A Survey of Large Language Models in Psychotherapy: Current Landscape and Future Directions

Hongbin Na[UTS]*, Yining Hua[Harvard]*, Zimu Wang[XJTLU]*,
Tao Shen[UTS], Beibei Yu[UTS], Lilin Wang[UPenn], Wei Wang[XJTLU], John Torous[Harvard], Ling Chen[UTS]

[UTS]Australian Artificial Intelligence Institute, University of Technology Sydney
[Harvard]Harvard University [XJTLU]Xi'an Jiaotong-Liverpool University [UPenn]University of Pennsylvania

hongbin.na@student.uts.edu.au, yininghua@g.harvard.edu
zimu.wang19@student.xjtlu.edu.cn

Abstract

Mental health is increasingly critical in contemporary healthcare, with psychotherapy demanding dynamic, context-sensitive interactions that traditional NLP methods struggle to capture. Large Language Models (LLMs) offer significant potential for addressing this gap due to their ability to handle extensive context and multi-turn reasoning. This review introduces a conceptual taxonomy dividing psychotherapy into interconnected stages–assessment, diagnosis, and treatment–to systematically examine LLM advancements and challenges. Our comprehensive analysis reveals imbalances in current research, such as a focus on common disorders, linguistic biases, fragmented methods, and limited theoretical integration. We identify critical challenges including capturing dynamic symptom fluctuations, overcoming linguistic and cultural biases, and ensuring diagnostic reliability. Highlighting future directions, we advocate for continuous multi-stage modeling, real-time adaptive systems grounded in psychological theory, and diversified research covering broader mental disorders and therapeutic approaches, aiming toward more holistic and clinically integrated psychotherapy LLMs systems.

1 Introduction

Mental health plays an increasingly critical role in current healthcare and social well-being. The high prevalence of common psychological disorders, such as depression and anxiety, has led to a growing demand for accessible and effective psychotherapy. The core of psychotherapy resides in *dynamic, contextual* interpersonal interactions—therapists should continuously assess and adjust their intervention strategies (Wampold and Imel, 2015) based on patients' emotional fluctuations, verbal expressions, and social backgrounds, fostering a strong therapeutic alliance (Stubbe, 2018) to achieve symptom resilience. This deep and flexible process contrasts sharply with traditional NLP, which is typically limited to static or single-task settings.

*Equal contribution.

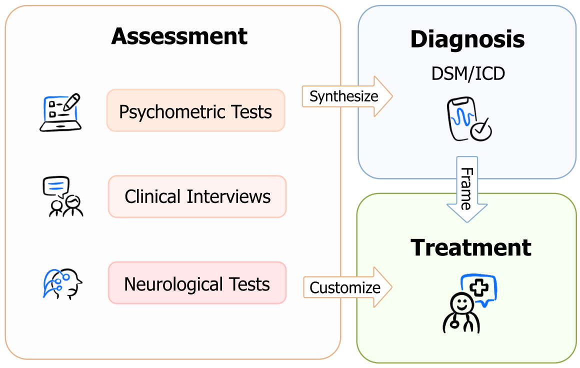


Figure 1: The dynamic and interrelated network among assessment, diagnosis, and treatment in psychotherapy.

Large language models (LLMs) offer a new perspective to addressing this challenge. By leveraging their capability to model extensive context and perform multi-turn reasoning (Wang et al., 2024f; Li et al., 2024b), LLMs can capture rich semantics and emotional signals in dialogues (Ma et al., 2025), enabling end-to-end language understanding and generation (Wang et al., 2024c; Qian et al., 2024). In assessment, LLMs can extract potential symptom cues from vague and fragmented expressions (Tu et al., 2024; Qiu et al., 2024). During diagnosis, they integrate subjective and objective patient information across multiple utterances (Chen et al., 2023a; Ren et al., 2024). In therapeutic interventions, they adapt conversational strategies based on patients' real-time feedback, enabling more flexible and human-like interactions compared to traditional scripted systems (Lee et al., 2024b,d). As a result, LLMs have the potential to surpass the conventional "discrete label recognition" paradigm, evolving toward a model of continuous, progressive clinical reasoning, enabling seamless connections

Findings of the Association for Computational Linguistics: ACL 2025, pages 7362–7376
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

across *assessment*, *diagnosis*, and *treatment*, aligning more closely with therapists' cognitive process and interaction flow.

However, existing research on applying LLMs in this field remains somewhat *fragmented*. Many studies have utilized LLMs for isolated tasks, such as depression detection (Yang et al., 2023; Bao et al., 2024) or diagnosis (Jiang et al., 2024c), regarding them as superior feature extractors. Another research line has focused on developing mental health counseling chatbots (Chen et al., 2023b; Zhang et al., 2024); however, these systems remain limited to partial assistance due to insufficient integration with clinical workflows. In other words, although LLMs hold the potential to span the entire continuum from assessment to intervention, they remain limited by the fragmented paradigms of traditional NLP, preventing them from fully leveraging their dynamic, contextual capabilities.

To address these gaps, we introduce the first *conceptual taxonomy* that divides the psychotherapy process into three interconnected dimensions: Assessment, Diagnosis, and Treatment, and systematically review recent advancements and critical challenges of applying LLMs at each stage. We provide an extensive analysis of the current landscape from multiple perspectives, including the distribution of research across different psychotherapy stages, the coverage of mental disorders, the diversity of linguistic resources, and the incorporation of psychotherapy theories. Moreover, we critically evaluate the fragmented nature of existing approaches, highlighting the inadequacies in capturing dynamic symptom representations, the inherent limitations due to linguistic resource biases and problematic translations, and the diagnostic risks affecting clinical acceptance. Building on these findings, we outline essential future directions, emphasizing the need for continuous multistage modeling for coherent patient state tracking, real-time adaptability grounded explicitly in psychological theory, and a broadened scope of mental disorders and therapeutic frameworks. Through this comprehensive review, we aim to offer detailed methodological insights, guiding future research efforts and facilitating the practical, continuous, and theoretically-grounded integration of LLMs across the full spectrum of psychotherapy.

Organization of This Survey. We present the first comprehensive survey of recent advancements in applying LLMs to psychotherapy. We introduce a conceptual taxonomy that organizes psychotherapy into three core components—Assessment, Diagnosis, and Treatment—and details their dynamic interrelations (Section §2). We review how LLMs are applied within these components, highlighting their roles in facilitating assessments, refining diagnostic processes, and enhancing treatment strategies (Section §3). We examine current research trends, including symptom and language coverage as well as the distribution of various models and techniques (Section §4). Finally, we discuss open challenges and outline promising directions for future work (Section §5).

2 Conceptual Taxonomy

To establish a standardized framework for understanding psychotherapy, we propose a hierarchical taxonomy aligned with the American Psychological Association (APA)'s tripartite model of psychotherapeutic processes[1]. As illustrated in Figure 1, this taxonomy organizes psychotherapy into three core components: (1) Assessment, (2) Diagnosis, and (3) Treatment, with dynamic interconnections[2]. Each component is detailed below.

2.1 Assessment

Definition. Psychological assessment constitutes the systematic collection and interpretation of data regarding an individual's cognitive, emotional, and behavioral functioning (Cohen et al., 1996; Kaplan and Saccuzzo, 2001). This process employs psychometric tests, structured clinical interviews, behavioral observations, and collateral information to establish a multidimensional profile of psychological states (Groth-Marnat, 2009).

Significance. As the foundational stage of psychotherapy, assessment provides the empirical basis for understanding a client's unique psychological landscape. It enables therapists to identify symptom patterns (Phillips et al., 2007), track temporal changes (Barkham et al., 1993), and contextualize subjective experiences within objective frameworks (Groth-Marnat, 2009). The continuous nature of psychological assessment allows for real-time adjustments to therapeutic strategies (Schiepek et al., 2016), ensuring interventions remain responsive to evolving client needs.

[1] https://www.apa.org/topics/psychotherapy

[2] Throughout this taxonomy, the terms *Assessment*, *Diagnosis*, and *Treatment* specifically refer to the three core components of psychotherapy.

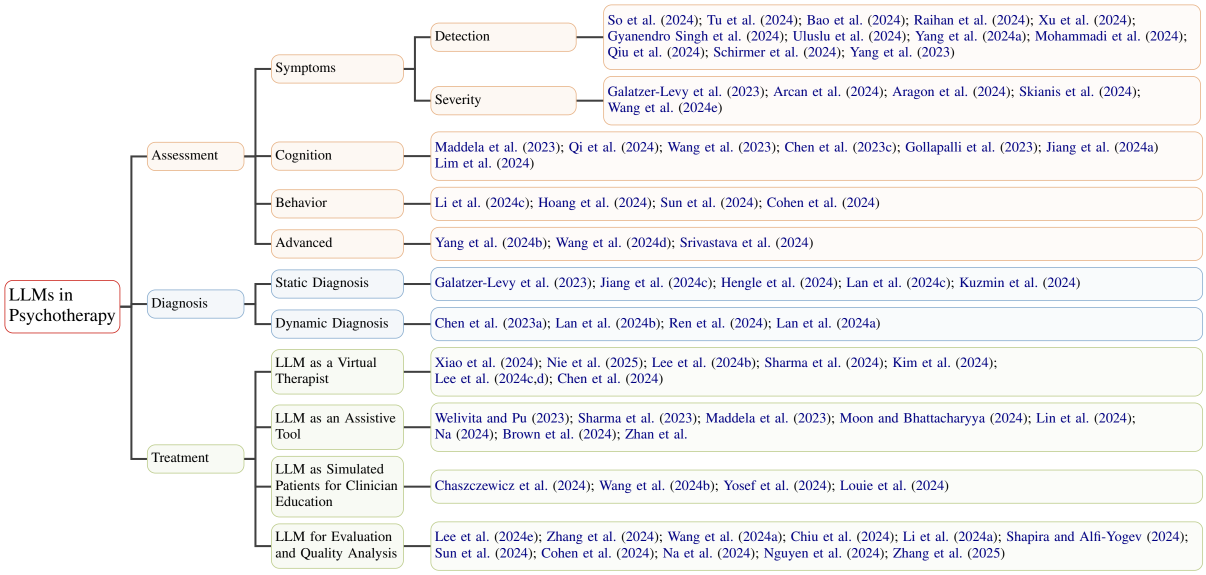


Figure 2: Taxonomy of Research on Large Language Models in Psychotherapy.

2.2 Diagnosis

Definition. Diagnosis represents the analytical process of categorizing psychological distress using established nosological systems such as the DSM-5 (American Psychiatric Association, 2022) and ICD-11 (World Health Organization, 2019). This involves differentiating normative emotional responses from pathological conditions while considering cultural (Teo, 2010) and developmental (Kawa and Giordano, 2012) variables that influence symptom manifestation.

Significance. Diagnosis serves as the conceptual bridge between assessment and treatment, providing a structured framework for intervention planning (Jensen-Doss and Hawley, 2011). By aligning clinical observations with standardized criteria, it enhances communication among professionals (Craddock and Mynors-Wallis, 2014) and facilitates evidence-based decision-making (American Psychiatric Association, 2006).

2.3 Treatment

Definition. Treatment includes evidence-based interventions designed to reduce psychological distress and improve functioning (American Psychiatric Association, 2006). These interventions work by building a therapeutic alliance (Elvins and Green, 2008), restructuring cognition (Ezawa and Hollon, 2023), and modifying behavior (Martin and Pear, 2019), all typically grounded in well-established theoretical orientations.

Significance. Treatment transforms the theories and information gleaned from assessment and diagnosis into practical interventions (Prochaska and Norcross, 2018) that directly address the client's psychological distress (Barlow, 2021) and foster personal growth (Lambert, 2013).

2.4 Interrelations

The taxonomy's components interact through three dynamic processes (see Figure 1) that define psychotherapy as a complex adaptive system:

Synthesizing (Assessment → Diagnosis) The dialectical integration of observational data with nosological frameworks enables diagnostic classifications to contextualize assessment findings, *synthesizing* the patient's various symptoms and behavioral patterns into a diagnostic result (Rencic et al., 2016).

Framing (Diagnosis → Treatment) Diagnosis functions as a *framing* mechanism, integrating complex and diverse symptoms into a coherent classification that establishes a clear blueprint for treatment (American Psychiatric Association, 2022).

Customization (Assessment → Treatment) A process where treatment plans are continuously *refined* based on assessment results, considering individual differences without being constrained by diagnostic labels, to enhance therapeutic effectivenesss (Waszczuk et al., 2017).

2.5 Scope of This Survey

Recent surveys at the intersection of artificial intelligence and mental health primarily cover broad NLP-driven interventions (Malgaroli et al., 2023) or generic AI applications in cognitive behavioral therapy (Jiang et al., 2024b), without specific emphasis on LLMs. Other reviews explicitly focusing on LLMs, such as the scoping review (Hua et al., 2024, 2025) and the overview of general opportunities and risks (Lawrence et al., 2024), examine general mental health rather than psychotherapy specifically. In contrast, our survey explicitly targets recent LLM applications within psychotherapy from the emergence of ChatGPT in late 2022 through October 2024, mainly including papers published in computational linguistics conferences and recent arXiv preprints. We adopt a slightly broad definition of LLMs, primarily including language models exceeding 7 billion parameters (Peng et al., 2023; Zhao et al., 2025). Using the APA's tripartite model as a foundation, we manually classify each paper according to psychotherapy-oriented components—Assessment, Diagnosis, and Treatment–clearly highlighting critical research gaps and future directions distinct from previous reviews.

3 LLMs in Psychotherapy

3.1 Assessment

Symptom Detection leverages LLMs to identify mental health conditions including depression, anxiety, PTSD, and suicidal ideation, demonstrating robust performance and multidimensional applicability across diverse scenarios. Yang et al. (2023) systematically evaluated GPT-3.5, InstructGPT3, and LLaMA models across 11 datasets, revealing that emotion-enhanced chain-of-thought prompting improves interpretability yet remains inferior to specialized supervised methods. So et al. (2024) achieved 70.8% zero-shot symptom retrieval accuracy in Korean psychiatric interviews using GPT-4 Turbo, while their fine-tuned GPT-3.5 attained 0.817 multi-label classification accuracy. Clinical applications show particular promise, as Tu et al. (2024) leveraged GPT-4 and Llama-2 to automate PTSD assessments through information extraction from 411 interviews, significantly enhancing diagnostic practicality.

Social media analysis benefits from approaches like Bao et al. (2024)'s interpretable depression detection framework, which demonstrated strong performance across Vicuna-13B and GPT-3.5 environments. Resource development advances include Raihan et al. (2024)'s *MentalHelp* dataset with 14 million instances, validated through GPT-3.5 zero-shot evaluations. For suicidal ideation monitoring, Gyanendro Singh et al. (2024) and Uluslu et al. (2024) achieved state-of-the-art evidence extraction in the CLPsych 2024 shared task through innovative prompting strategies. Open-source initiatives like *MentaLLaMA* by Yang et al. (2024a) and *Mental-LLM* by Xu et al. (2024) enable multi-symptom detection via instruction-tuned LLaMA variants, though Mohammadi et al. (2024)'s *Well-Dunn* framework reveals persistent gaps in GPT-family models' explanation consistency.

Cross-lingual adaptations include Qiu et al. (2024)'s *PsyGUARD* system based on fine-tuned CHATGLM2-6B for Chinese suicide risk assessment, while Schirmer et al. (2024) demonstrated domain-specific RoBERTa models outperforming GPT-4 in cross-domain PTSD pattern analysis, highlighting the critical balance between model specialization and interpretability.

Symptom Severity focuses on estimating the level of mental health condition intensity, particularly for depression, anxiety, and PTSD. Clinical evaluations reveal Med-PaLM 2's zero-shot depression scoring attains clinician-level alignment on interview data (Galatzer-Levy et al., 2023), though with limited PTSD generalizability. When benchmarked against specialized Transformers on DAIC-WOZ dataset (Gratch et al., 2014), ChatGPT and Llama-2 exhibit moderate efficacy (Arcan et al., 2024), suggesting domain-specific architectures retain advantages in structured assessments. Shifting attention to social media data, Aragon et al. (2024) proposed a pipeline that retrieves depression-relevant text, summarizes it according to the Beck Depression Inventory (BDI) (Jackson-Koku, 2016), and then utilizes LLMs to predict symptom severity, achieving performance similar to expert evaluations on certain measures. In a similar vein, Wang et al. (2024e) introduced an explainable depression detection system that leverages multiple open-source LLMs to generate BDI-based answers, reporting near state-of-the-art performance without additional training data. Cross-lingual extensions emerge through Skianis et al. (2024)'s framework enabling severity prediction across 6 languages and 2 mental conditions.

Cognition centers on identifying and understanding maladaptive thinking patterns, such as cogni-

Study	Text Granularity	Best Technique	NLP Task	Assessment Focus
		Symptom Detection		
Yang et al. (2023)	Single Post	Emotion Prompting	BC/MCC/EG	Multiple Symptoms
So et al. (2024)	Multi-turn Dialogue	Fine-Tuning	MLC/IE/SUM	Multiple Symptoms
Tu et al. (2024)	Multi-turn Dialogue	Few-Shot Prompting	MLC/IE/SUM	PTSD
Bao et al. (2024)	Single Post	Fine-Tuning	MLC/EG	Depression
Raihan et al. (2024)	Single Post	Few-Shot Prompting	MCC	Multiple Symptoms
Gyanendro Singh et al. (2024)	Posts From One User	Chain-of-Thought	IE/SUM	Suicidal Ideation
Uluslu et al. (2024)	Posts From One User	Role Prompting	IE/SUM	Suicidal Ideation
Yang et al. (2024a)	Single Post	Fine-Tuning	BC/MCC/EG	Multiple Symptoms
Xu et al. (2024)	Single Post	Fine-Tuning	BC/EG	Multiple Symptoms
Mohammadi et al. (2024)	Single Post	Few-Shot Prompting	MLC	Multiple Symptoms
Qiu et al. (2024)	Single Post	Fine-Tuning	MLC	Suicidal Ideation
Schirmer et al. (2024)	Single Post	Zero-Shot Prompting	BC	PTSD
		Symptom Severity		
Galatzer-Levy et al. (2023)	Multi-turn Dialogue	Zero-Shot Prompting	TR	Depression/PTSD
Arcan et al. (2024)	Multi-turn Dialogue	Zero-Shot Prompting	TR	Depression/Anxiety
Aragon et al. (2024)	Posts From One User	Zero-Shot Prompting	TR	Depression
Wang et al. (2024e)	Posts From One User	Zero-Shot Prompting	TR	Depression
Skianis et al. (2024)	Single Post	Zero-Shot Prompting	TR/MCC	Depression/Suicide
		Cognition		
Maddela et al. (2023)	Single Sentence	Few-Shot Prompting	MLC	Cognitive Distortions
Qi et al. (2024)	Single Post	Fine-Tuning	MLC	Cognitive Distortions
Wang et al. (2023)	Single Sentence	Few-Shot Prompting	MCC	Cognitive Distortions
Chen et al. (2023c)	Single-turn Dialogue	Zero-Shot Prompting	BC/MCC/EG	Cognitive Distortions
Gollapalli et al. (2023)	Single Post	Zero-Shot Prompting	MLC	Maladaptive Schemas
Jiang et al. (2024a)	Single Post	Zero-Shot Prompting	MCC/SUM	Cognitive Pathways
Lim et al. (2024)	Single-turn Dialogue	Multi-Agent Debate	MCC	Cognitive Distortions
		Behavior		
Li et al. (2024c)	Single Post	Zero-Shot Prompting	MLC/EG	Interpersonal Risk
Hoang et al. (2024)	Sentence From Dialogue	Few-Shot Prompting	MCC	MI-Adherent Behaviors
Sun et al. (2024)	Sentence From Dialogue	Zero-Shot Prompting	MCC	MI-Adherent Behaviors
Cohen et al. (2024)	Sentence From Dialogue	Zero-Shot Prompting	MCC	MI-Adherent Behaviors

Table 1: Comparison of Psychological Assessment Studies by Input Characteristics and Methodology. **MLC**: Multi-Label Classification, **IE**: Information Extraction, **SUM**: Summarization, **MCC**: Multi-Class Classification, **BC**: Binary Classification, **TR**: Text Regression, **EG**: Explanation Generation. Studies are categorized through text granularity, optimal technical approach (*Best Technique*), NLP task formulation, and specific assessment focus.

tive distortions and early maladaptive schemas, using LLMs. Maddela et al. (2023) introduced a cognitive distortion dataset and employed a few-shot strategy with GPT-3.5 to generate, classify, and reframe them, while Qi et al. (2024) constructed two Chinese social media benchmarks for cognitive distortion detection and suicidal risk assessment, demonstrating that fine-tuned LLMs are more closely than zero-/few-shot methods to supervised baselines. In a related effort, Wang et al. (2023) released the *C2D2* dataset containing 7,500 Chinese sentences with distorted thinking patterns. Expanding on detection methods, Chen et al. (2023c) proposed a *Diagnosis of Thought (DoT)* prompting approach for GPT-4 and ChatGPT, which breaks down patient utterances into factual versus subjective content and supports the generation of interpretable diagnostic reasoning. Beyond cognitive distortions, Gollapalli et al. (2023) investigated zero-shot approaches with GPT-3.5 to identify early maladaptive schemas in mental health forums, highlighting challenges in label interpretability and prompt sensitivity. Complementarily, Jiang et al. (2024a) presented a hierarchical classification and summarization pipeline to extract cognitive pathways from Chinese social media text, underscoring GPT-4's strong performance albeit with occasional hallucinations. Finally, Lim et al. (2024) introduced a multi-agent debate framework for cognitive distortion classification, reporting substantial gains in both accuracy and specificity by synthesizing multiple LLM opinions before forming a final verdict.

Behavior highlights how user actions–or in the case of Motivational Interviewing (MI), language itself–can serve as a measurable indicator of one's readiness for change. For instance, Li et al. (2024c) introduced the *MAIMS* framework, employing mental scales in a zero-shot setting to identify interpersonal risk factors on social media, thereby enhancing both interpretability and accuracy. In clinical dialogues, Hoang et al. (2024) demonstrated how LLMs can automatically detect a client's motivational direction (e.g., change versus sustain talk) and commitment level, offering valuable insights for MI-based interventions. Extending such analyses to bilingual settings, Sun et al. (2024) proposed the *BiMISC* dataset and prompt strategies that enable LLMs to code MI behaviors across multiple languages with expert-level performance. Lastly, Cohen et al. (2024) presented *MI-TAGS* for automated annotation of global MI scores, illustrating how context-sensitive modeling can approximate human annotations in psychotherapy transcripts.

Advanced research has evolved beyond foundational assessment tasks to emphasize novel methodological paradigms, bias mitigation, and domain-specific summarization frameworks. For instance, Yang et al. (2024b) introduced *PsychoGAT*—an interactive, game-based approach that transforms standardized psychometric instruments into engaging narrative experiences, improving psychometric reliability, construct validity, and user satisfaction when measuring constructs such as depression, cognitive distortions, and personality traits. In parallel, Wang et al. (2024d) systematically investigated potential biases in various LLMs across multiple mental health datasets, revealing that even high-performing models exhibit unfairness related to demographic factors. The authors proposed fairness-aware prompts to substantially reduce such biases without sacrificing predictive accuracy. Furthermore, Srivastava et al. (2024) presented the *PIECE* framework, which adopts a planning-based approach to domain-aligned counseling summarization, structuring and filtering conversation content before integrating domain knowledge.

3.2 Diagnosis

Static Diagnosis is based on a fixed set of data, typically derived from complete dialogues or social media posts. Galatzer-Levy et al. (2023) highlighted the effectiveness of Med-PaLM 2 in psychiatric condition assessment from patient interviews and clinical descriptions without specialized training. Similarly, Jiang et al. (2024c) showcased LLMs' superior performance on depression and anxiety detection on Russian datasets, particularly with noisy or small datasets. Hengle et al. (2024) evaluated PLMs and LLMs on multi-label classification in depression and anxiety, underscoring the ongoing challenges in applying LMs to mental health diagnostics. Besides, Lan et al. (2024c) introduced *DORIS*, a depression detection system integrating text embeddings with LLMs, utilizing symptom features, post-history, and mood course representations to make diagnostic predictions and generate explanatory outputs. Kuzmin et al. (2024) developed *ADOS-Copilot* for ASD diagnosis through diagnostic dialogues, employing In-context Enhancement, Interpretability Augmentation, and Adaptive Fusion based on real-world ADOS-2 clinic scenarios.

Dynamic Diagnosis involves real-time evaluation based on ongoing, interactive conversations between the patient and LLM, enabling more personalized and contextually relevant insights. Chen et al. (2023a) simulated psychiatrist-patient interactions with ChatGPT, in which the doctor chatbot focused on role, tasks, empathy, and questioning strategies, while the patient chatbot emphasized symptoms, language style, emotions, and resistance behaviors. Lan et al. (2024b) introduced the *Symptom-related and Empathy-related Ontology (SEO)*, grounded in DSM-5 and Helping Skills Theory, for depression diagnosis dialogues. Ren et al. (2024) dissected the doctor-patient relationship into psychologist's empathy and proactive guidance and introduced *WundtGPT* that integrated these elements. Lan et al. (2024a) further presented the *AMC*, a self-improving conversational agent system for depression diagnosis through simulated dialogues between patient and psychiatrist agents.

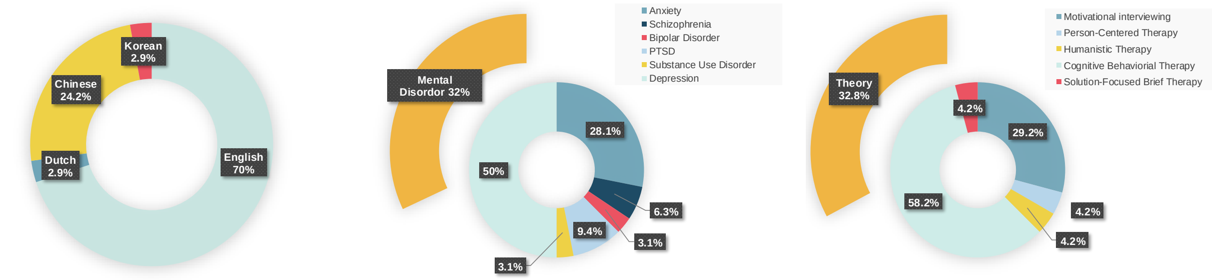
3.3 Treatment

LLM as a Virtual Therapist centers on leveraging LLMs to directly engage in therapeutic conversations, often adopting multi-turn dialogues that incorporate recognized psychotherapeutic frameworks. For instance, Xiao et al. (2024) proposed *HealMe* to facilitate cognitive reframing and empathetic support in line with established psychotherapy principles. Likewise, Nie et al. (2025) introduced *CaiTI*, a system embedded in everyday smart devices that conducts assessments of users' daily functioning and delivers psychotherapeutic interventions through adaptive dialogue flows. In a similar vein, Lee et al. (2024b) presented *CoCoA*, specializing in identifying and resolving cognitive distortions via dynamic memory mechanisms and CBT-based strategies, while Sharma et al. (2024) proposed a step-by-step approach guiding users to execute self-guided cognitive restructuring through multiple interactive sessions. Beyond standard CBT protocols, Kim et al. (2024) focused on aiding psychiatric patients in journaling their experiences, thereby offering richer clinical insights, whereas Lee et al. (2024c) developed a multi-round CBT dataset to refine LLMs for direct counseling-like interactions. Additionally, multi-agent frameworks like *MentalAgora* (Lee et al., 2024d) highlighted personalized mental health support by integrating multiple specialized agents, and Chen et al. (2024) further explored "mixed chain-of-psychotherapies" to combine various therapeutic methods, aiming to enhance the emotional support and customization delivered by chatbot interactions.

LLM as an Assistive Tool refrains from providing a holistic therapy role but instead offers targeted support such as rewriting suboptimal counselor responses, generating controlled reappraisal prompts, or aiding clinicians in specific tasks. For example, Welivita and Pu (2023) proposed to rewrite responses that violate MI principles into MI-adherent forms, ensuring more consistent therapeutic dialogue. Meanwhile, Sharma et al. (2023) and Maddela et al. (2023) focused on generating single-turn reframes of negative thoughts–often anchored in cognitive distortions–through controlled language attributes. On the detection side, Moon and Bhattacharyya (2024) built a multimodal pipeline to identify depression and provide CBT-style replies, albeit with an emphasis on technological assistance rather than full-fledged therapy. In the Chinese context, Lin et al. (2024) combined cognitive distortion detection with "positive reconstruction," demonstrating a single-round rewrite approach for negative or distorted statements, while Na (2024) showcased a structured Q&A format that offers professional yet succinct CBT-based responses. From a knowledge-distillation angle, Brown et al. (2024) demonstrated how smaller models could replicate GPT-4's MI-style reflective statements, and Zhan et al. introduced a lighter-weight framework *RESORT* to guide smaller LLMs toward effective cognitive reappraisal prompts, thus enabling broader accessibility of self-help tools.

LLM as Simulated Patients for Clinician Education pivots toward generating synthetic yet realistic patient behaviors or multi-level feedback to train or support mental health practitioners. For instance, Chaszczewicz et al. (2024) leveraged LLMs to deliver multi-tier feedback on novice peer counselors' conversational skills, significantly reducing the need for continuous expert oversight. Similarly, Wang et al. (2024b) introduced LLM-driven patient simulations that help trainees practice CBT core skills in a controlled, repeatable setup. In the realm of assessing therapy quality, Yosef et al. (2024) showcased a digital patient system to evaluate MI sessions, employing AI-generated transcripts to differentiate novice, intermediate, and expert therapeutic skill levels. Complementarily, Louie et al. (2024) offered *Roleplay-doh*, a pipeline wherein domain experts craft specialized principles that guide LLM-based role-playing agents, thereby providing customizable training for new therapists.

LLM for Evaluation and Quality Analysis targets the appraisal of therapy dialogue, counselor techniques, and treatment processes, typically without delivering direct interventions to clients. For instance, Lee et al. (2024e) augmented crisis counseling outcome prediction by fusing annotated counseling strategies with LLM-derived features, achieving substantially improved accuracy. In the Chinese context, Zhang et al. (2024) introduced *CPsyCoun*, employing reports-based dialogue reconstruction and automated evaluation to verify counseling realism and professionalism. Beyond single-session analyses, Wang et al. (2024a) used simulated clients to assess perceived therapy outcomes, while Chiu et al. (2024) created the *BOLT* framework for systematically comparing LLM-based therapy behaviors with high- and low-quality human sessions. Further extending to online counseling, Li et al. (2024a) proposed an LLM-based approach to measure therapeutic alliance, whereas Shapira and Alfi-Yogev (2024) delineated therapist self-disclosure classification as a new NLP task. In the MI domain, Sun et al. (2024) and Cohen et al. (2024) collected bilingual transcripts to systematically annotate therapist–client exchanges for behavior coding and global scores, respectively. Additionally, multi-session perspectives emerge in Na et al. (2024), who proposed *IPAEval* to track long-term progress from the client's viewpoint, and Nguyen et al. (2024) analyzed conversation redi-



(a) The Proportion of Language Coverage **(b)** The Proportion of Mental Disorder **(c)** The Proportion of Psychotherapy Theories

Figure 3: Distribution Analysis of The Current Landscape. Panel (a) indicates English is the predominant language (70%), with Chinese also represented (24.2%). Panel (b) shows that 32% of studies address mental disorders, with Depression and Anxiety being the most common topics within this group. Panel (c) reveals that 32.8% of studies incorporate psychotherapy theories, where CBT is the most frequently applied.

rection and its impact on patient–therapist alliance over multiple sessions. Finally, Iftikhar et al. (2024) and Zhang et al. (2025) explored the disparities between LLM- and human-led CBT sessions, highlighting gaps such as empathy and cultural nuance while also introducing *CBT-Bench* to probe LLMs' deeper psychotherapeutic competencies.

4 Current Landscape

4.1 Overview

Our survey encompasses a total of 69 studies in the field of LLMs in psychotherapy. Specifically, 33 studies address assessment, 9 focus on diagnosis, and 32 concentrate on treatment, with 5 studies overlapping across these dimensions. Approximately 74% of the studies employed commercial large language models, while about 77% used prompt-based techniques. This distribution highlights an imbalance in research focus across different stages of the psychotherapy process and reflects a heavy reliance on commercial models and prompt technologies.

Figure 3 presents a comprehensive analysis of the current research landscape in this field. Panel **(a)** reveals a significant linguistic bias in existing studies, with English-language corpora dominates. While there are limited studies involving Korean and Dutch languages, this highlights a substantial gap in multilingual research approaches. Panel **(b)** quantitatively demonstrates the distribution of mental health research focuses. Mental disorder-related studies constitute 32% of the total research corpus (represented by the orange outer ring). Within this subset, depression-focused research accounts for 50% of mental disorder studies, followed by anxiety-related research. This distribution indicates a concerning imbalance, where common conditions receive disproportionate attention while more complex disorders, such as bipolar disorder, remain understudied. The analysis of psychotherapy theories in panel **(c)** uncovers another critical gap in the field. Only 32.8% of the studies incorporate psychotherapy theories in their methodological approach. Notably, emerging therapeutic frameworks, such as humanistic therapy, are particularly underrepresented in current research applications.

4.2 Fragmented Approaches

Traditionally, LLM-based psychotherapy tools have addressed assessment, diagnosis, and treatment separately. Recently, a few studies have started to explore integrative approaches spanning multiple stages. Despite these emerging integrative efforts, the systems remain limited, typically addressing only two stages without achieving full continuity. Additionally, fragmentation occurs not only across these three dimensions but also at more granular levels; for example, some methodologies are narrowly focused on assessing single disorders (Tu et al., 2024; Bao et al., 2024), further limiting their applicability and integration potential in broader psychotherapy contexts.

4.3 Critical Issues and Risks

Dynamic Symptom Representation. Psychotherapy commonly involves shifting symptoms, comorbidities, and nuanced patient experiences, making static or single-label predictions insufficient. Current models fail to adequately capture multi-label conditions and temporal symptom fluctuations, leading to incomplete or inaccurate assessments (Lee et al., 2024a).

Linguistic Resource Bias and Translation Critique. Most psychotherapy-oriented LLMs are trained primarily on English datasets, with some researchers attempting to expand linguistic coverage through translation. However, recent studies highlight significant cultural specificity in mental health disorders (Watters, 2010; Abdelkadir et al., 2024), making direct translation of datasets unreliable for accurately capturing psychological nuances across different cultures.

Diagnostic Risks. Current approaches to automated diagnosis often struggle to gain acceptance among clinical practitioners due to concerns about reliability and patient safety. Despite the psychology community increasingly favoring transdiagnostic methods (Dalgleish et al., 2020), a segment of NLP researchers continues to emphasize diagnosis-specific studies, indicating a notable divergence in research priorities.

5 Future Directions

Continuous Multi-Stage Modeling. Psychotherapy inherently involves continuous interactions that progress through assessment, diagnosis, and intervention phases. Existing research indicates that several leading foundation models exhibit negligible hallucination issues in the medical domain (Kim et al., 2025), providing a promising foundation for integrating these stages, as minimizing hallucinations is crucial for ensuring the accuracy and reliability required for continuous patient state tracking across psychotherapy stages. Future models should aim for an evolving representation of patient states, ensuring consistency and coherence across the entire therapeutic process rather than handling stages as isolated segments.

Real-Time Adaptability Grounded in Psychological Theory. The development of real-time adaptive strategies represents a significant step beyond current static models. Current technical advancements, such as retrieval-augmented generation and long-context memory techniques (Jo et al., 2024), provide the necessary technical foundations for such strategies. Instead of simply reproducing patterns found in pre-collected dialogue datasets, future LLM applications should incorporate these advanced contextual memory mechanisms informed by established psychological theories. Such systems would dynamically interpret patient cues, adjusting interventions immediately in response to subtle shifts in emotional and cognitive states. This approach significantly surpasses mere language-style mimicry achieved through simple fine-tuning on existing datasets, enabling deeper, theoretically-informed therapeutic engagement.

Broadening Scope of Disorders and Therapeutic Approaches. Future research should prioritize diversification in terms of both disorders addressed and therapeutic methodologies employed. The current focus on common disorders like anxiety and depression has led to an imbalanced research landscape. There is a pressing need to incorporate underrepresented conditions, such as bipolar disorder and personality disorders, alongside a broader spectrum of psychotherapeutic frameworks, including psychodynamic (Shedler, 2010) and existential-humanistic approaches (Schneider and Krug, 2010). Such expansion would help address existing blind spots, contributing to a more inclusive and comprehensive application of LLMs in psychotherapy.

6 Conclusion

LLMs hold significant promise for revolutionizing psychotherapy by enhancing assessment, diagnosis, and treatment. However, the current landscape reveals critical limitations: research is often fragmented across these stages, exhibits notable biases in linguistic and disorder coverage, and underutilizes diverse psychotherapeutic theories. To overcome these challenges, future work must focus on creating continuous, multi-stage models that are grounded in psychological theory and capable of real-time adaptation to evolving patient states. Expanding the scope of addressed disorders and therapeutic modalities will be crucial for developing LLM-driven psychotherapy tools.

Limitations

We remind the readers that this survey has the following limitations: 1) The studies reviewed primarily focus on the application of LLMs in psychotherapy, and there may be relevant research in adjacent fields or interdisciplinary domains that was not included. 2) Due to the rapidly evolving nature of this area, some recent advancements may not be captured. The scope of this survey is limited to the available literature and may overlook emerging trends or unpublished findings. 3) While we provide a taxonomy of LLM applications in psychotherapy, this framework may not fully encompass the complexity of real-world clinical settings or the diverse range of therapeutic approaches currently in practice.

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.

American Psychiatric Association. 2006. Evidence-based practice in psychology. *The American Psychologist*, 61(4):271–285.

American Psychiatric Association. 2022. *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. text revision edition. American Psychiatric Publishing, Arlington, VA.

Mario Aragon, Javier Parapar, and David E Losada. 2024. Delving into the depths: Evaluating depression severity through BDI-biased summaries. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 12–22, St. Julians, Malta. Association for Computational Linguistics.

Mihael Arcan, David-Paul Niland, and Fionn Delahunty. 2024. An assessment on comprehending mental health through large language models. *Preprint*, arXiv:2401.04592.

Eliseo Bao, Anxo Pérez, and Javier Parapar. 2024. Explainable depression symptom detection in social media. *Health Information Science and Systems*, 12(1):47.

Michael Barkham, William B Stiles, and David A Shapiro. 1993. The shape of change in psychotherapy: longitudinal assessment of personal problems. *Journal of consulting and clinical psychology*, 61(4):667.

David H Barlow. 2021. *Clinical handbook of psychological disorders: A step-by-step treatment manual*. Guilford publications.

Andrew Brown, Jiading Zhu, Mohamed Abdelwahab, Alec Dong, Cindy Wang, and Jonathan Rose. 2024. Generation, distillation and evaluation of motivational interviewing-style reflections with a foundational language model. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1241–1252, St. Julian's, Malta. 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.

Siyuan Chen, Cong Ming, Zhiling Zhang, Yanyi Chen, Kenny Q. Zhu, and Mengyue Wu. 2024. Mixed chain-of-psychotherapies for emotional support chatbot. *Preprint*, arXiv:2409.19533.

Siyuan Chen, Mengyue Wu, Kenny Q. Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023a. Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *Preprint*, arXiv:2305.13614.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023b. SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183, Singapore. Association for Computational Linguistics.

Zhiyu Chen, Yujie Lu, and William Wang. 2023c. 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.

Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A computational framework for behavioral assessment of llm therapists. *Preprint*, arXiv:2401.00820.

Ben Cohen, Moreah Zisquit, Stav Yosef, Doron Friedman, and Kfir Bar. 2024. Motivational interviewing transcripts annotated with global scores. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11642–11657, Torino, Italia. ELRA and ICCL.

Ronald Jay Cohen, Mark E Swerdlik, and Suzanne M Phillips. 1996. *Psychological testing and assessment: An introduction to tests and measurement*. Mayfield Publishing Co.

Nick Craddock and Laurence Mynors-Wallis. 2014. Psychiatric diagnosis: impersonal, imperfect and important. *The British Journal of Psychiatry*, 204(2):93–95.

Tim Dalgleish, Melissa Black, David Johnston, and Anna Bevan. 2020. Transdiagnostic approaches to mental health problems: Current status and future directions. *Journal of consulting and clinical psychology*, 88(3):179.

Rachel Elvins and Jonathan Green. 2008. The conceptualization and measurement of therapeutic alliance: An empirical review. *Clinical psychology review*, 28(7):1167–1187.

Iony D Ezawa and Steven D Hollon. 2023. Cognitive restructuring and psychotherapy outcome: A meta-analytic review. *Psychotherapy*, 60(3):396.

Isaac R. Galatzer-Levy, Daniel McDuff, Vivek Natarajan, Alan Karthikesalingam, and Matteo Malgaroli. 2023. The capability of large language models to measure psychiatric functioning. *Preprint*, arXiv:2308.01834.

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.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Gary Groth-Marnat. 2009. *Handbook of psychological assessment*. John Wiley & Sons.

Loitongbam Gyanendro Singh, Junyu Mao, Rudra Mutalik, and Stuart E. Middleton. 2024. Extracting and summarizing evidence of suicidal ideation in social media contents using large language models. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 218–226, St. Julians, Malta. 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.

Van Hoang, Eoin Rogers, and Robert Ross. 2024. How can client motivational language inform psychotherapy agents? In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 23–40, St. Julians, Malta. Association for Computational Linguistics.

Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Hongbin Na, Yi han Sheu, Peilin Zhou, Lauren V. Moran, Sophia Ananiadou, Andrew Beam, and John Torous. 2024. Large language models in mental health care: a scoping review.

Yining Hua, Hongbin Na, Zehan Li, Fenglin Liu, Xiao Fang, David A. Clifton, and John Torous. 2025. A scoping review of large language models for generative tasks in mental health care. *NPJ Digital Medicine*, 8.

Zainab Iftikhar, Sean Ransom, Amy Xiao, and Jeff Huang. 2024. Therapy as an nlp task: Psychologists' comparison of llms and human peers in cbt. *Preprint*, arXiv:2409.02244.

Gordon Jackson-Koku. 2016. Beck depression inventory. *Occupational medicine*, 66(2):174–175.

Amanda Jensen-Doss and Kristin M Hawley. 2011. Understanding clinicians' diagnostic practices: Attitudes toward the utility of diagnosis and standardized diagnostic tools. *Administration and Policy in Mental Health and Mental Health Services Research*, 38:476–485.

Meng Jiang, Yi Jing Yu, Qing Zhao, Jianqiang Li, Changwei Song, Hongzhi Qi, Wei Zhai, Dan Luo, Xiaoqin Wang, Guanghui Fu, and Bing Xiang Yang. 2024a. Ai-enhanced cognitive behavioral therapy: Deep learning and large language models for extracting cognitive pathways from social media texts. *Preprint*, arXiv:2404.11449.

Meng Jiang, Qing Zhao, Jianqiang Li, Fan Wang, Tianyu He, Xinyan Cheng, Bing Xiang Yang, Grace WK Ho, and Guanghui Fu. 2024b. A generic review of integrating artificial intelligence in cognitive behavioral therapy. *arXiv preprint arXiv:2407.19422*.

Yi Jiang, Qingyang Shen, Shuzhong Lai, Shunyu Qi, Qian Zheng, Lin Yao, Yueming Wang, and Gang Pan. 2024c. Copiloting diagnosis of autism in real clinical scenarios via llms. *Preprint*, arXiv:2410.05684.

Eunkyung Jo, Yuin Jeong, SoHyun Park, Daniel A Epstein, and Young-Ho Kim. 2024. Understanding the impact of long-term memory on self-disclosure with large language model-driven chatbots for public health intervention. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Robert M Kaplan and Dennis P Saccuzzo. 2001. *Psychological testing: Principles, applications, and issues*. Wadsworth/Thomson Learning.

Shadia Kawa and James Giordano. 2012. A brief historicity of the diagnostic and statistical manual of mental disorders: Issues and implications for the future of psychiatric canon and practice. *Philosophy, Ethics, and Humanities in Medicine*, 7(1):2.

Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su-Woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. 2024. Mindfuldiary: Harnessing large language model to support psychiatric patients' journaling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, et al. 2025. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*.

Gleb Kuzmin, Petr Strepetov, Maksim Stankevich, Artem Shelmanov, and Ivan Smirnov. 2024. Mental disorders detection in the era of large language models. *Preprint*, arXiv:2410.07129.

Michael J Lambert. 2013. *Bergin and Garfield's handbook of psychotherapy and behavior change*. John Wiley & Sons.

Kunyao Lan, Bingrui Jin, Zichen Zhu, Siyuan Chen, Shu Zhang, Kenny Q. Zhu, and Mengyue Wu. 2024a. Depression diagnosis dialogue simulation: Self-improving psychiatrist with tertiary memory. *Preprint*, arXiv:2409.15084.

Kunyao Lan, Cong Ming, Binwei Yao, Lu Chen, and Mengyue Wu. 2024b. Towards reliable and empathetic depression-diagnosis-oriented chats. *Preprint*, arXiv:2404.05012.

Xiaochong Lan, Yiming Cheng, Li Sheng, Chen Gao, and Yong Li. 2024c. Depression detection on social media with large language models. *Preprint*, arXiv:2403.10750.

Hannah R Lawrence, Renee A Schneider, Susan B Rubin, Maja J Matarić, Daniel J McDuff, and Megan Jones Bell. 2024. The opportunities and risks of large language models in mental health. *JMIR Mental Health*, 11(1):e59479.

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.

Suyeon Lee, Jieun Kang, Harim Kim, Kyoung-Mee Chung, Dongha Lee, and Jinyoung Yeo. 2024b. Cocoa: Cbt-based conversational counseling agent using memory specialized in cognitive distortions and dynamic prompt. *Preprint*, arXiv:2402.17546.

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. 2024c. 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.

Yeonji Lee, Sangjun Park, Kyunghyun Cho, and JinYeong Bak. 2024d. Mentalagora: A gateway to advanced personalized care in mental health through multi-agent debating and attribute control. *Preprint*, arXiv:2407.02736.

Younghun Lee, Dan Goldwasser, and Laura Schwab Reese. 2024e. Towards understanding counseling conversations: Domain knowledge and large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2032–2047, St. Julian's, Malta. Association for Computational Linguistics.

Anqi Li, Yu Lu, Nirui Song, Shuai Zhang, Lizhi Ma, and Zhenzhong Lan. 2024a. Automatic evaluation for mental health counseling using llms. *Preprint*, arXiv:2402.11958.

Ruosen Li, Zimu Wang, Son Tran, Lei Xia, and Xinya Du. 2024b. Meqa: A benchmark for multi-hop event-centric question answering with explanations. In *Advances in Neural Information Processing Systems*, volume 37, pages 126835–126862. Curran Associates, Inc.

Wenyu Li, Yinuo Zhu, Xin Lin, Ming Li, Ziyue Jiang, and Ziqian Zeng. 2024c. Zero-shot explainable mental health analysis on social media by incorporating mental scales. In *Companion Proceedings of the ACM on Web Conference 2024*, WWW '24, page 959–962, New York, NY, USA. Association for Computing Machinery.

Sehee Lim, Yejin Kim, Chi-Hyun Choi, Jy-yong Sohn, and Byung-Hoon Kim. 2024. ERD: A framework for improving LLM reasoning for cognitive distortion classification. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 292–300, Mexico City, Mexico. Association for Computational Linguistics.

Shuya Lin, Yuxiong Wang, Jonathan Dong, and Shiguang Ni. 2024. Detection and positive reconstruction of cognitive distortion sentences: Mandarin dataset and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6686–6701, Bangkok, Thailand. Association for Computational Linguistics.

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.

Jiayuan Ma, Hongbin Na, Zimu Wang, Yining Hua, Yue Liu, Wei Wang, and Ling Chen. 2025. Detecting conversational mental manipulation with intent-aware prompting. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9176–9183, Abu Dhabi, UAE. Association for Computational Linguistics.

Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. 2023. Training models to generate, recognize, and reframe unhelpful thoughts. In *Proceedings of the 61st Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 13641–13660, Toronto, Canada. Association for Computational Linguistics.

Matteo Malgaroli, Thomas D Hull, James M Zech, and Tim Althoff. 2023. Natural language processing for mental health interventions: a systematic review and research framework. *Translational Psychiatry*, 13(1):309.

Garry Martin and Joseph J Pear. 2019. *Behavior modification: What it is and how to do it*. Routledge.

Seyedali Mohammadi, Edward Raff, Jinendra Malekar, Vedant Palit, Francis Ferraro, and Manas Gaur. 2024. WellDunn: On the robustness and explainability of language models and large language models in identifying wellness dimensions. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 364–388, Miami, Florida, US. Association for Computational Linguistics.

Palash Moon and Pushpak Bhattacharyya. 2024. We care: Multimodal depression detection and knowledge infused mental health therapeutic response generation. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 296–310, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLPAI).

Hongbin Na. 2024. CBT-LLM: A Chinese large language model for cognitive behavioral therapy-based mental health question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2930–2940, Torino, Italia. ELRA and ICCL.

Hongbin Na, Tao Shen, Shumao Yu, and Ling Chen. 2024. Multi-session client-centered treatment outcome evaluation in psychotherapy. *Preprint*, arXiv:2410.05824.

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.

Jingping Nie, Hanya (Vera) Shao, Yuang Fan, Qijia Shao, Haoxuan You, Matthias Preindl, and Xiaofan Jiang. 2025. Llm-based conversational ai therapist for daily functioning screening and psychotherapeutic intervention via everyday smart devices. *ACM Trans. Comput. Healthcare*. Just Accepted.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2023. When does in-context learning fall short and why? a study on specification-heavy tasks. *Preprint*, arXiv:2311.08993.

Michael Robert Phillips, Qijie Shen, Xiehe Liu, Sonya Pritzker, David Streiner, Ken Conner, and Gonghuan Yang. 2007. Assessing depressive symptoms in persons who die of suicide in mainland china. *Journal of Affective Disorders*, 98(1-2):73–82.

James O Prochaska and John C Norcross. 2018. *Systems of psychotherapy: A transtheoretical analysis*. Oxford University Press.

Hongzhi Qi, Qing Zhao, Jianqiang Li, Changwei Song, Wei Zhai, Dan Luo, Shuo Liu, Yi Jing Yu, Fan Wang, Huijing Zou, Bing Xiang Yang, and Guanghui Fu. 2024. Supervised learning and large language model benchmarks on mental health datasets: Cognitive distortions and suicidal risks in chinese social media. *Preprint*, arXiv:2309.03564.

Lu Qian, Yuqi Wang, Zimu Wang, Haiyang Zhang, Wei Wang, Ting Yu, and Anh Nguyen. 2024. Domain-specific guided summarization for mental health posts. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 150–159, Tokyo, Japan. Tokyo University of Foreign Studies.

Huachuan Qiu, Lizhi Ma, and Zhenzhong Lan. 2024. 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.

Nishat Raihan, Sadiya Sayara Chowdhury Puspo, Shafkat Farabi, Ana-Maria Bucur, Tharindu Ranasinghe, and Marcos Zampieri. 2024. MentalHelp: A multi-task dataset for mental health in social media. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11196–11203, Torino, Italia. ELRA and ICCL.

Chenyu Ren, Yazhou Zhang, Daihai He, and Jing Qin. 2024. Wundtgpt: Shaping large language models to be an empathetic, proactive psychologist. *Preprint*, arXiv:2406.15474.

Joseph Rencic, Steven J Durning, Eric Holmboe, and Larry D Gruppen. 2016. Understanding the assessment of clinical reasoning. *Assessing competence in professional performance across disciplines and professions*, pages 209–235.

Günter Schiepek, Wolfgang Aichhorn, Martin Gruber, Guido Strunk, Egon Bachler, and Benjamin Aas. 2016. Real-time monitoring of psychotherapeutic processes: concept and compliance. *Frontiers in psychology*, 7:604.

Miriam Schirmer, Tobias Leemann, Gjergji Kasneci, Jürgen Pfeffer, and David Jurgens. 2024. The language of trauma: Modeling traumatic event descriptions across domains with explainable AI. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13224–13242, Miami, Florida, USA. Association for Computational Linguistics.

Kirk J Schneider and Orah T Krug. 2010. *Existential-humanistic therapy*. American Psychological Association Washington, DC.

Natalie Shapira and Tal Alfi-Yogev. 2024. Therapist self-disclosure as a natural language processing task. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 61–73, St. Julians, Malta. 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 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Jonathan Shedler. 2010. The efficacy of psychodynamic psychotherapy. *American psychologist*, 65(2):98.

Konstantinos Skianis, John Pavlopoulos, and A. Seza Doğruöz. 2024. Severity prediction in mental health: Llm-based creation, analysis, evaluation of a novel multilingual dataset. *Preprint*, arXiv:2409.17397.

Jae-hee So, Joonhwan Chang, Eunji Kim, Junho Na, JiYeon Choi, Jy-yong Sohn, Byung-Hoon Kim, and Sang Hui Chu. 2024. Aligning large language models for enhancing psychiatric interviews through symptom delineation and summarization: Pilot study. *JMIR Form Res*, 8:e58418.

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.

Dorothy E Stubbe. 2018. The therapeutic alliance: The fundamental element of psychotherapy. *Focus*, 16(4):402–403.

Xin Sun, Jiahuan Pei, Jan de Wit, Mohammad Aliannejadi, Emiel Krahmer, Jos T.P. Dobber, and Jos A. Bosch. 2024. Eliciting motivational interviewing skill codes in psychotherapy with LLMs: A bilingual dataset and analytical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5609–5621, Torino, Italia. ELRA and ICCL.

Alan R. Teo. 2010. A new form of social withdrawal in japan: a review of hikikomori. *International Journal of Social Psychiatry*, 56(2):178–185. PMID: 19567455.

Sichang Tu, Abigail Powers, Natalie Merrill, Negar Fani, Sierra Carter, Stephen Doogan, and Jinho D. Choi. 2024. Automating PTSD diagnostics in clinical interviews: Leveraging large language models for trauma assessments. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 644–663, Kyoto, Japan. Association for Computational Linguistics.

Ahmet Yavuz Uluslu, Andrianos Michail, and Simon Clematide. 2024. Utilizing large language models to identify evidence of suicidality risk through analysis of emotionally charged posts. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 264–269, St. Julians, Malta. Association for Computational Linguistics.

Bruce E Wampold and Zac E Imel. 2015. *The great psychotherapy debate: The evidence for what makes psychotherapy work*. Routledge.

Bichen Wang, Pengfei Deng, Yanyan Zhao, and Bing Qin. 2023. 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.

Jiashuo Wang, Yang Xiao, Yanran Li, Changhe Song, Chunpu Xu, Chenhao Tan, and Wenjie Li. 2024a. Towards a client-centered assessment of llm therapists by client simulation. *Preprint*, arXiv:2406.12266.

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. 2024b. 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.

Yuqi Wang, Zimu Wang, Nijia Han, Wei Wang, Qi Chen, Haiyang Zhang, Yushan Pan, and Anh Nguyen. 2024c. Knowledge distillation from monolingual to multilingual models for intelligent and interpretable multilingual emotion detection. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 470–475, Bangkok, Thailand. Association for Computational Linguistics.

Yuqing Wang, Yun Zhao, Sara Alessandra Keller, Anne de Hond, Marieke M. van Buchem, Malvika Pillai, and Tina Hernandez-Boussard. 2024d. Unveiling and mitigating bias in mental health analysis with large language models. *Preprint*, arXiv:2406.12033.

Yuxi Wang, Diana Inkpen, and Prasadith Kirinde Gamaarachchige. 2024e. Explainable depression detection using large language models on social media data. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 108–126, St. Julians, Malta. Association for Computational Linguistics.

Zimu Wang, Lei Xia, Wei Wang, and Xinya Du. 2024f. Document-level causal relation extraction with knowledge-guided binary question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16944–16955, Miami, Florida, USA. Association for Computational Linguistics.

Monika A. Waszczuk, Mark Zimmerman, Camilo Ruggero, Kaiqiao Li, Annmarie MacNamara, Anna Weinberg, Greg Hajcak, David Watson, and Roman Kotov. 2017. What do clinicians treat: Diagnoses or symptoms? the incremental validity of a symptom-based, dimensional characterization of emotional disorders in predicting medication prescription patterns. *Comprehensive Psychiatry*, 79:80–88. Advances in Transdiagnostic Psychopathology Research.

Ethan Watters. 2010. *Crazy Like Us: The Globalization of the American Psyche*. Free Press.

Anuradha Welivita and Pearl Pu. 2023. Boosting distress support dialogue responses with motivational interviewing strategy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5411–5432, Toronto, Canada. Association for Computational Linguistics.

World Health Organization. 2019. International classification of diseases, eleventh revision (icd-11). Licensed under Creative Commons Attribution-NoDerivatives 3.0 IGO licence (CC BY-ND 3.0 IGO).

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.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(1).

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.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024a. Mentallama: Interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024*, WWW '24, page 4489–4500, New York, NY, USA. Association for Computing Machinery.

Qisen Yang, Zekun Wang, Honghui Chen, Shenzhi Wang, Yifan Pu, Xin Gao, Wenhao Huang, Shiji Song, and Gao Huang. 2024b. 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.

Stav Yosef, Moreah Zisquit, Ben Cohen, Anat Klomek Brunstein, Kfir Bar, and Doron Friedman. 2024. Assessing motivational interviewing sessions with AI-generated patient simulations. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 1–11, St. Julians, Malta. Association for Computational Linguistics.

Hongli Zhan, Allen Zheng, Yoon Kyung Lee, Jina Suh, Junyi Jessy Li, and Desmond Ong. Large language models are capable of offering cognitive reappraisal, if guided. In *First Conference on Language Modeling*.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024. CPsyCoun: A report-based multi-turn dialogue reconstruction and evaluation framework for Chinese psychological counseling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13947–13966, Bangkok, Thailand. 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. 2025. 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.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2025. A survey of large language models. *Preprint*, arXiv:2303.18223.