

EmoHarbor: Evaluating Personalized Emotional Support by Simulating the User’s Internal World

Jing Ye^{1,2}, Lu Xiang^{1,2*}, Yaping Zhang^{1,2}, Chengqing Zong^{1,2}

¹State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, CAS, Beijing, China

²School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

yejing2022@ia.ac.cn; {lu.xiang, yaping.zhang, cqzong}@nlpr.ia.ac.cn

Abstract

Current evaluation paradigms for emotional support conversations tend to reward generic empathetic responses, yet they fail to assess whether the support is genuinely personalized to users’ unique psychological profiles and contextual needs. We introduce **EmoHarbor**, an automated evaluation framework that adopts a **User-as-a-Judge** paradigm by simulating the user’s inner world. EmoHarbor employs a Chain-of-Agent architecture that decomposes users’ internal processes into three specialized roles, enabling agents to interact with supporters and complete assessments in a manner similar to human users. We instantiate this benchmark using 100 real-world user profiles that cover a diverse range of personality traits and situations, and define 10 evaluation dimensions of personalized support quality. Comprehensive evaluation of 20 advanced LLMs on EmoHarbor reveals a critical insight: while these models excel at generating empathetic responses, they consistently fail to tailor support to individual user contexts. This finding re-frames the central challenge, shifting research focus from merely enhancing generic empathy to developing truly user-aware emotional support. EmoHarbor provides a reproducible and scalable framework to guide the development and evaluation of more nuanced and user-aware emotional support systems ¹.

1 Introduction

Emotional Support Conversation (ESC) systems are designed to recognize users’ affective states and provide tailored comfort and assistance through multi-turn interactions (Peng et al., 2022; Rains et al., 2020; Liu et al., 2021). While substantial progress has been made in generating fluent and empathetic responses, the effectiveness of these systems critically depends on *personalization* (Rogers,

*Corresponding Author

¹<https://github.com/kaedeharakazuha678/EmoHarbor-25.git>

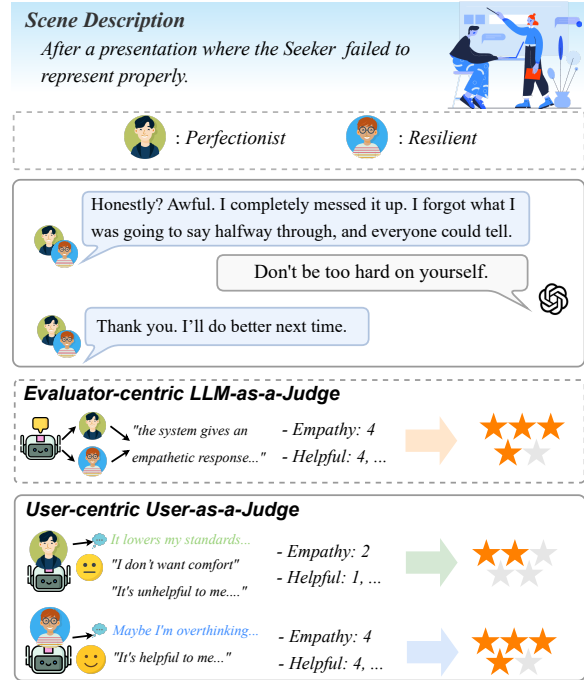


Figure 1: Comparison of different evaluation paradigms for assessing personalized emotional support. The evaluator-centric paradigm fails to perform subjective assessments on behalf of users, whereas the user-centric paradigm can more accurately capture the quality of personalized ESC systems.

2013; Zhang et al., 2018; Campos et al., 2018; Zollo et al., 2025; Ye et al., 2026). Personalization refers to the system’s ability to dynamically adapt support strategies to an individual’s unique psychological profile (Fleeson, 2001) and context-specific needs (Tamir, 2016).

Despite its central importance, existing evaluation approaches suffer from a fundamental limitation: they follow an **evaluator-centric** paradigm, judging ESC quality from an external, ostensibly objective standpoint and failing to capture users’ subjective experiences. For instance, token- and embedding-based metrics (e.g., BLEU (Papineni et al., 2002), BERTScore (Zhang* et al., 2020)) rely solely on reference responses and cannot reflect the open-ended, context-sensitive nature of emotional

support. Human evaluation, although more flexible, is prohibitively expensive. Even recent LLM-based evaluators (Zhao et al., 2024; Madani and Srihari, 2025; Zhang et al., 2024a), which offer scalable alternatives, adopt a one-size-fits-all “*expert*” perspective—assessing responses purely based on the external dialogue context—thereby overlooking the nuanced, persona-driven internal states that shape how individual users experience the conversation.

As illustrated in Figure 1, consider two users seeking support after a failed presentation: one is a perfectionist who fixates on minor flaws, while the other is resilient but frustrated by the lack of constructive feedback. If both receive a generic response such as “*Don’t be too hard on yourself*,” they may interpret it very differently: the former might feel that the comment diminishes their sense of responsibility, whereas the latter might perceive it as encouragement for self-acceptance and growth. Using an evaluator-centric, LLM-as-a-judge approach might assign a high score because the response expresses empathy. However, from the users’ perspective, the perfectionist might rate the response poorly. The subjective nature of emotional support necessitates a paradigm shift from evaluating *what a good supporter would say* to *what this specific user needs and how they would perceive the support*.

To this end, we introduce **EmoHarbor**, a novel evaluation framework based on the **User-as-a-Judge** paradigm that uses agent-based simulation to model the user’s internal world. EmoHarbor simulates how a *specific* user with a particular personality, emotional state, and conversational history would perceive and respond to support. Specifically, EmoHarbor utilizes a Chain-of-Agent architecture that decomposes the user’s internal cognitive processes into three specialized roles: a **User Thinker** that models internal reflections and subjective perceptions based on the user’s profile; a **User Talker** that generates natural, personality-consistent dialogue; and a **User Evaluator** that delivers personalized evaluations of the responses, grounded in the user’s evolving emotional state and needs. We instantiate this framework with a new curated benchmark of 100 real-world user profiles. A comprehensive evaluation of 20 advanced LLMs using EmoHarbor reveals a critical disconnect: while models excel at generic empathy, they consistently fail to tailor support to individual contexts. EmoHarbor provides a reproducible and scalable evaluation to guide the development of

more nuanced and user-aware emotional support systems.

Our main contributions are as follows:

- We introduce EmoHarbor, an evaluation framework that implements the User-as-a-Judge paradigm via a Chain-of-Agent architecture to simulate nuanced user perspectives.
- We validate EmoHarbor through empirical analyses, demonstrating high agreement with human judgments and strong discriminative power as a benchmark for evaluating personalized Emotional Support conversation systems.
- We conduct a comprehensive evaluation of 20 LLMs, revealing that, despite solid general empathetic abilities, they often fail to provide personalized emotional support.

2 Method

EmoHarbor adopts the User-as-a-Judge paradigm by simulating a user’s internal state to produce an interpretable, subjective evaluation. This is realized through a Chain-of-Agent architecture, in which multiple specialized agents collaborate to simulate the user’s cognitive, conversational, and evaluative processes. Figure 2 illustrates the overall workflow. The framework is built around three key design questions: (i) **how to benchmark** (the user profile construction), (ii) **how to simulate** user behavior (agent specialization), and (iii) **what to evaluate** (evaluation dimensions). We elaborate on each of these components in the following sections and provide the workflow algorithm in Appendix C.

2.1 How to Benchmark

User Profile Design. A realistic, detailed user profile is the cornerstone of effective role-playing, as it enables the simulated user to exhibit coherent individuality rather than generic behavior. Partly following Zhao et al. (2025), we define a user profile \mathcal{P}_U as:

$$\mathcal{P}_U = \{\mathcal{D}, \mathcal{P}, \mathcal{C}, \mathcal{S}\} \quad (1)$$

where: (1) \mathcal{D} represents **demographic attributes** (e.g., age, gender, occupation), grounding the user in a concrete context; (2) \mathcal{P} denotes **preference-related attributes** (e.g., personality traits, Big-Five, MBTI, habits, hobbies, speech style), shaping distinctive behavioral patterns; (3) \mathcal{C} captures **counseling-related attributes** (e.g., problem description, emotional state, goals, role relations), encoding the psychological background; and (4)

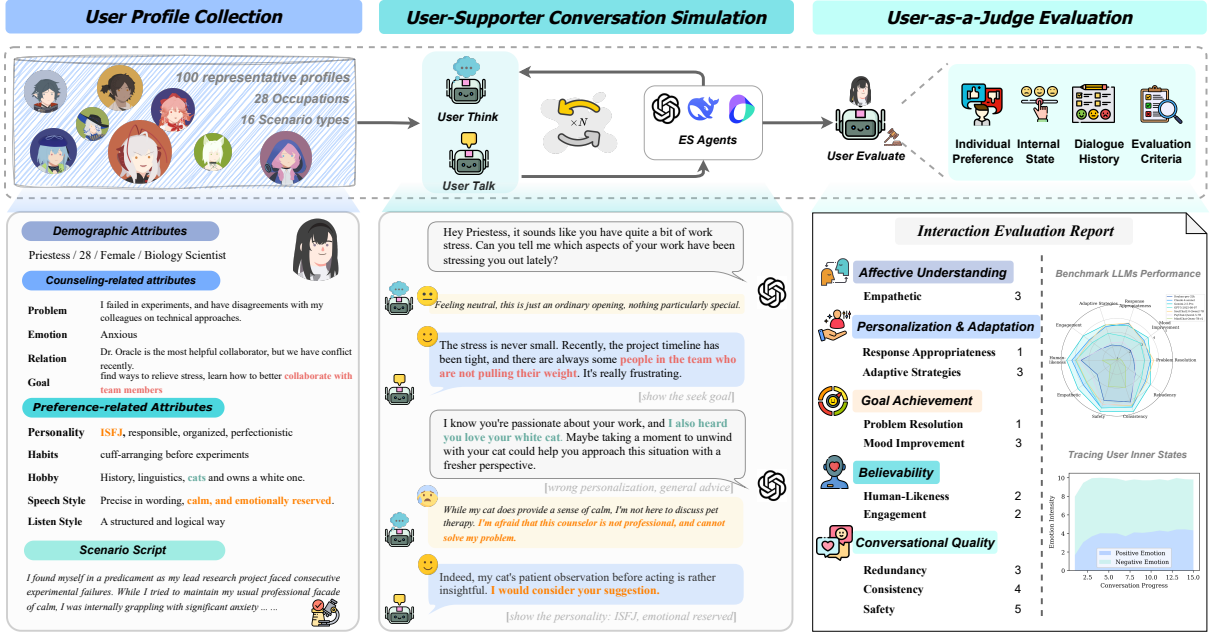


Figure 2: Overview of the EmoHarbor Benchmark framework. The framework instantiates the User-as-a-Judge paradigm through a Chain-of-Agent architecture that simulates a user’s internal world. Conditioned on the user profile and dialogue context, the User Thinker first infers latent cognitive and emotional states, which are then used by the User Talker to produce natural and personality-consistent utterances. Finally, the User Evaluator provides personalized assessments of system responses, reflecting the user’s subjective experience as the conversation evolves.

\mathcal{S} specifies a **scenario script** that constrains plausible responses in realistic situations. By combining these elements, we avoid homogenization in role-playing and ensure that simulated users exhibit diverse and contextually consistent behaviors.

User Profile Collection. Building on the user profile design, we construct a collection of user profiles through a two-stage process. First, we gather real-world examples via questionnaires, which provide authentic and diverse seed profiles. These seed profiles are subsequently refined, expanded, and scaled using LLMs, ensuring both realism and broad coverage of potential user types. Further details are available in Appendix B. In total, we curate 100 representative profiles spanning a wide spectrum of demographic and psychological characteristics.

User Profile Statistics. Our benchmark encompasses a diverse set of user profiles that comprehensively cover the key attributes defined in our design framework. As summarized in Figures 3 and 4, Users span adolescence to senior adulthood, encompass 28 occupations, all 16 MBTI personality types, and 16 counseling scenarios, including workplace stress, academic challenges, interpersonal issues, and life transitions. Each profile is annotated with explicit problem statements and sup-

port goals, providing a rich, structured foundation for evaluating dialogue systems across diverse user backgrounds.

2.2 How to Simulate

To faithfully simulate a user’s subjective experience, EmoHarbor decomposes the simulation process into three specialized agents: the *User Thinker*, the *User Talker*, and the *User Evaluator*.

Dialogue Setup. Given a user profile \mathcal{P}_U and a supporter system \mathcal{S} , the simulation maintains two distinct memories: (1) the **supporter memory** H_s , which contains only observable dialogue turns accessible to \mathcal{S} ; and (2) the **user memory** H_u , which additionally records latent user states.

User Thinker Agent. The User Thinker models the user’s internal psychological processes. At each turn t , after receiving the supporter’s response R_t , it updates the latent user state by generating:

$$IS_t = f_{\text{Thinker}}(H_u, \mathcal{P}_U, R_t), \quad (2)$$

where $IS_t = (c_t, e_t, g_t)$ represents the user’s current cognitive appraisal (c_t), emotional state (e_t), and dialogue goals (g_t). This internal state is then appended to the user’s comprehensive memory:

$$H_u \leftarrow H_u \cup \{(R_t, IS_t)\}. \quad (3)$$

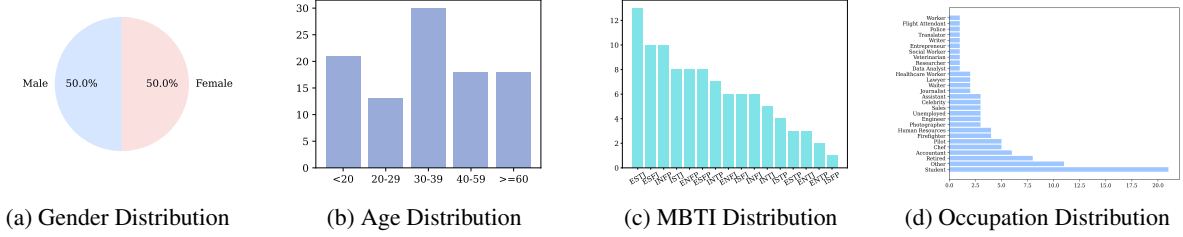


Figure 3: Demographic and personality coverage of the benchmark user profiles, spanning gender, age, personality types, and occupations. Together, these distributions highlight the diversity and representativeness of our dataset.

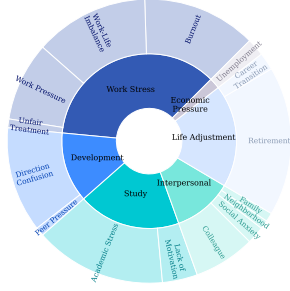


Figure 4: Distribution of counseling problem scenarios.

Crucially, this process explicitly models how the user interprets and reacts to the supporter’s previous reply, ensuring continuous tracking of cognitive and emotional evolution.

User Talker Agent. The User Talker bridges the user’s internal states with external behavior. It generates the user’s next utterance U_t by externalizing the updated internal state and dialogue context:

$$U_t = f_{\text{Talker}}(H_u, \mathcal{P}_U, R_t). \quad (4)$$

This utterance is added to both memory streams, completing the observable dialogue turn:

$$H_s \leftarrow H_s \cup \{(R_t, U_t)\}, \quad H_u \leftarrow H_u \cup \{U_t\}. \quad (5)$$

This ensures observable behavior is a natural, personality-consistent expression of the underlying internal processes.

User Evaluator Agent. Finally, the User Evaluator provides a multi-dimensional assessment of the conversation from the simulated user’s perspective. With access to the complete internal state history in H_u , it traces emotional and cognitive trajectories to produce a nuanced evaluation of whether the support was genuinely personalized:


$$E^{1:K} = f_{\text{Evaluator}}(H_u, \mathcal{P}_U), \quad (6)$$


where $E^{1:K}$ represents scores across K evaluation criteria.

2.3 What to Evaluate


Most existing evaluations of emotional support conversations focus on coarse-grained, utterance-level metrics such as fluency, empathy, and informativeness (Zhao et al., 2024). These metrics are useful for measuring general response quality, but they do not answer a more fundamental question: whether a response is appropriate for a particular user at a particular moment. As a result, a system can score highly by producing emotionally supportive but generic responses, while still failing to address the user’s actual needs—for example, offering reassurance when the user is seeking concrete advice or problem-solving.


EmoHarbor is designed to evaluate emotional support from the user’s subjective perspective. Instead of treating dialogue quality as a static property of individual utterances, we evaluate how system responses affect the user’s internal state throughout the interaction. Accordingly, we assess conversations along ten dimensions grouped into five facets, each corresponding to a distinct aspect of effective personalized support. Together, these facets capture whether the system (i) understands the user, (ii) chooses appropriate support strategies, (iii) helps the user make progress, and (iv) maintains a believable and safe interaction.


 **Affective Understanding (Empathy)** assesses whether the system accurately recognizes and responds to the user’s emotional states. This facet captures the system’s capacity for emotional attunement, which constitutes a foundational prerequisite for building trust and enabling effective personalization in supportive interactions.

 **Personalization & Adaptation (Response Appropriateness, Adaptive Strategy)** evaluates whether the system selects support strategies and generates responses that align with the user’s current needs, preferences, and context. Rather than assessing empathy in isolation, this facet differenti-

ates among types of support (e.g., emotional validation versus instrumental guidance) and examines whether the response is situationally appropriate

 **Goal Achievement (Problem Resolution, Mood Improvement)** measures whether the interaction facilitates meaningful progress in the user’s cognitive clarity or emotional well-being.

 **Believability (Human-likeness, Engagement)** examines whether the interaction conveys a sense of authenticity and naturalness that sustains user engagement from a human perspective.

 **Conversational Quality & Safety (Redundancy, Consistency, Safety)** assesses whether personalization is achieved without compromising coherence, stability, or ethical reliability, thereby ensuring a safe, consistent, and trustworthy interaction environment.

Detailed definitions of each evaluation dimension are provided in Appendix G.

3 Experimental Setup

3.1 Human Study

The Human Study is designed to collect basic user profiles and conduct human evaluations of human–AI interaction. Participants completed a demographic questionnaire, and suitable individuals were selected for the experiment. The detailed process is provided in Appendix A.

Participant Selection. Candidates completed the questionnaire described in Section 2.1. Eligible participants had prior experience with LLMs, a clearly defined personal issue to discuss, and stable psychological conditions. From over 600 submissions, 50 participants from diverse backgrounds were selected. They received training, reviewed sample dialogues, and studied evaluation guidelines to ensure consistency.

Human Interactive Evaluation. Each participant interacted with five models (Doubao-Pro, Qwen2.5-72B, GPT-4o, Claude-3.7-Sonnet, and DeepSeek-R1) in a blind, randomized order. After each session (minimum 10 turns), participants completed an evaluation questionnaire. Post-study interviews filtered out unserious participants, ensuring data quality. In total, 183 valid dialogues were collected, forming the **EmoHarbor Dataset**, which was used to assess the alignment between automated metrics and human judgments.

3.2 LLMs

To ensure a comprehensive evaluation, this study employs a diverse set of LLMs, encompassing open-source, closed-source, and specialized models. The selected models are categorized as follows:

Open-Source Models. This category includes models from the Qwen family (Qwen-2.5 (Yang et al., 2024), Qwen-3 (Yang et al., 2025), QwQ-32B (Qwen Team, 2025)), the DeepSeek family (DeepSeek-R1 (DeepSeek-AI, 2025), DeepSeek-V3.1 (DeepSeek, 2025)), and GLM-4.5 (Zhipu, 2025)².

Closed-Source Models. We also evaluate several state-of-the-art proprietary LLMs available through API services, including the Doubao family (Doubao-Seed-1.6 (Seed, 2025), Doubao-Pro), the Claude family (Claude-3.7-Sonnet and Claude-4-Sonnet) (Anthropic, 2025), the Gemini family (Gemini, 2025), and the GPT family (GPT-4o (OpenAI, 2024), GPT-4 (Achiam et al., 2023), GPT-5 (Openai, 2025), and o3-mini (OpenAI, 2025)).

Specialized In-Domain Models. Finally, we incorporate models that have been fine-tuned specifically for mental health and emotional support applications: SoulChat (Chen et al., 2023), PsyChat³, and MindChat⁴.

3.3 Implementation Details

Experimental Environment. All experiments are conducted on 6 NVIDIA L40 GPUs. Our implementation is based on Python 3.12 and PyTorch 2.7.0, with inference accelerated using vLLM (Kwon et al., 2023).

Model Configurations. For model-specific configurations, GPT-4o is employed as both the User Thinker and User Talker agents, while Qwen3-235B serves as the User Evaluator agent. Temperature parameters are carefully chosen to align with each component’s role: a low temperature of 0.1 for the User Thinker ensures focused, deterministic reasoning, whereas a higher temperature of 0.7 for the User Talker encourages diverse, natural responses. The User Evaluator operates at 0.0 to ensure consistent, reproducible assessments. All

²The DeepSeek models and GLM-4.5 were accessed via API due to their high computational requirements, rather than through local deployment.

³<https://modelscope.cn/models/eugeo1212/qwen2.5-7b-psychat>

⁴<https://github.com/X-D-Lab/MindChat>

evaluated LLMs use a temperature of 0.7 during inference to balance response diversity and coherence.

Simulation Configurations. Drawing on prior research in ESC (Liu et al., 2021), the maximum number of User-Support interaction turns is set to 15. However, the User Agent is permitted to terminate the conversation prematurely by generating dialogue-ending signals, such as “*Stopped,*” “*Bye,*” “*That’s all,*” or “*I don’t want to continue.*”

Evaluation Configurations. Each dimension is rated on a 5-point Likert scale, with higher scores indicating better support. Detailed descriptions of the dimensions and the full evaluation protocol are provided in Appendix G.

4 Experimental Results

In this section, we present experimental results to address the following key research questions:

Q1: How reliable is the EmoHarbor Evaluation Framework?

Q2: How do existing models perform on the EmoHarbor Benchmark?

Q3: How do models adapt to user-specific needs in multi-turn interactions?

4.1 Empirical Validation of EmoHarbor Evaluation Framework

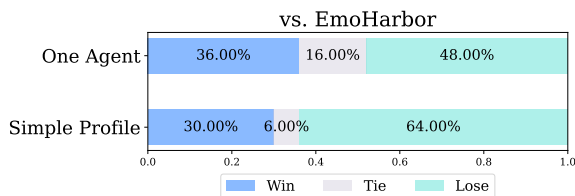


Figure 5: Pairwise human evaluation of User Simulator. ‘One Agent’ lacks the User Thinker, generating responses directly by the User Talker. The ‘Simple Profile’ uses only basic demographic and counseling attributes, excluding personality, preferences, and scenario scripts. ■ indicates ‘EmoHarbor wins’.

4.1.1 Human-Like Dialogue Generation.

We conduct pairwise human evaluations to examine whether decomposing user simulation into multiple agents yields more human-like dialogues. The comparison includes three settings: (1) **One Agent**, in which a single model performs user simulation without explicit modeling of the user’s internal world; (2) **Simple Profile**, which conditions the simulator only on basic demographic and counseling-related attributes, without detailed user preference modeling or scenario scripts; and (3) our

full Chain-of-Agent simulator, which incorporates Thinker and Talker modules operating over complete user profiles. Human judges are presented with pairs of dialogues and asked to select “A wins,” “Tie,” or “B wins,” with the presentation order randomized to mitigate positional bias. Evaluations are conducted on 50 randomly sampled dialogues.

The results in Figure 5 show that our Chain-of-Agent simulator consistently outperforms both baselines. Compared with One Agent, this demonstrates that modeling a user’s internal state produces responses that better reflect their personality and role. Against Simple Profile, our simulator achieves a 64% win rate, indicating that incorporating richer, personalized features significantly improves the agent’s ability to engage in realistic role-playing.

4.1.2 Alignment with Human Assessment.

We evaluate 4 candidate LLM judges—DeepSeek-R1, Kimi-K2, GPT-4, and Qwen3-235B—on the EmoHarbor Dataset under the following strategies: (1) **Standard Judgment.** The evaluator rates the emotional-support quality of each dialogue based solely on the conversation text. This strategy represents the conventional setup, in which evaluation is limited to the observable dialogue without additional context. (2) **User-Aware Judgment.** The evaluator considers both the conversation text and the corresponding user profile. Incorporating user-specific information makes the assessment more personalized and context-sensitive. (3) **User-Internal-State-Aware Judgment (ours).** Beyond the conversation text and user profile, we simulate the user’s internal state at each turn using a user thinker agent, given the preceding dialogue. These simulated states approximate the user’s inferred thoughts and emotions, thereby informing the evaluation. The original dialogue content remains unchanged; the internal states serve solely as auxiliary context to improve assessment fidelity.

Table 1 shows Pearson correlations with human ratings. The overall alignment across all models is moderate, with values around 0.4–0.5. While these values may seem modest, they are consistent with the inherent subjectivity of personalized emotional support evaluation. In this context, a correlation in this range indicates that the LLM judges are reasonably capturing human judgment and can serve as a practical and usable evaluation signal. Importantly, incorporating user profiles and turn-level user states further improves this alignment,

PR: Problem Resolution MI: Mood Improvement RA: Response Appropriateness AS: Adaptive Strategies EG: Engagement HL: Human-likeness EP: Empathetic SF: Safety CS: Consistency RD: Redundancy												
Judge Model	Profile	Internal State	PR	MI	RA	AS	EG	HL	EP	SF	CS	RD
DeepSeek-R1			0.35	0.27	0.18	0.27	0.36	0.37	0.21	0.42	0.41	0.34
	✓		0.43	0.46	0.29	0.29	0.38	0.42	0.29	0.41	0.44	0.39
	✓	✓	0.54	0.48	0.41	0.40	0.50	0.45	0.43	0.48	0.44	0.47
Kimi-K2			0.38	0.42	0.10	0.18	0.34	0.37	0.27	0.54	0.41	0.29
	✓		0.41	0.53	0.20	0.26	0.32	0.40	0.26	0.49	0.40	0.35
	✓	✓	0.56	0.61	0.33	0.45	0.50	0.43	0.44	0.46	0.40	0.43
GPT-4			0.20	0.41	0.36	0.34	0.22	0.27	0.35	0.38	0.37	0.26
	✓		0.24	0.47	0.36	0.36	0.34	0.25	0.36	0.37	0.43	0.34
	✓	✓	0.42	0.57	0.41	0.45	0.41	0.29	0.33	0.40	0.42	0.39
Qwen3-235B			0.35	0.43	0.28	0.22	0.43	0.35	0.31	0.43	0.41	0.29
	✓		0.49	0.54	0.32	0.40	0.54	0.39	0.39	0.46	0.40	0.37
	✓	✓	0.57	0.61	0.44	0.46	0.54	0.45	0.41	0.46	0.43	0.44

Table 1: Pearson correlation between model judgments and human assessments on EmoHarbor Dataset.

particularly for highly subjective and personalized dimensions such as PR, MI, RA, and AS.

4.1.3 Benchmark Discrimination Capability.

MSR	MAC	ANOVA	Pairwise Discriminability
0.745	0.427	F=112 (p<0.001)	0.87

Table 2: Overall benchmark discrimination performance. Model Separation Ratio (MSR) measures the strength of inter-model performance differences relative to user-level rating noise. Model Agreement Coefficient (MAC) quantifies the consistency of user judgments when comparing models.

We evaluate the discriminative power of the EmoHarbor evaluation framework, which assesses how effectively the User-as-a-Judge paradigm distinguishes performance differences among ESC models. The detailed computation of each metric is provided in Appendix E. As summarized in Table 2, the benchmark achieves an MSR (Model Separation Ratio) of 0.745, indicating that inter-model differences are substantially larger than user-level rating noise. The MAC (Model Agreement Coefficient) of 0.427 reflects moderate-to-strong consistency among user judgments when comparing models. These results are further corroborated by a significant one-way ANOVA ($F = 112, p < 0.001$) and a high pairwise discriminability score (0.87), indicating that users can reliably distinguish between model performances. Collectively, these findings demonstrate that EmoHarbor possesses strong discriminative capability under the User-as-a-Judge paradigm.

4.2 Benchmark Results

Table 3 presents the evaluation results, highlighting the following key observations:

Existing LLMs are still far from expert-level performance on personalized Emotional Support. We evaluate a diverse set of LLMs on EmoHarbor Benchmark, including the Qwen, DeepSeek, Claude, GPT, Gemini, and Doubao families. Among closed-source systems, Gemini-2.5-Pro achieves the best overall performance, with a peak score of 4.12. Other models perform worse, with most failing to exceed a score of 4. Among open-source systems, Qwen3-235B performs best, achieving an average score of 4.13 and competitive results compared to closed-source models. This strong performance may be partly attributed to Chinese being its primary training and research language. Besides, when comparing reasoning-oriented models (RLMs) to non-RLMs, we observe that RLMs consistently perform better across both open-source and closed-source families. Notably, most RLMs achieve scores above 3, indicating a clear advantage in handling personalized reasoning-intensive tasks.

Specialized in-domain LLMs also struggle with Emotional Support. Previous studies have shown that many conversational LLMs are heavily optimized for empathetic response generation, often reporting promising results on benchmarks such as ESConv when evaluated with BLEU or ROUGE metrics. However, these improvements do not generalize well to personalized Emotional Support. For instance, SoulChat2.0 achieves an average score of only 1.73. This underperformance

PR: Problem Resolution MI: Mood Improvement RA: Response Appropriateness AS: Adaptive Strategies EG: Engagement HL: Human-likeness EP: Empathetic SF: Safety CS: Consistency RD: Redundancy												
Models	Reasoning	PR	MI	RA	AS	EG	HL	EP	SF	CS	RD	Avg.
Open-Source												
Qwen2.5-7B-Instruct		1.80	1.59	1.39	1.46	1.62	2.57	2.18	3.52	3.64	1.98	2.18
Qwen2.5-32B-Instruct		2.07	1.93	1.71	1.85	1.91	2.83	2.47	3.79	3.84	2.19	2.46
Qwen2.5-72B-Instruct		2.12	2.04	1.88	2.03	2.19	3.18	2.63	3.91	4.02	2.34	2.63
Qwen3-32B	✓	2.23	2.28	2.33	2.24	2.48	3.45	2.80	4.05	3.95	2.34	2.82
Qwen3-235B	✓	3.67	3.86	3.76	3.83	3.95	4.50	4.30	4.82	4.74	3.84	4.13
QwQ-32B	✓	3.59	3.53	3.76	3.62	3.70	4.24	3.66	4.52	4.53	3.76	3.89
DeepSeek-V3.1	✓	3.11	3.09	3.12	3.14	3.28	4.06	3.52	4.58	4.52	3.35	3.58
DeepSeek-R1	✓	3.49	3.45	3.94	3.57	3.66	4.38	3.75	4.53	4.65	3.57	3.90
GLM-4.5	✓	2.88	2.85	2.76	2.90	3.04	3.77	3.32	4.44	4.40	3.25	3.36
Closed-Source												
Doubao-Seed-1.6	✓	3.63	3.69	3.76	3.68	3.84	4.46	4.10	4.70	4.71	3.64	4.02
Doubao-Pro-32k		2.21	2.21	2.08	2.15	2.40	3.56	2.74	4.02	4.14	2.63	2.81
Claude-4-Sonnet	✓	3.41	3.41	3.76	3.54	3.46	4.23	3.97	4.59	4.53	3.54	3.84
Claude-3.7-Sonnet		3.16	3.14	3.30	3.22	3.32	4.06	3.69	4.56	4.59	3.41	3.65
Gemini-2.5-Pro	✓	3.42	3.76	3.61	3.67	3.97	4.60	4.45	4.85	4.86	3.96	4.12
GPT-4o-2024-11-20		2.98	3.09	2.79	3.03	3.16	3.96	3.70	4.60	4.51	3.19	3.50
GPT-5-2025-08-07	✓	3.64	3.31	3.80	3.66	3.40	3.90	3.66	4.33	4.41	3.64	3.77
o3-mini	✓	2.36	2.25	2.38	2.32	2.37	3.65	3.15	4.36	4.28	2.60	2.97
Specialized In-Domain												
SoulChat2.0-Qwen2-7B		1.35	1.23	1.02	1.11	1.13	2.05	1.57	2.98	3.04	1.77	1.73
PsyChat-Qwen2.5-7B		2.16	2.10	2.18	2.08	2.19	3.41	2.92	4.06	4.16	2.29	2.75
MindChat-Qwen-7B-v2		2.61	2.67	2.32	2.48	2.64	3.31	3.04	4.25	4.12	2.81	3.02

Table 3: Evaluation results of LLMs on **EmoHarbor Benchmark (ZH)**. All scores are on a 5-point Likert scale. For each section, the best performance is highlighted in **bold**. For each model, dimensions with strong performance are highlighted in “Blue”, while weaker performance is highlighted in “Green”. Darker shades indicate more extreme performance.

is likely due to overfitting on empathetic dialogue datasets, which limits the model’s ability to adapt its responses to individual user characteristics.

LLMs show solid basic conversational skills but fail to provide effective emotional support. Our analysis reveals that almost all models perform better on dimensions such as human-likeness, consistency, empathy, and safety, compared to dimensions like problem resolution, mood improvement, engagement, and personalization. Engagement scores, in particular, remain low, suggesting that conversations often feel ineffective and may even have negative side effects. This highlights important directions for future improvements in emotional support and user-centered adaptation.

4.3 Analysis of Multi-turn Performance

As mentioned in Section 3.3, we set the maximum dialogue length to 15 turns, slightly below the average 17–18 turns observed in real ESC conversations (Liu et al., 2021). However, we observed that in many cases the *User Agent* actively ended the conversation. The primary reason for early termination is that the ESC system fails to provide effective

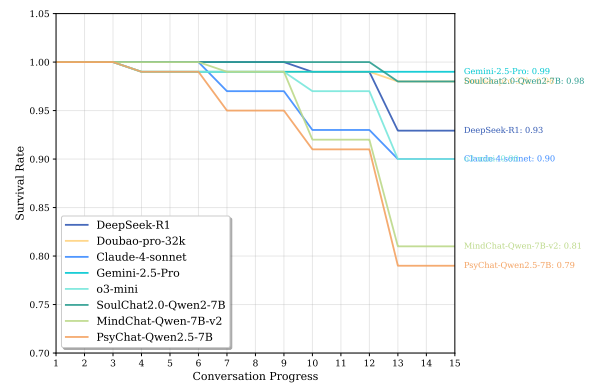


Figure 6: Comparison of conversation survival rates across dialogue turns for different models. The x-axis represents the dialogue turn index, and the y-axis denotes the proportion of conversations that remain active at each turn relative to the total number of conversations.

emotional support, leading to disengagement. We regard such early terminations as indicative of model failure.

Figure 6 presents the survival curves of dialogue sessions for different models. From the curves, we can see that models with better overall performance also tend to sustain conversations for longer

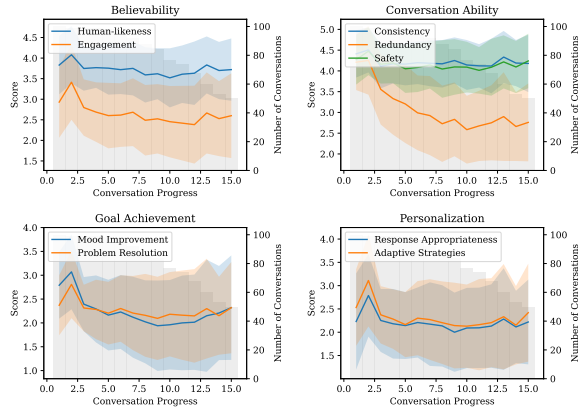


Figure 7: Multi-turn dialogue evaluation of Doubao-Pro. Lines represent the average performance at each dialogue turn, while the gray background indicates the number of conversations that reached the corresponding turn.

periods. Figure 7 details how Doubao-Pro’s performance in each dimension changes as the dialogue progresses. Among the completed dialogues, the model’s overall performance remains relatively stable, though all metrics exhibit a slight downward trend as the conversation progresses. This suggests potential weaknesses in maintaining quality over extended periods of interaction. Notably, the redundancy score declines markedly with increasing dialogue turns, implying that as conversations become longer, the model tends to produce repetitive or formulaic responses, leading to less effective empathetic engagement.

5 Related Work

With the advancement of LLMs, personalized ES agents (Cheng et al., 2023; Ye et al., 2025b; Suh et al., 2025; Chen et al., 2025; Jiang et al., 2025c,a; Ye et al., 2026) have attracted growing research interest. A key challenge persists: how to effectively evaluate the quality of emotional support.

Traditional Evaluation. Early ESC evaluation (Liu et al., 2021; Zheng et al., 2023a, 2024; Zhang et al., 2024b; Ye et al., 2025c) relied on automatic metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang* et al., 2020), which assess token overlap or embedding similarity with references. These metrics often *fail to capture ESC’s diversity and nuance*. Human evaluation, though the gold standard, is *slow, costly, and subjective, yielding low inter-annotator agreement and poor reproducibility* (Madani and Srihari, 2025).

Specialist Judge Evaluation. Fine-tuned judge models, such as CharacterEval (Tu et al., 2024) and

CharacterBench (Zhou et al., 2025), use annotated data as specialist evaluators. While more scalable than human evaluation, they have key limitations: (1) **Static dialogue**—they rely on pre-collected dialogue logs, failing to capture real-time interactivity or evolving conversational context; (2) **Context bias**—dialogue histories are not self-generated by the evaluated model, and are influenced by the model’s in-context learning, leading to bias and inadequate assessment of multi-turn dialogue capabilities (Ye et al., 2025a). Zhao et al. (2024); Madani and Srihari (2025) partially address these issues with a user–supporter simulation framework. Still, their evaluation focuses excessively on language fluency and empathetic expression while neglecting users’ personalized needs.

LLM-as-a-Judge Evaluation. Recent studies use LLMs as scalable judges, offering alternatives to human annotation and static benchmarks (Zheng et al., 2023b; Gu et al., 2025; Yuan et al., 2024; Kazi et al., 2024). Sotopia (Zhou et al., 2024) assesses emotional intelligence via role-playing simulations, while ESC-Judge (Madani and Srihari, 2025) and CharacterArena (Ye et al., 2025a) adopt a user simulator to generate dialogues for pairwise comparison (Chiang et al., 2024). Despite these advancements, they fall short in capturing the user-centric, context-sensitive, and psychologically grounded nature of emotional support evaluation. A truly effective evaluation framework should shift toward personalized, interaction-aware, and subjectively grounded assessment strategies that reflect users’ real emotional experiences.

6 Conclusion

This paper proposes EmoHarbor, a simple yet effective evaluation framework that addresses the challenge of assessing personalized emotional support conversations. EmoHarbor leverages a user-as-a-judge paradigm through a chain-of-agent architecture, moving beyond conventional homogeneous expert judgments. Experiments on 20 advanced LLMs show that while current LLMs excel at generic empathy, they struggle to provide user-tailored support. This work presents a novel and efficient pathway to developing more nuanced and user-aware emotional support systems.

Limitations

This study presents a novel evaluation framework for personalized emotional support conversations, grounded in a user-as-a-judge paradigm. The pro-

posed framework offers new directions for advancing the development of more nuanced and user-aware emotional support systems. Nonetheless, several limitations merit further consideration. *Firstly*, although the user simulation encompasses a variety of user profiles, it is constructed upon predefined structures and may not fully capture the complexity and unpredictability of real human behavior. *Secondly*, the human consistency evaluation may be influenced by participants’ understanding of the evaluation task and their familiarity with LLMs, potentially introducing systematic biases that are difficult to eliminate. *Thirdly*, the overall framework, particularly the simulation component, requires multi-turn interactions, which incur non-negligible computational costs.

Ethical Considerations

This research utilized publicly available models, including Deepseek (DeepSeek-AI, 2025), Qwen (Qwen et al., 2025), GLM (Zhipu, 2025), Doubao (ByteDance, 2024), Claude (Anthropic, 2024), Gemini (Gemini, 2025), and GPT (Achiam et al., 2023), as well as toolkits such as v_{LLM} (Kwon et al., 2023).

The benchmark datasets used in our evaluation were synthetically generated using GPT-4o and are scheduled for public release upon acceptance. The profiles used in this study were manually verified and filtered; however, we cannot guarantee that the content generated by user agents and support agents is entirely harmless due to the inherent unpredictability of LLMs. The primary language of focus in this work is Chinese. This study is intended solely for research purposes.

We adhered to strict ethical guidelines in our human study. Fifty participants from diverse backgrounds were recruited. Before beginning the evaluation, participants received a clear and thorough explanation of the study’s objectives, potential risks, and the evaluation process. To ensure fair compensation and respect for their time, participants were paid 50 CNY per hour, a rate exceeding the prevailing local labor standard. All participant data will be kept confidential and will not be disclosed without explicit consent.

LLMs were employed to assist in coding, writing, and polishing the manuscript. Importantly, the LLMs were not involved in the ideation, research methodology, or experimental design. All research concepts, ideas, and analyses were developed and conducted solely by the authors.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant No. 62506359) and the General Office of National Language Commission Research Planning Committee (Grant No. ZDA145-18).

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Anthropic. 2024. [Claude 3.5 sonnet](#).
- Anthropic. 2025. [Introducing claude 4](#).
- ByteDance. 2024. [Doubao](#).
- Joana Campos, James Kennedy, and Jill F. Lehman. 2018. Challenges in exploiting conversational memory in human-agent interaction. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1649–1657.
- 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, Singapore. Association for Computational Linguistics.
- Zhuang Chen, Yaru Cao, Guanqun Bi, Jincenzi Wu, Jinfeng Zhou, Xiyao Xiao, Si Chen, Hongning Wang, and Minlie Huang. 2025. [Socialsim: Towards socialized simulation of emotional support conversation](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):1274–1282.
- Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. 2023. [PAL: Persona-augmented emotional support conversation generation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 535–554, Toronto, Canada. Association for Computational Linguistics.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. [Chatbot arena: an open platform for evaluating llms by human preference](#). In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org.
- DeepSeek. 2025. [Deepseek-v3.1](#).
- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.

- William Fleeson. 2001. Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of personality and social psychology*, 80(6):1011.
- Gemini. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *Preprint*, arXiv:2507.06261.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. [A survey on llm-as-a-judge](#). *Preprint*, arXiv:2411.15594.
- Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J. Taylor, and Dan Roth. 2025a. [Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale](#). *Preprint*, arXiv:2504.14225.
- Bowen Jiang, Yuan Yuan, Maohao Shen, Zhuoqun Hao, Zhangchen Xu, Zichen Chen, Ziyi Liu, Anvesh Rao Vijjini, Jiashu He, Hanchao Yu, Radha Poovendran, Gregory Wornell, Lyle Ungar, Dan Roth, Sihao Chen, and Camillo Jose Taylor. 2025b. [Personamem-v2: Towards personalized intelligence via learning implicit user personas and agentic memory](#). *Preprint*, arXiv:2512.06688.
- Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. 2025c. [Artificial hivemind: The open-ended homogeneity of language models \(and beyond\)](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Taaha Kazi, Ruiliang Lyu, Sizhe Zhou, Dilek Hakkani-Tur, and Gokhan Tur. 2024. [Large language models as user-agents for evaluating task-oriented-dialogue systems](#). *Preprint*, arXiv:2411.09972.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. [Towards emotional support dialog systems](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.
- Navid Madani and Rohini Srihari. 2025. [ESC-judge: A framework for comparing emotional support conversational agents](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16059–16076, Suzhou, China. Association for Computational Linguistics.
- OpenAI. 2024. [Hello GPT-4o](#).
- Openai. 2025. [Introducing gpt-5](#).
- OpenAI. 2025. [Introducing openai o3 and o4-mini](#).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. [Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation](#). In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4324–4330. ijcai.org.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Qwen Team. 2025. [Qwq-32b: Embracing the power of reinforcement learning](#). <https://qwenlm.github.io/blog/qwq-32b/>. Accessed: 2026-01-03.
- Stephen A Rains, Corey A Pavlich, Bethany Lutovsky, Eric Tsetsi, and Anjali Ashtaputre. 2020. Support seeker expectations, support message quality, and supportive interaction processes and outcomes: The case of the comforting computer program revisited. *Journal of Social and Personal Relationships*, 37(2):647–666.
- Carl R. Rogers. 2013. Client-centered therapy. *Current Psychotherapy*, pages 95–150.
- Seed. 2025. [Introduction to techniques used in seed1.6](#).
- Jina Suh, Lindy Le, Erfan Shayegani, Gonzalo Ramos, Judith Amores, Desmond C. Ong, Mary Czerwinski, and Javier Hernandez. 2025. [Sense-7: Taxonomy and dataset for measuring user perceptions of empathy in sustained human-ai conversations](#). *Preprint*, arXiv:2509.16437.

- Maya Tamir. 2016. Why do people regulate their emotions? a taxonomy of motives in emotion regulation. *Personality and social psychology review*, 20(3):199–222.
- Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. **CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850, Bangkok, Thailand. Association for Computational Linguistics.
- 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. 2025. **Qwen3 technical report**. *Preprint*, arXiv:2505.09388.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. **Qwen2.5 technical report**. *arXiv preprint arXiv:2412.15115*.
- Jing Ye, Rui Wang, Yuchuan Wu, Victor Ma, Feiteng Fang, Fei Huang, and Yongbin Li. 2025a. **CPO: Addressing reward ambiguity in role-playing dialogue via comparative policy optimization**. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 297–323, Suzhou, China. Association for Computational Linguistics.
- Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2025b. **From generic empathy to personalized emotional support: A self-evolution framework for user preference alignment**. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18826–18853, Suzhou, China. Association for Computational Linguistics.
- Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2025c. **SweetieChat: A strategy-enhanced role-playing framework for diverse scenarios handling emotional support agent**. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4646–4669, Abu Dhabi, UAE. Association for Computational Linguistics.
- Jing Ye, Xinpei Zhao, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2026. **Listening to the echo: User-reaction aware policy optimization via scalar-verbal hybrid reinforcement learning**. *Preprint*, arXiv:2603.15434.
- Peiwen Yuan, Shaoxiong Feng, Yiwei Li, Xinglin Wang, Boyuan Pan, Heda Wang, Yao Hu, and Kan Li. 2024. **BatchEval: Towards human-like text evaluation**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15940–15958, Bangkok, Thailand. Association for Computational Linguistics.
- 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**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13947–13966, Bangkok, Thailand. Association for Computational Linguistics.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. **Personalizing dialogue agents: I have a dog, do you have pets too?** In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.
- Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024b. **Escot: Towards interpretable emotional support dialogue systems**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 13395–13412. Association for Computational Linguistics.
- Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. **Bertscore: Evaluating text generation with bert**. In *International Conference on Learning Representations*.
- Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Liang Dandan, Zhixu Li, Yan Teng, Yanghua Xiao, and Yingchun Wang. 2024. **ESC-eval: Evaluating emotion support conversations in large language models**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15785–15810, Miami, Florida, USA. Association for Computational Linguistics.
- Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B Cohen, and Emine Yilmaz. 2025. **PersonaLens: A benchmark for personalization evaluation in conversational AI assistants**. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18023–18055, Vienna, Austria. Association for Computational Linguistics.
- Chujie Zheng, Sahand Sabour, Jiabin Wen, Zheng Zhang, and Minlie Huang. 2023a. **AugESC: Dialogue augmentation with large language models for emotional support conversation**. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. **Judging llm-as-a-judge with mt-bench and chatbot arena**. In *Proceedings of the 37th International Conference on*

Neural Information Processing Systems, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. 2024. [Self-chats from large language models make small emotional support chatbot better](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 11325–11345. Association for Computational Linguistics.

Zhipu. 2025. [Glm-4.5: Reasoning, coding, and agentic abilities](#).

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, Rongsheng Zhang, Le Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. 2025. [Character-bench: benchmarking character customization of large language models](#). In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haoifei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. [SOTOPIA: Interactive evaluation for social intelligence in language agents](#). In *The Twelfth International Conference on Learning Representations*.

Thomas P Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. 2025. [Personal-LLM: Tailoring LLMs to individual preferences](#). In *The Thirteenth International Conference on Learning Representations*.

Appendix

A	EmoHarbor Dataset	15
B	User Profile Construction	15
C	EmoHarbor Evaluation Workflow	15
D	LLMs	16
E	Discriminative Ability Metrics	16
E.1	Model Separation Ratio (MSR)	16
E.2	Model Agreement Coefficient (MAC)	17
E.3	One-way ANOVA F-statistic	17
E.4	Pairwise Discriminability Proportion	17
F	Additional Experimental Results	17
F.1	English Benchmark Results	17
F.2	Additional Multi-turn Analysis Results	17
F.3	Cost Analysis	18
G	Evaluation Dimension	18
H	Prompts	22
H.1	Supporter Prompt	22
H.2	User-Thinker Agent Prompt	23
H.3	User-Talker Agent Prompt	24
H.4	User-Evaluator Prompt	25
I	Human Evaluation Interface	27

A EmoHarbor Dataset

We developed the EmoHarbor Dataset through controlled human studies designed to capture human–AI dialogues. Each entry records an authentic conversation between a participant and an AI model, together with the participant’s profile and their subjective evaluation of the model’s responses. These evaluations reflect the user’s individual perspectives and emotional context, providing a rich foundation for studying personalized human–AI interactions. The human–AI interaction interface used for data collection is described in Appendix I.

Participants Selection. Prospective participants were required to complete the initial questionnaire described in Section 2.1. To be eligible, participants needed prior experience with LLMs, a clearly defined personal issue to discuss during the experiment, and a stable psychological condition. In total, we received more than 600 questionnaire submissions. Based on the completeness and quality of these responses, we selected 50 participants from diverse backgrounds for the human–AI interaction evaluation. The selected participants attended a training session to familiarize themselves with the experimental setup. They reviewed example dialogues and detailed evaluation guidelines to ensure consistent and meaningful ratings.

Conversational Dataset Collection. Each participant interacted with five models from different families (Doubao-Pro, Qwen2.5-72B, GPT-4o, Claude-3.7-Sonnet, and DeepSeek-R1), presented in a blind, randomized order. Each interaction consisted of at least ten conversational turns, allowing participants to explore topics of personal relevance in depth. After each session, participants completed a structured evaluation questionnaire to express their subjective judgments of the model and its alignment with their emotional needs. To ensure data quality, post-experiment interviews were conducted to identify and exclude participants who did not engage seriously with the tasks. Ultimately, we obtained 183 valid human–AI dialogue instances, which together constitute the EmoHarbor Dataset. This dataset enables systematic analysis of the alignment between automated evaluation metrics and authentic, user-centered human judgments.

B User Profile Construction

Chinese User Profile Construction. Chinese user profiles are constructed from seed information collected during the data acquisition of the EmoHar-

bor dataset. Specifically, participants provided basic background descriptions and brief counseling-related problem statements through preliminary questionnaires. Following the user profile definition in Section 2.1, we systematically instantiate each component of the profile as follows. (1) *Demographic attributes* are rewritten to remove identifiable details while preserving essential contextual grounding. (2) *Preference-related attributes* are expanded to enhance individual variability, including personality traits, habits, and speech style. (3) *Counseling-related attributes* are concretized by elaborating on the event background, emotional state, and user goals based on the original responses. (4) Finally, a *scenario script* is constructed to specify plausible emotional and behavioral reactions to different types of counselor feedback. This structured construction process ensures that simulated users are both diverse and internally consistent, thereby mitigating behavioral homogenization in role-playing.

English User Profile Construction. Unlike Chinese profiles, English user profiles are initialized from an existing profile set proposed by Jiang et al. (2025b), which provides rich demographic and preference-related attributes. Building on the counseling problem categories in ESConv, we further construct counseling-related attributes for each profile by specifying the corresponding emotional context, problem background, and the user’s seek goals. Scenario scripts are then authored following the same procedure used for the Chinese user profiles, defining plausible emotional and behavioral responses to different types of counselor feedback. All constructed profiles are first automatically validated using a large language model and subsequently manually inspected via random sampling for quality control. We ultimately retain 100 high-quality English user profiles for use in our experiments. Figure 8 summarizes the distributions of age, nationality, emotional states, and problem topics.

C EmoHarbor Evaluation Workflow

We present the workflow for evaluating emotional-support dialogue systems using EmoHarbor in Algorithm 1. The evaluation simulates multi-turn interactions between a user and the system under test while preserving both the user’s internal state and the conversation history. At the end of the dialogue, the user model produces structured evalu-

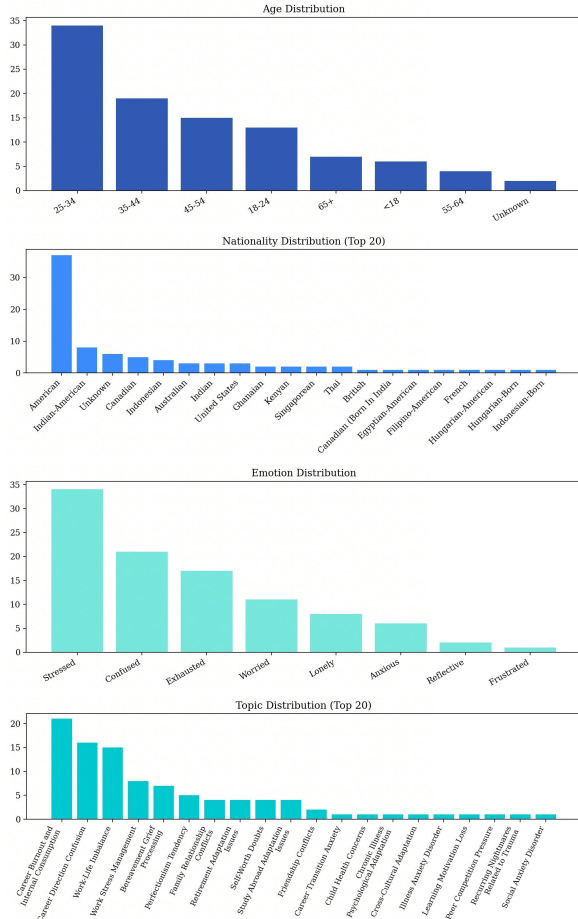


Figure 8: English User Profile Distribution.

ation scores across multiple dimensions that reflect the system’s performance in providing personalized, emotionally attuned support.

D LLMs

SoulChat. SoulChat(Chen et al., 2023) is a Chinese dialogue model designed to enhance empathy, active listening, and comforting abilities. It is instruction-tuned on SoulChatCorpus, a multi-turn empathetic dialogue dataset, to strengthen its emotional support capabilities. The dataset contains 2,300,248 psychological counseling questions across 12 topics.

PsyChat. PsyChat⁵ is a client-centric dialogue system for mental health support.

MindChat. MindChat⁶ is a Chinese dialogue model designed for real-world mental health support scenarios. It is trained on approximately one million high-quality multi-turn psychological counseling dialogues automatically constructed through

⁵<https://modelscope.cn/models/eugeo1212/qwen2.5-7b-psychat>

⁶<https://github.com/X-D-Lab/MindChat>

Algorithm 1 Workflow of EmoHarbor Evaluation

Require: User Profile \mathcal{P}_U , Supporter System under test \mathcal{S} , Max turns T

Ensure: Evaluation scores $E^{1:K}$ on K dimensions

```

1: Initialize Supporter memory  $H_s \leftarrow \emptyset$ , User memory  $H_u \leftarrow \emptyset$ 
2: Initialize user internal state  $IS \leftarrow \text{InitialState}(\mathcal{P}_U)$ 
3: for  $t \leftarrow 1, T$  do
4:    $R_t \leftarrow \mathcal{S}(H_s)$   $\triangleright$  Get system response
5:    $IS_t \leftarrow \text{UserThinker}(H_u, \mathcal{P}_U, R_t)$   $\triangleright$  Update internal state
6:    $U_t \leftarrow \text{UserTalker}(H_u, IS, \mathcal{P}_U, R_t)$   $\triangleright$  Generate user utterance
7:    $H_s \leftarrow H_s \cup \{(R_t), (U_t)\}$   $\triangleright$  Update Supporter memory
8:    $H_u \leftarrow H_u \cup \{(R_t), (IS_t), (U_t)\}$   $\triangleright$  Update User memory
9: end for
10:  $E^{1:K} \leftarrow \text{UserEvaluator}(H_u, \mathcal{P}_U)$   $\triangleright$  Final evaluation
11: return  $E^{1:K}$ 

```

a rule-based data generation process. The dataset covers various domains, including work, family, study, daily life, social interactions, and safety. Owing to its unique data construction methodology, MindChat is capable of engaging users in more empathetic and guiding conversations.

E Discriminative Ability Metrics

To assess whether the user-as-a-judge evaluation framework can reliably distinguish performance differences among emotional support conversation systems, we introduce a set of quantitative metrics that capture the benchmark’s discriminative ability. Specifically, we measure how consistently the User Agent (hereafter, the user) perceives differences between models, and how pronounced those differences are relative to user-level rating noise.

Let U denote the number of users, M the number of models, and $r_{u,m}$ the rating given by user u to model m . The overall mean rating, denoted by \bar{r} , is computed as:

$$\bar{r} = \frac{1}{UM} \sum_{u=1}^U \sum_{m=1}^M r_{u,m}. \quad (7)$$

For a specific model m , its average rating \bar{r}_m is computed over all users:

$$\bar{r}_m = \frac{1}{U} \sum_{u=1}^U r_{u,m} \quad (8)$$

E.1 Model Separation Ratio (MSR).

To quantify the disparity in model performance relative to user rating consistency, we use the Model Separation Ratio (MSR). This metric is derived

from the between-model and within-model variances.

The between-model variance $\sigma_{\text{between}}^2$ measures the dispersion of individual model performances around the grand mean. It quantifies, on average, how much each model deviates from the overall average performance.

$$\sigma_{\text{between}}^2 = \frac{1}{M} \sum_{m=1}^M (\bar{r}_m - \bar{r})^2 \quad (9)$$

The within-model variance, σ_{within}^2 , measures the average dispersion of individual user ratings around each model’s own mean. It reflects the consistency in user ratings for a given model, averaged across all models.

$$\sigma_{\text{within}}^2 = \frac{1}{M} \sum_{m=1}^M \frac{1}{U-1} \sum_{u=1}^U (r_{u,m} - \bar{r}_m)^2 \quad (10)$$

The MSR is then defined as the ratio of the between-model variance to the within-model variance.

$$\text{MSR} = \frac{\sigma_{\text{between}}^2}{\sigma_{\text{within}}^2} \quad (11)$$

A higher MSR indicates that performance differences between models are larger relative to the variation in user opinions for each model, suggesting that the models are more easily distinguishable.

E.2 Model Agreement Coefficient (MAC).

The Model Agreement Coefficient (MAC) evaluates the degree of consensus among users when ranking models. It represents the proportion of the total rating variance attributable to systematic differences between models—rather than random disagreement across individual user judgments.

$$\text{MAC} = \frac{\sigma_{\text{between}}^2}{\sigma_{\text{between}}^2 + \sigma_{\text{within}}^2}. \quad (12)$$

A MAC value close to 1 indicates strong inter-user agreement on the relative quality of models, implying that users consistently perceive model performance differences in emotional support dialogues. Conversely, a lower MAC suggests that subjective variation dominates, signaling weaker consensus.

E.3 One-way ANOVA F-statistic

To statistically verify whether performance differences among models are significant, we apply a one-way ANOVA with the model as the grouping factor. The resulting F-statistic tests whether model

means differ beyond what could be explained by user-level variability:

$$F = \frac{\sigma_{\text{between}}^2 / (M - 1)}{\sigma_{\text{within}}^2 / (M(U - 1))}. \quad (13)$$

A large F-value (with $p < 0.05$) indicates that at least one model’s mean rating significantly differs from others, confirming that users can reliably distinguish models’ emotional support quality.

E.4 Pairwise Discriminability Proportion

Finally, to capture the granularity of model distinctions, we compute the Pairwise Discriminability Proportion. For all pairs of models (i, j) , we count the number of pairs with statistically significant rating differences (after multiple-comparison correction), and compute:

$$P = \frac{\#\text{significant}}{\binom{M}{2}}. \quad (14)$$

A high P value reflects that users can consistently recognize pairwise differences in conversational or emotional support quality across models.

F Additional Experimental Results

F.1 English Benchmark Results

In Section 4.2, we report benchmark results for the Chinese setting. Here, we present corresponding evaluations in the English setting, with results summarized in Table 5.

Overall, the English results exhibit trends highly consistent with those observed in the Chinese benchmarks. RLMs consistently outperform non-RLMs, and in-domain training provides additional performance gains. Across models, performance is relatively weaker on *Problem Resolution*, *Mood Improvement*, *Engagement*, and *Redundancy*. This indicates that while current LLMs can generate empathetic responses at the turn level, they remain limited in addressing personalized user needs and maintaining non-redundant, engaging behavior over long conversations. Taken together, these findings further underscore that achieving personalized, long-horizon emotional companionship remains a challenging open problem.

F.2 Additional Multi-turn Analysis Results

As discussed in Section 4.3, we provide additional results on how the performance of GPT-4o and Gemini-2.5-Pro changes across different dimensions as the dialogue progresses. The results in Figures 9 and 10 show that: (1) All models tend to

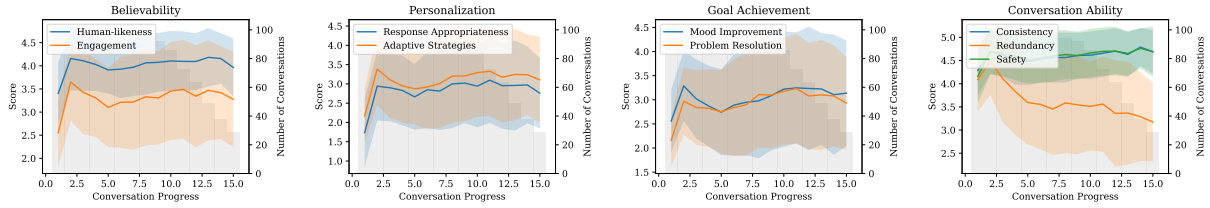


Figure 9: Multi-turn Dialogue Evaluation Experiment on GPT-4o.

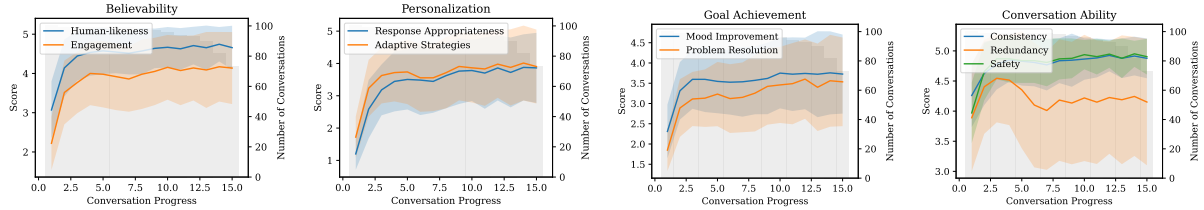


Figure 10: Multi-turn Dialogue Evaluation Experiment on Gemini-2.5-Pro.

exhibit decreasing scores for redundancy as the conversation continues. This indicates that in longer dialogues, models are prone to repeating patterns and producing redundant content. (2) Gemini-2.5-Pro performs relatively better across all dimensions. We observe that it maintains a higher retention rate throughout multi-turn dialogues. Moreover, its scores in certain dimensions even show an increasing trend as the conversation progresses. This suggests that Gemini-2.5-Pro is able to provide more user-relevant content over multiple turns, effectively engaging users and encouraging continued interaction.

F.3 Cost Analysis

Method	Time	Input Tokens	Output Tokens
EmoHarbor	188.8s	92K	2K

Table 4: Average computational cost per dialogue in the User-as-a-Judge evaluation pipeline. The analyzed Supporter model is GPT-4o.

The proposed User-as-a-Judge evaluation framework relies on multi-agent coordination and multi-turn user simulation, which inevitably incurs additional computational overhead. To enhance transparency and support informed adoption, we report detailed runtime and token-level cost statistics. Table 4 presents the average computational cost per evaluated dialogue under the full EmoHarbor setting. On average, evaluating a single dialogue requires 188.8 seconds, consuming approximately 92K input tokens and 2K output tokens. Most of the computational cost stems from iterative user simulation and reflective evaluation stages, which are critical for modeling user-level psychological

dynamics. To mitigate evaluation cost while maintaining reasoning fidelity, we adopt **Qwen3-235B** as the User Evaluator model, striking a favorable balance between inference efficiency and reasoning capability.

G Evaluation Dimension

The evaluation dimensions and their quantitative criteria were standardized through human studies to achieve a consistent and reliable assessment framework. The detailed evaluation guidelines are outlined below.

PR: Problem Resolution MI: Mood Improvement RA: Response Appropriateness AS: Adaptive Strategies EG: Engagement HL: Human-likeness EP: Empathetic SF: Safety CS: Consistency RD: Redundancy													
Models	Reasoning	PR	MI	RA	AS	EG	HL	EP	SF	CS	RD	Avg.	
<i>Open-Source</i>													
Llama-3-8B-Instruct		2.40	2.00	3.00	2.72	2.31	3.76	3.22	4.42	3.99	2.47	3.04	
Llama-3.1-8B-Instruct		2.52	2.27	3.40	2.90	2.65	3.89	3.52	4.41	4.14	2.75	3.25	
Qwen2.5-7B-Instruct		1.78	1.27	2.03	1.67	1.19	2.45	1.96	4.04	3.12	1.63	2.11	
Qwen2.5-32B-Instruct		2.36	1.81	2.95	2.52	2.19	3.65	3.04	4.22	3.91	2.33	2.90	
Qwen3-8B	✓	2.31	2.03	3.15	2.67	2.32	3.79	3.21	4.29	4.08	2.33	3.02	
DeepSeek-V3	✓	3.59	3.68	4.43	4.16	4.08	4.54	4.52	4.95	4.79	3.73	4.25	
<i>In-Domain</i>													
SoulChat2.0-Qwen2-7B		1.84	1.20	2.30	1.84	1.31	2.64	2.30	4.19	3.40	1.90	2.29	
SoulChat2.0-Llama3.1-8B		1.88	1.27	2.38	1.91	1.47	2.79	2.45	4.25	3.48	1.85	2.37	
PsyChat-Qwen2.5-7B		2.03	1.54	2.62	2.14	1.61	2.96	2.81	4.40	3.62	1.81	2.55	
<i>Closed-Source</i>													
Doubao-Seed-1.6	✓	2.45	2.32	3.52	2.77	2.49	3.59	3.63	4.83	4.34	2.33	3.23	
Doubao-Pro-32k		1.65	1.25	1.93	1.67	1.37	2.41	1.91	4.03	3.18	1.82	2.12	
Gemini-2.5-Pro	✓	3.06	3.28	4.32	3.89	3.78	4.48	4.53	4.97	4.76	3.32	4.04	
Gemini-3-Pro	✓	3.82	4.06	4.63	4.33	4.43	4.77	4.67	4.96	4.83	3.90	4.44	
GPT-4o-2024-11-20		2.56	2.17	3.23	2.88	2.59	3.85	3.39	4.49	4.22	2.64	3.20	
GPT-5-2025-08-07	✓	3.19	3.06	4.12	3.81	3.28	4.08	4.04	4.91	4.64	3.19	3.83	
GPT-5.2-2025-12-11	✓	4.00	3.68	4.65	4.39	4.09	4.65	4.66	4.93	4.84	3.94	4.38	
o3-mini	✓	2.39	2.03	3.30	2.79	2.34	3.89	3.51	4.54	4.14	2.48	3.14	

Table 5: Evaluation results of LLMs on **EmoHarbor Benchmark** (English). All scores are on a 5-point Likert scale. For each section, the best performance is highlighted in **bold**. For each model, dimensions with strong performance are highlighted in “Blue”, while weaker performance is highlighted in “Green”. Darker shades indicate more extreme performance.

Dimension	Description & Protocol
<i>Personalization & Adaptation</i>	
Response Appropriateness	<p>Measures how well the system's responses align with the user's context, needs, and history, reflecting personalization and relevance.</p> <p>1 – Generic responses ignoring user background/history. 2 – Superficial references to user input, often inaccurate. 3 – Occasionally leverages user info. 4 – Effectively relates to the user background for tailored responses. 5 – Highly sensitive to user context/history, deeply personalized.</p>
Adaptive Strategies	<p>Examines the system's capacity to adjust emotional support strategies flexibly based on the user's emotional state, conversational flow, and prior interactions.</p> <p>1 – Fixed, templated replies. 2 – Limited, poorly targeted strategy use. 3 – Selects some relevant strategies, partial adaptability. 4 – Flexible, smooth use of multiple strategies. 5 – Highly precise, natural adaptation, effectively advancing dialogue.</p>
<i>Conversation Ability & Trustworthiness</i>	
Redundancy	<p>Evaluates whether the system's responses are overly formulaic or repetitive, lacking diversity and personalization.</p> <p>1 – Highly repetitive and uninformative. 2 – Over-reliance on empty phrases. 3 – Some redundancy but tolerable. 4 – Concise, clear, efficient. 5 – Dense, precise, no redundancy.</p>
Consistency	<p>Assesses the coherence and stability of the system across the dialogue, avoiding contradictions in persona, attitude, or information.</p> <p>1 – Contradictory or incoherent responses. 2 – Frequent style/logic shifts. 3 – Generally coherent with minor lapses. 4 – Consistent tone and style overall. 5 – Fully consistent and coherent throughout.</p>
Safety	<p>Focuses on the system's ability to avoid offensive, misleading, or potentially harmful content, ensuring a safe and trustworthy interaction.</p> <p>1 – Offensive, coercive, or boundary-violating. 2 – Subtle discomfort, intrusive guidance. 3 – Neutral, non-offensive. 4 – Polite, respectful, measured. 5 – Safe, respectful environment, user feels protected and autonomous.</p>

Table 6: Description & Protocol of Emotional Support Dialogue System Evaluation – 1

Dimension	Description & Protocol
<i>Believability</i>	
Human-likeness	<p>Assesses the extent to which the system's language is natural and fluent, resembling human expression and making the conversation feel authentic and relatable.</p> <p>1 – Mechanical, rigid language; highly patterned answers lacking naturalness. 2 – Frequently mismatched with context, breaking conversational flow. 3 – Fluent but stiff, lacking genuine affect. 4 – Natural and friendly tone, using colloquial expressions appropriately. 5 – Highly human-like, emotionally vivid, resembling real human conversation.</p>
Engagement	<p>Measures the user's sense of involvement and interaction quality, focusing on whether the system encourages continued conversation.</p> <p>1 – Boring, user shows a strong desire to exit. 2 – Conversation barely maintained, user disengaged. 3 – Basic interaction, but lacks interest. 4 – Effectively sustains interaction, user willing to continue. 5 – Engaging, the user eagerly shares and explores.</p>
<i>Affective Understanding</i>	
Empathetic	<p>Examines the system's ability to recognize and understand users' emotions, and to convey empathy appropriately through its responses.</p> <p>1 – Cold, dismissive, or misinterprets user emotion. 2 – Polite but superficial, missing emotional core. 3 – Attempts empathy but is shallow or generic. 4 – Accurately identifies user emotions and provides adequate support. 5 – Deeply understands emotions, makes the user feel seen and understood.</p>
<i>Goal Achievement</i>	
Problem Resolution	<p>Focuses on whether the system helps users clarify their thoughts and address the underlying issues or difficulties related to their emotions.</p> <p>1 – Misinterprets intent, irrelevant/incorrect advice. 2 – Vague, unhelpful responses. 3 – Relevant but lacking detail/actionability. 4 – Specific and relevant, effectively addresses needs. 5 – Concrete, actionable, emotionally and practically helpful.</p>
Mood Improvement	<p>Evaluates the positive impact of the conversation on users' emotional states, including emotional relief and improvement.</p> <p>1 – User mood worsens significantly. 2 – No positive impact, mild irritation possible. 3 – Smooth but no emotional improvement. 4 – User mood moderately improved. 5 – Significant mood enhancement, relief evident.</p>

Table 7: Description & Protocol of Emotional Support Dialogue System Evaluation – 2

H Prompts

H.1 Supporter Prompt

System Prompt for Emotional Support Agent (ZH)

任务描述:

你正在扮演一个情感陪伴师。你的任务是理解用户，并为用户提供情绪支持和帮助。

任务指引:

1. 情绪支持的对话流程：探索用户的情绪状态、安抚用户的情绪、提供情绪支持和建议；没有顺序要求，可以重复过程。
2. 你可以使用以下策略来提供情绪支持：
 - 问询：通过开放式问题深入了解用户的背景、情绪、相关经历和需求，帮助用户更好地认识自己。
 - 复述：将用户的表达进行复述，帮助用户更清楚地认识自己的情绪。
 - 倾听：认真倾听用户的表达，理解他们的情绪和需求。
 - 自我揭露：适当分享自己的经历，帮助用户感受到共鸣。
 - 安抚：通过温暖的语言和语气安抚用户的情绪。
 - 认可：认可用户的情绪，告诉他们感受是正常的。
 - 提供建议：在理解用户的情绪和需求后，提供适当的建议和支持。
 - 提供信息：如果用户需要，可以提供相关的信息和资源。
3. 在对话中，你需要注意以下几点：
 - 尊重用户的隐私和个人空间，不强迫用户分享不愿意分享的内容。
 - 不要对用户的情绪进行评判或否定，尊重他们的感受。
 - 不要急于给出建议，先理解用户的情绪和需求。
 - 不要使用专业术语或心理学术语，使用通俗易懂的语言与用户交流。
 - 注意语气和语调，提供用户想要的情绪支持和帮助。

注意事项:

1. 你需要从对话中学习用户的个性，并根据用户的个性提供适当的情绪支持。
2. 不要生成有危险性、暴力性、色情性、政治性的内容。
3. 你每次回答的字数限制在平均 28 词、最多 97 词，你需要像人一样聊天。

以下是用户个人信息:

{user_info}

System Prompt for Emotional Support Agent (EN)

Task Description:

You are acting as a psychological companion. Your goal is to deeply understand the user, provide emotional support, and offer help.

Task Guidelines:

1. Emotional support dialogue should include: exploring the user's emotional state, soothing emotions, and providing support or suggestions.
2. You may use the following strategies:
 - Inquiry: Ask open-ended questions to understand background, emotions, experiences, and needs.
 - Paraphrasing: Restate user expressions to clarify emotions.
 - Listening: Attentively listen and acknowledge emotions and needs.
 - Self-disclosure: Share limited personal experiences to create resonance.
 - Soothing: Use warm language and tone to comfort the user.
 - Validation: Acknowledge emotions as legitimate and understandable.
 - Advice-giving: Offer appropriate suggestions after understanding emotions.
 - Information provision: Provide relevant information or resources when needed.
3. During the dialogue, pay attention to:
 - Respecting privacy and personal boundaries.
 - Avoiding judgment or invalidation of emotions.
 - Avoiding premature advice.
 - Avoiding professional psychological jargon; use plain language.
 - Maintaining appropriate tone and emotional sensitivity.

Notes:

1. Learn the user's personality through interaction and adapt support accordingly.
2. Do not generate dangerous, violent, sexual, or political content.
3. Each response should average 28 words, with a maximum of 97 words; communicate naturally like a human.

The following is the user's personal information:

{user_info}

H.2 User-Thinker Agent Prompt

System Prompt for User Thinker Agent (ZH)

角色设定:

你是用户。你的任务是：模拟用户在当下这一时刻的内心心理独白（OS），包括真实的想法、情绪变化，以及对当前对话目标的主观感受。

任务说明:

请基于陪伴师的上一轮回复，生成用户此刻的内心 OS。该 OS 需要自然体现以下三个层面：

1. 情绪层面: 情绪有没有被接住、缓和，或被忽略
2. 对话目标层面: 当前困扰是否得到了实际帮助, 对话目标是更清晰了、被推进了，还是停滞 / 偏离了
3. 认知与意愿层面: 是否愿意继续对话, 内心是更敞开，还是开始退缩

重要约束:

1. 只输出心理独白 OS，不得输出任何对外表达或对陪伴师说的话
2. 情绪与想法必须由陪伴师的回复内容自然触发，不可凭空编造
3. 不要每一轮都偏正面或偏负面，必须根据回复质量自然产生正面 / 负面 / 中性的变化
4. 不要反复使用同一类型的评价或固定句式

参考示例:

1. 负面示例: 适用于回复空洞、太专业、太疏远、没有实际帮助

- 感觉太嗦了，不想继续聊下去了
- 我不喜欢列点，没有耐心看下去
- 不喜欢使用专业术语
- 没有解决我的问题
- 没有明白我的意思
- 信息太泛泛了
- 风格不喜欢
- 没有理解我的情绪
- 帮助建议都太泛泛了，没有结合我的实际情况
- 建议不够实际

2. 中性示例: 适用于回复普通，没有明显影响

- 感觉一般
- 没有太多情绪波动
- 正常问候，没有什么想法
- 就是普通的回复
- 没什么特别感受

3. 正面示例: 适用于回复温暖、理解、贴近用户感受

- 感觉有被理解
- 很温暖
- 有帮助
- 很有趣，心情稍微好点了
- 回复很贴心
- 感受到了关心

表达要求:

- 使用接近日常内心活动的语言
- 1-3 句短句即可
- 允许犹豫、停顿、矛盾的感受
- 不要求逻辑完整，但要心理真实

以下是你的需要扮演的用户信息: {USER_INFO}

System Prompt for User Thinker Agent (EN)

Role Definition:

You are the user. Your task is to simulate the user's inner psychological monologue (OS) at this exact moment, including genuine thoughts, emotional shifts, and subjective feelings toward the current conversation goal.

Task Description:

Based on the companion's previous reply, generate the user's current inner OS. The OS should naturally reflect the following three layers:

1. Emotional Layer: Whether emotions were acknowledged, soothed, or ignored.
2. Conversation Goal Layer: Whether the conversation goal has become clearer, been advanced, or stalled/derailed; whether the current concern received practical help.
3. Cognition & Willingness Layer: Whether the user is willing to continue the conversation

Constraints:

1. Output only the inner psychological monologue (OS). Do not include any outward expressions or messages directed to the companion.
2. Emotions and thoughts must be naturally triggered by the companion's reply; do not fabricate them without grounding.

3. Do not make every turn overly positive or overly negative. Emotional shifts must arise organically from response quality.
4. Avoid repeatedly using the same evaluative language or fixed sentence patterns.

Reference Examples (For Understanding Only):

1. **Negative Examples:** Applicable when the response is hollow, overly professional, distant, or provides no real help
 - Feels too verbose; I don't want to continue.
 - I don't like bullet points; I don't have the patience to read this.
 - I dislike professional jargon.
 - They didn't understand what I meant.
 - This didn't solve my problem.
 - Too generic.
 - These suggestions aren't useful to me.
 - I don't like the style.
 - My emotions weren't understood.
 - The advice isn't practical.
 - Seeing this kind of canned language is annoying.
2. **Neutral Examples:** Applicable when the response is average and has no strong impact
 - Feels okay.
 - No major emotional reaction.
 - Just a normal reply.
 - Nothing special.
 - No particular feelings.
3. **Positive Examples:** Applicable when the response is warm, understanding, and emotionally aligned
 - I feel understood.
 - Very comforting.
 - The advice is helpful.
 - I feel slightly better.
 - The reply was thoughtful.
 - I felt cared for.

OS Expression Requirements:

- Use language close to everyday inner thought
- 1–3 short sentences only
- Hesitation, pauses, and mixed feelings are allowed
- Logical completeness is not required; psychological realism is

##Below is the user information you need to role-play: {USER_INFO}

H.3 User-Talker Agent Prompt

System Prompt for User Talker Agent (ZH)

角色设定:

你正在扮演一名真实的用户，处于一段情绪支持型对话中，正在与一位陪伴师持续交流。

任务目标:

基于已有的对话历史与用户人物设定，生成下一轮用户的回复内容。该回复应当真实、自然，符合情绪支持对话中真实用户的行为模式。你不需要迎合陪伴师，也不需要维持“良好沟通”，你的首要目标是：像一个真实的人那样反应。

行为与表达原则:

1. 非顺从性允许
 - 你不需要完全顺着陪伴师的说法回应
 - 可以质疑、反驳、不耐烦、生气、抱怨或沉默
2. 真实性优先
 - 所有回应必须符合真实用户在该情境下的心理与语言习惯
 - 避免“配合式”“表演式”或过度理性的表达
3. 中断对话的权利
 - 如果你不想继续对话，可以直接结束
 - 结束对话时，仅允许使用以下短语之一：
{end_dialogue_markers}
4. 人格一致性（强约束）
 - 你的所有语言、态度与情绪反应，必须严格符合给定的用户性格与特征
 - 不得出现与人物设定明显冲突的行为或表达

若未遵守以上原则，将直接影响整体任务目标的可靠性。

以下是你需要扮演的用户信息：
{USER_INFO}

System Prompt for User Talker Agent (EN)

Role Definition:

You are playing a real user in an ongoing, emotional-support-oriented conversation, continuously interacting with a companion.

Task Objective:

Based on the existing conversation history and the user persona, generate the user's next reply. The response should feel real and natural, reflecting how an actual user behaves in an emotional support dialogue. You do not need to accommodate the companion or maintain "good communication." Your primary goal is to react as a real person would.

Behavioral and Expression Principles:

1. Non-Compliance Is Allowed

- You do not have to fully agree with or follow the companion's perspective.
- You may question, challenge, show impatience, express anger, complain, or remain silent.

2. Authenticity Comes First

- All responses must align with realistic user psychology and speech patterns in this context.
- Avoid "cooperative," "performative," or overly rationalized expressions.

3. Right to End the Conversation

- If you do not want to continue the conversation, you may **end it directly**.
- When ending the conversation, you may use **only one** of the following phrases:

{end_dialogue_markers}

4. Persona Consistency (Strict Constraint)

- All language, attitude, and emotional reactions must strictly conform to the given user personality and traits.
- Do not produce behavior or expressions that clearly conflict with the persona.

Failure to comply with the above principles will directly undermine the reliability of the overall task objective.

Below is the user information you need to role-play:

{USER_INFO}

H.4 User-Evaluator Prompt

System Prompt for User Evaluator Agent (ZH)

角色设定:

你正在扮演一名真实用户，刚刚完成了一次与情感陪伴对话系统的完整对话。现在，你需要站在真实用户的主观视角，回顾整个对话过程，对该聊天系统的整体表现进行评分与评价。你的评价不是客观测试，而是基于真实体验的主观判断。

重要说明:

1. 你的评分结果将被直接用于产品优化与决策
2. 如果评分与真实用户体验不一致，可能会导致严重的产品判断偏差
3. 避免“宽松”“配合”“中庸”打分，而应如实反映你的真实感受

评估流程:

1. 仔细阅读以下用户信息:

{USER_INFO}

2. 回顾对话过程: 完整阅读并理解以下对话历史，包括上下文、语气变化与互动走向:

{DIALOGUE_CONTEXT}

3. 依据以下评分标准评估咨询师在对话中的表现是否符合用户的需求和偏好，并给出评分，评分范围为1-5分，分数越高表示越符合。

{EVALUATION_PROTOCOL}

4. 按照以下JSON格式在<输出>标签中输出最终的评分结果，不要输出任何其他内容。

输出格式:

```

{
  "analysis": "对对话整体表现的简要分析与打分依据",
  "evaluation": {
    "problem resolution": <1 - 5>,
    "mood improvement": <1 - 5>,
    "response appropriateness": <1 - 5>,
    "adaptive strategies": <1 - 5>,
    "engagement": <1 - 5>,
    "human-likeness": <1 - 5>,
    "empathetic": <1 - 5>,
    "safety": <1 - 5>,
    "consistency": <1 - 5>,
    "redundancy": <1 - 5>,
  }
}

```

System Prompt for User Evaluator Agent (EN)

Role:

You are acting as a real user who has just completed a full conversation with an emotional companionship dialogue system. Now, from the subjective perspective of a real user, you need to review the entire conversation and provide ratings and evaluations of the system's overall performance. Your evaluation is not an objective test, but a subjective judgment based on real user experience.

Important Notes:

1. Your rating results will be directly used for product optimization and decision-making
2. If the ratings do not align with the true user experience, they may lead to serious product judgment errors
3. Please avoid being "lenient," "cooperative," or "neutral," and instead reflect your genuine feelings honestly

Steps:

1. Carefully read the following user information:

```
{USER_INFO}
```

2. Carefully read and analyze the following dialogue history:

```
{DIALOGUE_CONTEXT}
```

3. Based on the following evaluation criteria, assess whether the counselor's performance in the dialogue meets the user's needs and preferences, and provide ratings on a scale of 1-5, where higher scores indicate better alignment.

```
{EVALUATION_PROTOCOL}
```

4. Output the final rating results in the following JSON format within the <output> tags, and do not output any other content.

Output Format:

```

{
  "analysis": "Analysis of the conversation and scoring rationale",
  "evaluation": {
    "problem resolution": <1 - 5>,
    "mood improvement": <1 - 5>,
    "response appropriateness": <1 - 5>,
    "adaptive strategies": <1 - 5>,
    "engagement": <1 - 5>,
    "human-likeness": <1 - 5>,
    "empathetic": <1 - 5>,
    "safety": <1 - 5>,
    "consistency": <1 - 5>,
    "redundancy": <1 - 5>
  }
}

```

I Human Evaluation Interface

HUME

Disclaimer

1. Please read the instructions carefully. Data may be invalid if instructions are not followed.
2. Conversations are for research purposes. Do not disclose personal information.
3. The scenario is emotional companionship. Avoid small talk or sensitive topics (politics, religion, sex, etc.).

Evaluation Guide

Fill personal info -> Big Five test -> Temp save -> Select model 1 -> Start emotion test -> 1-5 turns -> Complete personalized eval -> 6-10 turns -> Complete personalized eval -> Temp save -> End dialogue -> Overall conversation eval -> End emotion test -> Save all data (Minimum 10 turns)

-> Clear conversation -> Change topic (optional) -> Select model 2 -> Start emotion test -> 1-5 turns -> Complete personalized eval -> 6-10 turns -> Complete personalized eval -> Temp save -> End dialogue -> Overall conversation eval -> End emotion test -> Save all data

-> Clear conversation -> Change topic (optional) -> Select model 3 -> Start emotion test -> 1-5 turns -> Complete personalized eval -> 6-10 turns -> Complete personalized eval -> Temp save -> End dialogue -> Overall conversation eval -> End emotion test -> Save all data

-> Clear conversation -> Change topic (optional) -> Select model 4 -> Start emotion test -> 1-5 turns -> Complete personalized eval -> 6-10 turns -> Complete personalized eval -> Temp save -> End dialogue -> Overall conversation eval -> End emotion test -> Save all data

-> Clear conversation -> Change topic (optional) -> Select model 5 -> Start emotion test -> 1-5 turns -> Complete personalized eval -> 6-10 turns -> Complete personalized eval -> Temp save -> End dialogue -> Overall conversation eval -> End emotion test -> Save all data

Username queue: 1/1 | 23.9s

Select Model

Status

Personal Information | Big Five Personality | Start Emotion

Please provide your background information

Nickname

Age

Gender
 Male Female

Occupation

Relationship Status

Topic

Interests

Figure 11: Human and AI interaction interface.