's emotional and cognitive state. This setup reflects how clinicians choose among multiple interventions and tailor responses to patients' evolving needs.

They evaluated the tool via human-AI comparisons on transcribed therapy segments, with both therapists and the AI generating interventions that expert clinicians rated for theoretical appropriateness, contextual fit, and helpfulness. High inter-rater reliability (ICC and Cohen's κ) supported robustness, and AI interventions were generally clinically relevant and sometimes approached human quality, though they lacked the depth and personalization of experienced clinicians.

Applying the taxonomy shows strong construct validity and reliability, supported by theoretical grounding and high rater agreement. However, ecological validity across therapies, languages, and contexts, fairness across groups, and key implementation and maintenance issues, such as feasibility, clinician acceptance, ethical oversight, model stability, and unintended effects, were not addressed.

Using the taxonomy, future evaluations could go beyond expert ratings by testing validity across modalities, populations, and contexts, examining reliability across models and raters, and focusing on implementation through usability, clinician-patient co-creation, and ethical integration. Ongoing maintenance would monitor drift, bias, and unintended effects, including impacts on novice therapists, ensuring the tool remains theoretically sound, reliable, and clinically sustainable.

4.4 Study IV (*Intervention*): A clinically-grounded framework for evaluating LM-assisted cognitive restructuring

Sharma et al. (2023, 2024) conducted a multi-stage project to design, deploy, as well as evaluate a human-LM interaction tool for cognitive restructuring, a core Cognitive Behavioral Therapy (CBT) technique. Across all stages, the project integrated clinical validity, ecological evaluation, safety, and equity considerations.

The first stage defined and validated clinically meaningful AI objectives. Working with mental health professionals, the authors developed 7 linguistic attributes for reframing, including empathy, positivity, actionability, specificity, and addressing thinking traps. To ensure clinical validity, 600 reframes were collected and annotated by practitioners. A randomized field study ($N = 2{,}067$) on the Mental Health America platform showed users preferred highly empathic and specific reframes, while overly positive ones were less effective.

Under the implementation dimension, the framework was operationalized into an interactive LM-powered tool supporting users in cognitive restructuring. Co-designed with mental health professionals, it included safety mechanisms such as classification and rule-based filtering, IRB approval, and a user-reporting function, with flagged content (0.65%) confirming filter effectiveness. A large-scale field study on the MHA website ($N = 15{,}531$) evaluated user-reported outcomes, including emotional impact, therapeutic utility, and skill acquisition. The tool showed measurable benefits, with the majority of participants reporting reduced negative emotion and helpfulness of reframes for overcoming negative thoughts.

Under maintenance, the third stage assessed equity and found reduced effectiveness for adolescents aged 13-17. Targeted adaptations (simpler, more casual reframes) improved helpfulness in a follow-up trial without affecting other groups, demonstrating ongoing monitoring and refinement.

This case study shows how a clinically grounded, real-world evaluation framework centered on safety and equity can produce a tool with measurable utility. From a taxonomy perspective, it shows validity (clinically aligned outcomes), reliability (consistent effects), safety (content filtering and user flagging), equity (targeted improvements), and maintenance (iterative refinement). The tool has since been deployed by Mental Health America, serving over 160,000 users [7] .

4.5 Study V (*Information synthesis*): Hierarchical LLM-VAE tool for clinically meaningful timeline summarization

Song et al. (2024) proposed a hybrid tool that integrates hierarchical variational autoencoders (TH-VAEs) with LLMs to generate clinically meaningful summaries of long-term social media timelines. It produces two layers: a first-person evidence summary capturing subjective experiences, and a third-person clinical summary mapping these experiences to diagnostic indicators, interpersonal patterns, and moments of change. The goal is to help clinicians and researchers synthesize key information from longitudinal mental health data.

Evaluation of the tool integrated both automatic and expert-based components. Automatic metrics assessed meaning preservation, factual consistency, evidence appropriateness, coherence, and fluency. Clinical experts rated summaries for usefulness, diagnostic accuracy, and their ability to reflect dynamic psychological processes. Inter-rater agreement ensured the reliability of human judgments, and ablation studies tested the contribution of specific model components, such as keyphrase extraction and expert-informed prompting.

Applying the proposed taxonomy shows that the work addresses construct validity–alignment with clinical constructs–and criterion validity through correlations with expert judgments. It also touches on reliability via inter-rater agreement. However, because the tool was trained and evaluated only on social media data, ecological validity and clinical generalizability is limited. Although the peer-support platform provides authentic language, selective self-disclosure gives the model only a partial view of users' psychological states. The authors also acknowledge risks such as hallucinations, bias, and unsafe inferences, but do not systematically evaluate them, nor do they assess fairness across demographic or linguistic groups. Implementation

[7] https://screening.mhanational.org/changing-thoughts-with-an-ai-assistant/

and maintenance factors–such as clinical usability, practitioner acceptance, and long-term model stability–were likewise not examined.

Future work could extend evaluation to generalizability, usability, and sustainability. Ecological testing across cultures and clinical contexts, repeated assessments for reliability, clinician-focused implementation studies, and ongoing monitoring for drift or bias would support consistent performance, advancing the tool from proof of concept to a clinically robust, ethical, and sustainable mental health tool.

5 Moving Forward

Evaluation foundations and maturity pathways. Evaluation practices in AI for mental health remain concentrated in early-stage technical validation, with relatively few tools reaching implementation or maintenance. While this is typical for an emerging field, it motivates the need for explicit *minimum evaluation standards* appropriate for high-risk mental health contexts. Assessment tools should demonstrate convergent and discriminant validity with clinical constructs. Intervention tools should provide evidence of therapeutic benefit, safety, and acceptability, ideally supported by prospective or randomized evaluations (Hofmann and Weinberger, 2013; Cuijpers et al., 2019). Information synthesis tools should document measurable improvements in workflow, decision quality, or clinical comprehension.

A robust evaluation strategy requires a multilayered, standardized pipeline, in which evaluation depth increases with a tool's intended role and potential harm. We distinguish three maturity layers:

1. **Early maturity (exploratory):** Technical validation, including accuracy, robustness, and agreement with human annotations, typically using retrospective datasets. At this stage, evaluation supports feasibility assessment and hypothesis generation rather than clinical claims.

2. **Intermediate maturity (validation):** Human-centered evaluation, capturing expert judgment, usability, acceptability, and perceived clinical relevance, often through prospective or external validation and structured user studies.

3. **Advanced maturity (deployment):** Assessment of contextual and ecological characteristics, including workflow integration, feasibility across settings, long-term impact, equity, safety, and monitoring of failure modes over time.

This layered structure is particularly important in mental health settings, where concerns about invasiveness, reduced human oversight, and potential clinician deskilling are longstanding (Torous et al., 2019).

Safety, fairness, and adaptability. Safety and fairness require proactive, domain-specific protocols rather than retrospective checks. Because mental health involves power asymmetries and heightened risks of harm, AI support must be systematically stress-tested for hallucinations, inappropriate reassurance, and biased outputs. Fairness assessments should examine performance across demographic, cultural, and linguistic groups, acknowledging that fairness definitions entail unavoidable trade-offs (Kleinberg et al., 2016; Ryan et al., 2025). Notably, most of our case studies lacked explicit safety or fairness evaluations, highlighting a significant gap for future development. Recent advances in mental health science impose additional requirements for adaptability. Clinical theory is shifting from categorical diagnoses toward dimensional and dynamic frameworks, including network-based and dynamic-systems models that conceptualize mental health states as evolving systems of interacting components (Borsboom, 2017; Scheffer et al., 2024; Ong et al., 2025). AI support must therefore remain adaptable to evolving constructs and evidence, as theoretical advances directly shape evaluation targets, risk assessment, and patient safety.

Practical implications across maturity stages. We emphasize that the proposed taxonomy is maturity-aware rather than a uniform checklist: early exploratory systems are not expected to satisfy deployment-level criteria such as Maintenance or full Implementation. Instead, the framework clarifies which dimensions remain unaddressed and how evaluation expectations should scale with system claims and intended use. For researchers without clinical access or deployment resources, higher-level concerns can be partially approximated through structured patient simulations, scenario-based evaluations grounded in clinical guidelines, bias audits across demographic personas, expert-informed annotation protocols targeting construct validity, and rubric-based LLM-as-a-Judge assessments aligned with clinically meaningful criteria. While such technical proxies are not substitutes for real-world validation, they enable early-stage work to engage more explicitly with safety, validity, equity, and implementation considerations, and to calibrate claims appropriately to system maturity.

6 Conclusion

Current evaluation practices for AI in mental health are fragmented and often misaligned with clinical, social, and user-centered needs. By adopting an interdisciplinary framework and a taxonomy of assessment-, intervention-, and information synthesis-oriented tools, responsible evaluation can ensure AI is clinically meaningful, ethically grounded, and more likely to produce real-world impact in mental health care.

Limitations

This paper proposes a taxonomy and accompanying evaluation framework for mental health AI, but several boundaries of scope should be noted. The analysis is informed by a set of published case studies, which may not fully represent the breadth of ongoing work or emerging AI for mental health tools. The taxonomy and evaluation pathways are conceptual rather than empirically validated, and their applicability may vary across clinical, cultural, and linguistic contexts. Additionally, while we outline key evaluation principles, we do not provide detailed operational metrics, leaving room for future work to refine and adapt these ideas as the field continues to develop.

Acknowledgments

The authors acknowledge the support of Schloss Dagstuhl – Leibniz Center for Informatics through the Dagstuhl Seminar '25361: Natural Language Processing for Mental Health'.

This work was supported by the DYNAMIC Center, funded through the LOEWE program of the Hessian Ministry of Science and Arts (Grant No. LOEWE/1/16/519/03/09.001(0009)/98), and by the LOEWE Distinguished Chair Ubiquitous Knowledge Processing, LOEWE initiative, Hesse, Germany (Grant No. LOEWE/4a//519/05/00.002(0002)/81). Additional support was provided in part by the U.S. Centers for Disease Control and Prevention/National Institute for Occupational Safety and Health (CDC/NIOSH) under Grant U01 OH012476, and by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG; Grant Nos. 493169211 and 525286173). This work was also partly supported by the National Institutes of Health through Grants P50MH115838, R01MH117172, and R01MH135488, as well as by the American Foundation for Suicide Prevention and the Betty and Gordon Moore Foundation. Further support was provided by the Excellence Cluster EXC3066 The Adaptive Mind, the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 949944, INTEGRATOR), and the MUR FARE 2020 initiative (Grant Agreement Prot. R20YSMBZ8S, INDOMITA). One author is a member of the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis (BIDSA). This work was also supported by Keystone grant funding from Responsible AI UK (Grant No. EP/Y009800/1). This work was also supported in part by the Research Ireland Adapt Research Centre (grant 13/RC/2106_P2).

References

Nuredin Ali Abdelkadir, Charles Zhang, Ned Mayo, and Stevie Chancellor. 2024. Diverse perspectives, divergent models: Cross-cultural evaluation of depression detection on Twitter. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 672–680, Mexico City, Mexico. Association for Computational Linguistics.

Aakash Kumar Agarwal, Saprativa Bhattacharjee, Mauli Rastogi, Jemima S. Jacob, Biplab Banerjee, Rashmi Gupta, and Pushpak Bhattacharyya. 2025. ReDepress: A cognitive framework for detecting depression relapse from social media. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34652–34670, Suzhou, China. Association for Computational Linguistics.

Elham Aghakhani, Lu Wang, Karla T. Washington, George Demiris, Jina Huh-Yoo, and Rezvaneh Rezapour. 2025. From conversation to automation: Leveraging LLMs for problem-solving therapy analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25189–25207, Vienna, Austria. Association for Computational Linguistics.

Carlos Aguirre, Keith Harrigian, and Mark Dredze. 2021. Gender and racial fairness in depression research using social media. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2932–2949, Online. Association for Computational Linguistics.

Ankit Aich, Avery Quynh, Varsha Badal, Amy Pinkham, Philip Harvey, Colin Depp, and Natalie Parde. 2022. Towards intelligent clinically-informed language analyses of people with bipolar disorder and schizophrenia. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2871–2887, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mario Ezra Aragón, A. Pastor López-Monroy, Luis C. González, David E. Losada, and Manuel Montes-y Gómez. 2023. DisorBERT: A double domain adaptation model for detecting signs of mental disorders in social media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15305–15318, Toronto, Canada. Association for Computational Linguistics.

Thushari Atapattu, Menasha Thilakaratne, Duc Nhan Do, Mahen Herath, and Katrina E. Falkner. 2025. Exploring the role of mental health conversational agents in training medical students and professionals: A systematic literature review. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20785–20798, Vienna, Austria. Association for Computational Linguistics.

Simone Balloccu, Ehud Reiter, Karen Jia-Hui Li, Rafael Sargsyan, Vivek Kumar, Diego Reforgiato, Daniele Riboni, and Ondrej Dusek. 2024. Ask the experts: sourcing a high-quality nutrition counseling dataset through human-AI collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11519–11545, Miami, Florida, USA. Association for Computational Linguistics.

Yael Bar-Shachar, Dana Rafael, Anmol Goel, Ayal Klein, Iryna Gurevych, and Dana Atzil-Slonim. 2025. Evaluating the capabilities of large language models (llms) versus human therapists to generate personalized interventions. Preprint on the Open Science Framework.

Guanqun Bi, Zhuang Chen, Zhoufu Liu, Hongkai Wang, Xiyao Xiao, Yuqiang Xie, Wen Zhang, Yongkang Huang, Yuxuan Chen, Libiao Peng, and Minlie Huang. 2025. MAGI: Multi-agent guided interview for psychiatric assessment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24898–24921, Vienna, Austria. Association for Computational Linguistics.

Suhas Bn, Yash Mahajan, Dominik O. Mattioli, Andrew M. Sherrill, Rosa I. Arriaga, Christopher Wiese, and Saeed Abdullah. 2025a. The pursuit of empathy: Evaluating small language models for PTSD dialogue support. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30888–30910, Suzhou, China. Association for Computational Linguistics.

Suhas Bn, Dominik O. Mattioli, Andrew M. Sherrill, Rosa I. Arriaga, Christopher Wiese, and Saeed Abdullah. 2025b. How real are synthetic therapy conversations? evaluating fidelity in prolonged exposure dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20986–20995, Suzhou, China. Association for Computational Linguistics.

Rishi Bommasani. 2023. Evaluation for change. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8227–8239, Toronto, Canada. Association for Computational Linguistics.

Denny Borsboom. 2017. A network theory of mental disorders. *World psychiatry*, 16(1):5–13.

Layla Bouzoubaa, Elham Aghakhani, Max Song, Quang Trinh, and Shadi Rezapour. 2024. 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.

Greg Buda, Ignacio J. Tripodi, Margaret Meagher, and Elizabeth A. Olson. 2024. Crisis counselor language and perceived genuine concern in crisis conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7149–7160, Miami, Florida, USA. Association for Computational Linguistics.

Mohit Chandra, Siddharth Sriraman, Gaurav Verma, Harneet Singh Khanuja, Jose Suarez Campayo, Zihang Li, Michael L. Birnbaum, and Munmun De Choudhury. 2025. Lived experience not found: LLMs struggle to align with experts on addressing adverse drug reactions from psychiatric medication use. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11083–11113, Albuquerque, New Mexico. Association for Computational Linguistics.

Alicja Chaszczewicz, Raj Shah, Ryan Louie, Bruce Arnow, Robert Kraut, and Diyi Yang. 2024. Multi-level feedback generation with large language models for empowering novice peer counselors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4130–4161, Bangkok, Thailand. Association for Computational Linguistics.

Mingyu Chen, Jingkai Lin, Zhaojie Chu, Xiaofen Xing, Yirong Chen, and Xiangmin Xu. 2025a. CATCH: A novel data synthesis framework for high therapy fidelity and memory-driven planning chain of thought in AI counseling. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10254–10286, Suzhou, China. Association for Computational Linguistics.

Siyuan Chen, Meilin Wang, Minghao Lv, Zhiling Zhang, Juqianqian Juqianqian, Dejiyangla Dejiyangla, Yujia

Peng, Kenny Zhu, and Mengyue Wu. 2024a. Mapping long-term causalities in psychiatric symptomatology and life events from social media. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5472–5487, Mexico City, Mexico. Association for Computational Linguistics.

Siyuan Chen, Zhiling Zhang, Mengyue Wu, and Kenny Zhu. 2023a. Detection of multiple mental disorders from social media with two-stream psychiatric experts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9071–9084, Singapore. Association for Computational Linguistics.

Yujia Chen, Changsong Li, Yiming Wang, Tianjie Ju, Qingqing Xiao, Nan Zhang, Zifan Kong, Peng Wang, and Binyu Yan. 2025b. MIND: Towards immersive psychological healing with multi-agent inner dialogue. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9380–9413, Suzhou, China. Association for Computational Linguistics.

Zhiyu Chen, Yujie Lu, and William Wang. 2023b. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304, Singapore. Association for Computational Linguistics.

Zhuang Chen, Jiawen Deng, Jinfeng Zhou, Jincenzi Wu, Tieyun Qian, and Minlie Huang. 2024b. Depression detection in clinical interviews with LLM-empowered structural element graph. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8181–8194, Mexico City, Mexico. Association for Computational Linguistics.

Zhuohao Chen, Nikolaos Flemotomos, Zac Imel, David Atkins, and Shrikanth Narayanan. 2022. Leveraging open data and task augmentation to automated behavioral coding of psychotherapy conversations in low-resource scenarios. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5787–5795, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. 2023. PAL: Persona-augmented emotional support conversation generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 535–554, Toronto, Canada. Association for Computational Linguistics.

Ronald Jay Cohen, Pamela Montague, Linda Sue Nathanson, and Mark E Swerdlik. 1988. *Psychological testing: An introduction to tests & measurement*. Mayfield Publishing Co.

David A Cook and Thomas J Beckman. 2006. Current concepts in validity and reliability for psychometric instruments: theory and application. *The American journal of medicine*, 119(2):166–e7.

Pablo Cruz-Gonzalez, Aaron Wan-Jia He, Elly PoPo Lam, Ingrid Man Ching Ng, Mandy Wingman Li, Rangchun Hou, Jackie Ngai-Man Chan, Yuvraj Sahni, Nestor Vinas Guasch, Tiev Miller, and 1 others. 2025. Artificial intelligence in mental health care: a systematic review of diagnosis, monitoring, and intervention applications. *Psychological medicine*, 55:e18.

Pim Cuijpers, Mirjam Reijnders, and Marcus JH Huibers. 2019. The role of common factors in psychotherapy outcomes. *Annual review of clinical psychology*, 15(1):207–231.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, and 1 others. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4079–4095, Toronto, Canada. Association for Computational Linguistics.

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

Steffen T Eberhardt, Antonia Vehlen, Jana Schaffrath, Brian Schwartz, Tobias Baur, Dominik Schiller, Tobias Hallmen, Elisabeth André, and Wolfgang Lutz. 2025. Development and validation of large language model rating scales for automatically transcribed psychological therapy sessions. *Scientific Reports*, 15(1):29541.

Aparna Elangovan, Ling Liu, Lei Xu, Sravan Babu Bodapati, and Dan Roth. 2024. ConSiDERS-the-human evaluation framework: Rethinking human evaluation for generative large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1137–1160, Bangkok, Thailand. Association for Computational Linguistics.

Yi Feng, Jiaqi Wang, Wenxuan Zhang, Zhuang Chen, Shen Yutong, Xiyao Xiao, Minlie Huang, Liping Jing, and Jian Yu. 2025. Reframe your life story: Interactive narrative therapist and innovative moment assessment with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24495–24520, Suzhou, China. Association for Computational Linguistics.

Matthew Flathers, Bridget Dwyer, Eden Rozenblit, and John Torous. 2025. Contextualizing clinical benchmarks: a tripartite approach to evaluating llm-based tools in mental health settings. *Journal of Psychiatric Practice®*, 31(6):294–301.

Saadia Gabriel, Isha Puri, Xuhai Xu, Matteo Malgaroli, and Marzyeh Ghassemi. 2024. Can AI relate: Testing large language model response for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2206–2221, Miami, Florida, USA. Association for Computational Linguistics.

Muskan Garg, Amirmohammad Shahbandegan, Amrit Chadha, and Vijay Mago. 2023. An annotated dataset for explainable interpersonal risk factors of mental disturbance in social media posts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11960–11969, Toronto, Canada. Association for Computational Linguistics.

Bhagesh Gaur, Karan Gupta, Aseem Srivastava, Manish Gupta, and Md Shad Akhtar. 2025. Assess and prompt: A generative RL framework for improving engagement in online mental health communities. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18102–18118, Suzhou, China. Association for Computational Linguistics.

Soumitra Ghosh, Gopendra Vikram Singh, Shambhavi Shambhavi, Sabarna Choudhury, and Asif Ekbal. 2025. Just a scratch: Enhancing LLM capabilities for self-harm detection through intent differentiation and emoji interpretation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27428–27445, Vienna, Austria. Association for Computational Linguistics.

Evangelia Gogoulou, Magnus Boman, Fehmi Ben Abdesslem, Nils Hentati Isacsson, Viktor Kaldo, and Magnus Sahlgren. 2021. Predicting treatment outcome from patient texts:the case of Internet-based cognitive behavioural therapy. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 575–580, Online. Association for Computational Linguistics.

Sujatha Gollapalli, Beng Ang, and See-Kiong Ng. 2023. Identifying Early Maladaptive Schemas from mental health question texts. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11832–11843, Singapore. Association for Computational Linguistics.

Zhuojun Gu, Katarina Kjell, H Andrew Schwartz, and Oscar Kjell. 2025. Natural language response formats for assessing depression and worry with large language models: A sequential evaluation with model pre-registration. *Assessment*, page 10731911251364022.

Tobias Hallmen, Dominik Schiller, Antonia Vehlen, Steffen Eberhardt, Tobias Baur, Daksitha Withanage Don, Wolfgang Lutz, and Elisabeth André. 2025. Discover: a data-driven interactive system for comprehensive observation, visualization, and exploration of human behavior. *Frontiers in Digital Health*, Volume 7 - 2025.

Sarthak Harne, Monjoy Narayan Choudhury, Madhav Rao, T K Srikanth, Seema Mehrotra, Apoorva Vashisht, Aarushi Basu, and Manjit Singh Sodhi. 2024. CASE: Efficient curricular data pre-training for building assistive psychology expert models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15769–15778, Miami, Florida, USA. Association for Computational Linguistics.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do models of mental health based on social media data generalize? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3774–3788, Online. Association for Computational Linguistics.

Kilichbek Haydarov, Youssef Mohamed, Emilio Goldenhersch, Paul OCallaghan, Li-jia Li, and Mohamed Elhoseiny. 2025. Towards AI-assisted psychotherapy: Emotion-guided generative interventions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32724–32743, Suzhou, China. Association for Computational Linguistics.

Amey Hengle, Atharva Kulkarni, Shantanu Deepak Patankar, Madhumitha Chandrasekaran, Sneha D'silva, Jemima S. Jacob, and Rashmi Gupta. 2024. Still not quite there! evaluating large language models for comorbid mental health diagnosis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16698–16721, Miami, Florida, USA. Association for Computational Linguistics.

Anthony Hills, Talia Tseriotou, Xenia Miscouridou, Adam Tsakalidis, and Maria Liakata. 2024. Exciting mood changes: A time-aware hierarchical transformer for change detection modelling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12526–12537, Bangkok, Thailand. Association for Computational Linguistics.

Stefan G Hofmann and Joel Weinberger. 2013. *The art and science of psychotherapy*. Routledge.

Emma Holdsworth, Erica Bowen, Sarah Brown, and Douglas Howat. 2014. Client engagement in psychotherapeutic treatment and associations with client characteristics, therapist characteristics, and treatment factors. *Clinical Psychology Review*, 34(5):428–450.

Simin Hong, Jun Sun, and Hongyang Chen. 2025. Third-person appraisal agent: Simulating human emotional reasoning in text with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23684–23701, Suzhou, China. Association for Computational Linguistics.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Jiyue Jiang, Sheng Wang, Qintong Li, Lingpeng Kong, and Chuan Wu. 2023. A cognitive stimulation dialogue system with multi-source knowledge fusion for elders with cognitive impairment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10628–10640, Toronto, Canada. Association for Computational Linguistics.

Swanie Juhng, Matthew Matero, Vasudha Varadarajan, Johannes Eichstaedt, Adithya V Ganesan, and H. Andrew Schwartz. 2023. Discourse-level representations can improve prediction of degree of anxiety. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1500–1511, Toronto, Canada. Association for Computational Linguistics.

Migyeong Kang, Goun Choi, Hyolim Jeon, Ji Hyun An, Daejin Choi, and Jinyoung Han. 2024. CURE: Context- and uncertainty-aware mental disorder detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17924–17940, Miami, Florida, USA. Association for Computational Linguistics.

Mina Kian, Kaleen Shrestha, Katrin Fischer, Xiaoyuan Zhu, Jonathan Ong, Aryan Trehan, Jessica Wang, Gloria Chang, Séb Arnold, and Maja Mataric. 2025. Using linguistic entrainment to evaluate large language models for use in cognitive behavioral therapy. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7724–7743, Albuquerque, New Mexico. Association for Computational Linguistics.

Hyunjong Kim, Suyeon Lee, Yeongjae Cho, Eunseo Ryu, Yohan Jo, Suran Seong, and Sungzoon Cho. 2025a. KMI: A dataset of Korean motivational interviewing dialogues for psychotherapy. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10803–10828, Albuquerque, New Mexico. Association for Computational Linguistics.

Juhee Kim, Chunghu Mok, Jisun Lee, Hyang Sook Kim, and Yohan Jo. 2025b. Dialogue systems for emotional support via value reinforcement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28733–28766, Vienna, Austria. Association for Computational Linguistics.

Jun Seo Kim and Hye Hyeon Kim. 2025. KoACD: The first Korean adolescent dataset for cognitive distortion analysis via role-switching multi-LLM negotiation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22050–22078, Suzhou, China. Association for Computational Linguistics.

Subin Kim, Hoonrae Kim, Heejin Do, and Gary Lee. 2025c. Multimodal cognitive reframing therapy via multi-hop psychotherapeutic reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4863–4880, Albuquerque, New Mexico. Association for Computational Linguistics.

Subin Kim, Hoonrae Kim, Jihyun Lee, Yejin Jeon, and Gary Lee. 2025d. MIRROR: Multimodal cognitive reframing therapy for rolling with resistance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14851–14880, Suzhou, China. Association for Computational Linguistics.

Oscar NE Kjell, Sverker Sikström, Katarina Kjell, and H Andrew Schwartz. 2022. Natural language analyzed with ai-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific reports*, 12(1):3918.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Neema Kotonya and Francesca Toni. 2024. Towards a framework for evaluating explanations in automated fact verification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16364–16377, Torino, Italia. ELRA and ICCL.

Raja Kumar, Kishan Maharaj, Ashita Saxena, and Pushpak Bhattacharyya. 2024. Mental disorder classification via temporal representation of text. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10901–10916, Miami, Florida, USA. Association for Computational Linguistics.

Gleb Kuzmin, Petr Strepetov, Maksim Stankevich, Natalia Chudova, Artem Shelmanov, and Ivan Smirnov. 2025. Exploring large language models for detecting mental disorders. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34523–34547, Suzhou, China. Association for Computational Linguistics.

Xiaochong Lan, Zhiguang Han, Yiming Cheng, Li Sheng, Jie Feng, Chen Gao, and Yong Li. 2025. Depression detection on social media with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2155–2171.

Rosa Lavelle-Hill, Gavin Smith, Hannah Deininger, and Kou Murayama. 2025. An explainable artificial intelligence handbook for psychologists: Methods, opportunities, and challenges. *Psychological Methods*.

Andrew Lee, Jonathan K. Kummerfeld, Larry An, and Rada Mihalcea. 2021. Micromodels for efficient, explainable, and reusable systems: A case study on mental health. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4257–4272, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daeun Lee, Hyolim Jeon, Sejung Son, Chaewon Park, Ji hyun An, Seungbae Kim, and Jinyoung Han. 2024a. Detecting bipolar disorder from misdiagnosed major depressive disorder with mood-aware multi-task learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4954–4970, Mexico City, Mexico. Association for Computational Linguistics.

Daeun Lee, Soyoung Park, Jiwon Kang, Daejin Choi, and Jinyoung Han. 2020. Cross-lingual suicidal-oriented word embedding toward suicide prevention. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2208–2217, Online. Association for Computational Linguistics.

Gyeongeun Lee, Zhu Wang, Sathya N. Ravi, and Natalie Parde. 2025. From heart to words: Generating empathetic responses via integrated figurative language and semantic context signals. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4490–4502, Vienna, Austria. Association for Computational Linguistics.

Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyong-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024b. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274, Miami, Florida, USA. Association for Computational Linguistics.

Anqi Li, Yu Lu, Nirui Song, Shuai Zhang, Lizhi Ma, and Zhenzhong Lan. 2024. Understanding the therapeutic relationship between counselors and clients in online text-based counseling using LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1280–1303, Miami, Florida, USA. Association for Computational Linguistics.

Anqi Li, Lizhi Ma, Yaling Mei, Hongliang He, Shuai Zhang, Huachuan Qiu, and Zhenzhong Lan. 2023. Understanding client reactions in online mental health counseling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10358–10376, Toronto, Canada. Association for Computational Linguistics.

Tong Li, Shu Yang, Junchao Wu, Jiyao Wei, Lijie Hu, Mengdi Li, Derek F. Wong, Joshua R. Oltmanns, and Di Wang. 2025. Can large language models identify implicit suicidal ideation? an empirical evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18392–18413, Suzhou, China. Association for Computational Linguistics.

Inna Lin, Lucille Njoo, Anjalie Field, Ashish Sharma, Katharina Reinecke, Tim Althoff, and Yulia Tsvetkov. 2022. Gendered mental health stigma in masked language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2152–2170, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Inna Lin, Ashish Sharma, Christopher Rytting, Adam Miner, Jina Suh, and Tim Althoff. 2024. IMBUE: Improving interpersonal effectiveness through simulation and just-in-time feedback with human-language model interaction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 810–840, Bangkok, Thailand. Association for Computational Linguistics.

Shir Lissak, Nitay Calderon, Geva Shenkman, Yaakov Ophir, Eyal Fruchter, Anat Brunstein Klomek, and Roi Reichart. 2024. The colorful future of LLMs: Evaluating and improving LLMs as emotional supporters for queer youth. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2040–2079, Mexico City, Mexico. Association for Computational Linguistics.

Siyang Liu, Bianca Brie, Wenda Li, Laura Biester, Andrew Lee, James Pennebaker, and Rada Mihalcea. 2025. Eeyore: Realistic depression simulation via expert-in-the-loop supervised and preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13750–13770, Vienna, Austria. Association for Computational Linguistics.

Siyang Liu, Naihao Deng, Sahand Sabour, Yilin Jia, Minlie Huang, and Rada Mihalcea. 2023. Task-adaptive tokenization: Enhancing long-form text generation efficacy in mental health and beyond. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15264–15281, Singapore. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create LLM-simulated patients via eliciting and adhering to principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10570–10603, Miami, Florida, USA. Association for Computational Linguistics.

Daniel Lozoya, Alejandro Berazaluce, Juan Perches, Eloy Lúa, Mike Conway, and Simon D'Alfonso. 2024. Generating mental health transcripts with SAPE (Spanish adaptive prompt engineering). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5096–5113, Mexico City, Mexico. Association for Computational Linguistics.

Wolfgang Lutz, Brian Schwartz, Antonia Vehlen, Steffen T Eberhardt, and Jaime Delgadillo. 2025. Advances in personalization of psychological interventions. *World Psychiatry*, 24(3):343.

Wolfgang Lutz, Antonia Vehlen, and Brian Schwartz. 2024. Data-informed psychological therapy, measurement-based care, and precision mental health. *Journal of Consulting and Clinical Psychology*, 92(10):671.

Minghao Lv, Siyuan Chen, Haoan Jin, Minghao Yuan, Qianqian Ju, Yujia Peng, Kenny Q. Zhu, and Mengyue Wu. 2025. Tracking life's ups and downs: Mining life events from social media posts for mental health analysis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6950–6965, Vienna, Austria. Association for Computational Linguistics.

Aaron Lyon, Sean A Munson, Madhu Reddy, Stephen M Schueller, Elena Agapie, Svetlana Yarosh, Alex Dopp, Ulrica von Thiele Schwarz, Gavin Doherty, Andrea K Graham, and 1 others. 2023. Bridging hci and implementation science for innovation adoption and public health impact. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Zafarullah Mahmood, Soliman Ali, Jiading Zhu, Mohamed Abdelwahab, Michelle Yu Collins, Sihan Chen, Yi Cheng Zhao, Jodi Wolff, Osnat C. Melamed, Nadia Minian, Marta Maslej, Carolynne Cooper, Matt Ratto, Peter Selby, and Jonathan Rose. 2025. A fully generative motivational interviewing counsellor chatbot for moving smokers towards the decision to quit. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25008–25043, Vienna, Austria. Association for Computational Linguistics.

Patrick Meyer. 2010. *Understanding measurement: reliability*. Oxford University Press.

Do June Min, Verónica Pérez-Rosas, Kenneth Resnicow, and Rada Mihalcea. 2022. PAIR: Prompt-aware margIn ranking for counselor reflection scoring in motivational interviewing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 148–158, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kshitij Mishra, Priyanshu Priya, Manisha Burja, and Asif Ekbal. 2023a. e-THERAPIST: I suggest you to cultivate a mindset of positivity and nurture uplifting thoughts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13952–13967, Singapore. Association for Computational Linguistics.

Kshitij Mishra, Priyanshu Priya, and Asif Ekbal. 2023b. PAL to lend a helping hand: Towards building an emotion adaptive polite and empathetic counseling conversational agent. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12254–12271, Toronto, Canada. Association for Computational Linguistics.

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.

Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. 2022. Improving the generalizability of depression detection by leveraging clinical questionnaires. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8446–8459, Dublin, Ireland. Association for Computational Linguistics.

Viet Cuong Nguyen, Mohammad Taher, Dongwan Hong, Vinicius Konkolics Possobom, Vibha Thirunellayi Gopalakrishnan, Ekta Raj, Zihang Li, Heather J. Soled, Michael L. Birnbaum, Srijan Kumar, and Munmun De Choudhury. 2025a. Do large language models align with core mental health counseling competencies? In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7488–7511, Albuquerque, New Mexico. Association for Computational Linguistics.

Vivian Nguyen, Sang Min Jung, Lillian Lee, Thomas D. Hull, and Cristian Danescu-Niculescu-Mizil. 2024.

Taking a turn for the better: Conversation redirection throughout the course of mental-health therapy. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9507–9521, Miami, Florida, USA. Association for Computational Linguistics.

Vivian Nguyen, Lillian Lee, and Cristian Danescu-Niculescu-Mizil. 2025b. Hanging in the balance: Pivotal moments in crisis counseling conversations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29801–29817, Vienna, Austria. Association for Computational Linguistics.

Clarissa W. Ong, Hiba Arnaout, Kate Sheehan, Estella Fox, Eugen Owtscharow, and Iryna Gurevych. 2025. Using large language models to create personalized networks from therapy sessions. *Preprint*, arXiv:2512.05836.

Sungjoon Park, Kiwoong Park, Jaimeen Ahn, and Alice Oh. 2020. Suicidal risk detection for military personnel. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2523–2531, Online. Association for Computational Linguistics.

Priyanshu Priya, Gopendra Singh, Mauajama Firdaus, Jyotsna Agrawal, and Asif Ekbal. 2024. On the way to gentle AI counselor: Politeness cause elicitation and intensity tagging in code-mixed Hinglish conversations for social good. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4678–4696, Mexico City, Mexico. Association for Computational Linguistics.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024a. SMILE: Single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 615–636, Miami, Florida, USA. Association for Computational Linguistics.

Huachuan Qiu and Zhenzhong Lan. 2025. PsyDial: A large-scale long-term conversational dataset for mental health support. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21624–21655, Vienna, Austria. Association for Computational Linguistics.

Huachuan Qiu, Lizhi Ma, and Zhenzhong Lan. 2024b. PsyGUARD: An automated system for suicide detection and risk assessment in psychological counseling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4581–4607, Miami, Florida, USA. Association for Computational Linguistics.

Jiahao Qiu, Yinghui He, Xinzhe Juan, Yimin Wang, Yuhan Liu, Zixin Yao, Yue Wu, Xun Jiang, Ling Yang, and Mengdi Wang. 2025a. EmoAgent: Assessing and safeguarding human-AI interaction for mental health safety. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11752–11767, Suzhou, China. Association for Computational Linguistics.

Wenyu Qiu, Yuxiong Wang, Jiajun Tan, Hanchao Hou, Qinda Liu, Wei Yao, and Shiguang Ni. 2025b. DeepWell-adol: A scalable expert-based dialogue corpus for adolescent positive mental health and wellbeing promotion. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12797–12821, Suzhou, China. Association for Computational Linguistics.

Federico Ravenda, Seyed Ali Bahrainian, Andrea Raballo, Antonietta Mira, and Noriko Kando. 2025. Are LLMs effective psychological assessors? leveraging adaptive RAG for interpretable mental health screening through psychometric practice. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8975–8991, Vienna, Austria. Association for Computational Linguistics.

Sandeep Reddy. 2024. Generative ai in healthcare: an implementation science informed translational path on application, integration and governance. *Implementation Science*, 19(1):27.

Maor Reuben, Ortal Slobodin, Idan-Chaim Cohen, Aviad Elyashar, Orna Braun-Lewensohn, Odeya Cohen, and Rami Puzis. 2025. Assessment and manipulation of latent constructs in pre-trained language models using psychometric scales. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2433–2444, Vienna, Austria. Association for Computational Linguistics.

Cecil R Reynolds and RA Livingston. 2021. *Mastering modern psychological testing*. Springer.

Gony Rosenman, Talma Hendler, and Lior Wolf. 2024. LLM questionnaire completion for automatic psychiatric assessment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 403–415, Miami, Florida, USA. Association for Computational Linguistics.

Seamus Ryan, Wanling Cai, Robert Bowman, and Gavin Doherty. 2025. Fairness challenges in the design of machine learning applications for healthcare. *ACM Trans. Comput. Healthcare*, 6(4).

Archie Sage, Jeroen Keppens, and Helen Yannakoudakis. 2025. A survey of cognitive distortion detection and classification in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14884–14899, Suzhou, China. Association for Computational Linguistics.

Tulika Saha, Saichethan Reddy, Anindya Das, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A shoulder to cry on: Towards a motivational virtual assistant for assuaging mental agony. In *Proceedings of the 2022 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, pages 2436–2449, Seattle, United States. Association for Computational Linguistics.

Msvpj Sathvik, Zuhair Hasan Shaik, and Vivek Gupta. 2025. M-help: Using social media data to detect mental health help-seeking signals. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22510–22520, Suzhou, China. Association for Computational Linguistics.

Ramit Sawhney, Harshit Joshi, Lucie Flek, and Rajiv Ratn Shah. 2021a. PHASE: Learning emotional phase-aware representations for suicide ideation detection on social media. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2415–2428, Online. Association for Computational Linguistics.

Ramit Sawhney, Harshit Joshi, Rajiv Ratn Shah, and Lucie Flek. 2021b. Suicide ideation detection via social and temporal user representations using hyperbolic learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2176–2190, Online. Association for Computational Linguistics.

Marten Scheffer, Claudi L Bockting, Denny Borsboom, Roshan Cools, Clara Delecroix, Jessica A Hartmann, Kenneth S Kendler, Ingrid van de Leemput, Han LJ Van Der Maas, Egbert van Nes, and 1 others. 2024. A dynamical systems view of psychiatric disorders—theory: a review. *JAMA psychiatry*, 81(6):618–623.

Raj Sanjay Shah, Lei Xu, Qianchu Liu, Jon Burnsky, Andrew Bertagnolli, and Chaitanya Shivade. 2025. TN-eval: Rubric and evaluation protocols for measuring the quality of behavioral therapy notes. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 179–199, Vienna, Austria. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9977–10000, Toronto, Canada. Association for Computational Linguistics.

Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, Theresa Nguyen, and Tim Althoff. 2024. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–29.

Heereen Shim. 2021. Development of conversational AI for sleep coaching programme. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 121–128, Online. Association for Computational Linguistics.

Jaemin Shin, Hyungjun Yoon, Seungjoo Lee, Sungjoon Park, Yunxin Liu, Jinho Choi, and Sung-Ju Lee. 2023. FedTherapist: Mental health monitoring with user-generated linguistic expressions on smartphones via federated learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11971–11988, Singapore. Association for Computational Linguistics.

Gopendra Vikram Singh, Sai Vardhan Vemulapalli, Mauajama Firdaus, and Asif Ekbal. 2024. Deciphering cognitive distortions in patient-doctor mental health conversations: A multimodal LLM-based detection and reasoning framework. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22546–22570, Miami, Florida, USA. Association for Computational Linguistics.

Khushboo Singh, Vasudha Varadarajan, Adithya V. Ganesan, August Håkan Nilsson, Nikita Soni, Syeda Mahwish, Pranav Chitale, Ryan L. Boyd, Lyle Ungar, Richard N. Rosenthal, and H. Andrew Schwartz. 2025. Systematic evaluation of auto-encoding and large language model representations for capturing author states and traits. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18955–18973, Vienna, Austria. Association for Computational Linguistics.

Karan Singla, Zhuohao Chen, David Atkins, and Shrikanth Narayanan. 2020. Towards end-2-end learning for predicting behavior codes from spoken utterances in psychotherapy conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3797–3803, Online. Association for Computational Linguistics.

Hoyun Song, Huije Lee, Jisu Shin, Sukmin Cho, Changgeon Ko, and Jong C. Park. 2025a. Does rationale quality matter? enhancing mental disorder detection via selective reasoning distillation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21738–21756, Vienna, Austria. Association for Computational Linguistics.

Hoyun Song, Jisu Shin, Huije Lee, and Jong Park. 2023. A simple and flexible modeling for mental disorder detection by learning from clinical questionnaires.

In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12190–12206, Toronto, Canada. Association for Computational Linguistics.

Jiayu Song, Mahmud Elahi Akhter, Dana Atzil-Slonim, and Maria Liakata. 2025b. Temporal reasoning for timeline summarisation in social media. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28085–28101, Vienna, Austria. Association for Computational Linguistics.

Jiayu Song, Jenny Chim, Adam Tsakalidis, Julia Ive, Dana Atzil-Slonim, and Maria Liakata. 2024. Combining hierachical VAEs with LLMs for clinically meaningful timeline summarisation in social media. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14651–14672, Bangkok, Thailand. Association for Computational Linguistics.

Robert L Spitzer, Kurt Kroenke, Janet BW Williams, Patient Health Questionnaire Primary Care Study Group, and 1 others. 1999. Validation and utility of a self-report version of prime-md: the phq primary care study. *jama*, 282(18):1737–1744.

Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. 2006. A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of internal medicine*, 166(10):1092–1097.

Aseem Srivastava, Smriti Joshi, Tanmoy Chakraborty, and Md Shad Akhtar. 2024. Knowledge planning in large language models for domain-aligned counseling summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17775–17789, Miami, Florida, USA. Association for Computational Linguistics.

Daniela Teodorescu, Tiffany Cheng, Alona Fyshe, and Saif Mohammad. 2023. Language and mental health: Measures of emotion dynamics from text as linguistic biosocial markers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3117–3133, Singapore. Association for Computational Linguistics.

Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Transactions on Computer-Human Interaction*, 27(5).

Roberto Tornero-Costa, Antonio Martinez-Millana, Natasha Azzopardi-Muscat, Ledia Lazeri, Vicente Traver, and David Novillo-Ortiz. 2023. Methodological and quality flaws in the use of artificial intelligence in mental health research: systematic review. *JMIR Mental Health*, 10(1):e42045.

John Torous, Hannah Wisniewski, Bruce Bird, Elizabeth Carpenter, Gary David, Eduardo Elejalde, Dan Fulford, Synthia Guimond, Ryan Hays, Philip Henson, and 1 others. 2019. Creating a digital health smartphone app and digital phenotyping platform for mental health and diverse healthcare needs: an interdisciplinary and collaborative approach. *Journal of Technology in Behavioral Science*, 4(2):73–85.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022. Identifying moments of change from longitudinal user text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4647–4660, Dublin, Ireland. Association for Computational Linguistics.

Talia Tseriotou, Adam Tsakalidis, Peter Foster, Terence Lyons, and Maria Liakata. 2023. Sequential path signature networks for personalised longitudinal language modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5016–5031, Toronto, Canada. Association for Computational Linguistics.

Adithya V Ganesan, Matthew Matero, Aravind Reddy Ravula, Huy Vu, and H. Andrew Schwartz. 2021. Empirical evaluation of pre-trained transformers for human-level NLP: The role of sample size and dimensionality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4515–4532, Online. Association for Computational Linguistics.

Vasudha Varadarajan, Sverker Sikström, Oscar Kjell, and H. Andrew Schwartz. 2024. ALBA: Adaptive language-based assessments for mental health. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2466–2478, Mexico City, Mexico. Association for Computational Linguistics.

Krishnapriya Vishnubhotla, Daniela Teodorescu, Mallory J Feldman, Kristen Lindquist, and Saif M. Mohammad. 2024. Emotion granularity from text: An aggregate-level indicator of mental health. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19168–19185, Miami, Florida, USA. Association for Computational Linguistics.

Hanna Wallach, Meera Desai, A Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P Alex Dow, and 1 others. 2025. Position: Evaluating generative ai systems is a social science measurement challenge. *arXiv preprint arXiv:2502.00561*.

Bichen Wang, Pengfei Deng, Yanyan Zhao, and Bing Qin. 2023a. C2D2 dataset: A resource for the cognitive distortion analysis and its impact on mental health. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10149–10160, Singapore. Association for Computational Linguistics.

Liying Wang, Tanmay Bhanushali, Zhuoran Huang, Jingyi Yang, Sukriti Badami, and Lisa Hightow-Weidman. 2025a. Evaluating generative ai in mental health: systematic review of capabilities and limitations. *JMIR mental health*, 12(1):e70014.

Ming Wang, Peidong Wang, Lin Wu, Xiaocui Yang, Daling Wang, Shi Feng, Yuxin Chen, Bixuan Wang, and Yifei Zhang. 2025b. AnnaAgent: Dynamic evolution agent system with multi-session memory for realistic seeker simulation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23221–23235, Vienna, Austria. Association for Computational Linguistics.

Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. 2024. PATIENT-ψ: Using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12772–12797, Miami, Florida, USA. Association for Computational Linguistics.

Xiaoyi Wang, Jiwei Zhang, Guangtao Zhang, and Honglei Guo. 2025c. Feel the difference? a comparative analysis of emotional arcs in real and LLM-generated CBT sessions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19999–20017, Suzhou, China. Association for Computational Linguistics.

Zhong-Ling Wang, Po-Hsien Huang, Wen-Yau Hsu, and Hen-Hsen Huang. 2023b. Self-adapted utterance selection for suicidal ideation detection in lifeline conversations. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1436–1446, Dubrovnik, Croatia. Association for Computational Linguistics.

Jason Wei, Kelly Finn, Emma Templeton, Thalia Wheatley, and Soroush Vosoughi. 2021. Linguistic complexity loss in text-based therapy. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4450–4459, Online. Association for Computational Linguistics.

Mengxi Xiao, Qianqian Xie, Ziyan Kuang, Zhicheng Liu, Kailai Yang, Min Peng, Weiguang Han, and Jimin Huang. 2024. HealMe: Harnessing cognitive reframing in large language models for psychotherapy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1707–1725, Bangkok, Thailand. Association for Computational Linguistics.

Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin, and Xiangmin Xu. 2025. PsyDT: Using LLMs to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1081–1115, Vienna, Austria. Association for Computational Linguistics.

Yangyang Xu, Jinpeng Hu, Zhuoer Zhao, Zhangling Duan, Xiao Sun, and Xun Yang. 2025. MultiAgentESC: A LLM-based multi-agent collaboration framework for emotional support conversation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4665–4681, Suzhou, China. Association for Computational Linguistics.

Shweta Yadav, Cornelia Caragea, Chenye Zhao, Naincy Kumari, Marvin Solberg, and Tanmay Sharma. 2023. Towards identifying fine-grained depression symptoms from memes. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8890–8905, Toronto, Canada. Association for Computational Linguistics.

Chenghao Yang, Yudong Zhang, and Smaranda Muresan. 2021. Weakly-supervised methods for suicide risk assessment: Role of related domains. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1049–1057, Online. Association for Computational Linguistics.

Jiamin Yang and David Jurgens. 2024. Modeling empathetic alignment in conversation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3127–3148, Mexico City, Mexico. Association for Computational Linguistics.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, Singapore. Association for Computational Linguistics.

Qisen Yang, Zekun Wang, Honghui Chen, Shenzhi Wang, Yifan Pu, Xin Gao, Wenhao Huang, Shiji Song, and Gao Huang. 2024. PsychoGAT: A novel psychological measurement paradigm through interactive fiction games with LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14470–14505, Bangkok, Thailand. Association for Computational Linguistics.

Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang, Phey Ling Kit, Nicholas Gabriel Lim, Cameron Tan Shi Ern, and Ee-Peng Lim. 2025a. CAMI: A counselor agent supporting motivational interviewing through state inference and topic exploration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21037–21081, Vienna, Austria. Association for Computational Linguistics.

Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang, Nicholas Gabriel Lim, Cameron Tan Shi Ern, Phey Ling Kit, Jenny Giam Xiuhui, John Pinto, and Ee-Peng Lim. 2025b. Consistent client simulation for motivational interviewing-based counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20959–20998, Vienna, Austria. Association for Computational Linguistics.

Binwei Yao, Chao Shi, Likai Zou, Lingfeng Dai, Mengyue Wu, Lu Chen, Zhen Wang, and Kai Yu. 2022. D4: a Chinese dialogue dataset for depression-diagnosis-oriented chat. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2438–2459, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sourabh Zanwar, Xiaofei Li, Daniel Wiechmann, Yu Qiao, and Elma Kerz. 2023a. What to fuse and how to fuse: Exploring emotion and personality fusion strategies for explainable mental disorder detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8926–8940, Toronto, Canada. Association for Computational Linguistics.

Sourabh Zanwar, Daniel Wiechmann, Yu Qiao, and Elma Kerz. 2023b. SMHD-GER: A large-scale benchmark dataset for automatic mental health detection from social media in German. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1526–1541, Dubrovnik, Croatia. Association for Computational Linguistics.

Wei Zhai, Nan Bai, Qing Zhao, Jianqiang Li, Fan Wang, Hongzhi Qi, Meng Jiang, Xiaoqin Wang, Bing Xiang Yang, and Guanghui Fu. 2025. MentalGLM series: Explainable large language models for mental health analysis on Chinese social media. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13599–13614, Suzhou, China. Association for Computational Linguistics.

Wei Zhai, Hongzhi Qi, Qing Zhao, Jianqiang Li, Ziqi Wang, Han Wang, Bing Yang, and Guanghui Fu. 2024. Chinese MentalBERT: Domain-adaptive pre-training on social media for Chinese mental health text analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10574–10585, Bangkok, Thailand. Association for Computational Linguistics.

Enshi Zhang and Christian Poellabauer. 2025. Mitigating interviewer bias in multimodal depression detection: An approach with adversarial learning and contextual positional encoding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12169–12188, Suzhou, China. Association for Computational Linguistics.

Linhai Zhang, Ziyang Gao, Deyu Zhou, and Yulan He. 2025a. Explainable depression detection in clinical interviews with personalized retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9927–9944, Vienna, Austria. Association for Computational Linguistics.

Mian Zhang, Xianjun Yang, Xinlu Zhang, Travis Labrum, Jamie C. Chiu, Shaun M. Eack, Fei Fang, William Yang Wang, and Zhiyu Chen. 2025b. CBT-bench: Evaluating large language models on assisting cognitive behavior therapy. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3864–3900, Albuquerque, New Mexico. Association for Computational Linguistics.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2023. Ask an expert: Leveraging language models to improve strategic reasoning in goal-oriented dialogue models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6665–6694, Toronto, Canada. Association for Computational Linguistics.

Qiyang Zhang, Renwen Zhang, Yiying Xiong, Yuan Sui, Chang Tong, and Fu-Hung Lin. 2025c. Generative ai mental health chatbots as therapeutic tools: Systematic review and meta-analysis of their role in reducing mental health issues. *Journal of Medical Internet Research*, 27:e78238.

Xiangyu Zhang, Hexin Liu, Kaishuai Xu, Qiquan Zhang, Daijiao Liu, Beena Ahmed, and Julien Epps. 2024. When LLMs meets acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 146–158, Miami, Florida, USA. Association for Computational Linguistics.

Xiangyu Zhang, Hexin Liu, Qiquan Zhang, Beena Ahmed, and Julien Epps. 2025d. SpeechT-RAG: Reliable depression detection in LLMs with retrieval-augmented generation using speech timing information. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10019–10030, Vienna, Austria. Association for Computational Linguistics.

Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Zhu. 2022. Symptom identification for interpretable detection of multiple mental disorders on social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9970–9985, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xinzhe Zheng, Sijie Ji, Jiawei Sun, Renqi Chen, Wei Gao, and Mani Srivastava. 2025. ProMind-LLM: Proactive mental health care via causal reasoning with sensor data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20150–20171, Vienna, Austria. Association for Computational Linguistics.

Jinfeng Zhou, Yuxuan Chen, Jianing Yin, Yongkang Huang, Yihan Shi, Xikun Zhang, Libiao Peng, Rongsheng Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. 2025a. Crisp: Cognitive restructuring of negative thoughts through multi-turn supportive dialogues. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32462–32491, Suzhou, China. Association for Computational Linguistics.

Naitian Zhou, David Bamman, and Isaac L. Bleaman. 2025b. Culture is not trivia: Sociocultural theory for cultural NLP. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25869–25886, Vienna, Austria. Association for Computational Linguistics.

A Details about the surveyed papers

The full list of surveyed papers with the observed practices are in Tables 3–16. Psychologically grounded metrics are based on criteria derived from psychological theory, clinical research, or expert input. AI/NLP metrics, by contrast, focus on computational performance (e.g., accuracy, BLEU, ROUGE) and are largely agnostic to psychological or therapeutic soundness.

The tasks addressed in these papers are listed in Table 17.

Some papers examine mental disorders at a broad level, while others focus on specific diagnosed conditions. Among those that target specific conditions, the following named disorders are examined: *Anxiety, Depression, Suicide ideation, Cognitive distortions, Post-Traumatic Stress Disorder (PTSD), Bipolar Disorder, Schizophrenia, Self-harm, Anorexia, Trauma, Stress, Attention-Deficit/Hyperactivity Disorder (ADHD), Obsessive-Compulsive Disorder (OCD), Panic, and Addiction*.

Paper	Evaluation metrics	Human evaluation	Expert evaluators	Evaluation guidelines provided	Discuss limitations of evaluation
Linguistic Complexity Loss in Text-Based Therapy (Wei et al., 2021)	AI/NLP metrics	No	No	N. A.	Yes
Depression Detection in Clinical Interviews with LLM-Empowered Structural Element Graph (Chen et al., 2024b)	AI/NLP metrics	No	No	N. A.	No
An Annotated Dataset for Explainable Interpersonal Risk Factors of Mental Disturbance in Social Media Posts (Garg et al., 2023)	AI/NLP metrics	No	No	N. A.	No
Lived Experience Not Found: LLMs Struggle to Align with Experts on Addressing Adverse Drug Reactions from Psychiatric Medication Use (Chandra et al., 2025)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
Weakly-Supervised Methods for Suicide Risk Assessment: Role of Related Domains (Yang et al., 2021)	AI/NLP metrics	No	No	N. A.	Yes
Generating Mental Health Transcripts with SAPE (Spanish Adaptive Prompt Engineering) (Lozoya et al., 2024)	Psychologically grounded metrics	Yes	Yes	No	Yes
Suicidal Risk Detection for Military Personnel (Park et al., 2020)	AI/NLP metrics	No	No	N. A.	No
C2D2 Dataset: A Resource for the Cognitive Distortion Analysis and Its Impact on Mental Health (Wang et al., 2023a)	AI/NLP metrics	No	No	N. A.	Yes
Taking a turn for the better: Conversation redirection throughout the course of mental-health therapy (Nguyen et al., 2024)	Psychologically grounded metrics	Yes	No	No	Yes
Towards Intelligent Clinically-Informed Language Analyses of People with Bipolar Disorder and Schizophrenia (Aich et al., 2022)	AI/NLP metrics	No	No	N. A.	Yes

Table 3: List of surveyed papers *(part 1 of 14)*.

Paper	Evaluation metrics	Human evaluation	Expert evaluators	Evaluation guidelines provided	Discuss limitations of evaluation
PAL: Persona-Augmented Emotional Support Conversation Generation (Cheng et al., 2023)	Psychologically grounded metrics	Yes	No	Yes	Yes
Towards end-2-end learning for predicting behavior codes from spoken utterances in psychotherapy conversations (Singla et al., 2020)	AI/NLP metrics	No	No	N. A.	No
Detecting Bipolar Disorder from Misdiagnosed Major Depressive Disorder with Mood-Aware Multi-Task Learning (Lee et al., 2024a)	AI/NLP metrics	No	No	N. A.	Yes
Towards Emotional Support Dialog Systems (Liu et al., 2021)	Psychologically grounded metrics	Yes	No	Yes	Yes
Mental Disorder Classification via Temporal Representation of Text (Kumar et al., 2024)	AI/NLP metrics	No	No	N. A.	No
Discourse-Level Representations can Improve Prediction of Degree of Anxiety (Juhng et al., 2023)	AI/NLP metrics	No	No	N. A.	Yes
Ask an Expert: Leveraging Language Models to Improve Strategic Reasoning in Goal-Oriented Dialogue Models (Zhang et al., 2023)	Psychologically grounded metrics	Yes	No	Yes	Yes
Empowering Psychotherapy with Large Language Models: Cognitive Distortion Detection through Diagnosis of Thought Prompting (Chen et al., 2023b)	AI/NLP metrics	No	No	N. A.	Yes
Identifying Moments of Change from Longitudinal User Text (Tsakalidis et al., 2022)	AI/NLP metrics	Yes	No	Yes	No
Multi-Level Feedback Generation with Large Language Models for Empowering Novice Peer Counselors (Chaszczewicz et al., 2024)	Psychologically grounded metrics	Yes	Yes	Yes	Yes

Table 4: List of surveyed papers *(part 2 of 14)*.

Paper	Evaluation metrics	Human evaluation	Expert evaluators	Evaluation guidelines provided	Discuss limitations of evaluation
Using Linguistic Entrainment to Evaluate Large Language Models for Use in Cognitive Behavioral Therapy (Kian et al., 2025)	AI/NLP metrics	No	No	N. A.	Yes
Identifying Early Maladaptive Schemas from Mental Health Question Texts (Gollapalli et al., 2023)	AI/NLP metrics	No	No	N. A.	Yes
Understanding the Therapeutic Relationship between Counselors and Clients in Online Text-based Counseling using LLMs (Li et al., 2024)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
Diverse Perspectives, Divergent Models: Cross-Cultural Evaluation of Depression Detection on Twitter (Abdelkadir et al., 2024)	AI/NLP metrics	No	No	N. A.	No
PHASE: Learning Emotional Phase-aware Representations for Suicide Ideation Detection on Social Media (Sawhney et al., 2021a)	AI/NLP metrics	No	No	N. A.	Yes
Improving the Generalizability of Depression Detection by Leveraging Clinical Questionnaires (Nguyen et al., 2022)	AI/NLP metrics	No	No	N. A.	Yes
What to Fuse and How to Fuse: Exploring Emotion and Personality Fusion Strategies for Explainable Mental Disorder Detection (Zanwar et al., 2023a)	AI/NLP metrics	No	No	N. A.	No
Do Large Language Models Align with Core Mental Health Counseling Competencies? (Nguyen et al., 2025a)	AI/NLP metrics	No	No	N. A.	Yes
Cross-Lingual Suicidal-Oriented Word Embedding toward Suicide Prevention (Lee et al., 2020)	AI/NLP metrics	No	No	N. A.	No
Understanding Client Reactions in Online Mental Health Counseling (Li et al., 2023)	AI/NLP metrics	No	No	N. A.	Yes

Table 5: List of surveyed papers *(part 3 of 14)*.

Paper	Evaluation metrics	Human evaluation	Expert evaluators	Evaluation guidelines provided	Discuss limitations of evaluation
IMBUE: Improving Interpersonal Effectiveness through Simulation and Just-in-time Feedback with Human-Language Model Interaction (Lin et al., 2024)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
PsychoGAT: A Novel Psychological Measurement Paradigm through Interactive Fiction Games with LLM Agents (Yang et al., 2024)	Psychologically grounded metrics	Yes	Yes	No	Yes
ALBA: Adaptive Language-Based Assessments for Mental Health (Varadarajan et al., 2024)	AI/NLP metrics	No	No	N. A.	Yes
Ask the experts: sourcing a high-quality nutrition counseling dataset through Human-AI collaboration (Balloccu et al., 2024)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
Cognitive Reframing of Negative Thoughts through Human-Language Model Interaction (Sharma et al., 2023)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support (Sharma et al., 2020)	AI/NLP metrics	No	No	N. A.	No
A Shoulder to Cry on: Towards A Motivational Virtual Assistant for Assuaging Mental Agony (Saha et al., 2022)	AI/NLP metrics	Yes	No	No	No
Knowledge-enhanced Mixed-initiative Dialogue System for Emotional Support Conversations (Deng et al., 2023)	Psychologically grounded metrics	Yes	No	No	Yes
Language and Mental Health: Measures of Emotion Dynamics from Text as Linguistic Biosocial Markers (Teodorescu et al., 2023)	AI/NLP metrics	No	No	N. A.	Yes
CURE: Context- and Uncertainty-Aware Mental Disorder Detection (Kang et al., 2024)	AI/NLP metrics	No	No	N. A.	Yes

Table 6: List of surveyed papers *(part 4 of 14)*.

Paper	Evaluation metrics	Human evaluation	Expert evaluators	Evaluation guidelines provided	Discuss limitations of evaluation
Still Not Quite There! Evaluating Large Language Models for Comorbid Mental Health Diagnosis (Hengle et al., 2024)	AI/NLP metrics	No	No	N. A.	Yes
A Cognitive Stimulation Dialogue System with Multi-source Knowledge Fusion for Elders with Cognitive Impairment (Jiang et al., 2023)	Psychologically grounded metrics	Yes	No	No	No
KMI: A Dataset of Korean Motivational Interviewing Dialogues for Psychotherapy (Kim et al., 2025a)	Psychologically grounded metrics	Yes	Yes	Yes	No
Combining Hierachical VAEs with LLMs for clinically meaningful timeline summarisation in social media (Song et al., 2024)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
A Simple and Flexible Modeling for Mental Disorder Detection by Learning from Clinical Questionnaires (Song et al., 2023)	AI/NLP metrics	No	No	N. A.	Yes
On the Way to Gentle AI Counselor: Politeness Cause Elicitation and Intensity Tagging in Code-mixed Hinglish Conversations for Social Good (Priya et al., 2024)	AI/NLP metrics	No	No	N. A.	Yes
D4: a Chinese Dialogue Dataset for Depression-Diagnosis-Oriented Chat (Yao et al., 2022)	Psychologically grounded metrics	Yes	Yes	No	Yes
Task-Adaptive Tokenization: Enhancing Long-Form Text Generation Efficacy in Mental Health and Beyond (Liu et al., 2023)	Psychologically grounded metrics	Yes	Yes	Yes	No
Gendered Mental Health Stigma in Masked Language Models (Lin et al., 2022)	AI/NLP metrics	No	No	N. A.	Yes
LLM Questionnaire Completion for Automatic Psychiatric Assessment (Rosenman et al., 2024)	AI/NLP metrics	No	No	N. A.	Yes

Table 7: List of surveyed papers *(part 5 of 14)*.

Paper	Evaluation metrics	Human evaluation	Expert evaluators	Evaluation guidelines provided	Discuss limitations of evaluation
Emotion Granularity from Text: An Aggregate-Level Indicator of Mental Health (Vishnubhotla et al., 2024)	AI/NLP metrics	No	No	N. A.	Yes
Leveraging Open Data and Task Augmentation to Automated Behavioral Coding of Psychotherapy Conversations in Low-Resource Scenarios (Chen et al., 2022)	AI/NLP metrics	No	No	N. A.	No
CASE: Efficient Curricular Data Pre-training for Building Assistive Psychology Expert Models (Harne et al., 2024)	AI/NLP metrics	No	No	N. A.	No
Do Models of Mental Health Based on Social Media Data Generalize? (Harrigian et al., 2020)	AI/NLP metrics	No	No	N. A.	No
Self-Adapted Utterance Selection for Suicidal Ideation Detection in Lifeline Conversations (Wang et al., 2023b)	AI/NLP metrics	No	No	N. A.	No
Can AI Relate: Testing Large Language Model Response for Mental Health Support (Gabriel et al., 2024)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
Suicide Ideation Detection via Social and Temporal User Representations using Hyperbolic Learning (Sawhney et al., 2021b)	AI/NLP metrics	No	No	N. A.	Yes
PAIR: Prompt-Aware margIn Ranking for Counselor Reflection Scoring in Motivational Interviewing (Min et al., 2022)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
Crisis counselor language and perceived genuine concern in crisis conversations (Buda et al., 2024)	AI/NLP metrics	No	No	N. A.	Yes
Empirical Evaluation of Pre-trained Transformers for Human-Level NLP: The Role of Sample Size and Dimensionality (V Ganesan et al., 2021)	AI/NLP metrics	No	No	N. A.	No

Table 8: List of surveyed papers *(part 6 of 14)*.

Paper	Evaluation metrics	Human evaluation	Expert evaluators	Evaluation guidelines provided	Discuss limitations of evaluation
Roleplay-doh: Enabling Domain-Experts to Create LLM-simulated Patients via Eliciting and Adhering to Principles (Louie et al., 2024)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
Exciting Mood Changes: A Time-aware Hierarchical Transformer for Change Detection Modelling (Hills et al., 2024)	AI/NLP metrics	Yes	No	Yes	No
SMILE: Single-turn to Multi-turn Inclusive Language Expansion via ChatGPT for Mental Health Support (Qiu et al., 2024a)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
PsyGUARD: An Automated System for Suicide Detection and Risk Assessment in Psychological Counseling (Qiu et al., 2024b)	AI/NLP metrics	No	No	N. A.	No
FedTherapist: Mental Health Monitoring with User-Generated Linguistic Expressions on Smartphones via Federated Learning (Shin et al., 2023)	AI/NLP metrics	No	No	N. A.	Yes
Modeling Empathetic Alignment in Conversation (Yang and Jurgens, 2024)	AI/NLP metrics	No	No	N. A.	Yes
Towards Interpretable Mental Health Analysis with Large Language Models (Yang et al., 2023)	AI/NLP metrics	Yes	No	Yes	No
Multimodal Cognitive Reframing Therapy via Multi-hop Psychotherapeutic Reasoning (Kim et al., 2025c)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
SMHD-GER: A Large-Scale Benchmark Dataset for Automatic Mental Health Detection from Social Media in German (Zanwar et al., 2023b)	AI/NLP metrics	No	No	N. A.	No
Towards Identifying Fine-Grained Depression Symptoms from Memes (Yadav et al., 2023)	AI/NLP metrics	Yes	No	No	Yes

Table 9: List of surveyed papers *(part 7 of 14)*.

Paper	Evaluation metrics	Human evaluation	Expert evaluators	Evaluation guidelines provided	Discuss limitations of evaluation
Mapping Long-term Causalities in Psychiatric Symptomatology and Life Events from Social Media (Chen et al., 2024a)	AI/NLP metrics	No	No	N. A.	No
Symptom Identification for Interpretable Detection of Multiple Mental Disorders on Social Media (Zhang et al., 2022)	AI/NLP metrics	No	No	N. A.	Yes
Gender and Racial Fairness in Depression Research using Social Media (Aguirre et al., 2021)	AI/NLP metrics	No	No	N. A.	Yes
CBT-Bench: Evaluating Large Language Models on Assisting Cognitive Behavior Therapy (Zhang et al., 2025b)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
e-THERAPIST: I suggest you to cultivate a mindset of positivity and nurture uplifting thoughts (Mishra et al., 2023a)	Psychologically grounded metrics	Yes	No	No	No
Knowledge Planning in Large Language Models for Domain-Aligned Counseling Summarization (Srivastava et al., 2024)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
PATIENT-ψ: Using Large Language Models to Simulate Patients for Training Mental Health Professionals (Wang et al., 2024)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
Deciphering Cognitive Distortions in Patient-Doctor Mental Health Conversations: A Multimodal LLM-Based Detection and Reasoning Framework (Singh et al., 2024)	AI/NLP metrics	Yes	No	Yes	Yes
Sequential Path Signature Networks for Personalised Longitudinal Language Modeling (Tseriotou et al., 2023)	AI/NLP metrics	Yes	No	Yes	No
DisorBERT: A Double Domain Adaptation Model for Detecting Signs of Mental Disorders in Social Media (Aragón et al., 2023)	AI/NLP metrics	No	No	N. A.	No

Table 10: List of surveyed papers *(part 8 of 14)*.

Paper	Evaluation metrics	Human evaluation	Expert evaluators	Evaluation guidelines provided	Discuss limitations of evaluation
Cactus: Towards Psychological Counseling Conversations using Cognitive Behavioral Theory (Lee et al., 2024b)	Psychologically grounded metrics	Yes	Yes	Yes	No
Development of Conversational AI for Sleep Coaching Programme (Shim, 2021)	AI/NLP metrics	No	No	N. A.	No
Social Biases in NLP Models as Barriers for Persons with Disabilities (Hutchinson et al., 2020)	AI/NLP metrics	No	No	N. A.	Yes
Chinese MentalBERT: Domain-Adaptive Pre-training on Social Media for Chinese Mental Health Text Analysis (Zhai et al., 2024)	AI/NLP metrics	No	No	N. A.	No
Micromodels for Efficient, Explainable, and Reusable Systems: A Case Study on Mental Health (Lee et al., 2021)	AI/NLP metrics	No	No	N. A.	No
HealMe: Harnessing Cognitive Reframing in Large Language Models for Psychotherapy (Xiao et al., 2024)	Psychologically grounded metrics	Yes	Yes	Yes	No
PAL to Lend a Helping Hand: Towards Building an Emotion Adaptive Polite and Empathetic Counseling Conversational Agent (Mishra et al., 2023b)	Psychologically grounded metrics	Yes	Yes	No	No
Predicting Treatment Outcome from Patient Texts:The Case of Internet-Based Cognitive Behavioural Therapy (Gogoulou et al., 2021)	AI/NLP metrics	No	No	N. A.	Yes
Detection of Multiple Mental Disorders from Social Media with Two-Stream Psychiatric Experts (Chen et al., 2023a)	AI/NLP metrics	No	No	N. A.	No
When LLMs Meets Acoustic Landmarks: An Efficient Approach to Integrate Speech into Large Language Models for Depression Detection (Zhang et al., 2024)	AI/NLP metrics	No	No	N. A.	No

Table 11: List of surveyed papers *(part 9 of 14)*.

Paper	Evaluation metrics	Human evaluation	Expert evaluators	Evaluation guidelines provided	Discuss limitations of evaluation
The Colorful Future of LLMs: Evaluating and Improving LLMs as Emotional Supporters for Queer Youth (Lissak et al., 2024)	Psychologically grounded metrics	Yes	No	Yes	No
Decoding the Narratives: Analyzing Personal Drug Experiences Shared on Reddit (Bouzoubaa et al., 2024)	AI/NLP metrics	No	No	N. A.	Yes
A Fully Generative Motivational Interviewing Counsellor Chatbot for Moving Smokers Towards the Decision to Quit (Mahmood et al., 2025)	Psychologically grounded metrics	Yes	No	Yes	Yes
PsyDial: A Large-scale Long-term Conversational Dataset for Mental Health Support (Qiu and Lan, 2025)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
SpeechT-RAG: Reliable Depression Detection in LLMs with Retrieval-Augmented Generation Using Speech Timing Information (Zhang et al., 2025d)	AI/NLP metrics	No	No	N. A.	No
DeepWell-Adol: A Scalable Expert-Based Dialogue Corpus for Adolescent Positive Mental Health and Wellbeing Promotion (Qiu et al., 2025b)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
Hanging in the Balance: Pivotal Moments in Crisis Counseling Conversations (Nguyen et al., 2025b)	Psychologically grounded metrics	No	No	N. A.	Yes
Assess and Prompt: A Generative RL Framework for Improving Engagement in Online Mental Health Communities (Gaur et al., 2025)	AI/NLP metrics	Yes	No	No	Yes
AnnaAgent: Dynamic Evolution Agent System with Multi-Session Memory for Realistic Seeker Simulation (Wang et al., 2025b)	Psychologically grounded metrics	No	No	N. A.	Yes
Tracking Life's Ups and Downs: Mining Life Events from Social Media Posts for Mental Health Analysis (Lv et al., 2025)	Psychologically grounded metrics	No	No	N. A.	Yes

Table 12: List of surveyed papers *(part 10 of 14)*.

Paper	Evaluation metrics	Human evaluation	Expert evaluators	Evaluation guidelines provided	Discuss limitations of evaluation
Can Large Language Models Identify Implicit Suicidal Ideation? An Empirical Evaluation (Li et al., 2025)	Psychologically grounded metrics	Yes	Yes	Yes	No
Eeyore: Realistic Depression Simulation via Expert-in-the-Loop Supervised and Preference Optimization (Liu et al., 2025)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
Dialogue Systems for Emotional Support via Value Reinforcement (Kim et al., 2025b)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
MultiAgentESC: A LLM-based Multi-Agent Collaboration Framework for Emotional Support Conversation (Xu et al., 2025)	Psychologically grounded metrics	Yes	Yes	Yes	No
MAGI: Multi-Agent Guided Interview for Psychiatric Assessment (Bi et al., 2025)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
Systematic Evaluation of Auto-Encoding and Large Language Model Representations for Capturing Author States and Traits (Singh et al., 2025)	AI/NLP metrics	No	No	N. A.	No
The Pursuit of Empathy: Evaluating Small Language Models for PTSD Dialogue Support (Bn et al., 2025a)	Psychologically grounded metrics	Yes	No	Yes	Yes
Just a Scratch: Enhancing LLM Capabilities for Self-harm Detection through Intent Differentiation and Emoji Interpretation (Ghosh et al., 2025)	AI/NLP metrics	No	No	N. A.	Yes
Temporal reasoning for timeline summarisation in social media (Song et al., 2025b)	Psychologically grounded metrics	Yes	Yes	Yes	No
MIRROR: Multimodal Cognitive Reframing Therapy for Rolling with Resistance (Kim et al., 2025d)	Psychologically grounded metrics	Yes	Yes	Yes	Yes

Table 13: List of surveyed papers *(part 11 of 14)*.

Paper	Evaluation metrics	Human evaluation	Expert evaluators	Evaluation guidelines provided	Discuss limitations of evaluation
Are LLMs effective psychological assessors? Leveraging adaptive RAG for interpretable mental health screening through psychometric practice (Ravenda et al., 2025)	AI/NLP metrics	No	No	N. A.	No
ReDepress: A Cognitive Framework for Detecting Depression Relapse from Social Media (Agarwal et al., 2025)	AI/NLP metrics	No	No	N. A.	No
MentalGLM Series: Explainable LLMs for Mental Health Analysis on Chinese Social Media (Zhai et al., 2025)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
Towards AI-Assisted Psychotherapy: Emotion-Guided Generative Interventions (Haydarov et al., 2025)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
Mitigating Interviewer Bias in Multimodal Depression Detection: An Approach with Adversarial Learning and Contextual Positional Encoding (Zhang and Poellabauer, 2025)	AI/NLP metrics	No	No	N. A.	Yes
Explainable Depression Detection in Clinical Interviews with Personalized Retrieval-Augmented Generation (Zhang et al., 2025a)	AI/NLP metrics	No	No	N. A.	No
From Heart to Words: Generating Empathetic Responses via Integrated Figurative Language and Semantic Context Signals (Lee et al., 2025)	Psychologically grounded metrics	Yes	No	Yes	Yes
Reframe Your Life Story: Interactive Narrative Therapist and Innovative Moment Assessment with Large Language Models (Feng et al., 2025)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
From Conversation to Automation: Leveraging LLMs for Problem-Solving Therapy Analysis (Aghakhani et al., 2025)	AI/NLP metrics	No	No	N. A.	Yes
Feel the Difference? A Comparative Analysis of Emotional Arcs in Real and LLM-Generated CBT Sessions (Wang et al., 2025c)	Psychologically grounded metrics	No	No	N. A.	Yes

Table 14: List of surveyed papers *(part 12 of 14)*.

Paper	Evaluation metrics	Human evaluation	Expert evaluators	Evaluation guidelines provided	Discuss limitations of evaluation
CAMI: A Counselor Agent Supporting Motivational Interviewing through State Inference and Topic Exploration (Yang et al., 2025a)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
Consistent Client Simulation for Motivational Interviewing-based Counseling (Yang et al., 2025b)	Psychologically grounded metrics	Yes	Yes	Yes	No
Does Rationale Quality Matter? Enhancing Mental Disorder Detection via Selective Reasoning Distillation (Song et al., 2025a)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
Crisp: Cognitive Restructuring of Negative Thoughts through Multi-turn Supportive Dialogues (Zhou et al., 2025a)	Psychologically grounded metrics	Yes	No	Yes	Yes
ProMind-LLM: Proactive Mental Health Care via Causal Reasoning with Sensor Data (Zheng et al., 2025)	Psychologically grounded metrics	Yes	Yes	No	Yes
EmoAgent: Assessing and Safeguarding Human-AI Interaction for Mental Health Safety (Qiu et al., 2025a)	Psychologically grounded metrics	No	No	N. A.	Yes
MIND: Towards Immersive Psychological Healing with Multi-Agent Inner Dialogue (Chen et al., 2025b)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
PsyDT: Using LLMs to Construct the Digital Twin of Psychological Counselor with Personalized Counseling Style for Psychological Counseling (Xie et al., 2025)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
Third-Person Appraisal Agent: Simulating Human Emotional Reasoning in Text with Large Language Models (Hong et al., 2025)	Psychologically grounded metrics	Yes	No	Yes	No
CATCH: A Novel Data Synthesis Framework for High Therapy Fidelity and Memory-Driven Planning Chain of Thought in AI Counseling (Chen et al., 2025a)	Psychologically grounded metrics	Yes	Yes	Yes	No

Table 15: List of surveyed papers *(part 13 of 14)*.

Paper	Evaluation metrics	Human evaluation	Expert evaluators	Evaluation guidelines provided	Discuss limitations of evaluation
Assessment and manipulation of latent constructs in pre-trained language models using psychometric scales (Reuben et al., 2025)	Psychologically grounded metrics	No	No	N. A.	No
How Real Are Synthetic Therapy Conversations? Evaluating Fidelity in Prolonged Exposure Dialogues (Bn et al., 2025b)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
KoACD: The First Korean Adolescent Dataset for Cognitive Distortion Analysis via Role-Switching Multi-LLM Negotiation (Kim and Kim, 2025)	Psychologically grounded metrics	Yes	Yes	Yes	Yes
Exploring Large Language Models for Detecting Mental Disorders (Kuzmin et al., 2025)	AI/NLP metrics	No	No	N. A.	Yes
M-Help: Using Social Media Data to Detect Mental Health Help-Seeking Signals (Sathvik et al., 2025)	AI/NLP metrics	No	No	N. A.	Yes

Table 16: List of surveyed papers *(part 14 of 14)*.

Application Type	**Tasks**
Assessment	Anxiety detection; Depression detection; Classification of interpersonal risk factors; Adverse drug reactions detection; Suicide risk detection; Cognitive distortion detection; Detection of schizophrenia disorders; Detecting bipolar disorder; Mental disorder classification; Predicting degree of anxiety; Detection of moments of change; Maladaptive schema detection; Cross-cultural evaluation of depression detection; Psychological profile generation; Measuring emotion granularity from text to detect mental health conditions; Detecting mood changes in social media users over time; Automatic detection of mental health conditions from social media posts in German; Identifying depression symptoms from memes; Multimodal LLM-based cognitive distortions detection; Personalized mood change detection from users' online text over time; Predicting treatment outcome in internet-based therapy; Classification of Reddit drug-use narratives into psychologically and socially meaningful categories; Chinese language model for psychological text analysis on social media; Evaluating how well depression detection models generalize across social media platforms; Identifying social biases toward disability in NLP models; Analyze fairness and bias in depression detection models on social media across gender and racial groups
Intervention	Emotional support conversation generation; Using entrainment in CBT; Nutrition counseling; Synthetic dialogue generation for elders with cognitive impairment; Mental illness conditioned motivational dialogue generation; Generating motivational interviewing dialogues; Cognitive reframing; AI-assisted multimodal therapy; Evaluating how well large language models can assist cognitive behavioral therapy; Developing dialogue system for mental health support; Structured, empathetic cognitive reframing in psychotherapy; LLMs as emotional supporters for queer youth; Generating synthetic therapy transcripts; Enhancing long-form text generation for psychological question-answering; Evaluating whether LLMs can provide ethical, empathetic, and theory-grounded responses for mental health support
Information synthesis	Analysis of quality of therapy conversations; Behavior code prediction; Understanding the therapeutic relationship between counselors and clients; Evaluating LLM alignment with counseling competencies; Enhancing interpersonal skills; Analysis of client reactions in online mental health counseling; Understanding empathy in mental health support text; Clinically meaningful timeline summarisation in social media; Politeness and intensity tagging in conversations; Teaching AI to automatically label behaviors in therapy conversations using small amounts of data; Scoring counselor responses for reflective listening in motivational interviewing; Creating realistic AI-simulated patients for counselor training; Counseling summarization; Patient simulation for training therapists

Table 17: Overview of the diverse tasks addressed in the surveyed papers.