

# PsychEval: A Multi-Session and Multi-Therapy Benchmark for High-Realism AI Psychological Counselor

Qianjun Pan<sup>1\*</sup>, Junyi Wang<sup>1\*</sup>, Jie Zhou<sup>1†</sup>, Yutao Yang<sup>1</sup>, Junsong Li<sup>1</sup>, Kaiyin Xu<sup>1</sup>,  
Yougen Zhou<sup>1</sup>, Yihan Li<sup>1</sup>, JingYuan Zhao<sup>1</sup>, Qin Chen<sup>1†</sup>,  
Ningning Zhou<sup>2</sup>, Kai Chen<sup>3</sup>, Liang He<sup>1</sup>,

<sup>1</sup>School of Computer Science and Technology, East China Normal University,

<sup>2</sup>School of Psychology and Cognitive Science, East China Normal University,

<sup>3</sup>Shanghai AI Laboratory

## Abstract

To develop a reliable AI for psychological assessment, we introduce PsychEval, a multi-session, multi-therapy, and highly realistic benchmark designed to address three key challenges: **1) Can we train a highly realistic AI counselor?** Realistic counseling is a longitudinal task requiring sustained memory and dynamic goal tracking. We propose a multi-session benchmark (spanning 6-10 sessions across three distinct stages) that demands critical capabilities such as memory continuity, adaptive reasoning, and longitudinal planning. The dataset is annotated with extensive professional skills, comprising over 677 meta-skills and 4577 atomic skills. **2) How to train a multi-therapy AI counselor?** While existing models often focus on a single therapy, complex cases frequently require flexible strategies among various therapies. We construct a diverse dataset covering five therapeutic modalities alongside an integrative therapy with a unified three-stage clinical framework across six core psychological topics. **3) How to systematically evaluate an AI counselor?** We establish a holistic evaluation framework with 18 therapy-specific and therapy-shared metrics across Client-Level and Counselor-Level dimensions. We also construct over 2,000 diverse client profiles. Extensive experimental analysis fully validates the superior quality and clinical fidelity of our dataset. Our datasets and evaluation framework are available at <https://github.com/ECNU-ICALK/PsychEval>.

## 1 Introduction

The intersection of Artificial Intelligence and mental healthcare has evolved from a theoretical curiosity into a burgeoning subfield of computational psychiatry (Na et al., 2025; Hua et al., 2025). Historically, computer-mediated therapy was confined to rule-based systems like ELIZA, which relied on

rigid pattern matching and lacked genuine semantic understanding. The advent of Large Language Models (LLMs) (Liu et al., 2024; Brown et al., 2020) has fundamentally transformed this landscape: trained on internet-scale corpora and fine-tuned via reinforcement learning, modern LLMs exhibit emergent capabilities not only to parse complex syntax but also to generate responses demonstrating empathy, reasoning, and context retention (Fu et al., 2023; Liu et al., 2023a). This shift enables the potential to democratize psychological support through scalable, autonomous agents (Nguyen et al., 2025; Lai et al., 2023).

However, transitioning from general-purpose chatbots to clinically reliable AI counselors poses significant challenges. Therapeutic alliance, which depends on trust, confidentiality, and ethical navigation, requires more than fluency. Generic LLMs often lack clinical alignment, risking hallucinations, toxic positivity, and critical safety failures in crisis situations. To mitigate these risks, recent research has introduced specialized models such as SoulChat (Chen et al., 2023), which optimizes for emotional resonance, and PsyLLM (Hu et al., 2025), which incorporates diagnostic reasoning. Furthermore, advanced frameworks like HealMe (Xiao et al., 2024) and SMILE (Qiu et al., 2024) employ specialized architectural constraints and safety modules to ensure adherence to clinical guidelines.

Despite these advances, a critical gap remains between current research and real-world psychological counseling. **First**, real counseling is not a static QA task but a dynamic, longitudinal process that requires memory continuity across sessions, tracking of evolving goals, and progression through distinct stages—from case conceptualization to intervention and consolidation (Fishman, 2013). **Second**, effective counseling is rarely one-size-fits-all; clinicians draw on diverse theoretical modalities (e.g., CBT, Psychoanalysis, Humanistic) (Behan, 2022; Schiepek and Pincus, 2023; Zarbo et al., 2016) and

<sup>1</sup>These authors contributed equally to this work.

<sup>2</sup>Corresponding authors, {jzhou, qchen}@cs.ecnu.edu.cn

often integrate approaches tailored to each client. **Third**, systematic evaluation methods are lacking. Standard NLP metrics like BLEU and ROUGE are widely criticized in clinical contexts for capturing only surface lexical overlap, not therapeutic quality (Liu et al., 2023b). They fail to penalize “over-validation” (indiscriminate agreement with distorted thoughts) or assess the pragmatic appropriateness of interventions (Iftikhar et al., 2025).

In contrast, existing AI mental health benchmarks primarily assess single-turn dialogues, measuring generic empathy or safety in isolation. Datasets like PsyQA (Sun et al., 2021) offer valuable resources for advisory capabilities but consist largely of fragmented, single-session interactions or synthetic chat logs lacking the “arc of therapy.” For instance, they typically ignore the temporal progression of a client’s symptoms and the gradual deepening of the therapeutic bond over weeks. Such static evaluations fail to capture therapy’s temporal dynamics (i.e., how clients evolve) and the theoretical versatility needed for complex cases. Consequently, an AI excelling on single-turn empathy benchmarks may fail catastrophically when managing long-term therapeutic alliances or adhering to specific clinical frameworks.

To bridge the gap between static benchmarking and dynamic clinical reality, we introduce PsychEval, a comprehensive, multi-session, and multi-therapy benchmark designed to rigorously assess AI counselors. Unlike previous datasets, PsychEval simulates the full trajectory of counseling. It incorporates a structured three-stage clinical flow: *Case Conceptualization*, *Core Intervention*, and *Consolidation*, spanning 6 to 10 sessions per client. To ensure theoretical depth, we construct a diverse dataset covering five major therapeutic modalities and an Integrative approach, grounded in over 2,000 rigorous client profiles across six core psychological topics.

Through PsychEval, we aim to answer three fundamental research questions:

**RQ1: Can we train a highly realistic AI counselor?** To capture therapy’s longitudinal nature, we ground training in high-fidelity, empirically sourced case reports rather than synthetic chat logs and integrate a hierarchical skill taxonomy that maps strategic meta-skills to atomic actions, providing cognitive scaffolding for sustained logical coherence and precise goal-tracking across multi-session trajectories.

**RQ2: How to train a multi-school AI coun-**

**selor?** To enable the mastery of distinct therapeutic orientations, we construct a comprehensive dataset encompassing five specific therapies alongside an integrative category, covering a diverse array of counseling topics. Crucially, we employ a unified three-stage clinical framework to model these varied data sources. This methodological design ensures that the agent can flexibly adapt to specific theoretical demands while adhering to a consistent structural progression across all sessions.

**RQ3: How to systematically evaluate an AI counselor?** To move assessment beyond generic perplexity metrics, we establish a holistic evaluation framework operationalized across two key dimensions: *Client-Level* simulation fidelity and *Counselor-Level* clinical proficiency. Crucially, we design a hybrid metric set comprising therapy-specific indicators and shared indicators, ensuring a rigorous standard for versatility and specialization.

Finally, we construct and release our PsychEval. Comparative analysis with existing benchmarks shows that PsychEval achieves the highest quality, session depth, and closest alignment with real-world counseling dynamics to date. Crucially, it extends beyond a static evaluation testbed to serve as a high-fidelity reinforcement learning environment, simulating diverse and evolving client states with rich reward functions to enable self-evolutionary training of AI agents and provide a foundational ecosystem for developing clinically responsible, continuously improving AI counselors.

## 2 Background

### 2.1 Different Psychological Therapies

Over the past century, psychological counseling has evolved from singular theoretical origins into a multifaceted discipline, shaped by diverging and converging philosophical and scientific paradigms that have given rise to distinct therapeutic approaches. In constructing our benchmark, we primarily incorporate the following five major therapies: Psychodynamic (Summers et al., 2024), Behavioral (Wolpe, 1990; Wilson, 2005), Cognitive-Behavioral (Rothbaum et al., 2000; Hofmann et al., 2012), Humanistic-Existential (Farber, 2010), and Postmodernist (Flaskas, 2003), alongside an **Integrative** therapy. The descriptions of these therapies are provided in Section B in the Appendix.

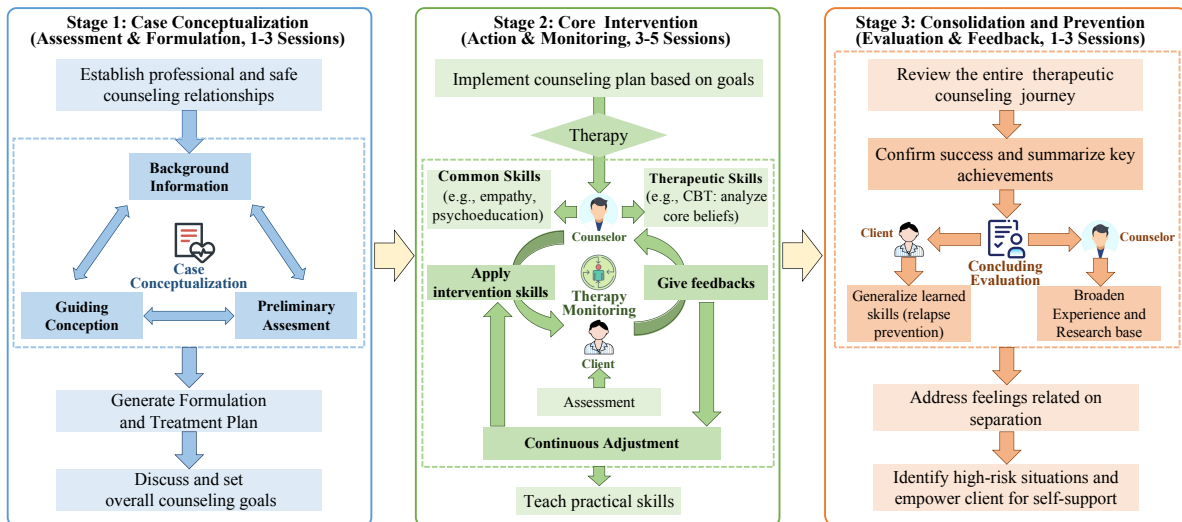


Figure 1: The Unified Flow of Psychological Counseling.

## 2.2 Unified Counseling Workflow

Although different therapeutic approaches vary in their theories, techniques, and focal points, they exhibit common structural differentiations as the client’s problem emerges and evolves. Generally, the current counseling process can be divided into three distinct phases: evaluation, intervention, and consolidation (Peterson, 1991; Fishman, 2013). To ensure reproducibility and scientific rigor, we introduce a generalized framework that designates three phases: Case Conceptualization, Core Intervention, and Consolidation and Prevention (See Fig. 1).

**Stage 1: Case Conceptualization (Assessment & Formulation)** This foundational phase centers on the client and typically spans 1–3 sessions. Its goals are to establish a secure therapeutic alliance and conduct a comprehensive assessment via intake interviews, gathering background, history, and symptom data. Guided by a theoretically informed conception rooted in clinical experience and research, the counselor synthesizes this information into an individualized case formulation and treatment plan, a personalized hypothesis about the etiology of the client’s difficulties that shapes the intervention trajectory.

**Stage 2: Core Intervention (Action & Monitoring)** This central phase implements the course of therapy over 3–5 sessions, focusing on facilitating change through therapy-specific techniques (e.g., cognitive restructuring in CBT). Crucially, ongoing therapy monitoring creates a continuous feedback loop: if outcomes are unsatisfactory, the process loops back to reassess rapport, collect new data, or refine the case formulation. Thus, intervention is

not rigid but a dynamic cycle of hypothesis testing and adaptation to the client’s real-time progress.

**Stage 3: Consolidation and Prevention (Evaluation & Feedback)** The final 1–3 sessions focus on concluding evaluation to solidify gains and prepare for termination. Counselor and client collaboratively review the therapeutic journey to confirm intervention success. This evaluation both helps the client generalize skills to daily life (supporting relapse prevention) and provides the counselor with feedback to refine their internal guiding conceptions. The ultimate aim is to empower the client to sustain well-being independently.

## 3 Benchmark Construction

To bridge the gap between current AI capabilities and the complexity of real-world clinical practice, we construct a high-fidelity, multi-session benchmark. This benchmark integrates diverse therapeutic therapies to simulate the timely, accurate, and genuine adjustments required in psychological counseling. We employ advanced LLMs (e.g., GPT-5) to synthesize and reconstruct high-quality dialogues derived from empirically grounded clinical case reports. The construction pipeline proceeds through three primary phases: data collection, structured case extraction, and multi-stage dialogue construction.

### 3.1 Data Collection

To guarantee data authenticity and reliability, we curate a corpus of clinical case reports sourced from authoritative psychology journals (See Section E.3 in Appendix). Unlike unstructured tran-

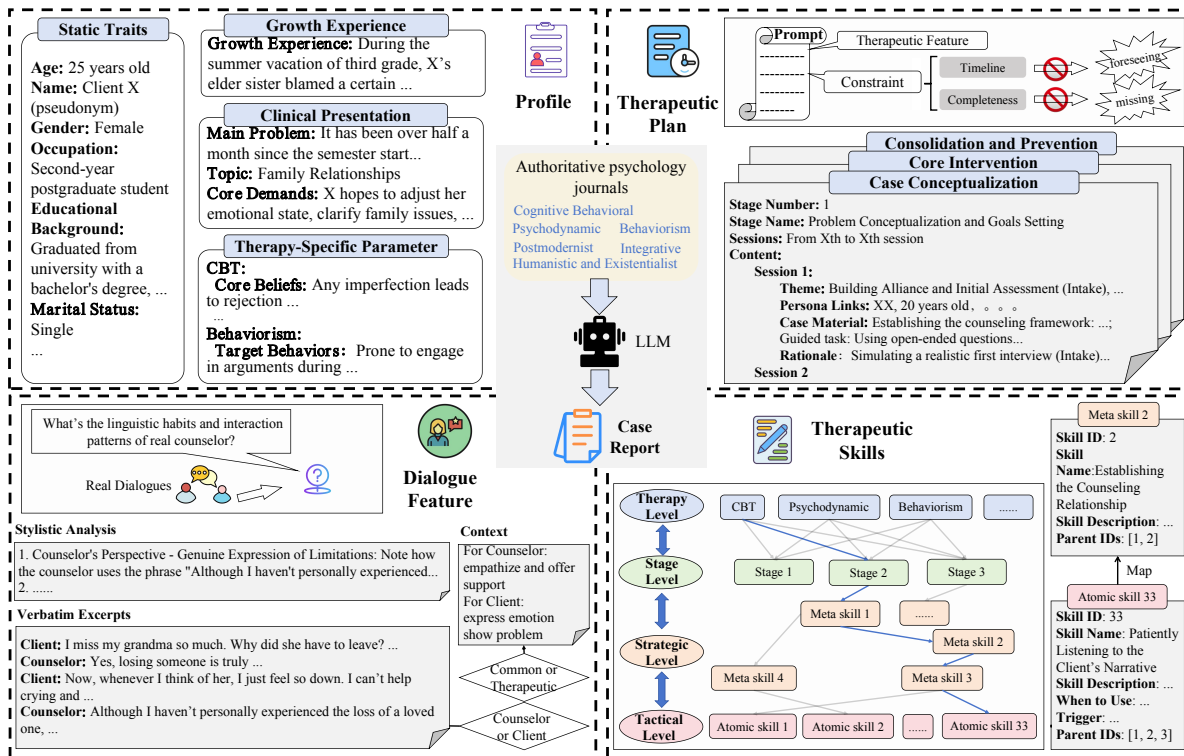


Figure 2: The Flow of Structured Case Extraction.

scripts from online consultation platforms (e.g., *Yidianling*), these peer-reviewed reports offer verified clinical frameworks, distinct theoretical orientations, and detailed longitudinal trajectories. Crucially, they originate from documented, real-world counseling sessions, thereby ensuring high ecological validity. Following rigorous cleaning and classification, the final dataset comprises 369 case reports. The distribution encompasses five major therapies: Cognitive Behavioral Therapy accounts for 148 reports, while Psychoanalysis, Humanistic, and Postmodernist approaches each contribute 50. Additionally, the dataset includes 43 reports for Behaviorism and 28 for Integrative Therapy.

### 3.2 Structured Case Extraction

While LLMs possess strong document analysis capabilities, directly instructing them to reconstruct full counseling sessions from raw reports often yields suboptimal results, characterized by hallucinations, loss of clinical nuance, and flattened therapeutic arcs lacking cross-session continuity. To mitigate this, we implement an intermediate structured extraction phase. We convert raw text into a standardized schema consisting of four modules: Client Profile, Therapeutic Plan, Dialogue Features, and Therapeutic Skills (Fig. 2). This structured approach ensures the LLM captures the full granularity of the clinical context.

**Client Profile** The Profile module constructs a structured persona integrating four dimensions: 1) Static Traits (Appendix Fig. 6), encompassing demographics, medical history, and linguistic features; 2) Clinical Presentation (Appendix Fig. 5), specifying the presenting problem, core demands, and counseling topics; 3) Growth Experience (Appendix Fig. 7), highlighting worldview-shaping life events; and 4) Therapy-Specific Parameters, extracting therapy-unique elements like core beliefs (CBT) or target behaviors (Behaviorism) to ensure professional distinctiveness.

**Therapeutic Plan** As a strategic roadmap, this module generates a three-phase multi-session plan via the Profile and original report. While allowing flexible session counts, it enforces two constraints: the Timeline Constraint ensures chronological progression without foreseeing future events, while the Completeness Constraint permits logical inference of missing details. Each session plan is defined by five fields: *Theme*, *Persona Links* (activated profile data), *Case Material* (concrete quotes or tasks), *Rationale* (clinical reasoning), and *Psychoeducation Skills* (concepts for the client to learn).

**Dialogue Features** To capture the micro-level dynamics of real counseling, this module extracts verbatim dialogue segments from the reports to serve as few-shot exemplars. This ensures the gen-

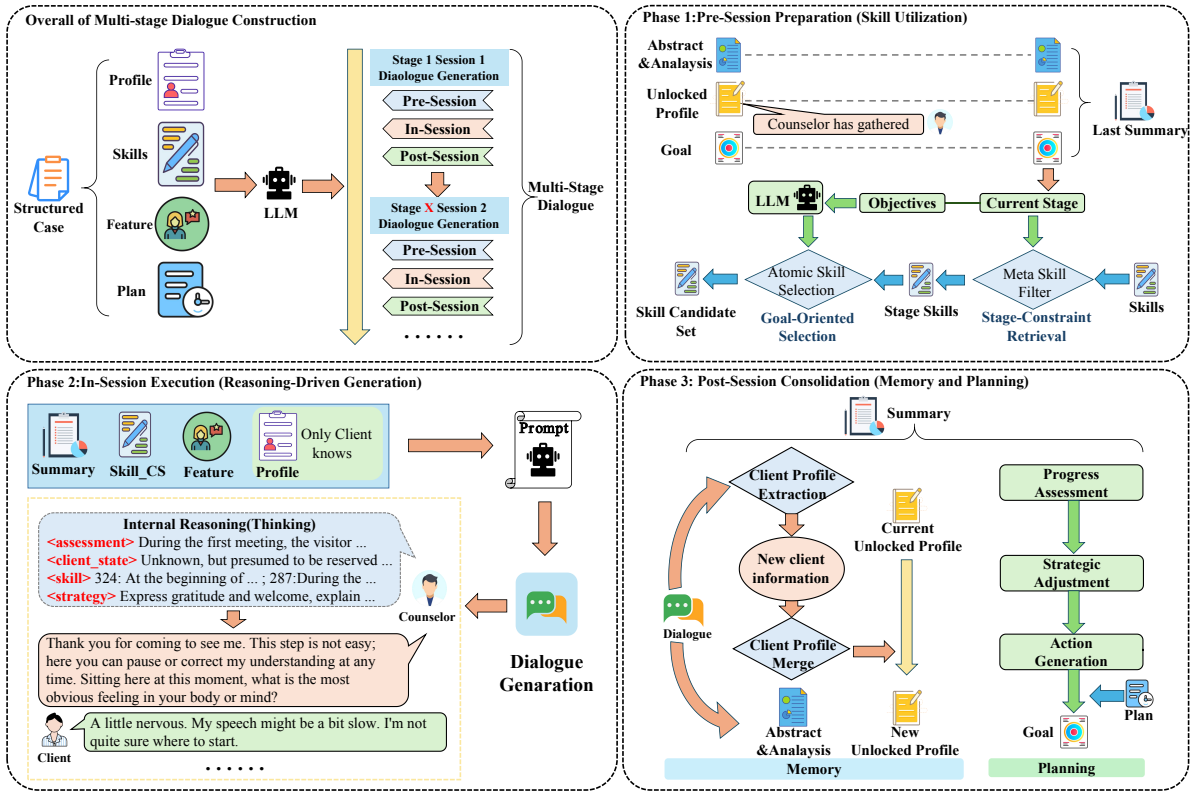


Figure 3: The skill-informed generative pipeline for multi-stage dialogue construction.

erated text moves beyond generic tones to reflect specific linguistic habits and interaction patterns. This module defines the Context (e.g., homework assignment), provides Verbatim Excerpts of 3–5 turn exchanges from the original report, and includes Stylistic Analysis instructions regarding sentence construction, lexical choice, and turn-taking habits derived from the real-world samples.

**Therapeutic Skills** To bridge abstract theory with concrete action, we construct a hierarchical taxonomy enabling coarse-to-fine reasoning. Organized as a tree structure, Meta Skills (Strategic/Branch nodes) define therapeutic intent, while Atomic Skills (Tactical/Leaf nodes) define specific executable behaviors. **1) Meta Skills** derived from authoritative textbooks (e.g., *CBT: Basics and Beyond*) represent high-level competencies. The extraction pipeline utilizes GPT-5 to parse digitized texts into a structured format, where each entry includes a Skill ID, Name, Description, and Parent IDs to maintain hierarchical lineage. **2) Atomic Skills** represent granular verbal techniques extracted from practical clinical guides and transcripts. Unlike broad strategies, these are mapped directly to Meta Skill nodes to ensure theoretical alignment. To support dynamic decision-making, the schema is enriched with *When to Use* (con-

textual timing) and *Trigger* (specific client cues) alongside standard identifiers.

### 3.3 Multi-stage Dialogue Construction

Following case extraction, we develop a Skill-Informed Generative Pipeline (Fig. 3) to ensure realism and longitudinal continuity. The process comprises three phases: Pre-Session Preparation, In-Session Execution, and Post-Session Consolidation. This structure ensures interventions are grounded in assessment and maintains consistency across the counseling arc.

**Phase 1: Pre-Session Preparation (Skill Utilization)** To define the intervention space, we propose a Coarse-to-Fine Skill Retrieval mechanism. First, Stage-Constraint Retrieval filters Meta Skills relevant to the current Therapeutic Plan phase. Then, functioning as a selector, GPT-5 identifies pertinent Atomic Skills based on session goals to construct a precise skill candidate set, ensuring the model possesses the necessary strategic toolkit without navigating the entire taxonomy.

**Phase 2: In-Session Execution (Reasoning-Driven Generation)** Simulating human cognition, we employ Chain-of-Thought (CoT) (Wei et al., 2022) to generate an Internal Reasoning Trace prior to each utterance. This trace models

four components: 1) Assessment (evaluating client state/subtext); 2) State Tracking (monitoring the alliance); 3) Skill Selection (choosing the optimal Atomic Skill); and 4) Strategy Formulation (determining tone/pacing). Decoupling reasoning from response ensures every interaction remains strategically grounded in clinical theory.

**Phase 3: Post-Session Consolidation (Memory and Planning)** Ensuring cross-session consistency, this dual-mechanism module handles: 1) Information Synthesis, compressing the session into a *Clinical Summary* and updating the *Client Profile* with new disclosures; and 2) Trajectory Refinement, evaluating efficacy against the *Therapeutic Plan* to generate the Next Session Plan. Through *Progress Assessment*, the system determines whether to advance phases or schedule reinforcement, explicitly defining objectives and target skills for the subsequent encounter.

## 4 Holistic Evaluation Framework

To objectively quantify clinical and technical efficacy beyond traditional lexical metrics (e.g., BLEU), which fail to capture therapeutic intentionality, we establish a Holistic Evaluation Framework. Adopting an external supervisory paradigm, this methodology operationalizes validated psychometric instruments for automated LLM-based assessment. It integrates surrogate supervision to distinguish between universal and therapy-specific competencies, while systematically quantifying longitudinal therapeutic progress through multi-session trajectory analysis. We operationalize established psychometric instruments as compact observer-rated proxies of counseling quality, covering alliance, procedural helpfulness, genuineness, symptom burden, affective shift, and therapy-specific fidelity. Due to space limitations, we provide the full construct definitions, score mappings, and prompt templates in C.

### 4.1 Counselor-Level Evaluation

The counselor-level evaluation assesses clinical proficiency, focusing on the delivery of theoretically grounded and ethically sound interventions. Adopting an external supervisory paradigm, we evaluate the agent’s longitudinal coherence and adherence to therapy-specific technical protocols.

**Therapy-shared Metrics** Therapy-shared metrics assess universal counseling competencies

across orientations. Specifically, WAI measures goal–task–bond alignment, HTAIS (Li et al., 2022) evaluates the helpfulness and procedural integrity of interventions, and RRO (Békés et al., 2025) captures counselor genuineness and realism. We additionally include a customized rubric covering clinical perception, longitudinal strategy, therapeutic depth, and ethical safety. For readability, we report aggregated scale scores.

**Therapy-specific Metrics** These metrics assess theoretical fidelity across orientations using specialized scales. We employ the Cognitive Therapy Rating Scale (CTRS) for CBT competencies (e.g., Socratic questioning) and the Psychodynamic Supervision Checklist (PSC) for insight into unconscious conflicts. Furthermore, the Therapist Empathy Scale (TES) and Therapeutic Focus Scale (EFT-TFS) measure empathic resonance and affective processing in Humanistic/Behavioral approaches, while the Motivational Interviewing Treatment Integrity (MITI) scale gauges change discourse evocation in Postmodern therapies.

### 4.2 Client-Level Evaluation

The Client-Level evaluation focuses on simulation fidelity, assessing whether simulated clients maintain clinical consistency and authentic psychological shifts. We employ an LLM-based supervisor to convert qualitative dialogue into quantifiable clinical data via standardized scoring.

**Therapy-shared Metrics** Therapy-shared client metrics capture symptom burden, affective fluctuation, relational authenticity, and session outcome. Specifically, SCL-90 measures broad symptom severity, PANAS captures short-term affective change, RRO reflects perceived dyadic authenticity, and SRS measures session alliance and satisfaction. Unless otherwise noted, higher scores indicate better outcomes, except for symptom-burden measures such as SCL-90, BDI-II, and IPO.

**Therapy-specific Metrics** These metrics validate theoretical alignment using five specialized instruments. We employ the Beck Depression Inventory-II (BDI-II) to verify cognitive symptom reduction in CBT and the Inventory of Personality Organization (IPO) to track personality integration in Psychodynamic therapy. The State-Trait Anxiety Inventory (STAI) and Client-Centered Therapy Criteria (CCT) measure anxiety and self-acceptance in Behavioral and Humanistic approaches, respec-

Table 1: Statistical information of our PsychEval benchmark. #AvgSess and #AvgTurns denote the average sessions per sample and average turns per session, respectively. #AvgWords and #AvgWords<sub>w/o t</sub> are the average words per turn with and without thinking. #MetaSkill and #AtomicSkill indicate the counts of meta- and atomic-level skills.

Therapy	#Sample	#AvgSess	#AvgTurns	#AvgWords <sub>w/o t</sub>	#AvgWords	#MaxSess	#MetaSkill	#AtomicSkill
Cognitive-Behavioral	148	7.5	22.0	119.9	371.7	10	103	1368
Psychodynamic	50	8.0	22.0	86.4	334.7	10	210	1144
Behavioral	43	7.2	28.0	70.3	306.0	9	156	854
Humanistic-Existential	50	7.7	25.1	81.6	326.7	10	136	752
Postmodernist	50	7.5	26.7	73.0	302.1	10	72	459
Integrative	28	7.7	25.9	75.4	308.3	10	-	-
Total	369	7.6	24.1	92.9	336.2	10	677	4577

Table 2: Comparison with existing benchmarks on key characteristics. Reasoning denotes turn-level reasoning; Evaluation refers to a comprehensive assessment framework; Integrative indicates support for integrative therapy.

Dataset	Reasoning	Evaluation	MultiSess	MultiTherapy	Integrative	MultiStage	Skill	OpenSource
CounselingBench (Nguyen et al., 2025)	✗	✗	✗	✓	✗	✗	✗	✓
PsyDTCorpus (Xie et al., 2025)	✗	✗	✗	✗	✗	✗	✗	✓
CACTUS (Lee et al., 2024)	✗	✗	✗	✗	✗	✗	✗	✓
PsyDial (Qiu and Lan, 2025)	✗	✗	✗	✓	✗	✗	✗	✓
SimPsyDial (Qiu and Lan, 2024)	✗	✗	✗	✓	✗	✗	✗	✓
Cpsycoun (Zhang et al., 2024a)	✗	✗	✗	✓	✗	✗	✗	✓
MindChat (Xin Yan, 2023)	✗	✗	✗	✗	✗	✗	✗	✗
SMILE (Qiu et al., 2024)	✗	✗	✗	✗	✗	✗	✗	✓
MusPsy (Wang et al., 2025)	✗	✗	✓	✗	✗	✗	✗	✗
Psy-Insight (Chen et al., 2025)	✗	✗	✓	✓	✗	✗	✗	✓
PsychEval (Our)	✓	✓	✓	✓	✓	✓	✓	✓

tively, while SFBT Indicators monitor solution-oriented discourse in Postmodern interventions.

## 5 Analysis of Benchmark

### 5.1 Statistic Information

Table 1 presents the statistical profile of PsychEval, a benchmark distinguished by unprecedented longitudinal depth and high-fidelity information density. It comprises 369 cases across five therapeutic genres, breaking the "single-session" barrier of prior works with an average of 7.6 sessions per case (up to 10). This multi-session structure provides the temporal span essential for evaluating memory continuity and long-term planning. Interactions exhibit professional-grade density, averaging 24.1 turns per session and 94.7 words per turn, with realistic variation across genres: *Cognitive-Behavioral Therapy* has the highest volume (119.9 words/turn) due to its psychoeducational nature, while *Behavioral* approaches remain concise (~70 words/turn), reflecting authentic clinical practices.

Moreover, PsychEval sets a benchmark for interpretable counseling through a hierarchical taxonomy of 677 Meta and 4,577 Atomic skills (Fig. 4 in Appendix), enabling navigation from strategic planning to tactical intervention. As shown in Table 8 (Appendix), the framework’s structural integrity is validated by its alignment with a profes-

sional three-stage clinical model: metric distributions mirror authentic therapeutic dynamics, from high personalization in intake to sustained duration and material richness during core intervention.

### 5.2 Comparison with Existing Benchmark

To contextualize PsychEval’s contribution, we compare it multidimensionally against representative datasets. As shown in Tables 2 and 3, PsychEval shifts focus from large-scale but shallow interactions to high-fidelity, longitudinal, and skill-aware assessment.

**First**, it sets a new standard in longitudinal depth and information density. While most existing datasets are limited to single-session interactions and low verbal complexity (e.g., *ESConv*: 16.3 words/turn), PsychEval provides an unprecedented average of 7.6 sessions per case and 92.9 words per turn. This extended temporal span, coupled with rigorous “chit-chat” filtering, captures the evolving dynamics of real-world therapy rather than static snapshots. **Second**, PsychEval is the most functionally complete benchmark to date. As Table 2 shows, it is the only open-source dataset that simultaneously supports turn-level reasoning, multi-stage clinical workflows, and hierarchical skill annotations. Unlike single-modality or supervision-lacking datasets, it integrates five distinct therapeutic genres plus an integrative ap-

Table 3: Comparison with existing benchmarks in terms of statistical information. † indicates the number of topics present in the dataset, but no topic label is provided for each individual sample.

Dataset	#Sample	#Topic	#AvgSess	#Therapy	#AvgWords	#AvgTurn	#TotalSess	#TotalTurn
PsyDTCorpu(Xie et al., 2025)	4760	12	1	1	44.8	18.1	4760	86054
CACTUS (Lee et al., 2024)	31577	9†	1	1	26.4	15.6	31577	491316
PsyDial (D4) (Qiu and Lan, 2025)	2382	60†	1	3	32.4	37.8	2382	90033
SimPsyDial (Qiu and Lan, 2024)	1000	60†	1	3	62.5	12.9	1000	12948
Cpsycoun (Zhang et al., 2024a)	3084	9	1	7	39.0	8.0	3084	24808
SMILE (Qiu et al., 2024)	55165	60†	1	-	81.3	5.8	55165	318395
ESConv (Liu et al., 2021)	1300	12	1	-	16.3	14.5	1300	18864
MusPsy (Wang et al., 2025)	1400	-	6.2	1	27.0	28.6	-	-
Psy-Insight <sub>(en)</sub> (Chen et al., 2025)	114	11	4.56	-	17.1	6.2	520	3202
Psy-Insight <sub>(cn)</sub> (Chen et al., 2025)	75	11	-	-	41.6	6.9	431	2995
PsychEval (Our)	369	6	7.6	5	92.9	24.1	2798	67314

Table 4: Data quality of our benchmark in terms of counselor-level metrics.

	Therapy-Shared				Therapy-Specific				
	HTAIS	RRO	WAI	Custom	CTRS	PSC	TES	EFT-TFS	MITI
CACTUS (Lee et al., 2024)	4.92	7.12	6.08	3.64	1.41	-	-	-	-
Cpsycoun (Zhang et al., 2024a)	4.74	6.70	5.43	3.94	1.30	2.87	2.87	1.04	2.94
PsyDial (Qiu and Lan, 2025)	4.25	6.39	4.15	4.48	4.47	-	5.71	-	-
PsyDTCorpus (Xie et al., 2025)	5.50	7.44	6.16	4.40	1.35	-	-	-	-
SimPsyDial (Qiu and Lan, 2024)	4.55	6.02	4.87	4.10	1.96	3.10	3.52	-	-
SMILE (Qiu et al., 2024)	5.32	7.07	5.37	4.13	-	-	-	-	-
Psy-Insight <sub>(en)</sub> (Chen et al., 2025)	5.02	6.72	4.97	4.13	4.32	4.26	4.19	1.85	-
Psy-Insight <sub>(cn)</sub> (Chen et al., 2025)	2.78	5.01	2.25	2.45	4.40	2.99	2.24	1.35	-
ESConv (Liu et al., 2021)	4.60	6.79	4.66	4.19	-	-	3.26	-	-
PsychEval (Our)	<b>6.40</b>	<b>7.64</b>	<b>7.26</b>	<b>7.36</b>	<b>9.19</b>	<b>7.27</b>	<b>7.36</b>	<b>3.14</b>	<b>5.88</b>

Table 5: Adjacent difference of our benchmark in terms of client-level metrics.

	Therapy-Shared				Therapy-Specific				
	RRO	Panas	SCL-90↓	SRS	BDI-II ↓	IPO ↓	CCT	SFBT	STAI
Psy-Insight <sub>(en)</sub> (Chen et al., 2025)	<b>0.07</b>	0.07	-0.01	0.03	-0.10	0.04	-0.19	-0.13	-
PsychEval (Our)	<b>0.07</b>	<b>0.17</b>	<b>-0.06</b>	<b>0.08</b>	<b>-0.45</b>	<b>-0.09</b>	<b>0.14</b>	<b>0.16</b>	<b>0.28</b>

proach. This combination of longitudinal fidelity and structured expert annotations advances the field beyond basic dialogue modeling toward skill-aware, planning-capable AI counselors.

### 5.3 Data Quality Assessment

To rigorously validate the clinical fidelity of PsychEval, we conducted a comparative quality assessment against eight established benchmarks over psychological metrics. As shown in Tables 4, 9, and 5, PsychEval demonstrates superior performance across both counselor and client dimensions.

In Table 4, PsychEval excels in both *Therapy-Shared* and *Therapy-Specific* dimensions, outperforming datasets such as *PsyDTCorpus* and *Cpsycoun*. It achieves SOTA scores in Working Alliance (WAI: 7.26) and Helper Skills (HTAIS: 6.40), and shows especially strong therapy-specific adherence, with CTRS at 9.19 and PSC at 7.27, nearly double prior models like *PsyDial* and *SimPsyDial*. This confirms that PsychEval captures both empathetic

tone and precise clinical interventions.

For client realism and therapeutic effectiveness, we evaluate post-session psychological states and longitudinal changes. Table 9 (Appendix) shows PsychEval clients report the highest satisfaction and alliance (SRS: 8.29) and the largest reduction in symptom distress (SCL-90: 1.48). Crucially, authentic therapy is defined by change over time. Table 5 presents adjacent differences (session  $t + 1$  minus session  $t$ ) as proxies for progress: PsychEval exhibits a coherent “healing” trajectory—symptom measures (BDI-II, IPO) decrease significantly (-0.45, -0.09), while positive capacities (CCT, SFBT) increase (+0.14, +0.16). In contrast, *Psy-Insight* shows counter-therapeutic trends: IPO rises (+0.04), and CCT/SFBT decline (-0.19, -0.13). Thus, PsychEval is the only benchmark that authentically models real-world psychological evolution and positive transformation.

Table 6: Human evaluation. Overall and Process mean the Overall Goal and Process Detail Consistency.

Evaluator	Sessions	Coherence	Depth	Overall	Process
Human Experts	231	8.32	7.76	8.43	8.03
LLM Judge	231	8.18	7.72	8.32	7.93

## 5.4 Human Expert Validation

To complement the model-based assessment in Section 5.3 and further validate the clinical fidelity of PsychEval, we conduct human expert validation on a stratified random subset of the benchmark. Specifically, we sample 231 sessions ( $\approx 1/12$  of all benchmark sessions) from 30 clients, covering all six therapies, and invite two experts in psychological counseling to independently perform blind review. Each session is rated on four clinically grounded dimensions—*Coherence*, *Depth*, *Overall Goal Consistency*, and *Process Detail Consistency*—which evaluate the internal logical consistency of the dialogue, the extent to which the interaction reaches core emotional and cognitive material, adherence to staged treatment goals, and fidelity to intervention procedures. See Appendix F for human evaluation instructions, and ethical protocols.

As shown in Table 6, human experts assign consistently high scores across all four dimensions, with averages of 8.32 for *Coherence*, 7.76 for *Depth*, 8.43 for *Overall Goal Consistency*, and 8.03 for *Process Detail Consistency*. The model-based evaluator shows a highly similar pattern, with corresponding scores of 8.18, 7.72, 8.32, and 7.93. These results indicate that the reconstructed sessions preserve both the macro-level therapeutic trajectory and the micro-level procedural structure of the source case reports, while also suggesting that the proposed LLM-based evaluation framework produces judgments broadly consistent with expert assessment.

We further examine the reliability of expert assessment and the alignment between human and automated evaluation. As shown in Table 7, the weighted kappa between the two experts reaches 0.83 on average, indicating substantial inter-rater reliability across the four dimensions. Agreement between expert and model-based judgments is also high, with average weighted kappas of 0.83 for H1 vs. LLM and 0.80 for H2 vs. LLM. Taken together, these results show that PsychEval maintains high clinical fidelity under expert review, while the proposed LLM-based evaluation framework remains

Table 7: The Weighted Kappa Scores between two human experts and LLM. H1 and H2 indicate Human1 and Human2, LLM indicates GPT-5.

Agreement Pair	Coherence	Depth	Overall	Process	Average
H1 vs. H2	0.68	0.86	0.73	0.88	0.83
H1 vs. LLM	0.76	0.81	0.79	0.83	0.83
H2 vs. LLM	0.59	0.87	0.70	0.85	0.80

well aligned with human judgment, supporting its use for large-scale benchmark analysis.

## 6 Conclusions and Future Work

We introduce PsychEval, a benchmark and training ecosystem that aligns AI capabilities with professional psychological assessment demands. Built from 369 authentic clinical case reports, it captures therapy’s longitudinal complexity with an average of 7.6 sessions per case. PsychEval combines a three-stage clinical framework with a hierarchical taxonomy of over 4,500 atomic skills, enabling AI agents to perform coarse-to-fine clinical reasoning. It significantly outperforms existing benchmarks, achieving a CTRS score of 9.19 and a WAI score of 7.26. Beyond evaluation, PsychEval serves as a high-fidelity reinforcement learning environment for training clinically responsible, logically coherent, and ethically aligned AI counselors. Future work will use PsychEval to develop a self-evolving AI counselor for autonomous, adaptive mental health care.

### Limitations

PsychEval has several limitations. First, it focuses on text-only counseling interactions and therefore does not capture non-verbal cues such as prosody, facial expression, or body language. Second, because the benchmark is reconstructed from ethically shareable journal case reports, it underrepresents extreme or high-risk scenarios such as acute suicidality and highly resistant clients. Third, the current benchmark is grounded in the Chinese linguistic and socio-cultural context, and may not transfer directly to other cultural settings. These limitations motivate future work on multimodal counseling, high-risk scenario modeling, and cross-cultural adaptation.

### Acknowledgements

This research is funded by the National Nature Science Foundation of China (No.62477010, No.62577022 and No.62307028), Shanghai Science and Technology Innovation Action Plan

(No.24YF2710100), and CIPS-SMP-Zhipu Large Model Fund.

## References

- Anonymous. 2025. From empathy to action: Benchmarking LLMs in mental health with mentalbench-10 and a novel cognitive-affective evaluation approach. In *Submitted to ACL Rolling Review - May 2025*. Under review.
- Douglas Behan. 2022. Do clients train therapists to become eclectic and use the common factors? a qualitative study listening to experienced psychotherapists. *BMC psychology*, 10(1):183.
- Vera Békés, Daniel S Spina, Katie Aafjes-van Doorn, Bernard S Gorman, Karl Stukenberg, and Sherwood Waldron. 2025. Development of a new observer-rated measure to assess the real relationship in psychotherapy sessions. *Psychotherapy Research*, 35(5):807–817.
- Sorio Boit and Rajvardhan Patil. 2025. A prompt engineering framework for large language model-based mental health chatbots: Conceptual framework. *JMIR Mental Health*, 12(1):e75078.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Keqi Chen, Zekai Sun, Yuhua Wen, Huijun Lian, Yingming Gao, and Ya Li. 2025. Psy-insight: Explainable multi-turn bilingual dataset for mental health counseling. *arXiv preprint arXiv:2503.03607*.
- Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. 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.
- Eugene W Farber. 2010. Humanistic–existential psychotherapy competencies and the supervisory process. *Psychotherapy: Theory, Research, Practice, Training*, 47(1):28.
- Daniel B Fishman. 2013. The pragmatic case study method for creating rigorous and systematic, practitioner-friendly research. *Pragmatic Case Studies in Psychotherapy*, 9(4):403–425.
- Carmel Flaskas. 2003. *Family therapy beyond postmodernism: Practice challenges theory*. Routledge.
- Guanghui Fu, Qing Zhao, Jianqiang Li, Dan Luo, Changwei Song, Wei Zhai, Shuo Liu, Fan Wang, Yan Wang, Lijuan Cheng, and 1 others. 2023. Enhancing psychological counseling with large language model: A multifaceted decision-support system for non-professionals. *arXiv preprint arXiv:2308.15192*.
- Qiming Guo, Jinwen Tang, Wenbo Sun, Haoteng Tang, Yi Shang, and Wenlu Wang. 2024. Soullmate: An adaptive llm-driven system for advanced mental health support and assessment, based on a systematic application survey. *arXiv preprint arXiv:2410.11859*.
- Guangzeng Han, Weisi Liu, Xiaolei Huang, and Brian Borsari. 2024. Chain-of-interaction: Enhancing large language models for psychiatric behavior understanding by dyadic contexts. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pages 392–401. IEEE.
- Ayaan Haque, Viraj Reddi, and Tyler Giallanza. 2021. Deep learning for suicide and depression identification with unsupervised label correction. In *International Conference on Artificial Neural Networks*, pages 436–447. Springer.
- Stefan G Hofmann, Anu Asnaani, Imke JJ Vonk, Alice T Sawyer, and Angela Fang. 2012. The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive therapy and research*, 36(5):427–440.
- He Hu, Yucheng Zhou, Juzheng Si, Qianning Wang, Hengheng Zhang, Fuji Ren, Fei Ma, Laizhong Cui, and Qi Tian. 2025. Beyond empathy: Integrating diagnostic and therapeutic reasoning with large language models for mental health counseling. *arXiv preprint arXiv:2505.15715*.
- Yining Hua, Hongbin Na, Zehan Li, Fenglin Liu, Xiao Fang, David Clifton, and John Torous. 2025. A scoping review of large language models for generative tasks in mental health care. *npj Digital Medicine*, 8(1):230.
- Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2024. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.
- Zainab Iftikhar, Amy Xiao, Sean Ransom, Jeff Huang, and Harini Suresh. 2025. How llm counselors violate ethical standards in mental health practice: A practitioner-informed framework. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 1311–1323.
- Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.
- Max Lamparth, Declan Grabb, Amy Franks, Scott Gershan, Kaitlyn N Kunstman, Aaron Lulla, Monika Drummond Roots, Manu Sharma, Aryan

- Shrivastava, Nina Vasan, and 1 others. 2025. Moving beyond medical exam questions: A clinician-annotated dataset of real-world tasks and ambiguity in mental healthcare. *arXiv preprint arXiv:2502.16051*.
- Suyeon Lee, Sunghwan Mac Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, and 1 others. 2024. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274.
- Wenyu Li, Yinuo Zhu, Xin Lin, Ming Li, Ziyue Jiang, and Ziqian Zeng. 2024. Zero-shot explainable mental health analysis on social media by incorporating mental scales. In *Companion Proceedings of the ACM Web Conference 2024*, pages 959–962.
- Xiaobao Li, Hong Gu, Xudong Zhao, Fazhan Chen, and Liang Liu. 2022. Development of a measure quantifying helpful psychotherapy interventions: The helpful therapeutic attitudes and interventions scale. *Frontiers in Psychiatry*, 13:1023346.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023a. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.
- 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.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruo Chen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.
- 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. *arXiv preprint arXiv:2502.11095*.
- Viet Cuong Nguyen, Mohammad Taher, Dongwan Hong, Vinicius Konkolics Possobom, Vibha Thirunellai Gopalakrishnan, Ekta Raj, Zihang Li, Heather J Soled, Michael L Birnbaum, Srijan Kumar, and 1 others. 2025. Do large language models align with core mental health counseling competencies? In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7488–7511.
- Donald R Peterson. 1991. Connection and disconnection of research and practice in the education of professional psychologists. *American Psychologist*, 46(4):422.
- Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024. 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.
- Huachuan Qiu and Zhenzhong Lan. 2024. Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions. *arXiv preprint arXiv:2408.15787*.
- 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.
- Barbara Olasov Rothbaum, Elizabeth A Meadows, Patricia Resick, and David W Foy. 2000. Cognitive-behavioral therapy.
- Günter Schiepek and David Pincus. 2023. Complexity science: A framework for psychotherapy integration. *Counselling and Psychotherapy Research*, 23(4):941–955.
- Sahar Khan Sherwani, Zarak Khan, Jim Samuel, Rajiv Kashyap, and Kushal Gunavantkumar Patel. 2025. Eshro: An innovative evaluation framework for ai-driven mental health chatbots. *Available at SSRN 5254332*.
- Richard F Summers, Jacques P Barber, and Sigal Zilcha-Mano. 2024. *Psychodynamic therapy*. Guilford Publications.
- Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. Psyqa: A chinese dataset for generating long counseling text for mental health support. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1489–1503.
- Thomas Yu Chow Tam, Sonish Sivarajkumar, Sumit Kapoor, Alisa V Stolyar, Katelyn Polanska, Karleigh R McCarthy, Hunter Osterhoudt, Xizhi Wu, Shyam Visweswaran, Sunyang Fu, and 1 others. 2024. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ digital medicine*, 7(1):258.
- Elsbeth Turcan and Kathleen McKeown. 2019. Dreddit: A reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107.

- Junzhe Wang, Bichen Wang, Xing Fu, Yixin Sun, Yanyan Zhao, and Bing Qin. 2025. Psychological counseling cannot be achieved overnight: Automated psychological counseling through multi-session conversations. *arXiv preprint arXiv:2506.06626*.
- Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Jiayin Zhi, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate Hardy, Hong Shen, and 1 others. 2024. Patient-{\Psi}: Using large language models to simulate patients for training mental health professionals. *arXiv preprint arXiv:2405.19660*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- G Terence Wilson. 2005. Behavior therapy.
- Joseph Wolpe. 1990. *The practice of behavior therapy*. Pergamon press.
- Mengxi Xiao, Qianqian Xie, Ziyang 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.
- 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.
- Dong Xue\* Xin Yan. 2023. Mindchat: Psychological large language model. <https://github.com/X-D-Lab/MindChat>.
- Cristina Zarbo, Giorgio A. Tasca, Francesco Cattafi, and Angelo Compare. 2016. [Integrative psychotherapy works](#). *Frontiers in Psychology*, Volume 6 - 2015.
- Min Zeng. 2025. Psychobench: Evaluating the psychology intelligence of large language models. *arXiv e-prints*, pages arXiv–2510.
- Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024a. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. *arXiv preprint arXiv:2405.16433*.
- 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.
- Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2024b. Safety-bench: Evaluating the safety of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15537–15553.

## A Related Work

### A.1 LLMs for Psychological Counseling

The application of Large Language Models (LLMs) in mental health has evolved from generic chat systems to specialized agents capable of therapeutic nuance. Early efforts primarily focused on enhancing emotional resonance and empathy. For instance, SoulChat (Chen et al., 2023) fine-tuned models on a large-scale empathetic dialogue dataset, significantly outperforming base models in providing emotional support. Similarly, the SMILE framework (Qiu et al., 2024) synthesized multi-turn dialogues to overcome data scarcity, enabling models to better handle therapeutic flow.

Beyond basic empathy, recent research aims to integrate specific therapeutic frameworks and clinical reasoning. HealMe (Xiao et al., 2024) employs specific prompting strategies to guide patients in cognitive reframing, demonstrating the potential of LLMs in delivering Cognitive Behavioral Therapy (CBT). To enhance clinical rigor, PsyLLM (Lai et al., 2023) incorporates diagnostic reasoning (aligned with DSM-5) and therapeutic strategies (e.g., ACT) via a "Chain of Thought" approach, addressing the critical need for long-term continuity, SouLLMate (Guo et al., 2024) utilizes a dual-memory system with Retrieval-Augmented Generation (RAG), while the Chain-of-interaction framework (Han et al., 2024) models the iterative counselor-client exchange to improve engagement.

However, most existing models are either confined to a single therapeutic approach (predominantly CBT) or lack the flexibility to switch strategies based on client needs. Our work addresses this by training agents not just in one therapy, but across five distinct therapeutic schools and an integrative approach.

### A.2 Psychological Benchmarks

Benchmarking in mental health has shifted from simple classification tasks to complex competency

evaluations. Traditional benchmarks like Dreddit (Turcan and McKeown, 2019) and SDCNL (Haque et al., 2021) treated mental health analysis as binary classification (e.g., stress detection), which fails to capture clinical nuance.

Subsequent benchmarks moved toward Question-Answering (QA) and Knowledge Assessment. PsyQA (Sun et al., 2021) provided a large-scale dataset for single-turn advisory capabilities. To assess professional competency, recent works like PsychoBench (Huang et al., 2024) and PsychCounsel-Bench (Zeng, 2025) evaluate LLMs against U.S. National Counselor Certification Exam (NCE) standards, ensuring models possess theoretical knowledge. Similarly, (Nguyen et al., 2025) aligns evaluation with competencies required of aspiring counselors.

More advanced benchmarks focus on Multi-turn Dialogue and Safety. ESConv (Liu et al., 2021) annotates emotional support strategies in conversations. CBT-BENCH (Zhang et al., 2025) specifically evaluates capabilities in CBT sessions. Regarding safety, SafetyBench (Zhang et al., 2024b) and MENTAT (Lamparth et al., 2025) test decision-making in crisis scenarios, while MentalBench-10 (Anonymous, 2025) differentiates between logical safety and affective resonance.

Despite these advancements, a critical gap remains: existing benchmarks are either static (exams), single-turn, or limited to specific modalities. There is a lack of a unified benchmark that integrates assessment, diagnosis, and treatment into a continuous, multi-session evaluation framework (Na et al., 2025). PsychEval fills this void by simulating the full longitudinal counseling process across diverse therapeutic schools.

### A.3 Psychological Evaluation Frameworks

Evaluating the output of Mental Health LLMs (MH-LLMs) is notoriously difficult. Traditional NLP metrics (e.g., BLEU, ROUGE) are widely considered inadequate as they fail to capture therapeutic quality (Liu et al., 2023b). Consequently, the field is moving toward multidimensional and model-based frameworks.

Recent frameworks prioritize Safety and Ethics. MIND-SAFE (Boit and Patil, 2025) proposes a layered architecture assessing risk detection and therapeutic adherence. ESHRO (Sherwani et al., 2025) offers a quantifiable metric combining empathy, safety, and quality. Crucially, research by Iftikhar et al. (Iftikhar et al., 2025) highlights the risk of

"over-validation," where models indiscriminately agree with distorted thoughts, necessitating metrics that penalize clinically inappropriate agreement.

Methodologically, evaluation is transitioning from costly human ratings to Simulation and LLM-as-a-Judge. The QUEST framework (Tam et al., 2024) standardizes subjective human ratings. Meanwhile, simulation-based approaches (Wang et al., 2024) use LLMs to role-play patients with specific disorders (e.g., depression) to test "therapist" models in controlled environments, often utilizing scales like PHQ-9 to quantify improvement (Li et al., 2024).

Building on these innovations, PsychEval introduces a comprehensive evaluation system that triangulates Client-Level (simulation feedback) and Counselor-Level (clinical adherence) metrics to ensure a holistic assessment of AI counselors.

## B Different Psychological Therapies

Psychological counseling has evolved over a century from nascent, singular theoretical prototypes into a robust, multi-faceted discipline. Throughout this evolution, diverse philosophical frameworks and scientific paradigms have converged and diverged, giving rise to numerous theoretical systems. These systems have gradually differentiated and iterated into distinct therapeutic schools. In constructing our benchmark, we primarily incorporate the following five major therapies: Psychodynamic (Summers et al., 2024), Behavioral (Wolpe, 1990; Wilson, 2005), Cognitive-Behavioral (Rothbaum et al., 2000; Hofmann et al., 2012), Humanistic-Existential (Farber, 2010), and Postmodernist (Flaskas, 2003), alongside an Integrative therapy.

**Psychodynamic Therapy** Originating from the pioneering work of Sigmund Freud, this therapy delves into the subconscious psychological processes. It emphasizes the profound impact of early childhood experiences, internal conflicts (e.g., among the id, ego, and superego), and defense mechanisms on adult behavior, emotions, and personality. The core premise is that psychological distress is rooted in unconscious dynamics unknown to the individual. By bringing these unconscious conflicts into conscious awareness, the therapy aims to alleviate symptoms and promote personality integration. Representative therapies include classic Psychoanalysis and the subsequently developed Psychodynamic Psychotherapy.

**Behaviorism Therapy** Emerging in the early 20th century with figures like John B. Watson and B.F. Skinner, this therapy advocates that psychology should focus on observable and measurable behaviors. It posits that both adaptive and maladaptive behaviors are not symptoms of internal disorders but are acquired and maintained through environmental learning mechanisms, specifically classical and operant conditioning. Consequently, therapy focuses on behavioral analysis to identify and modify the contingencies between environmental stimuli and problematic behaviors, thereby extinguishing maladaptive patterns and reinforcing adaptive ones. Key therapies include Systematic Desensitization, Exposure Therapy, and Aversion Therapy.

**Cognitive Behavioral Therapy** Building upon behaviorism and integrating cognitive psychology, this therapy was founded by Aaron Temkin Beck and Albert Ellis. Its central tenet is that an individual's emotions and behaviors are influenced not by events themselves, but by their cognitive appraisal—interpretations and beliefs (e.g., "automatic thoughts" and "core beliefs") regarding those events. Psychological distress arises from distorted or irrational cognitive patterns. The goal is to help clients identify, evaluate, and restructure these negative cognitions through structured interventions and behavioral experiments, leading to emotional and behavioral change. Key therapies include Cognitive Behavioral Therapy (CBT), Rational Emotive Behavior Therapy (REBT), Dialectical Behavior Therapy (DBT), and Acceptance and Commitment Therapy (ACT).

**Humanistic and Existentialist Therapy** Centered on the theories of Carl Rogers and Viktor Frankl, this therapy emphasizes subjective experience, free will, self-actualization, and the search for meaning. It posits that every individual possesses the internal resources and potential for growth. The counselor's role is to cultivate a therapeutic relationship characterized by genuineness, empathy, and unconditional positive regard. This supportive environment empowers clients to explore their authentic selves, realize their potential, and confront existential themes such as freedom, responsibility, isolation, and mortality. Representative therapies include Client-Centered Therapy, Existential Therapy, and Gestalt Therapy.

**Postmodernist Therapy** Influenced heavily by social constructionism, this relatively modern orientation features figures like Michael White. It fundamentally challenges the assumption of "objective truth" or "universal psychological reality" held by traditional schools. Instead, it argues that identity, problems, and "truth" are socially, linguistically, and culturally constructed narratives. Therapy shifts from diagnosing internal pathology to a collaborative co-creation process. By "deconstructing" dominant problem-saturated stories, counselors help clients discover overlooked positive exceptions and "alternative stories," enabling them to rewrite their life narratives. Key therapies include Narrative Therapy and Solution-Focused Brief Therapy (SFBT).

**Integrative Therapy** The Integrative Therapy is not a specific school but a holistic framework that synthesizes theories and techniques from diverse therapeutic traditions to address the multifaceted needs of clients. Recognizing that no single theory can explain all human complexities, this therapy moves beyond "schoolism." It typically operates through *technical eclecticism* (selecting the best techniques for a specific problem without adhering to their theoretical origin) or *theoretical integration* (synthesizing concepts from psychoanalysis, CBT, etc., into a coherent framework). In our benchmark, the Integrative counselor flexibly employs strategies—such as combining empathetic listening (Humanistic) with cognitive restructuring (CBT)—to provide the most effective, personalized intervention for the client's unique context.

## C Evaluations

To objectively quantify the clinical efficacy and technical performance of the proposed framework in simulating psychological interventions, we establish a Holistic Evaluation Framework. This framework addresses the limitations of traditional NLP metrics, such as BLEU or ROUGE, which rely on lexical overlap and fail to account for therapeutic intentionality, emotional regulation, and clinical coherence. Our approach adopts an external supervisory paradigm, operationalizing validated psychometric instruments into the automated assessment of LLM-based agents.

The contributions of this evaluation framework are categorized into three primary components:

- **Integration of Validated Clinical Instruments:**

We have operationalized a series of authoritative psychometric scales—including the SCL-90 for symptom severity and the Working Alliance Inventory (WAI) for therapeutic bonding—into structured evaluation protocols. By utilizing advanced LLMs as surrogate expert supervisors, we provide observer-rated assessments consistent with professional clinical supervision.

- **Bifurcated Metric System for Universal and Modality-Specific Competencies:** Recognizing the theoretical diversity in counseling, our framework distinguishes between *Therapy-shared Metrics* and *Therapy-specific Metrics*. The former evaluates universal clinical factors such as rapport building and empathy, while the latter assesses the technical precision of interventions unique to specific modalities, such as cognitive restructuring in CBT or the analysis of defense mechanisms in Psychodynamic therapy.
- **Quantification of Longitudinal Therapeutic Progress:** To capture the dynamic nature of counseling, we utilize the multi-session structure of our dataset to implement a longitudinal tracking mechanism. By calculating the differential scores ( $\Delta$  Score) between consecutive sessions, we quantify the trajectory of client improvement and the counselor's capacity for strategic, long-term treatment planning over time.

### C.1 Counselor-Level Evaluation

The counselor-level evaluation focuses on Clinical Proficiency, assessing the AI agent's capacity to deliver theoretically grounded and ethically sound interventions. By adopting an external supervisory paradigm, we evaluate whether the counselor maintains longitudinal clinical coherence and adheres to the technical requirements of specific therapy.

**Therapy-shared Metrics** The therapy-shared metrics evaluate universal clinical competencies that are essential across all therapeutic orientations. These instruments provide a benchmark for assessing the fundamental proficiency of the AI counselor:

- **WAI (Working Alliance Inventory — Counselor Form):** As a primary predictor of therapeutic outcomes, the WAI measures the counselor's ability to establish a therapeutic alliance across three dimensions: agreement on Goals, collaboration on Tasks, and the formation of an emotional Bond.

- **HTAIS (Li et al., 2022) (Helping Transaction Audit Inventory Scale):** Utilizing a supervisory auditing perspective, this scale evaluates the counselor's facilitative behaviors and the procedural integrity of the helping transaction, assessing the quality of interactional sequences and the counselor's ability to maintain an effective helping process.
- **RRO (Békés et al., 2025) (Real Relationship - Observer):** Consistent with the client-level assessment, this scale evaluates the counselor's Genuineness and Realism, measuring whether the AI counselor demonstrates authentic engagement that adheres to professional standards.
- **Customized:** To address the unique technical requirements of LLM-based counselors, we developed a multidimensional assessment encompassing four core domains: 1) **Clinical Perception:** Evaluates the precision of emotional state recognition and the delivery of accurate empathic reflections. 2) **Intervention Strategy:** Measures the maintenance of longitudinal memory, dynamic adaptation to evolving client needs, and the consistency of evidence-based intervention outputs across multiple sessions. 3) **Conversational Therapeutic Depth:** Assesses the coherence of the dialogue and the counselor's ability to advance therapeutic exploration while managing ambiguous or complex client inputs. 4) **Clinical Ethics and Safety:** Evaluates the identification of crisis markers, adherence to professional boundaries, and cultural sensitivity to ensure non-maleficence.

**Therapy-specific Metrics** To evaluate the Theoretical Fidelity and technical execution of the AI counselor within specific orientations, we incorporate specialized clinical rating scales. The supervisor scrutinizes high-level intervention skills against the established technical standards of each modality:

- **CTRS (Cognitive Therapy Rating Scale — Specific to CBT):** Used to quantify proficiency in Cognitive Behavioral Therapy. The evaluation focuses on structured elements, including agenda setting, the application of Socratic questioning for guided discovery, the identification of automatic thoughts, and the formulation of behavioral homework.

- **PSC (Psychodynamic Supervision Checklist — Specific to Psychodynamic):** Operationalized to assess the quality of psychodynamic interventions. The assessment targets the counselor’s insight into unconscious conflicts and defense mechanisms, as well as clinical sensitivity in managing transference and counter-transference phenomena.
- **TES (Therapist Empathy Scale — Specific to Humanistic):** Measures the depth of empathic resonance, evaluating whether the counselor accurately identifies and reflects latent emotional experiences, a prerequisite for a growth-promoting therapeutic climate.
- **EFT-TFS (Therapeutic Focus Scale — Specific to BT):** Assesses the counselor’s efficacy in facilitating affective processing. It measures the ability to guide the client through emotional transformation and maintain focus on the depth of emotional experiencing.
- **MITI (Motivational Interviewing Treatment Integrity — Specific to Postmodern):** Employed to evaluate technical proficiency in fostering collaboration and strategically evoking change-oriented discourse while maintaining a non-directive yet purposeful clinical stance.

## C.2 Client-Level Evaluation

The Client-Level evaluation focuses on Simulation Fidelity, measuring the extent to which the simulated client maintains clinical consistency and exhibits authentic psychological shifts. We employ an LLM-based supervisor (e.g., Deepseek V3.1) to perform standardized scoring on session transcripts, converting qualitative dialogue into quantifiable clinical data.

**Therapy-shared Metrics** To assess the client’s internal state and the universal aspects of the therapeutic dyad, we utilize four established psychometric instruments:

- **SCL-90 (Symptom Checklist-90):** Utilized as the primary instrument for psychopathological profiling. It evaluates the client across nine symptom dimensions to verify the consistency of the simulated clinical phenotype. A longitudinal reduction in SCL-90 scores across 10 sessions indicates overall therapeutic improvement.
- **PANAS (Positive and Negative Affect Schedule):** Tracks acute affective transitions within and between sessions. By measuring Positive Affect (PA) and Negative Affect (NA), we determine whether the counselor’s interventions effectively facilitate shifts in the client’s immediate emotional state.
- **RRO (Real Relationship - Observer):** Assesses the perceived authenticity of the therapeutic bond. It measures *Genuineness* and *Realism* to evaluate whether the simulated client perceives the AI counselor as an authentic interlocutor, which is a prerequisite for clinical work.
- **SRS (Session Rating Scale):** Serves as an outcome monitor at the conclusion of each session. It assesses the working alliance across four domains (Relationship, Goals and Topics, Approach, and Overall satisfaction), enabling the identification and analysis of relationship ruptures.

**Therapy-specific Metrics** To evaluate the precision of simulated client responses to specialized interventions, we incorporate five modality-specific instruments. These metrics validate whether the client exhibits clinical developments aligned with the theoretical mechanisms of the specific therapy:

- **BDI-II (Specific to CBT):** Measures cognitive symptoms of depression, such as hopelessness and self-criticism. It verifies whether the client achieves a reduction in symptom severity through the cognitive restructuring of automatic thoughts.
- **IPO (Inventory of Personality Organization — Specific to Psychodynamic):** Evaluates changes in identity consolidation, defense mechanisms, and reality testing. This serves as an indicator of personality integration, which is a primary objective of long-term psychodynamic intervention.
- **CCT (Client-Centered Therapy Criteria — Specific to Humanistic):** Assesses the client’s self-acceptance and congruence, validating progress fostered by the counselor’s provision of unconditional positive regard.
- **SFBT Indicators (Specific to Postmodern):** Monitors the frequency of solution-oriented discourse relative to problem-description. It assesses the client’s capacity to identify exceptions to distress and construct goal-directed narratives.
- **STAI (State-Trait Anxiety Inventory — Specific to Behaviorism):** Distinguishes between

state anxiety (situational) and trait anxiety (enduring), providing a precise measurement of improvements in response to specific behavioral triggers or desensitization protocols.

## D Statistic Information

Furthermore, PsychEval establishes a new standard for interpretable counseling through its massive, hierarchical skill annotation system (Figure 4). As detailed in the table, we define a comprehensive taxonomy comprising **677 Meta Skills** (strategic level) and **4,577 Atomic Skills** (tactical level). This granularity is particularly evident in complex modalities like the *Psychodynamic* approach, which alone features 210 meta and 1,144 atomic skills to capture the nuance of unconscious interpretation. This rich supervision signal empowers the training of AI counselors capable of sophisticated reasoning, navigating seamlessly from high-level therapeutic strategies to precise, turn-level verbal interventions.

Table 8 quantitatively validates the structural integrity of PsychEval, demonstrating its alignment with a professional three-stage clinical framework: *Intake & Conceptualization*, *Core Intervention*, and *Consolidation & Termination*. The distribution of metrics across these stages mirrors authentic therapeutic dynamics. **Stage 1** exhibits the highest density of personalization (Avg. P-Links: **7.47**), reflecting the intensive information gathering and rapport building required at the onset of therapy. **Stage 2** forms the backbone of the treatment, spanning the longest duration (Avg. Sessions: **4.13**) and maintaining high intervention richness (Avg. C-Materials: **5.92**), which indicates a sustained focus on deep therapeutic work and skill application. Finally, **Stage 3** facilitates structured closure.

## E Experimental Settings and Dataset Construction Details

### E.1 Model Configurations

For generating clinical dialogues, we employed GPT-5 with high reasoning effort to accurately follow complex clinical psychology instructions while preserving professional terminology and capturing the natural linguistic variability of clients. To balance evaluation effectiveness with computational cost, we selected DeepSeek-V3.1 as our primary evaluation model (LLM-as-a-Judge), owing to its strong performance in Chinese linguistic comprehension and logical reasoning. Due to

budget constraints, we uniformly sampled 1,000 instances from each existing benchmark for comparative analysis.

### E.2 Data Reconstruction and Quality Control

We implemented a rigorous, multi-stage quality control process to transform static clinical case reports into high-fidelity therapeutic dialogues. First, a timeline constraint was algorithmically enforced during generation to ensure the AI counselor only accessed information disclosed up to the current session, eliminating any “anticipatory” knowledge and preserving temporal realism. Second, we extracted a Global Plan from each original case and used it to pre-generate detailed, session-level agendas that guided the AI’s therapeutic focus, maintaining goal continuity and clinical plausibility across long dialogues. Finally, qualified psychology researchers conducted random audits of the generated dialogues, evaluating clinical logic, role adherence, and annotation accuracy; any deficient segments were refined through iterative regeneration or manual correction until they met professional standards—ensuring the dataset’s clinical validity and pedagogical robustness.

### E.3 Authoritative Psychology Resources

A substantial portion of our case reports are derived from two distinguished, peer-reviewed Chinese psychology journals, ensuring that the data adheres to rigorous ethical standards and scientific scrutiny: *Psychological Monthly*<sup>1</sup> and *Theory and Practice of Psychological Counseling*<sup>2</sup>. In addition to empirical case reports, psychology textbooks played a crucial role in our work. They not only helped us extract meta-level therapeutic skills but also served as foundational references for prompt design. Key texts include *The System of Counseling and Therapies*<sup>3</sup>, *Theory and Practice of Counseling & Psychotherapy (Eighth Edition)*<sup>4</sup>, *Cognitive Behavior Therapy: Basics and Beyond (Third Edition)*<sup>5</sup>, *Existential-Humanistic Therapy*<sup>6</sup>, *Behavioral Treatment*<sup>7</sup>, *Psychodynamic Formulation*<sup>8</sup>, *More Than Miracles—The State of the Art*

<sup>1</sup><https://www.xlykzz.com/CN/home>

<sup>2</sup><https://www.sciscanpub.com/journals/tppc>

<sup>3</sup><https://book.douban.com/subject/5979615/>

<sup>4</sup><https://book.douban.com/subject/4179683/>

<sup>5</sup><https://book.douban.com/subject/36815051/>

<sup>6</sup><https://book.douban.com/subject/26304954/>

<sup>7</sup><https://book.douban.com/subject/20494848/>

<sup>8</sup><https://book.douban.com/subject/26327172/>

Table 8: Quantitative Characteristics of the Structured Global Plans and Session Goals. Avg. Sessions and Avg. Objectives denote the average number of sessions per sample and the average number of session objectives, respectively. Avg. P-Links and Avg. C-Materials indicate the average number of persona links and clinical materials per session, representing the density of personalization and intervention richness.

Therapeutic Phase / Stage	Avg. Sessions (per Case)	Avg. P-Links (per Session)	Avg. C-Materials (per Session)	Avg. Objectives (per Session)
<b>Overall (Total)</b>	7.58	4.65	5.73	7.58
Stage 1: Intake & Conceptualization	1.84	7.47	5.91	8.56
Stage 2: Core Intervention	4.13	3.98	5.92	7.51
Stage 3: Consolidation & Termination	1.62	3.14	5.03	6.64

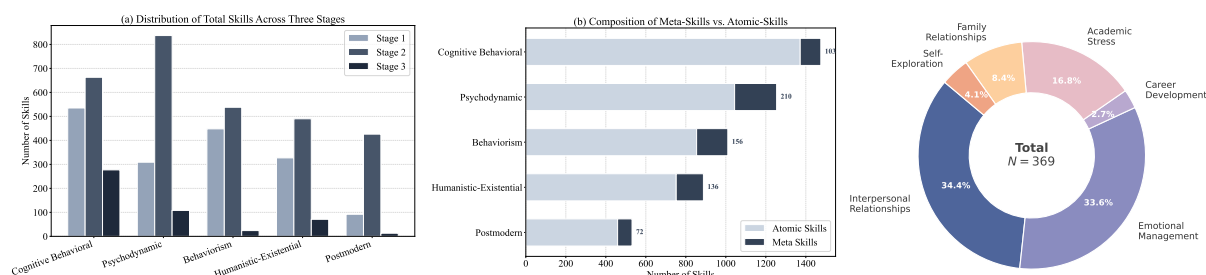


Figure 4: Statistical information of skills and topics.

Table 9: Data quality of our benchmark in terms of client-level metrics.

	Therapy-Shared				Therapy-Specific				
	RRO	Panas	SCL-90↓	SRS	BDI-II↓	IPO↓	CCT	SFBT	STAI
CACTUS (Lee et al., 2024)	6.37	4.92	3.09	7.47	1.84	-	-	-	-
Cpsycoun (Zhang et al., 2024a)	6.15	4.69	2.94	6.83	1.20	<b>0.64</b>	4.04	5.24	-
PsyDial (Qiu and Lan, 2025)	6.36	4.12	3.23	6.99	2.33	-	6.36	-	-
PsyDTCorpus (Xie et al., 2025)	6.94	4.83	3.06	8.01	1.79	-	-	-	-
SimPsyDial (Qiu and Lan, 2024)	6.11	4.72	2.97	7.63	1.63	1.27	4.85	-	-
SMILE (Qiu et al., 2024)	6.52	4.83	3.04	6.80	-	-	-	-	-
Psy-Insight <sub>(en)</sub> (Chen et al., 2025)	6.09	4.90	2.82	5.38	0.50	2.51	5.95	3.92	-
Psy-Insight <sub>(en)</sub> (Chen et al., 2025)	4.98	5.03	2.75	4.91	<b>0.38</b>	1.17	2.90	1.73	-
ESConv (Liu et al., 2021)	6.49	4.68	3.07	6.36	-	-	4.66	-	-
PsychEval (Our)	<b>6.70</b>	<b>5.27</b>	<b>1.48</b>	<b>8.29</b>	1.87	2.17	<b>8.41</b>	<b>7.84</b>	<b>5.57</b>

of Solution-Focused Brief Therapy<sup>9</sup>.

## F Human Expert Validation Details

We recruited graduate students with foundational knowledge in psychology to conduct the human evaluation. Prior to the task, all annotators were fully informed of the research background, detailed annotation rules, and potential emotional risks associated with clinical data. All participants were fairly compensated for their effort in accordance with local wage standards.

<sup>9</sup><https://book.douban.com/subject/26761621/>

### An Example of Clinical Presentation

**Main Problem:** 开学半个多月以来，因家庭矛盾激化陷入焦虑和抑郁情绪，偶尔会独自落泪，感觉自己过往调节情绪的方式皆失效。自觉目前自己在家庭关系、婚恋态度、人际关系、自我成长及生涯发展方面均存在问题，而这些问题的源头是自己的家庭矛盾。(It has been over half a month since the semester started. Due to the intensification of family conflicts, X has fallen into anxiety and depression, occasionally shedding tears alone. X feels that all her previous ways of regulating her emotions have become ineffective. X is aware that she currently has problems in family relationships, attitudes toward marriage and romance, interpersonal relationships, self-growth, and career development, and the source of these problems is her family conflicts )

**Topic:** 家庭关系(Family Relationships)

**Core Demands:** 希望调整情绪状态，厘清家庭问题，让自己不再受家庭问题的困扰如此之深，能够恢复正常的学习和社交。(X hopes to adjust her emotional state, clarify family issues, so that she is no longer so deeply troubled by family problems and can return to normal studying and socializing. )

Figure 5: An Example of Clinical Presentation.

### An Example of Static Traits

**Age:** 25岁(25 years old)

**Name:** 来访者X (化名) (Client X (pseudonym))

**Gender:** 女(Female)

**Occupation:** 研究生二年级(Second-year postgraduate student)

**Educational Background:** 本科毕业，二战考研成功，现为研究生二年级，成绩较好。(Graduated from university with a bachelor's degree, successfully passed the postgraduate entrance examination for the second time, currently in the second year of postgraduate study, with good academic performance)

**Marital Status:** 单身(Single)

**Family Status:** 来自农村，汉族，五口之家，姐妹3个，排行老二；父母关系不睦、常争吵；母亲脾气暴躁、对来访者较依赖；父亲对家庭冷漠、婚内私生活混乱；当地及原生家庭重男轻女思想严重；与姐姐关系亲密但有冲突；妹妹15岁，厌学情绪严重。(From a rural area, a family of five including three sisters, X is the second child, Han ethnicity. The relationship between X's parents is not harmonious and they often quarrel. X's mother has a bad temper and is rather dependent on her. The father is indifferent to the family and has a chaotic private life. There is a serious preference for sons over daughters in the local area and in X's original family. X has a close but conflicting relationship with her elder sister. X's younger sister is 15 years old and has a serious aversion to studying)

**Social Status:** 现居某县城；有一些朋友、老师能够提供支持；平日每周会与同学相约打1~2次羽毛球，开学两周以来曾中断，后恢复。(Currently residing in a county town; There are some friends and teachers who can offer support. On weekdays, X always play badminton with her classmates 1 to 2 times, which was interrupted for two weeks after the start of the new term and then resumed.)

**Medical History:** 无重大躯体疾病，排除器质性病变；无药物服用经历，排除药物滥用；近日入睡困难、常感疲惫，注意力和记忆力下降。(No major physical diseases or organic lesions; No history of taking medication, ruling out drug abuse. Recently, X have had difficulty falling asleep, an often feel tired, with her attention and memory declined)

**Language Features:** 语速平缓，表达清晰，逻辑清楚；谈吐有礼貌且克制，端坐，略有拘谨；求助动机强，配合度高。(Gentle speaking speed, Clear and Logical expression. X's speech is polite and restrained, and the sitting posture is slightly reserved. Strong motivation for seeking help and high cooperation)

Figure 6: An Example of Static Traits.

### An Example of Growth Experience

**Growth Experience:**

小学三年级暑假，X姐姐把某件坏事说成是X干的，X因此被母亲打骂得很重”，X至今仍深感痛苦(During the summer vacation of third grade, X's elder sister blamed a certain misdeed on her, and she was severely beaten and scolded by her mother because of it. X still feels deeply pained by this to this day )

X父亲回乡后对家庭疏离，对X和姐姐也变得冷漠，母亲愈发急躁，经常打骂X，X常感委屈和被忽视(After X's father returned to his hometown, he became distant from the family and indifferent toward her and her sister. X's mother grew increasingly irritable, often beating and scolding her, making her frequently feel wronged and neglected. ) X姐姐辍学后打工寄钱回家使X意识到学习好可以获得父母的关注，认为对家庭“有用”才能被看到(After her elder sister dropped out of school and sent money home from work, X realized that doing well in studies could gain her parents' attention, believing that being "useful" to the family was the way to be seen. )

X妹妹出生后她母亲的情绪好了一些，但之后对妹妹的偏爱愈发明显，使X感到被再次剥夺(After the younger sister was born, X's mother's mood improved somewhat, but later, the increasing favoritism toward the younger sister made X feel deprived once again )

X本科毕业后考研失败在家备考一年，经常听到父母吵架和小妹的哭泣，受家庭环境影响十分压抑、痛苦(After graduating with her bachelor's degree, X failed the postgraduate entrance exam and spent a year studying at home to retake it. During this time, X often heard her parents arguing and her younger sister crying, feeling extremely suppressed and painful due to such family environment. )

X二战考研成功后父母对自己态度转变，父亲到处宣传女儿考上了研究生(After successfully passing the postgraduate entrance exam on her second attempt, X's parents' attitude toward her changed. Her father began telling everywhere that his daughter had been admitted to graduate school. )

Figure 7: An Example of Growth Experience.

### An Example of Meta Skill

**Skill ID:** 3

**Skill Name:** 建立咨询关系(Establishing the Counseling Relationship)

**Skill Description:** 通过表达理解与共情、耐心倾听、尊重来访者感受并处理来访者不满或顾虑，提供温暖的人性关怀，从而获取信任并巩固咨询关系——这是影响咨询效果的关键因素之一，尤其在初期必须着力建立。(By expressing understanding and empathy, listening patiently, respecting the client's feelings, and addressing any dissatisfaction or concerns, the counselor provides warm, humanistic care to gain trust and consolidate the counseling relationship. This is one of the critical factors influencing counseling outcomes and must be prioritized, especially in the initial stages.)

**Parent IDs:** [1, 2, 3]

Figure 8: An Example of Meta Skill.

### An Example of Atomic Skill

**Skill ID:** 33

**Skill Name:** 耐心倾听来访者叙述(Patiently Listening to the Client's Narrative)

**Skill Description:** 在会谈中，咨询师应耐心倾听来访者的叙述，避免打断，提供对话空间，展示对来访者的关注与尊重。(During the session, the counselor should patiently listen to the client's narrative, avoid interruptions, provide space for dialogue, and demonstrate attention and respect for the client.)

**When to Use:** 在会谈的初期阶段，当来访者开始分享个人问题和情感时，咨询师需给予充分的倾听。(In the initial stages of the session, when the client begins to share personal issues and emotions, the counselor must listen attentively.)

**Trigger:** 来访者开始详细描述自己的问题或情感，特别是在展示情绪波动时。(The client begins to describe their problems or emotions in detail, particularly when exhibiting emotional fluctuations.)

**Parent IDs:** [1, 2, 3, 33]

Figure 9: An Example of Atomic Skill.

### An Example of Dialogue

**Counselor:** <think><assessment>初次会面，来访者可能紧张、回避目光。我需要用温和的开场与节奏选择来降低警觉，营造安全感。</assessment><client\_state>轻到中度紧张、谨慎但愿意配合。</client\_state><skill>32:在会谈开始时，通过语言表达对来访者情绪的理解，并展示共情，例如通过表达对来访者困难的理解，缓解其焦虑或紧张情绪。; 35:通过言语和非言语方式表达对来访者的关怀和支持，增强来访者的安全感。</skill><strategy>简短问候与情绪验证，邀请以舒适节奏开展，减少压力。</strategy></think>你好，欢迎你来。我们可以按你觉得舒服的节奏聊，如果有任何时候想停一下或换个话题，都没问题。此刻坐在这里，你的感觉更像是有点紧张，还是比较平静? (<think><assessment>During the initial meeting, the client may feel nervous and avoid eye contact. I need to use a gentle opening and pace my approach to lower their alertness and create a sense of safety.</assessment><client\_state>Mild to moderate tension, cautious but willing to cooperate.</client\_state><skill>32: At the beginning of the session, verbally acknowledge and empathize with the client's emotional state—for example, by expressing understanding of their difficulties to alleviate anxiety or nervousness. 35: Convey care and support through both verbal and nonverbal means to enhance the client's sense of safety.</skill><strategy>Offer a brief greeting and emotional validation, and invite the client to proceed at a comfortable pace to reduce pressure.</strategy></think>Hello, welcome. We can go at whatever pace feels comfortable for you, and if at any point you'd like to pause or switch topics, that's completely fine. Right now, sitting here—would you say you're feeling a bit nervous, or more on the calmer side?)

**Client:** 有点紧张吧... 第一次来，不太会说。(A little nervous, I guess... It's my first time here, and I'm not really sure how to talk about things.)

**Counselor:** ...

Figure 10: An Example of Dialogue.