EmoRes: Toward Adaptive Psychological Support via User-Agnostic Benchmark and Topic-Mining Agent

Zhengwei Zou[1,*], **Xuanming Jiang**[1,2,*], **Baoyi An**[2,*]
Dingyu Nie[2], **Zhengxing Fang**[2], **Qingyu Liu**[3], **Xueming Qian**[1,4]
Guoshuai Zhao[1,4,†], **Zhongyu Yang**[2,†]
[1]Xi'an Jiaotong University
[2]Xi'an Jiyun Technology Co., Ltd., [3]Shenzhen MSU-BIT University
[4]Shaanxi Yulan Jiuzhou Intelligent Optoelectronic Technology Co., Ltd.
guoshuai.zhao@xjtu.edu.cn, yangzhy21@gmail.com

Abstract

Large language models exhibit significant potential for psychological support, yet they often generate fragmented and emotionally inconsistent dialogues that lack the therapeutic structure necessary for reliable assessment. To address these issues, we introduce **VeilEval**, a clinically grounded and privacy-preserving benchmark equipped with interpretable metrics for evaluating multi-turn psychological dialogues. Furthermore, we propose Emotion-Resonance (**EmoRes**), a multi-agent framework that boosts psychological reasoning via a Topic-Mining Emotional Agent and a multi-perspective Self-Reflection Agent, thereby jointly improving topic continuity, emotional coherence, and clinical interpretability. Experiments demonstrate that EmoRes achieves up to $\sim 3\times$ improvement over strong baselines on VeilEval, with its effectiveness further validated by ablation studies and human evaluations.

1 Introduction

> *The most important instrument ... is ... the ability to* ***listen to the patient's story***.
>
> – Charon (2008)

The relentless pace of modern life has exacerbated psychological distress globally (Counts et al., 2023; Firose et al., 2025), urgently necessitating accessible and scalable mental health support. While Large Language Models (LLMs) show promise for empathetic assistance (Lai et al., 2023; Yang et al., 2025b), they often generate fragmented, superficial dialogues lacking therapeutic coherence and emotional depth (Beatty et al., 2022). These limitations highlight a critical need for dialogue systems capable of sustaining stable emotional continuity and clinically grounded reasoning throughout prolonged psychological conversations.

Figure 1: **Comparison of Baseline (left) and EmoRes (right).** EmoRes turns fragmented exchanges into emotionally coherent and reflective dialogue, addressing the challenge of sustaining therapeutic depth across turns. **Legend:** Formulaic generic response; Guided narrative exploration; Contextualized, tailored response.

Recent studies have primarily adopted recursive context concatenation to preserve multi-turn conversational continuity. As illustrated in Figure 1, while retaining superficial linguistic fluency, such methods manifest inherent systemic drawbacks: (i) conversational stagnation induced by template-based repetition (Colombo et al., 2019); (ii) fragmented emotional synthesis, incapable of capturing long-range emotional cues (Chu et al., 2024); and (iii) rigid stylistic conformity, resulting in perfunctorily empathetic yet detached responses (Qiu and Lan, 2024; Tu et al., 2024). Thus, high-quality psychological dialogue generation necessitates not only linguistic fluency but also adaptive topic steering and consistent emotional resonance.

[1]* These authors contributed equally.
[2]† Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 8575–8586

To bridge this gap, we propose Emotion-Resonance (EmoRes), a framework grounded in Solution-Focused Brief Therapy (SFBT) (Bannink, 2007) that prioritizes goal-oriented reflection and emotional reframing. EmoRes operationalizes these clinical principles via two tightly coupled agents engineered to sustain thematic coherence and emotional depth. The Topic-Mining Emotional Agent (TEAgent) dynamically identifies and tracks evolving conversational themes through context-sensitive transitions, ensuring each turn remains clinically relevant and progress-oriented; in parallel, the Self-Reflection Agent iteratively evaluates and refines responses from both counselor and user perspectives, reinforcing emotional alignment and interpretive consistency throughout the entire dialogue. Together, these mechanisms transform fragmented exchanges into reflective, empathetic, and therapeutically grounded conversations.

Moreover, a bottleneck in this field is that existing benchmarks typically rely on small-scale or ethically sensitive datasets, hindering public accessibility and reproducible evaluation (Garg et al., 2023). To overcome these constraints, we introduce **VeilEval**, which builds clinically grounded and systematically structured synthetic profiles, thus preserving therapeutic realism without compromising user privacy. This user-agnostic design is tailored to enable interpretable and scalable assessment of psychological dialogue systems while ensuring both privacy compliance and reproducibility.

Experiments demonstrate that EmoRes consistently outperforms strong baselines in Psychological Health Support (PHS) dialogue generation. Ablation studies further confirm that removing any agent deteriorates performance, underscoring the importance of coordinated framework design.

Our main contributions are as follows:

- We introduce two complementary agents that enhance psychological dialogue quality, topic alignment, and empathy consistency.
- We propose VeilEval, a clinically grounded and privacy-preserving benchmark tailored for the reproducible, ethical, and user-agnostic evaluation of psychological support systems.
- We present EmoRes, a therapy-grounded framework that improves emotional engagement and topic coherence, achieving $\sim 3\times$ performance gains over open-source and commercial baseline LLMs on bilingual VeilEval.

2 Related Work

Solution-Focused Brief Therapy. SFBT is a strengths-oriented approach proposed by de Shazer and Berg in the 1980s (Franklin et al., 2011). Unlike problem-centered therapies, SFBT facilitates structured and goal-directed dialogue focused on solution-building rather than problem analysis (Franklin et al., 2017, 2019). Its brevity and outcome orientation enable measurable progress within a few sessions, making it suitable for scalable computational modeling. However, it has rarely been implemented in LLMs, leaving this structured and iterative dialogue process underexplored in LLM-based PHS systems.

Psychological LLMs. LLM-based systems have advanced the fields of psychological assessment, diagnosis, and counseling (Wu et al., 2025; Lee et al., 2024a; Qiu et al., 2024; Chen et al., 2023). Specifically, assessment models identify mental cues from fragmented dialogues (Tu et al., 2024; Raihan et al., 2024), whereas diagnostic models integrate subjective narratives with clinical data (Hengle et al., 2024; Lan et al., 2024). During treatment, moreover, adaptive models adjust strategies in real time (Xiao et al., 2024; Nie et al., 2025), yet they still suffer from fragmented context and inadequate emotional coherence, constraining therapeutic realism and continuity. Ultimately, integrating psychological principles with structured and emotion-aware reasoning remains an open challenge.

Multi-Agent Systems. Beyond standalone models, multi-agent frameworks facilitate structured collaboration among specialized agents for real-world tasks (Yang et al., 2026a,c,b; Liu et al., 2026b). Recent designs adopt hierarchical message pools (Hu et al., 2025; Chen and Sun, 2025) or debate-based roles (Li et al., 2023; Xu et al., 2025) to enhance reasoning stability and factual accuracy. In psychological contexts, MentalAgora (Lee et al., 2024b) coordinates virtual counselors for personalized therapy, and Mind (Chen et al., 2025) employs empathic, critical, and reconstructive agents for inner dialogue. Despite these significant advances, their loosely coupled interactions often diminish emotional focus and erode therapeutic structure.

In contrast, our EmoRes framework introduces a one-to-one coordination paradigm between a Topic-Mining Emotional Agent and a Self-Reflection Agent, sustaining coherent topic transitions and consistent emotional depth across dialogue turns.

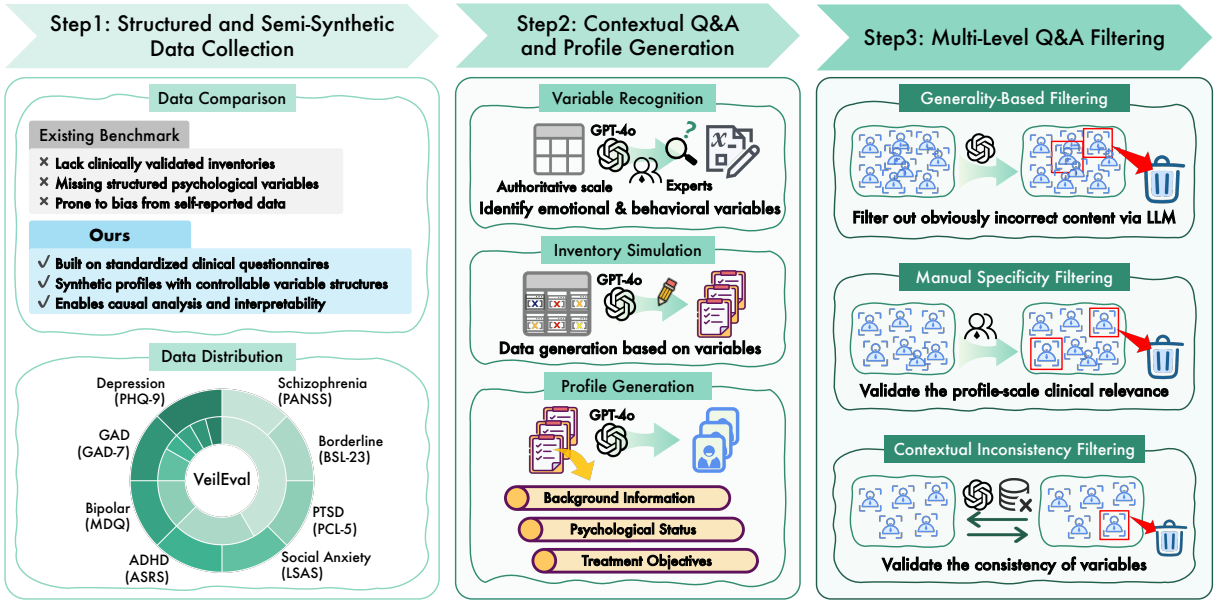

Figure 2: **VeilEval Construction.** ***(i) Data Collection:*** Variables are extracted from validated clinical inventories to synthesize relationally consistent tables simulating clinical assessments. ***(ii) Profile Generation:*** Tables are enriched with user profiles through variable-aware prompting and controlled sampling to preserve psychological patterns. ***(iii) Filtering:*** Q&A pairs are filtered by predefined criteria for generality, clinical relevance, and contextual coherence.

3 Benchmark

3.1 Task Definition

Let the user input sequence be denoted as $U = (u_1, \cdots, u_T)$ and the corresponding model response sequence as $R = (r_1, \cdots, r_T)$. T denotes the conversation length. The dialogue history up to turn $t-1$ is denoted as $H_{t-1} = (u_1, r_1, \cdots, u_{t-1}, r_{t-1})$.

The Doctor Agent generates the next response conditioned on the dialogue history and the accumulated topic states: $r_t = \mathcal{D}(H_{t-1}, S_{\leq t})$, $S_{\leq t} = (S_1, \cdots, S_t)$. At each turn t, TEAgent produces a latent state S_t defined as:

$$S_t = \mathcal{A}(H_{t-1}, u_t), \tag{1}$$

where $S_t = \{(\tau_{t,i}, \sigma_{t,i})\}$ consists of topic labels $\tau_{t,i}$ and their corresponding importance weights $\sigma_{t,i}$, which represent the latent emotional and thematic state inferred from the dialogue.

The dialogue generation process is defined as:

$$P(R \mid U) = \prod_{t=1}^{T} P(r_t \mid r_{<t}, u_{\leq t}), \tag{2}$$

which captures the sequential dependency between user inputs and model responses.

To enhance response quality and consistency, the generated response is iteratively refined by a recursive reflection mechanism $\mathcal{R}$ as follows:

$$R_t^{(i+1)} = \mathcal{R}(R_t^{(i)}, H_{t-1}, S_{\leq t}), \tag{3}$$

until convergence or maximum iteration count k is reached, producing the final response $R_t^* = R_t^{(k)}$. Eq. (3) enables the model to revisit earlier reasoning and align its feedback with therapeutic intent.

3.2 VeilEval Construction

A key challenge in PHS is the lack of standardized, scalable, and privacy-preserving evaluation. Existing datasets often rely on real user data, posing ethical risks and impeding reproducibility. To mitigate this issue, VeilEval is proposed as an automated benchmark that evaluates psychological dialogue systems without relying on personal data, while maintaining clinical realism and interpretability.

Benchmark objectives. As illustrated in Figure 2, VeilEval pursues two main objectives. First, it facilitates multi-turn, psychologically grounded evaluation using clinically validated metrics. Second, it simulates real-world conversational variability, including evasive self-reporting and idealized responses characteristic of therapeutic dialogue.

Data collection and profile generation. Structured variables are extracted from validated psychological inventories that encompass a broad spectrum of mental constructs. Synthetic tables are populated with values that preserve clinical dependencies while ensuring full anonymity. Each inventory serves as an independent evaluation track, enabling fine-grained analysis of disorder-specific reasoning within a unified benchmarking framework.

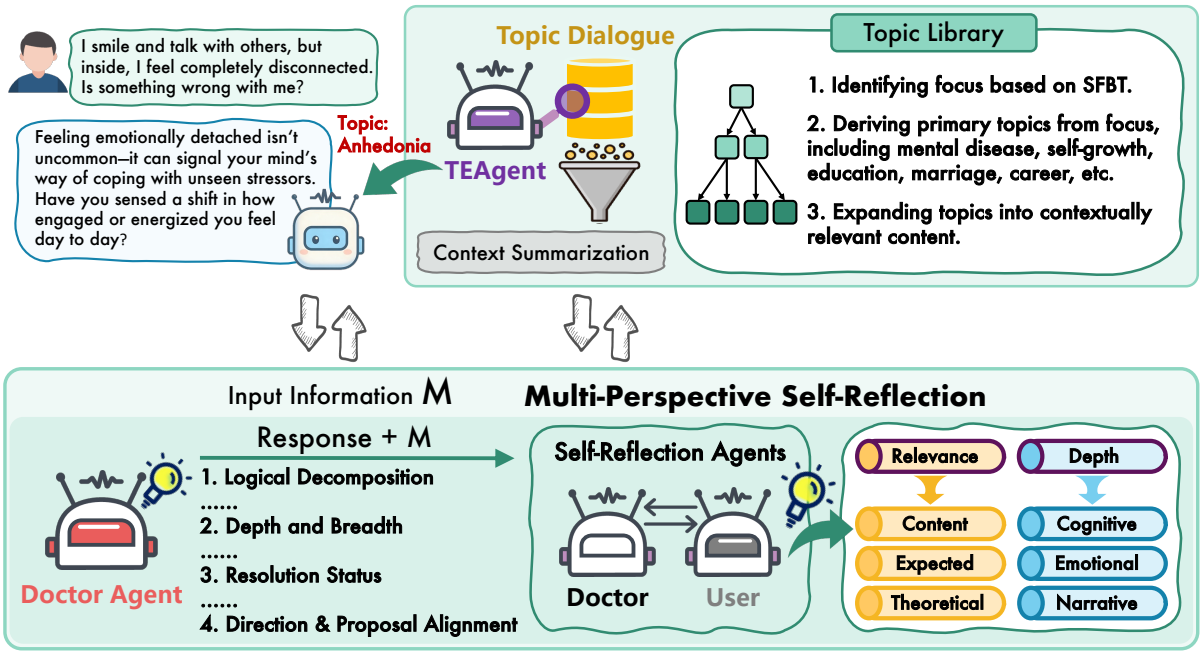

Figure 3: **Overview of EmoRes Framework.** (1) The **TEAgent** identifies diagnostic topics via a hierarchical Topic Library and Context Summarization module. (2) The **Doctor Agent** generates supportive responses conditioned on extracted topics and emotional cues. (3) The **Self-Reflection Agent** revises generated responses from dual clinician (doctor)–user perspectives, thus improving empathic coherence and emotional alignment.

Filtering and quality control. To enhance contextual realism, user profiles consistent with the tabular schema are generated by GPT-4o, incorporating psychosocial and behavioral details. Next, all Q&A pairs are filtered through multi-stage screening, via a combination of automatic and manual validation to ensure generality, clinical relevance, and cross-variable coherence. The resulting dataset constitutes a reliable and ethically grounded resource for evaluating psychological dialogue systems.

4 Methodology

4.1 Motivation

Existing LLM-driven counseling systems suffer from three critical limitations: (1) premature intervention, in which models hastily offer suggestions without sufficient scaffolding to elicit user self-disclosure (Xiao et al., 2024); (2) theoretical misalignment, where systems either lack rigorous psychological grounding or adopt multi-session CBT protocols that do not align with single-session interaction settings (Xiao et al., 2024); and (3) opaque reasoning, providing descriptive yet clinically unvalidated rationales for professional practitioners.

To tackle these challenges, EmoRes incorporates the goal-oriented principles of SFBT into a multi-agent architecture that provides structured, interpretable, and psychologically grounded supportive dialogue. The following sections elaborate the operational mechanisms of each agent.

4.2 Doctor Agent

The Doctor Agent generates supportive, contextually adaptive responses based on thematic and affective cues extracted by the TEAgent. By integrating key topics and latent affective signals, it produces psychologically consistent replies that align with user concerns, steering the dialogue toward empathic engagement and goal-directed reflection.

4.3 Topic-Mining Emotional Agent

TEAgent guides dialogue progression and maintains engagement through a set of three coordinated modules that seamlessly blend SFBT principles with adaptive topic control.

(1) Topic-driven reasoning. TEAgent is initially instantiated with a role prompt embedding both counselor identity and SFBT guidelines. Subsequently, it structures the dialogue as a tree-structured topic graph, where each node represents a goal-aligned topic abstraction. By dynamically expanding subtopics from a clinically curated topic library and adapting to real-time conversational signals, TEAgent sustains coherent and progress-oriented interactions.

(2) Information mining. For each active topic, TEAgent strategically extracts salient affective and factual cues, structuring them into compact contextual packages. These packages anchor the Doctor Agent's reasoning, ensuring responses remain both factually consistent and psychologically valid.

Dataset	Models	Methods	Content Quality		Informativeness				Average
			E-A	*Relevance*	*Repetition*	*Helpfulness*	*Breadth*	*Depth*	
Supervised Fine-Tuning Models									
VeilEval$_{EN}$	**Crispers-7B**	Fine-tuning	6.51	8.00	5.17	7.87	6.65	6.09	6.72
	Falcon-7B-FT	Fine-tuning	7.77	8.86	5.90	8.85	7.77	7.81	7.83
	PsychoCounsel-8B	Fine-tuning	7.80	8.77	6.19	8.73	7.57	7.24	7.72
VeilEval$_{CH}$	**MeChat**	Fine-tuning	5.25	7.24	4.89	7.00	5.46	4.55	5.73
	EmoLLM	Fine-tuning	6.70	8.05	5.52	8.05	6.76	6.12	6.87
	SimpsyBot	Fine-tuning	5.54	7.59	5.09	7.43	5.94	5.13	6.12
Open-Source Models									
VeilEval$_{EN}$	**Qwen3-8B**	Raw	2.23	3.10	3.01	2.98	3.48	3.58	3.08
		+ EmoRes	7.45 (+5.22)	8.37 (+5.27)	6.82 (+3.81)	8.82 (+5.84)	7.98 (+4.50)	8.83 (+5.25)	8.04 (+4.96)
	LLaMA-3.1-8B	Raw	2.20	2.92	2.84	2.80	3.37	3.55	2.95
		+ EmoRes	6.69 (+4.49)	7.93 (+5.01)	6.18 (+3.34)	8.57 (+5.77)	7.62 (+4.25)	8.68 (+5.13)	7.61 (+4.66)
	DeepSeek-R1-8B	Raw	2.35	3.22	2.97	3.41	3.60	3.67	3.20
		+ EmoRes	6.77 (+4.42)	8.22 (+5.00)	6.51 (+3.54)	8.81 (+5.40)	7.96 (+4.36)	8.77 (+5.10)	7.84 (+4.64)
VeilEval$_{CH}$	**Qwen3-8B**	Raw	2.60	3.37	3.29	2.93	3.40	3.64	3.20
		+ EmoRes	7.55 (+4.95)	8.49 (+5.12)	6.44 (+3.15)	8.74 (+5.81)	7.63 (+4.23)	8.71 (+5.07)	7.93 (+4.73)
	LLaMA-3.1-8B	Raw	2.34	3.12	2.98	3.07	3.53	3.70	3.12
		+ EmoRes	6.61 (+4.27)	7.91 (+4.79)	5.95 (+2.97)	8.46 (+5.39)	7.34 (+3.81)	8.44 (+4.74)	7.45 (+4.33)
	DeepSeek-R1-8B	Raw	2.40	3.19	3.05	3.50	3.60	3.65	3.23
		+ EmoRes	6.29 (+3.89)	7.85 (+4.66)	5.53 (+2.48)	8.49 (+4.99)	7.44 (+3.84)	8.23 (+4.58)	7.30 (+4.07)
Commercial Models									
VeilEval$_{EN}$	**GPT-4o-mini**	Raw	2.20	3.02	2.83	3.09	3.47	3.57	3.03
		+ EmoRes	7.14 (+4.94)	8.27 (+5.25)	6.69 (+3.86)	8.77 (+5.68)	7.92 (+4.45)	8.84 (+5.27)	7.94 (+4.91)
	DeepSeek-V3	Raw	2.20	3.10	2.88	3.39	3.56	3.49	3.10
		+ EmoRes	8.14 (+5.94)	8.70 (+5.60)	7.85 (+4.97)	9.08 (+5.69)	8.66 (+5.10)	9.55 (+6.06)	8.67 (+5.57)
	ChatGLM4-plus	Raw	2.23	3.10	2.88	3.53	3.63	3.54	3.15
		+ EmoRes	7.58 (+5.35)	8.53 (+5.43)	7.01 (+4.13)	8.97 (+5.44)	8.06 (+4.43)	8.95 (+5.41)	8.19 (+5.04)
VeilEval$_{CH}$	**GPT-4o-mini**	Raw	2.22	2.97	2.82	3.30	3.40	3.52	3.04
		+ EmoRes	6.46 (+4.24)	7.96 (+4.99)	5.79 (+2.97)	8.43 (+5.13)	7.43 (+4.03)	8.37 (+4.85)	7.40 (+4.36)
	DeepSeek-V3	Raw	2.26	3.21	2.80	3.41	3.54	3.45	3.11
		+ EmoRes	8.40 (+6.14)	8.88 (+5.67)	8.13 (+5.33)	9.10 (+5.69)	8.72 (+5.18)	9.66 (+6.21)	8.81 (+5.70)
	ChatGLM4-plus	Raw	2.41	3.22	3.15	3.41	3.49	3.55	3.21
		+ EmoRes	7.40 (+4.99)	8.35 (+5.13)	6.51 (+3.36)	8.75 (+5.34)	7.72 (+4.23)	8.78 (+5.23)	7.92 (+4.71)

Table 1: **Psychological Inquiry Quality Comparison Based on VeilEval.** The content in "( )" indicates the metric gains of models using the EmoRes framework compared to raw models. *E-A* stands for *Emotional Alignment*.

(3) Context summarization. To preserve long-form continuity within context limits, TEAgent periodically generates condensed summaries that retain affective tone and core intent, enabling seamless topic transitions and stable therapeutic focus.

4.4 Self-Reflection Agent

This agent reinforces topic consistency and psychological depth by iteratively assessing the Doctor Agent's outputs along two dimensions:

(1) Relevance. Evaluating adherence to natural dialogue flow and user-centered concerns, sustaining robust contextual consistency.

(2) Depth. Quantifying the level of psychological insight, exploring internal conflicts, beliefs, and formative experiences that promote self-reflection.

Evaluation is performed from complementary viewpoints: (i) from the doctor's perspective, it inspects logical rigor, affective appropriateness, and alignment with SFBT therapeutic principles; (ii) from the user's perspective, it focuses on empathic resonance, clarity, and practical applicability.

By integrating these assessments, the agent iteratively refines responses to ensure clinical validity and empathic coherence, yielding cohesive and emotionally resonant therapeutic dialogues.

5 Experiments

5.1 Implementation Details

All LLM components are executed via zero-shot prompting with GPT-4o-mini. GPT-4o-mini is also adopted for the TEAgent owing to its superior reasoning capability, while unmodified raw outputs are employed for other modules. All experiments use a temperature of 0.5 for reproducibility. Each response is rated on a 1–10 scale under standardized LLM-as-a-judge protocols with GPT-4o.

In addition to *Relevance* and *Depth*, we introduce four metrics to quantify empathetic engagement and psychological reasoning as follows:

- ***Emotional Alignment*:** evaluates affective congruence, ranging from weak emotional matching to responses that facilitate emotional awareness and regulation.
- ***Repetition*:** measures response diversity and informativeness, penalizing redundant or mechanical generation prevalent in LLMs.
- ***Breadth*:** captures cognitive comprehensiveness by uniting emotional, behavioral, and contextual factors in a cohesive interpretation.

Modules				Content Quality		Informativeness				Average
Multi-Agent	*Summary*	*Self-Reflection*	*Topic-Selection*	*E-A*	*Relevance*	*Repetition*	*Helpfulness*	*Breadth*	*Depth*	
				2.87	3.87	2.87	3.47	3.13	3.67	3.31
✔				7.13 (+4.26)	8.27 (+4.40)	6.23 (+3.36)	8.53 (+5.06)	7.20 (+4.07)	8.63 (+4.96)	7.67 (+4.36)
✔	✔			7.10 (+4.23)	8.27 (+4.40)	6.00 (+3.13)	8.47 (+5.00)	7.30 (+4.17)	8.70 (+5.03)	7.64 (+4.33)
✔		✔		7.63 (+4.76)	8.53 (+4.66)	6.97 (+4.10)	8.70 (+5.23)	7.47 (+4.34)	8.77 (+5.10)	8.01 (+4.70)
✔			✔	7.27 (+4.40)	8.37 (+4.50)	5.70 (+2.83)	8.73 (+5.26)	7.40 (+4.27)	8.70 (+5.03)	7.70 (+4.39)
✔	✔	✔		7.83 (+4.96)	8.50 (+4.63)	7.00 (+4.13)	8.80 (+5.33)	7.37 (+4.24)	8.83 (+5.16)	8.06 (+4.75)
✔		✔	✔	7.60 (+4.73)	8.60 (+4.73)	6.53 (+3.66)	8.60 (+5.13)	7.57 (+4.44)	8.87 (+5.20)	7.96 (+4.65)
✔	✔		✔	7.27 (+4.40)	8.47 (+4.60)	6.30 (+3.43)	8.73 (+5.26)	7.33 (+4.20)	8.77 (+5.10)	7.81 (+4.50)
✔	✔	✔	✔	7.77 (+4.90)	8.67 (+4.80)	7.30 (+4.43)	8.80 (+5.33)	7.70 (+4.57)	8.97 (+5.30)	8.20 (+4.89)

Table 2: **Ablation Study.** The impact of individual modules relative to the baseline model (GPT-4o-mini).

- ***Helpfulness*:** assesses if responses provide practical, personalized guidance conducive to emotional relief and adaptive coping.

5.2 Baseline Selection

We evaluate representative LLMs for PHS on Veil-Eval across three categories as follows:

- **Supervised Fine-Tuned Models:** including Crispers-7B (Zhou et al., 2025), Falcon-7B-FT (Brahma, 2024), PsychoCounsel-8B (Zhang et al., 2025), MeChat (Qiu et al., 2024), EmoLLM (Team, 2024), and SimpsyBot (Qiu and Lan, 2024), all trained on psychological dialogue data.

- **Open-Source Foundation Models:** including Qwen3-8B (Yang et al., 2025a), LLaMA-3.1-8B (Touvron et al., 2024), and DeepSeek-R1-8B (DeepSeek-AI, 2025), which exhibit strong general language capabilities, but lack task-specific optimization for PHS.

- **Commercial Models:** including GPT-4o-mini (OpenAI, 2022), DeepSeek-V3 (DeepSeek-AI, 2024), and ChatGLM4-plus (GLM et al., 2024), which deliver fluent mainstream conversational performance with varying psychological reasoning capacities.

5.3 Main Results

As shown in Table 1, EmoRes enhances both open-source and commercial systems, raising average scores from ∼3.1 to 7.3-8.8 on bilingual Veil-Eval. Among open-source models, Qwen3-8B with EmoRes achieves the best performance (7.99 on average), followed by LLaMA-3.1-8B (7.53) and DeepSeek-R1-8B (7.57). Commercial models show larger gains: DeepSeek-V3 reaches 8.74 (+5.64), GPT-4o-mini reaches 7.67 (+4.64), and ChatGLM4-plus reaches 8.06 (+4.88), confirming strong generalization across model architectures.

Models	GPT-4o-mini	Gemini-2.5-Flash-Lite	Claude-Haiku-3
Open-Source Models			
Qwen3-8B	3.14	4.62	3.17
	7.99 (+4.85)	**7.79** (+3.17)	**7.31** (+4.14)
LLaMA-3.1-8B	3.04	4.48	4.29
	7.53 (+4.49)	**6.99** (+2.51)	**7.96** (+3.67)
DeepSeek-R1-8B	3.19	4.61	4.21
	7.57 (+4.38)	**6.65** (+2.04)	**8.08** (+3.87)
Commercial Models			
GPT-4o-mini	3.04	4.58	4.30
	7.67 (+4.63)	**7.38** (+2.80)	**7.94** (+3.64)
DeepSeek-V3	3.11	5.32	4.72
	8.74 (+5.63)	**7.09** (+1.77)	**8.71** (+3.99)
ChatGLM4-plus	3.18	4.58	3.61
	8.06 (+4.88)	**7.69** (+3.11)	**7.63** (+4.02)

Table 3: **Bias Analysis.** Average scores of all metrics in Table 1 across three independent evaluators, with **boldface** denoting models enhanced by EmoRes.

Metric trends are consistent across settings. For example, *Emotional Alignment* rises by 4.9, *Helpfulness* rises by 5.5, and *Depth* rises by 5.2. Compared with fine-tuned systems, baselines with EmoRes remain >1 points higher on average, showing that structured reasoning and self-reflection can compete with or exceed data-intensive optimization. Qualitatively, generated responses become more reflective and contextually grounded, sustaining topic focus while deepening psychological insight.

These performance improvements alleviate the core limitations identified earlier: (i) paced exploration alleviates premature intervention, (ii) topic grounding stabilizes inferential consistency, and (iii) reflective evaluation enhances transparency and clinical interpretability in practice.

5.4 Ablation Studies

As shown in Table 2, the Multi-Agent framework raises the average (3.31→7.67), reflecting improved multi-turn structuring. Summary slightly lowers the score (7.67→7.64), but enhances the effective information density (*Breadth* 7.20→7.30, *Depth* 8.63→8.70). Self-Reflection increases the average to 8.01, with gains in *E-A* (7.13→7.63) and *Relevance* (8.27→8.53), suggesting stronger con-

Models	*Insight.*	*E-S*	*M-L*	Average
Evaluator: GPT-4o-mini				
Qwen3-8B	3.68	4.00	3.85	3.84
	8.20$_{(+4.52)}$	**9.15**$_{(+5.15)}$	**8.54**$_{(+4.69)}$	**8.63**$_{(+4.79)}$
LLaMA-3.1-8B	3.58	3.98	3.92	3.83
	7.92$_{(+4.34)}$	**8.97**$_{(+4.99)}$	**8.38**$_{(+4.46)}$	**8.42**$_{(+4.59)}$
DeepSeek-V3	3.66	4.22	4.18	4.02
	8.88$_{(+5.22)}$	**9.64**$_{(+5.42)}$	**8.67**$_{(+4.49)}$	**9.06**$_{(+5.04)}$
ChatGLM4-plus	3.72	3.99	4.11	3.94
	8.30$_{(+4.58)}$	**9.17**$_{(+5.18)}$	**8.68**$_{(+4.57)}$	**8.72**$_{(+4.78)}$
Evaluator: Claude-Haiku-3				
Qwen3-8B	3.78	4.68	4.23	4.23
	8.51$_{(+4.73)}$	**8.99**$_{(+4.31)}$	**8.87**$_{(+4.64)}$	**8.79**$_{(+4.56)}$
LLaMA-3.1-8B	3.78	4.76	4.25	4.27
	8.35$_{(+4.57)}$	**8.88**$_{(+4.12)}$	**8.63**$_{(+4.38)}$	**8.62**$_{(+4.35)}$
DeepSeek-V3	3.94	5.16	5.26	4.79
	9.12$_{(+5.18)}$	**9.27**$_{(+4.11)}$	**9.13**$_{(+3.87)}$	**9.18**$_{(+4.39)}$
ChatGLM4-plus	3.83	4.76	4.34	4.31
	8.65$_{(+4.82)}$	**9.02**$_{(+4.26)}$	**8.97**$_{(+4.63)}$	**8.88**$_{(+4.57)}$

Table 4: **Results for User-Centric Metrics,** including *Insightfulness* (*Insight.*), *Emotional Safety* (*E-S*), *Motivational Lift* (*M-L*), and mean scores over VeilEval$_{\text{EN/CH}}$, with **boldface** denoting EmoRes-enhanced results.

textual alignment. Topic Selection yields smaller gains (7.67→7.70), mainly improving *Helpfulness* (8.53→8.73) via more solution-focused responses.

Table 2 also reveals the necessity of collaborative coordination: excluding any module degrades performance, most severely when Self-Reflection is removed, causing a 0.39 drop in the average score.

These findings confirm that EmoRes depends on synergistic cooperation among agents: (i) routing organizes conversational flow, (ii) summarization preserves theory-consistent transitions, and (iii) reflection enforces relevance and depth, collectively converting fragmented exchanges into coherent and psychologically grounded dialogues.

5.5 Bias Analysis

Table 3 assesses the robustness of EmoRes across different evaluators. Across three distinct LLM judges (GPT-4o-mini (OpenAI, 2022), Gemini-2.5-Flash-Lite (Cloud, 2025), and Claude-Haiku-3 (Anthropic, 2024)), EmoRes consistently surpasses vanilla models, with average improvements ranging from +1.7 to +5.6 points. Open-source backbones such as Qwen3-8B and LLaMA-3.1-8B exhibit stable cross-evaluator gains, whereas commercial systems including DeepSeek-V3 and ChatGLM4-plus achieve the highest and second-highest average absolute scores, respectively (8.18 and 7.79).

These findings confirm that the performance gains of EmoRes are not evaluator-dependent, indicating minimal evaluation bias and strong generalization across widely-used judging paradigms.

Models	Methods	Mean	StdDev	t-test p	Wilcoxon p
Qwen3-8B	Raw	3.21	0.40	< 0.01	< 0.01
	+ EmoRes	8.16	0.54		
GPT-4o-mini	Raw	3.23	0.33	< 0.01	< 0.01
	+ EmoRes	7.87	0.65		

Table 5: **Statistical Comparison on VeilEval Benchmark.** StdDev stands for standard deviation. p-values from paired one-sided t-test and Wilcoxon signed-rank test are reported at significance level $\alpha = 5\%$.

5.6 User-Centric Analysis

To complement the core metrics in VeilEval, we introduce three user-centric indicators that assess deeper dimensions of therapeutic quality: (i) ***Insightfulness*** captures how effectively a response promotes self-reflection and cognitive exploration; (ii) ***Emotional Safety*** gauges whether the wording preserves empathy while avoiding emotional harm; and (iii) ***Motivational Lift*** measures the extent to which a reply inspires constructive, goal-oriented engagement. Together, these indicators extend evaluation beyond surface-level linguistic relevance toward psychological realism and behavioral impact.

Table 4 summarizes results across two evaluators (GPT-4o-mini and Claude-Haiku-3). EmoRes consistently improves performance in both English and Chinese settings, with average gains of ∼4.6. In particular, DeepSeek-V3 achieves the highest average score (9.12), with the two largest increases occur in *Emotional Safety* (+5.42) and *Insightfulness* (+5.22), showing that EmoRes produces emotionally secure yet motivating responses. Overall, these findings indicate that EmoRes enhances not only empathy and coherence, but also alignment with nuanced cognitive and affective states, maintaining consistency across evaluators and languages.

5.7 Statistical Significance Study

To validate the robustness of the observed improvements, we conduct ten independent runs with distinct random seeds under identical experimental settings. Paired one-sided t-tests and Wilcoxon signed-rank tests are performed with a significance level of $\alpha = 0.05$ (Jiang et al., 2025b). The null hypothesis states that EmoRes performs equally or worse than the baseline models, while the alternative hypothesis posits superior performance.

As shown in Table 5, EmoRes achieves substantially higher mean scores: 7.87 for GPT-4o-mini and 8.16 for Qwen3-8B, compared to 3.23 and 3.21 for their vanilla counterparts. Although the multi-agent setting slightly increases variance, both models produce small p-values (<0.05).

These results statistically confirm that EmoRes delivers significant and reproducible performance gains, which arise from structured agent coordination rather than random variation in practice.

5.8 Rationality Analysis

We further analyze the theoretical and practical rationale underlying VeilEval and EmoRes.

Why SFBT-based EmoRes is effective. SFBT offers a future-oriented, solution-driven framework that naturally aligns with concise, goal-directed LLM interactions. By prioritizing reflective dialogue over diagnostic depth, it supports coherent topic progression and psychologically grounded reasoning in single-session scenarios, ensuring interpretability and consistent therapeutic focus.

Why a multi-agent design is adopted. Psychological dialogue generation requires emotional sensitivity, structured reasoning, and interpretability–objectives difficult to unify in a single model. EmoRes distributes these responsibilities across three specialized agents: TEAgent extracts affective and topical cues, the Doctor Agent generates clinically consistent responses, and the Self-Reflection Agent assesses relevance and depth from dual perspectives. This collaborative scheme stabilizes conversational flow, improves emotional precision, and enhances interpretability. Ablation studies verify that removing any agent significantly impairs coherence and alignment, demonstrating that the multi-agent architecture is indispensable for psychologically consistent dialogue generation.

Why VeilEval adopts a user-agnostic design. Real user data raises inevitable privacy risks, annotation noise, and excessive behavioral heterogeneity that hinder reproducibility. VeilEval replaces such data with synthetic yet clinically valid profiles derived from standardized psychological inventories, achieving therapeutic realism without ethical compromises. Therefore, this controlled variability enables a scalable, consistent, and fair evaluation of psychological dialogue systems.

6 Human Evaluation

To validate the real-world efficacy of EmoRes in psychological dialogue systems, we conducted a human evaluation using GPT-4o-mini outputs. Ten topics were randomly sampled from 160 scenarios, and for each topic, baseline and EmoRes outputs were paired and randomly arranged in an A/B format. Each pair was rated by 100 trained annotators

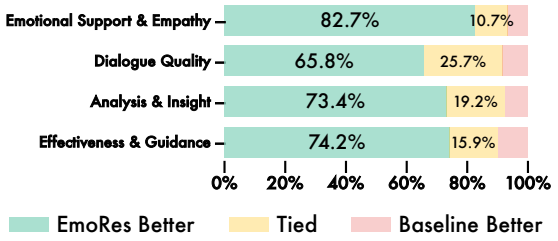

Figure 4: **Human Evaluation.** Comparison of EmoRes and the baseline model (GPT-4o-mini) across four dimensions of psychological dialogue quality.

with verified backgrounds in psychology or counseling, who follow standardized annotation guidelines and calibration samples to ensure consistent evaluation criteria. Annotators classify Response A as *Better*, *Worse*, or *Tied* across four dimensions: (1) *Emotional Support and Empathy*, (2) *Dialogue Quality* (fluency and coherence), (3) *Analysis and Insight*, and (4) *Effectiveness and Guidance*. All presentation orders were randomized, and inter-rater agreement was monitored to ensure reliability.

As shown in Figure 4, EmoRes surpasses the baseline across all dimensions, attaining an 82.7% preference rate for *Emotional Support and Empathy*, and over 73% for *Effectiveness and Guidance* and *Analysis and Insight*. The narrower advantage in *Dialogue Quality* mainly stems from some annotators favoring the concise, structured response style typical of conventional psychological counseling. These results verify that EmoRes consistently generates more empathetic, insightful, and actionable responses that align with human judgment.

7 Conclusion

Current psychological dialogue systems suffer from insufficient clinical grounding, fragmented reasoning, and unstable therapeutic efficacy. Thus, we introduce EmoRes, a multi-agent framework integrating topic-guided reasoning and self-reflective assessment to ensure conversational coherence, empathetic engagement, and clinical validity in PHS. On the privacy-preserving VeilEval benchmark, EmoRes achieves consistent and significant gains across open-source, commercial, and fine-tuned models, exhibiting strong robustness and transferability without task-specific retraining. Ablation and human evaluations validate that each agent provides distinct yet complementary strengths, enabling interpretable and psychologically consistent reasoning. These findings position EmoRes as a feasible solution for building clinically reliable LLM-based systems for accessible PHS.

Discussion. While existing psychological LLMs rely on dialogue analysis for affective reasoning, advances in real-time and user-agnostic physiological signal processing (Cui et al., 2026; Jiang et al., 2026; Liu et al., 2024; Can and Ersoy, 2023), combined with LLMs' increasing ability to decode subjects' affective variations and basic action intentions (Gu et al., 2026; Jiang et al., 2025a; d'Ascoli et al., 2025; Liu et al., 2026a), may enable future PHS systems to transcend text-only paradigms and achieve groundbreaking cross-modal progress.

8 Limitations

This section critically addresses the limitations of the current research.

Model Choice for Dialogue Construction and Evaluation. This study uses GPT-4o-mini as the user-side model for dialogue construction, while multiple auxiliary models are integrated with the proposed EmoRes framework for response generation. GPT-4o-mini is also employed as an evaluator. LLM selection significantly impacts multiple dimensions of generated dialogues, such as content and style. Furthermore, the choice of evaluation models may bias the interpretation of final results to varying degrees. In subsequent research, we intend to explore alternative LLMs as both generative sources and evaluation benchmarks to improve dialogue diversity and reliability.

Dialogue Quality and Human Evaluation Bias. Although we have employed LLMs and manual efforts for extensive dialogue quality evaluation and refinement, the model focuses primarily on easily measurable metrics (e.g., repetitiveness and topic relevance). Meanwhile, human reviewers, including experts and students from diverse social backgrounds, may introduce unintended biases, despite comprehensive pre-review training.

Psychological Change Capture Bias. Although clinical metrics are used to assess user responses, their reliability in detecting fleeting, nuanced psychological status changes during short interactions remains unproven. This highlights the need for rigorous empirical evaluation of these metrics and identifies avenues to optimize VeilEval's analytical framework, thereby improving its ability to capture such brief, subtle psychological fluctuations.

Increased Response Latency. The multi-agent coordination in EmoRes incurs additional computational overhead relative to backbone models, compromising its real-time performance.

9 Ethical Considerations

Data Usage and Privacy Protection. The VeilEval benchmark constructed in this study contains no identifiable personal information or sensitive content, retaining only anonymized multi-turn dialogues for psychological assessment and model training. Following the principle of minimal data necessity, only data directly relevant to research objectives was collected, with multi-layered encryption protocols implemented to prevent unauthorized access during storage and transmission (Hu et al., 2023). In addition, all data governance procedures were subject to specialized information security audits to mitigate leakage risks.

Participant Rights and Informed Consent. A large group of participants was recruited for human evaluations. All signed informed consent forms acknowledging their interactions with AI-based emotional support models and their unconditional right to withdraw without penalty. This study adheres to the American Psychological Association's Right to Withdraw principle, with data recording terminated immediately after dialogue completion. Participants' emotion scales and reflections were anonymized to prevent traceability.

Model Safety and Risk Mitigation. Multi-tiered safety filters were incorporated during training and deployment, including sensitive content detection and dialogue redirection for high-risk topics. The model activates safety protocols (e.g., recommending professional help) when detecting self-injury cues or severe mental health crises. This study makes no claims of therapeutic efficacy; instead, it positions the system as an auxiliary assessment tool with explicit limitations in high-risk scenarios.

Fairness and Bias Mitigation. Models may exhibit biased judgments across different populations due to structural data biases. To address this, diverse and balanced datasets were constructed, and performance disparities across subgroups were systematically evaluated. Regular third-party fairness audits are advised to mitigate systemic bias and prevent the amplification of social inequalities.

Future Compliance and Social Responsibility. Future work will further enhance data governance and algorithmic transparency while developing cross-cultural ethical guidelines. Therefore, we advocate for standardized safety metrics and ethical reporting norms in psychological LLM research to ensure legally and morally robust AI-assisted psychological health support applications.

10 Acknowledgments

This project was jointly supported by the National Natural Science Foundation of China (Grant No. 62372364) and the Technical Innovation Guidance Plan of Shaanxi Province, China (Grant No. 2024QCY-KXJ-199). We sincerely appreciate the valuable participation of all volunteers in benchmark construction and human evaluation, whose dedicated contributions have significantly strengthened the real-world relevance of this work.

References

Anthropic. 2024. Claude 3 haiku: our fastest model yet. https://www.anthropic.com/news/claude-3-haiku. Accessed: 2025-10-07.

Fredrike P Bannink. 2007. Solution-focused brief therapy. *Journal of Contemporary Psychotherapy*, 37(2):87–94.

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

Arun Brahma. 2024. Finetuning of falcon-7b llm using qlora on mental health conversational dataset.

Yekta Said Can and Cem Ersoy. 2023. Smart Affect Monitoring With Wearables in the Wild: An Unobtrusive Mood-Aware Emotion Recognition System . *IEEE Transactions on Affective Computing*, 14(04):2851–2863.

Rita Charon. 2008. *Narrative medicine: Honoring the stories of illness*. Oxford University Press.

Kai Chen and Zebing Sun. 2025. Deeppsy-agent: A stage-aware and deep-thinking emotional support agent system. *Preprint*, arXiv:2503.15876.

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

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

KuanChao Chu, Yi-Pei Chen, and Hideki Nakayama. 2024. Cohesive conversations: Enhancing authenticity in multi-agent simulated dialogues. In *Proceedings of the 1st Conference on Language Modeling (COLM)*, Philadelphia, PA, USA. OpenReview.

Google Cloud. 2025. Gemini 2.5 flash-lite. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash-lite. Accessed: 2025-10-07.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3734–3743, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicholas Z. Counts, David E. Bloom, and Neal Halfon. 2023. Psychological distress as a systemic economic risk in the usa. *Nature Mental Health*, 1:950–955.

Wenhui Cui, Christopher Michael Sandino, Hadi Pouransari, Ran Liu, Juri Minxha, Ellen L Zippi, Erdrin Azemi, and Behrooz Mahasseni. 2026. Embridge: Enhancing gesture generalization from emg signals through cross-modal representation learning. In *The Fourteenth International Conference on Learning Representations*, Rio de Janeiro, Brazil. OpenReview.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Stéphane d'Ascoli, Corentin Bel, Jérémy Rapin, Hubert Banville, Yohann Benchetrit, Christophe Pallier, and Jean-Rémi King. 2025. Towards decoding individual words from non-invasive brain recordings. *Nature Communications*, 16(1):10521.

Musthafa Mohamed Firose, Bogahawaththage Nishadi Madushika Chathurangi, and Imriyas Kamardeen. 2025. Work stress among construction professionals during an economic crisis: a case study of sri lanka. *Smart and Sustainable Built Environment*.

C. Franklin, K. W. Bolton, and S. Guz. 2019. Solution-focused brief family therapy. In B. H. Fiese, M. Celano, K. Deater-Deckard, E. N. Jouriles, and M. A. Whisman, editors, *APA handbook of contemporary family psychology: Family therapy and training*, pages 139–153. American Psychological Association.

C. Franklin, A. Zhang, A. S. Froerer, and S. K. Johnson. 2017. Solution focused brief therapy: A systematic review and meta-summary of process research. *Journal of Marital and Family Therapy*, 43(1):16–30.

Cynthia Franklin, Terry S. Trepper, Eric E. McCollum, and Wallace J. Gingerich. 2011. *Solution-Focused Brief Therapy: A Handbook of Evidence-Based Practice*. Oxford University Press.

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

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Wei Gu, Luo Tianming, Qiran Zhang, Mohan Ye, Xiao Shen, Wenxin Chen, Yunhuan Li, Yichen Zhang, Jing Hong, Bao-liang Lu, and 1 others. 2026. Cerebragloss: Instruction-tuning a large vision-language model for fine-grained clinical eeg interpretation. In *The Fourteenth International Conference on Learning Representations*, Rio de Janeiro, Brazil. OpenReview.

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.

Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, and Ping Luo. 2025. HiAgent: Hierarchical working memory management for solving long-horizon agent tasks with large language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32779–32798, Vienna, Austria. Association for Computational Linguistics.

Songbo Hu, Han Zhou, Mete Hergul, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Ivan Vulić, and Anna Korhonen. 2023. Multi3woz: A multilingual, multi-domain, multi-parallel dataset for training and evaluating culturally adapted task-oriented dialog systems. *Transactions of the Association for Computational Linguistics*, 11:1396–1415.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. 2025a. NeuroLM: A universal multitask foundation model for bridging the gap between language and EEG signals. In *The Thirteenth International Conference on Learning Representations*, Singapore. OpenReview.

Xuanming Jiang, Baoyi An, Zhengwei Zou, Dingyu Nie, Jialie Shen, Xueming Qian, and Guoshuai Zhao. 2025b. Ear with eye: Lightweight multimodal audio-visual network inspired by bionic structures. In *Proceedings of the 33rd ACM International Conference on Multimedia*, page 1346–1355, New York, NY, USA. Association for Computing Machinery.

Xuanming Jiang, Dingyu Nie, Baoyi An, Yuzhe Zheng, Yichuan Mao, Jialie Shen, Xueming Qian, Zhiwen Jin, Wei Lan, and Guoshuai Zhao. 2026. Whole-field action sensing via wearable single-channel emg sensors and resource-efficient motion network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(21):17508–17516.

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. *Preprint*, arXiv:2307.11991.

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

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. 2024a. 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. 2024b. Mentalagora: A gateway to advanced personalized care in mental health through multi-agent debating and attribute control. *Preprint*, arXiv:2407.02736.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, pages 51991–52008, Red Hook, NY, USA. Curran Associates, Inc.

Ruoqi Liu, Yuelin Bai, Xiang Yue, and Ping Zhang. 2026a. Teaching multimodal llms to comprehend 12-lead electrocardiographic images. *npj Digital Medicine*.

Yunsong Liu, Zunamys I Carrero, Xiaofeng Jiang, Dyke Ferber, Georg Wölflein, Li Zhang, Sanddhya Jayabalan, Tim Lenz, Zhouguang Hui, and Jakob Nikolas Kather. 2026b. Benchmarking large language model-based agent systems for clinical decision tasks. *npj Digital Medicine*.

Zengding Liu, Chen Chen, Jiannong Cao, Minglei Pan, Jikui Liu, Nan Li, Fen Miao, and Ye Li. 2024. Large language models for cuffless blood pressure measurement from wearable biosignals. In *Proceedings of*

the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, BCB '24, New York, NY, USA. Association for Computing Machinery.

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 Transactions on Computing for Healthcare*.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. Accessed: 2025-03-04.

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, Miami, Florida, USA. Association for Computational Linguistics.

Huachuan Qiu and Zhenzhong Lan. 2024. Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions. *Preprint*, arXiv:2408.15787.

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.

EmoLLM Team. 2024. Emollm: Reinventing mental health support with large language models.

Hugo Touvron and 1 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Shenghan Wu, Yimo Zhu, Wynne Hsu, Mong-Li Lee, and Yang Deng. 2025. From personas to talks: Revisiting the impact of personas on LLM-synthesized emotional support conversations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5439–5453, Suzhou, China. Association for Computational Linguistics.

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

Ancheng Xu, Di Yang, Renhao Li, Jingwei Zhu, Minghuan Tan, Min Yang, Wanxin Qiu, Mingchen Ma, Haihong Wu, Bingyu Li, and 1 others. 2025. Autocbt: An autonomous multi-agent framework for cognitive behavioral therapy in psychological counseling. *Preprint*, arXiv:2501.09426.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Zhongyu Yang, Wei Pang, and Yingfang Yuan. 2026a. Xr: Cross-modal agents for composed image retrieval. In *Proceedings of the ACM Web Conference 2026*, WWW '26, page 2071–2082, New York, NY, USA. Association for Computing Machinery.

Zhongyu Yang, Junhao Song, Siyang Song, Wei Pang, and Yingfang Yuan. 2025b. MERMAID: Multi-perspective self-reflective agents with generative augmentation for emotion recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24650–24666, Suzhou, China. Association for Computational Linguistics.

Zhongyu Yang, Zuhao Yang, Shuo Zhan, Tan Yue, Wei Pang, and Yingfang Yuan. 2026b. Svagent: Storyline-guided long video understanding via cross-modal multi-agent collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Denver CO, United States. IEEE.

Zhongyu Yang, Yingfang Yuan, Xuanming Jiang, Baoyi An, and Wei Pang. 2026c. Inex: Hallucination mitigation via introspection and cross-modal multi-agent collaboration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(35):29829–29837.

Mian Zhang, Shaun M Eack, and Zhiyu Zoey Chen. 2025. Preference learning unlocks llms' psychocounseling skills. *Preprint*, arXiv:2502.19731.

Jinfeng Zhou, Yuxuan Chen, Jianing Yin, Yongkang Huang, Yihan Shi, Xikun Zhang, Libiao Peng, Rongsheng Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. 2025. Crisp: Cognitive restructuring of negative thoughts through multi-turn supportive dialogues. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32474–32503, Suzhou, China. Association for Computational Linguistics.