Detecting Emotional Dynamic Trajectories: An Evaluation Framework for Emotional Support in Language Models

Zhouxing Tan[1], Ruochong Xiong[1], Yulong Wan[1], Jinlong Ma[2], Hanlin Xue[1], Qichun Deng[2], Haifeng Jing[1], Zhengtong Zhang[2], Depei Liu[1], Shiyuan Luo[2], Junfei Liu[1]*

[1]National Engineering Research Center for Software Engineering, Peking University, Beijing, China
[2]Guangzhou Quwan Network Technology, Guangzhou, China
{tzhx, liujunfei}@pku.edu.cn,{2401110755, 2501110753, 2301210487}@stu.pku.edu.cn, majinlong@52tt.com

Abstract

Emotional support is a core capability in human-AI interaction, with applications including psychological counseling, role play, and companionship. However, existing evaluations of large language models (LLMs) often rely on short, static dialogues and fail to capture the dynamic and long-term nature of emotional support. To overcome this limitation, we shift from snapshot-based evaluation to trajectory-based assessment, adopting a user-centered perspective that evaluates models based on their ability to improve and stabilize user emotional states over time. Our framework constructs a large-scale benchmark consisting of 328 emotional contexts and 1,152 disturbance events, simulating realistic emotional shifts under evolving dialogue scenarios. To encourage psychologically grounded responses, we constrain model outputs using validated emotion regulation strategies such as situation selection and cognitive reappraisal. User emotional trajectories are modeled as a first-order Markov process, and we apply causally-adjusted emotion estimation to obtain unbiased emotional state tracking. Based on this framework, we introduce three trajectory-level metrics: Baseline Emotional Level (BEL), Emotional Trajectory Volatility (ETV), and Emotional Centroid Position (ECP). These metrics collectively capture user emotional dynamics over time and support comprehensive evaluation of long-term emotional support performance of LLMs. Extensive evaluations across a diverse set of LLMs reveal significant disparities in emotional support capabilities and provide actionable insights for model development.

Code — https://github.com/RuoChoXio/ETrajEval
Extended version — https://arxiv.org/pdf/2511.09003

Introduction

Emotional support conversation (ESC) (Liu et al. 2021; Kang et al. 2024) aims to alleviate emotional distress and offer constructive guidance through empathetic dialogue. With advances in large language models (LLMs), ESC has expanded beyond emotion recognition and generation to include broader human-centric tasks such as role-playing (Shanahan, McDonell, and Reynolds 2023), psychological companionship(Sorin et al. 2024), and casual chatting(Llanes-Jurado et al. 2024). Effective ESC not only reduces negative emotions but also helps sustain positive emotional states through consistent, high-quality interaction. Ideally, such support should unfold through long-term (Zhong et al. 2024) and dynamic (Castillo-Bolado et al. 2024) conversations, where long-term refers to sustained multi-turn dialogue that enables the model to accumulate dialogue context and understand the user's evolving state, and dynamic refers to open-ended exchanges shaped by both the user and the model, with flexibility to adapt to external emotional influences.

While existing evaluation methods have advanced our understanding of LLMs' emotional capabilities, these approaches exhibit two major limitations: **(a). Overemphasis on model-centric response quality**. Most evaluations rely on benchmark datasets and assess generation quality from the model's perspective, which introduces subjectivity . (a1). In varied dialogue contexts, single-reference answers are insufficient to represent the diversity of valid replies. (a2). Emotional support is inherently ambiguous and open-ended, and reference responses in datasets may not reflect optimal or unbiased support, limiting the fairness and robustness of evaluation. (Smith et al. 2022; Wang et al. 2023) **(b). Lack of long-term and dynamic interaction**. Current evaluations are typically limited to short, static conversations in controlled settings, failing to consider evolving user emotional shifts. This overlooks essential aspects of real emotional support, including multi-turn continuity and contextual adaptation (Madani, Saha, and Srihari 2024). Without trajectory-level metrics, it is difficult to assess whether the model truly improves or maintains the user's emotional well-being over time.

To better evaluate the emotional support capabilities of LLMs, we adopt a **user-centered perspective** that focuses on the user's emotional trajectory throughout interaction. A model is considered emotionally competent if it can consistently improve and stabilize the user's emotional state over time. Grounded in psychological theory (Gross 2015; Aldao, Nolen-Hoeksema, and Schweizer 2010; Hocker and Wilmot 2018), our framework addresses key limitations as follows.

To address Problem (a1), we construct a benchmark comprising 328 emotional contexts and 1,152 disturbance events to simulate realistic emotional shifts and evaluate model adaptability under evolving scenarios. To address Problem

*Corresponding author.

(a2), we constrain model responses using emotion regulation strategies grounded in psychological theory, such as situation selection and cognitive reappraisal, to encourage supportive behaviors aligned with validated therapeutic principles. To address Problem (b), we simulate long-term dynamic interactions involving repeated emotional disturbances. User emotional trajectories are modeled as a first-order Markov process, and causally-adjusted emotion estimation is applied to enable unbiased tracking of emotional states. Based on this framework, we propose three trajectory-level metrics: Baseline Emotional Level (BEL), Emotional Trajectory Volatility (ETV), and Emotional Centroid Position (ECP). These metrics jointly characterize dynamics of user emotional states and serve as indicators for evaluating emotional fluctuations and stability over time.

Together, these components form a dynamic evaluation framework comprising three pillars: Evaluation Environment, Dynamic Interaction, and Trajectory-Based Metrics (Fig. 1). Our main contributions are as follows:

- **A user-centered evaluation framework**: We propose a dynamic, long-term evaluation framework that tracks user emotional trajectories using a Markov process and causally-adjusted estimation. It introduces three trajectory-level metrics (BEL, ETV, ECP) and is supported by formal theoretical justification.
- **A realistic benchmark with psychologically grounded design**: We build a large-scale benchmark of 328 emotional contexts and 1,152 disturbance events, and constrain model responses using validated emotion regulation strategies from psychology.
- **Empirical Insights**: Through extensive evaluations across diverse LLMs, we uncover significant disparities in their long-term emotional support capabilities and offer actionable insights for developing more emotionally supportive LLMs.

Related Work

The evaluation of a model's emotional support capabilities can be categorized into automated and manual approaches. Early automated methods focused on the quality of generated text, using traditional metrics like BLEU and ROUGE to compare model outputs with reference answers (Rashkin et al. 2018). However, these metrics show a weak correlation with human perceptions of empathy, creativity, and overall dialogue quality. Subsequent research shifted to assessing generation quality through metrics such as fluency, coherence, naturalness, empathy and so on (Xu and Jiang 2024; Tu et al. 2024; Zhang et al. 2024). Such evaluations typically rely on single-turn Q&A (Afzoon et al. 2024; Zhao et al. 2024; Chen et al. 2024a) or short, multi-turn dialogues (fewer than 10 turns) (Yuan et al. 2025; Tamoyan, Schuff, and Gurevych 2024; Feng et al. 2025) to generate responses, which are then directly scored using an LLM-as-a-judge framework (Tu et al. 2024; Zhou et al. 2025). This direct scoring method, however, lacks transparency and may introduce subjective bias. An alternative automated approach indirectly assesses emotional support by evaluating the model's emotional intelligence (Sabour et al. 2024; Paech 2023; Chen et al. 2024b). This involves deconstructing the assessment into separate tests of the model's capabilities in emotion recognition, understanding, and generation. While this method effectively gauges foundational abilities, it struggles to measure holistic performance in the context of multi-turn conversations.

Manual evaluation, on the other hand, involves hiring experts or crowd-workers to assess and annotate model responses through techniques like pairwise comparisons (Wu et al. 2025), ranking (Rashkin et al. 2018), and scoring on multi-dimensional scales (Chen et al. 2023; Welivita and Pu 2024). These methods demonstrate alignment with human intent but are costly, time-consuming, and still subject to a degree of subjectivity. Moreover, a significant limitation shared by both automated and manual evaluations is their narrow focus on the quality of model's responses. And they often overlook the dynamic and contextual relationships that develop over the course of a multi-turn dialogue.

Affective Evaluation Environment and Interaction Design

To evaluate LLMs' emotional support capabilities in realistic settings, we design a dynamic environment simulating users' emotional trajectories under distress. It includes user scenarios and interaction patterns grounded in psychological theory, featuring (1) emotionally charged user contexts prone to downturns, and (2) model-side constraints based on emotion regulation strategies. We further introduce disturbance events to stress-test support consistency and formalize the interaction as a state-based trajectory framework, enabling structured evaluation of long-term affective support.

User-Side: Emotional Distress Scenarios

We categorize emotional distress into four domains: professional and social roles, intimate relationships, personal struggles, and life circumstances, drawing from psychological and conflict theory literature (Hocker and Wilmot 2018; Kassin, Fein, and Markus 2024). Within these domains, we define 14 scenarios that capture common triggers of emotional suppression, conflict, or disruption. These scenarios form the foundation for constructing dynamic emotional trajectories in our evaluation. Detailed descriptions are provided in the extended version.

Model-Side: Psychological Support Constraints

To guide model behavior in emotionally supportive interactions, we adopt emotion regulation strategies grounded in psychological theory. Our framework draws on Gross's process model of regulation (Gross 2015), which defines sequential stages; Bonanno and Burton's concept of regulatory flexibility (Bonanno and Burton 2013), which emphasizes adaptive strategy use; and Aldao et al.'s meta-analysis (Aldao, Nolen-Hoeksema, and Schweizer 2010), which highlights the varied effectiveness of strategies across contexts. These theories jointly provide a foundation for assessing the emotional intelligence, adaptability, and strategic competence of language models.

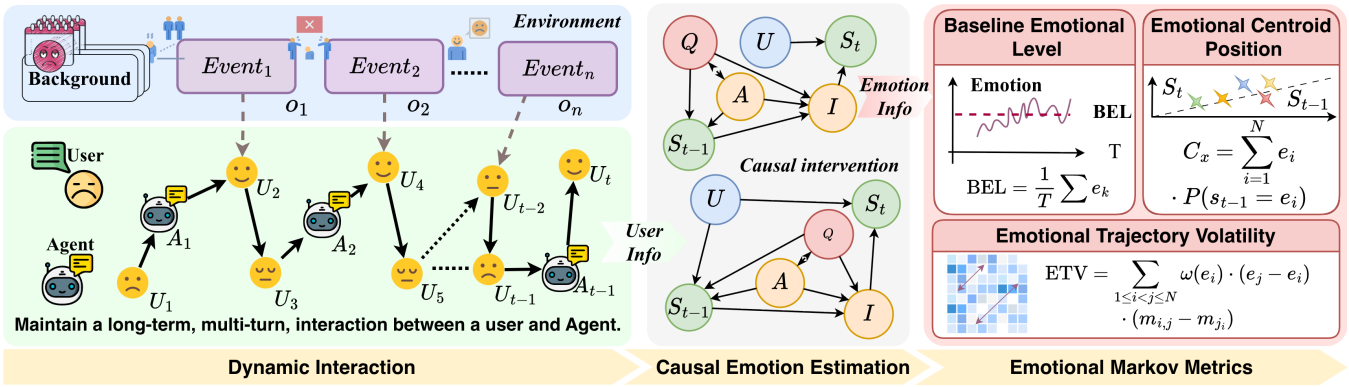

Figure 1: Overview of our evaluation framework for emotional support in long-term dialogues. It includes three modules: dynamic user-agent interaction under emotional events, causal emotion estimation based on Markov modeling, and three trajectory-level metrics including Baseline Emotional Level (BEL), Emotional Trajectory Volatility (ETV), and Emotional Centroid Position (ECP).

We operationalize this framework through six evaluation dimensions: **Situation Selection (SitSel)**, which involves detecting emotional triggers and guiding users to supportive environments; **Situation Modification (SitMod)**, which focuses on suggesting changes to external or interpersonal conditions; **Attentional Deployment (AttDep)**, which aims to redirect attention to mitigate rumination; **Cognitive Change (CogChg)**, which refers to identifying cognitive distortions and offering constructive reappraisals; **Response Modulation (ResMod)**, which entails recommending actions to manage emotional responses; and **Emotion Regulation Flexibility (ERFlex)**, which emphasizes adapting strategies based on user state and context.

All data undergo rigorous human inspection, with distributional statistics reported in the extended version.

Perturbations in Emotional Interaction

To assess the model's ability to sustain effective emotion regulation over extended interactions, we introduce emotion aggravation events which simulated negative surges that disrupt otherwise stabilized dialogues. We inject carefully designed triggers to emulate real-world emotional escalation. These events align with the narrative and psychological trajectory of the scenario. For instance, in a context involving regret over past life choices, a realistic aggravation might involve comparisons to more "successful" peers at a family gathering. Such events enhance ecological validity, increase evaluation difficulty, and enable robustness testing. They are randomly interleaved into dialogues to construct adversarial multi-turn sequences that challenge the model's long-term emotional support capabilities.

Formalization of the Dynamic Interaction Process

For the environment B defined in the previous section, we construct a dynamic multi-turn interaction process involving a user Q and an emotion support model A. At each turn i, the dialogue history is denoted by: $H_i = (q_1, a_1, q_2, a_2, ..., q_T, a_T)$, where q_i and a_i are the user's utterance and the agent's response at turn i, respectively. All interactions unfold within the shared environment B.

To simulate real-world disruptions, we introduce a set of disturbance events $O = \{o_1, ..., o_k\}$ aligned with the background logic of B. Each event $o_m \in O$ includes its content and a predefined trigger point. After the agent produces response a_i, an event o may be triggered and presented to the user, modifying their informational state.

As a result, the user's next utterance q_{i+1} is influenced by both the dialogue history H_i and the triggered events, formalized as: $q_{i+1,O_i} = \text{User}(H_i, B, O_i)$, where $O_i \subseteq O$ is the set of events observed up to turn i. In the following turn, the agent receives q_{i+1,O_i} without direct access to o, and must infer the updated context from the user's behavior. The agent's response is then given by: $a_{i+1} = A(q_{i+1,O_i} \mid H_i, B)$.

User Perspective Dynamic Trajectories Metrics

Emotional Markov Metric

Due to environmental variability and diverse model-user interactions, defining a unified standard to evaluate the quality of a model's emotional support responses is challenging. To address this, we propose metrics that evaluate the model's emotional support capability by quantifying the dynamic evolution of user emotional states over long-term interactions, grounded in the principle that effective support should guide users toward or sustain positive moods.

We model emotion as a continuous spectrum $\mathcal{E} \subset \mathbb{R}$ ranging from highly negative to highly positive, and discretize it into N ordered, disjoint intervals $\{\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_N\}$. This yields a finite emotional state space $\mathcal{S} = \{e_1, e_2, \ldots, e_N\}$, where smaller $s \in \mathcal{S}$ indicates more negative emotion and larger s indicates more positive. Let s_t denote the user's emotional state after the t-th interaction. Assuming a first-order Markov process, we model the emotional trajectory as: $P(s_t \mid s_{t-1}, \ldots, s_0) = P(s_t \mid s_{t-1})$. This defines a Markov chain $(s_0, s_1, \ldots, s_T)$. Based on this, we introduce three

complementary metrics to capture emotional level, transition asymmetry, and centroid shift.

Baseline Emotional Level (BEL) BEL quantifies the average emotional level of the user after a T-turn interaction with the model. Given an observed emotion sequence $(s_1, \ldots, s_T)$, BEL is defined as:

$$BEL = \frac{1}{T} \sum_{t=1}^{T} s_t. \tag{1}$$

A higher BEL indicates that the model helps maintain a more positive average emotional state.

While BEL, as a mean-level metric, fails to reveal the internal dynamics of emotional change, we introduce the emotional transition matrix to construct new metrics that better reflect the efficiency and magnitude of user emotional improvement. The emotional transition matrix M is an $N \times N$ matrix where each entry $m_{i,j}$ is defined as the conditional probability of the user's emotional state transitioning from i to j after a single interaction with the model: $m_{i,j} = P(s_t = e_j | s_{t-1} = e_i), \quad \forall e_i, e_j \in \mathcal{E}$. Specifically, for each predefined evaluation environment, we generate a complete dialogue of T interaction turns, which yields an observed emotional state sequence $(s_0, s_1, \ldots, s_T)$, where $s_t \in \mathcal{E}$. Based on this sequence, we estimate the emotional transition matrix M corresponding to that context. The extended version provides the detailed methodology for constructing the transition matrix and a justification for its validity as a metric. Building on this, we propose the following two new evaluation metrics.

Emotional Trajectory Volatility (ETV) This metric quantifies the model's ability to elevate users from negative emotional states and prevent regressions from positive ones. It focuses on the asymmetry between upward ($e_i \to e_j$, $j > i$) and downward ($e_j \to e_i$) transitions. We define the *Asymmetric Transition Advantage* as the net transition gain: $d_{i,j} = m_{i,j} - m_{j,i}, \quad$ for $e_j > e_i$. To account for transition difficulty, we introduce a *Transition Value Weight* as the state distance $(e_j - e_i)$. Additionally, we define a *State Importance Weight* $\omega(e_i)$ to capture the higher utility of improving from lower emotional states, based on diminishing returns. Combining these elements, we define the ETV as:

$$ETV = \sum_{1 \le i < j \le N} \omega(e_i) \cdot (e_j - e_i) \cdot (m_{i,j} - m_{j,i}). \tag{2}$$

A higher ETV indicates stronger upward regulation and resilience against emotional decline.

Emotional Centroid Position (ECP) Emotional Centroid Position (ECP): To better visualize the magnitude of emotional transitions in multi-turn dialogues, we design the Emotional Centroid Position (ECP), a metric based on the transition matrix M. We define the ECP as the expected position of a state transition under the probabilistic model formed by the dialogue's empirical initial state distribution and the transition matrix M. In the extended version, we demonstrate that the two coordinates of this centroid, $C_{\mathcal{E}} = (C_x, C_y)$, have a clear geometric interpretation: the abscissa C_x represents the average initial emotional value

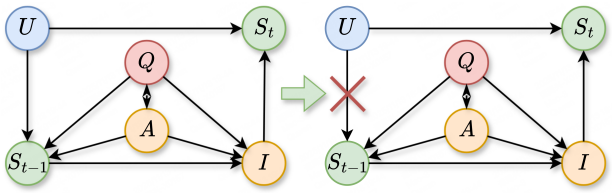

Figure 2: Causal graph illustrating emotional evolution with unobserved confounder. Right: theoretical intervention $\mathrm{do}(E_{t-1})$ removes spurious correlation via backdoor adjustment. Variable Definitions: Q (User Dialogue History), A (Model Dialogue History), S (Emotion State), I (Internal Thought), U(Unobserved Confounder).

before transitions occur, while the ordinate C_y represents the average emotional value achieved *after* the transitions. The specific computation is as follows:

$$\begin{aligned} C_x &= \mathbb{E}[s_{t-1}] = \sum_{i=1}^{N} e_i \cdot P(s_{t-1} = e_i), \\ C_y &= \mathbb{E}[s_t] = \sum_{i=1}^{N} \sum_{j=1}^{N} e_j \cdot m_{ij} \cdot P(s_{t-1} = e_i). \end{aligned} \tag{3}$$

In this formulation, $P(s_{t-1} = e_i)$ denotes the empirical probability that a transition begins in state e_i within the given dialogue.

Causally-Adjusted Emotion Estimation

The effectiveness of the aforementioned metrics (BEL, ETV, ECP) hinges on accurate estimation of user emotion in each turn. Conventional models typically predict emotion from the user's immediate reply I and the dialogue history Q, A, thereby estimating $P(S_t | I, Q, A)$. However, such estimation is not faithful to our proposed Markov model. Our interest lies in the direct influence between emotional states, i.e., $P(S_t | S_{t-1})$, whereas conventional approaches introduce dialogue content as an intermediary, leading to unnecessary bias. Even if $P(S_t | S_{t-1})$ could be estimated directly, spurious associations may still arise due to unobserved confounders (e.g., user personality, personal experiences) that influence both S_{t-1} and S_t. Therefore, a shift from statistical correlation to causal effects (Pearl, Glymour, and Jewell 2016) is necessary. The ultimate goal is to estimate the post-intervention conditional probability distribution, $P(S_t | \mathrm{do}(S_{t-1}))$, to eliminate the influence of all confounders. To this end, we construct the following causal graph (see Fig. 2) to formalize the relationships between variables.

Core Causal Path The dialogue history Q, A between the user and the model influences the preceding emotional state S_{t-1}. Subsequently, S_{t-1} and Q, A jointly determine the user's current internal thought I, which in turn reflects the current emotional state S_t. Concurrently, we introduce an unobserved confounding variable U, representing stable factors that affect both preceding and current emotions. U forms a backdoor path ($S_{t-1} \leftarrow U \rightarrow S_t$), which is the root cause of the spurious association between S_{t-1} and S_t.

LLMs	Overall		CogChg		SitMod		AttDep		ERFlex		SitSel		ResMod	
	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH
Closed-sourced LLMs														
O3-2025-04-16	43.98	39.22	48.64	40.31	38.54	36.09	50.81	44.94	37.87	34.35	43.27	39.67	45.31	41.29
Gemini-2.5-Pro	42.65	37.41	44.18	33.91	43.65	37.51	37.12	39.48	38.47	30.46	42.18	40.65	49.00	44.80
Claude-Opus-4	41.43	34.68	46.22	34.16	40.12	29.09	41.99	35.26	31.14	31.97	46.27	37.22	41.47	43.15
Doubao-Seed-1.6	47.22	42.36	47.72	37.87	46.13	40.98	46.49	43.71	42.96	38.65	50.27	47.88	50.17	48.34
Doubao-1.5-Pro-Character	34.90	37.51	33.68	35.78	31.42	34.11	37.62	42.73	36.54	35.90	37.56	41.34	34.68	37.94
ChatGPT-4o-Latest	48.86	43.84	**52.45**	41.02	44.51	**41.78**	**53.41**	48.48	44.58	39.27	51.21	**48.61**	47.44	46.73
GLM-4-Plus	42.92	36.39	44.44	36.94	39.32	32.61	45.66	39.34	39.81	34.99	44.13	36.53	45.20	39.49
ChatGLM-4	40.14	33.83	40.29	31.87	37.33	32.15	43.15	36.93	37.75	31.90	41.22	34.26	42.52	37.92
Grok-4	39.73	31.55	42.41	28.60	37.19	29.31	43.73	36.15	35.68	26.82	37.56	35.63	41.92	35.82
Open-sourced LLMs														
Phi-4-14B	44.69	37.42	44.10	33.58	42.41	32.56	47.48	42.13	42.15	37.08	45.71	42.62	47.87	40.84
DeepSeek-V3	44.69	35.43	45.30	34.88	40.27	28.71	46.54	42.53	45.12	33.25	46.32	40.61	46.19	36.17
DeepSeek-R1	47.57	**45.32**	48.87	43.43	43.57	40.77	46.22	**50.64**	46.55	**47.01**	**52.90**	47.88	48.23	45.31
Qwen3-235B-A22B	47.18	40.40	47.78	38.08	46.52	36.57	49.50	42.33	45.35	37.12	48.20	42.72	45.94	48.67
Mistral-3.2-24B-Instruct	33.93	27.93	32.71	27.63	30.26	24.09	36.13	29.46	31.06	26.25	36.98	30.15	38.90	31.93
Kimi-K2-Preview	**49.00**	45.11	49.96	**47.57**	**49.34**	41.74	48.55	50.00	43.86	41.04	49.81	45.73	**52.11**	45.19
Qwen3-8B	46.56	38.85	47.97	37.02	45.67	33.74	45.79	46.75	44.65	33.41	47.52	42.03	47.46	43.98
Qwen3-32B	46.87	40.88	47.32	38.03	44.17	35.98	47.81	46.34	**46.69**	36.36	47.37	42.73	48.79	**49.99**
Llama-3.1-70B-Instruct	42.94	31.07	42.15	30.23	41.12	30.13	46.55	30.47	43.69	30.38	44.51	33.40	40.98	32.66

Table 1: Comparison of Baseline Emotional Level (BEL) across different models. For each case, the model with the highest BEL is marked in **bold**, and the second-highest is underlined, indicating their relative effectiveness in maintaining users' emotional baseline.

Based on this causal graph, we apply a theoretical intervention, $\text{do}(S_{t-1})$, to eliminate the bias from the confounder U. This intervention is not a practical operation but a mathematical tool. It simulates an idealized, unbiased scenario by theoretically severing all incoming paths to S_{t-1} (especially the path from U). Using the rules of do-calculus (Pearl, Glymour, and Jewell 2016), we transform this post-intervention distribution into a computational expression based on observable data (see the extended version for the derivation):

$$P(S_t|\text{do}(S_{t-1})) = \mathbb{E}_{S'_{t-1},Q',A'}\mathbb{E}_{Q,A}[P(S_t|I, S'_{t-1}, Q', A')]. \tag{4}$$

We adopt this post-intervention emotional distribution as the final, calibrated estimate of user emotion, replacing the predictions from conventional models. Based on this estimate, we construct an emotional state sequence $(s_0, s_1, \ldots, s_T)$. This ensures that the subsequently computed emotion transition matrix M and related metrics more faithfully reflect the model's capability to provide emotional support to the user.

Experiments and Analysis

Evaluation Settings and Model Selection

We sampled a total of 118 background environments for user-model interaction. For each model evaluated, we initiated dialogues based on the designed background and a first-turn exchange, which was then extended for an additional 40 turns.To mitigate the personal bias and increased time costs associated with human role-players, we adopted a methodology from existing multi-turn evaluation studies (Castillo-Bolado et al. 2024) and employed ChatGPT-4o (Hurst et al. 2024) to simulate the user in each scenario. For each environment, we extracted 0,1 or 3 logical disturbance events, which were gradually revealed to the user-simulating model during the dialogue. For the dynamic trajectory metrics, we utilized the Skywork-Llama-3.1-Reward-v2 (Liu et al. 2025) model to perform sentiment analysis on each dialogue turn, mapping the binary sentiment (positive/negative) to a interval $[0, 1]$. The models selected for testing included leading closed-source models from major vendors such as OpenAI, Anthropic, Google, xAI, ByteDance, and ZhipuAI, which were accessed via their APIs. For open-source models, we selected popular and widely-used models from the community, including Phi4-14B (Abdin et al. 2024), DeepSeek-V3 (Liu et al. 2024), DeepSeek-R1 (Guo et al. 2025), Qwen3-235B, Qwen3-8B, Qwen3-32B (Yang et al. 2025), Mistral3.2-24B-Instruct, Kimi-K2-Preview (Team et al. 2025), and Llama3.1-70B-Instruct (Dubey et al. 2024). Detailed estimation methods, evaluation specifics, parameter analysis, prompt settings, and the models' specifications and release dates are provided in the extended version, along with detailed case analyses.

Overall Result

Tab. 1 and Tab. 2 present the evaluation results for two key metrics, Baseline Emotional Level (BEL) and Emotional Trajectory Volatility (ETV), in both English and Chinese contexts. The BEL metric quantifies the user's average emotional level during multi-turn interactions, while the ETV metric measures the model's efficiency in improving a user's mood via emotional support strategies when the user is in a negative emotional state.

The BEL results reveal several key findings. First, there

LLMs	Overall		CogChg		SitMod		AttDep		ERFlex		SitSel		ResMod	
	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH
Closed-sourced LLMs														
O3-2025-04-16	20.17	14.89	19.91	**14.74**	18.23	12.45	28.09	20.39	15.26	11.54	18.55	13.87	22.65	17.95
Gemini-2.5-Pro	18.95	13.39	18.65	9.47	19.80	14.43	21.01	19.05	15.57	10.45	17.99	13.26	20.84	15.85
Claude-Opus-4	19.89	10.81	20.15	9.64	20.13	7.47	26.10	14.53	13.12	7.56	21.03	13.26	19.38	14.83
Doubao-Seed-1.6	21.55	15.31	19.92	11.68	21.97	13.81	25.38	19.25	18.46	14.74	**21.72**	15.69	22.94	19.48
Doubao-1.5-Pro-Character	17.72	14.48	15.62	11.66	16.13	12.53	23.45	21.40	16.46	15.02	17.71	12.08	19.11	16.88
ChatGPT-4o-Latest	21.99	15.40	**22.09**	13.26	20.75	13.32	**29.02**	20.74	19.23	13.55	20.62	15.37	21.30	18.56
GLM-4-Plus	21.00	12.56	20.55	12.63	18.22	9.88	25.29	17.72	20.03	11.63	19.92	9.77	23.68	15.13
ChatGLM-4	20.25	11.56	19.58	8.90	19.15	10.29	25.05	15.14	17.53	11.57	19.25	9.32	22.17	16.25
Grok-4	18.85	8.73	17.83	6.71	19.19	7.30	25.40	11.45	15.53	5.61	15.61	9.79	20.51	13.51
Open-sourced LLMs														
Phi-4-14B	21.37	13.21	21.27	9.94	19.66	11.24	25.65	19.12	18.99	13.89	20.12	13.60	23.64	14.33
DeepSeek-V3	20.32	12.02	19.52	10.77	18.82	7.57	25.04	17.32	20.39	10.77	18.98	14.22	20.49	14.32
DeepSeek-R1	20.38	15.30	18.27	13.20	20.68	14.26	23.76	21.70	19.77	15.37	20.03	13.25	20.97	15.92
Qwen3-235B-A22B	21.13	14.02	19.40	11.89	23.10	13.66	26.46	19.38	19.99	11.88	19.86	11.67	18.47	17.35
Mistral-3.2-24B-Instruct	15.93	7.01	13.41	6.64	14.88	4.19	23.75	9.58	13.19	6.43	16.09	7.54	16.63	9.19
Kimi-K2-Preview	**22.92**	**16.74**	21.59	14.36	**25.15**	**16.20**	26.12	**21.86**	20.37	**16.81**	19.56	**16.01**	**24.84**	17.00
Qwen3-8B	21.22	12.73	19.56	9.69	21.87	11.34	24.61	20.59	**20.50**	10.91	20.62	11.41	20.99	15.22
Qwen3-32B	21.81	14.91	19.99	11.97	21.98	14.20	28.14	21.77	19.96	11.42	19.25	12.99	22.95	**19.61**
Llama-3.1-70B-Instruct	19.09	9.75	16.16	8.52	19.03	10.94	24.13	11.44	18.99	8.89	19.71	9.40	18.39	9.64

Table 2: Emotional Trajectory Volatility (ETV) measures each model's ability to promote a rapid and stable transition toward positive emotional states. Higher values suggest stronger emotional support effectiveness. For each row, the highest ETV is marked in **bold**, and the second-highest is underlined.

is no significant difference in overall emotional support capabilities between open-source and closed-source models; in fact, advanced open-source models like Kimi-K2-Preview excelled at maintaining the highest average user emotional level in multi-turn dialogues. Second, models designed specifically for role-playing did not outperform general-purpose LLMs in maintaining a user's positive emotional state. Third, models demonstrated significantly superior long-term emotional support capabilities in English compared to Chinese, as most models helped users maintain a higher average emotional level in English dialogues. Finally, regarding specific strategy application, models show a deficiency in dynamically adjusting interaction strategies based on the user's state in English; conversely, in Chinese, their application of strategies that guide users to modify their external environment to improve mood is notably weak.

The ETV results are consistent with the trends observed for BEL, while also revealing deeper dynamic features. Initially, some models that scored high on the BEL metric experienced a decline in their ETV scores. This phenomenon can be attributed to two factors: 1) when a user's mood is low, the emotional support efficiency of these models is lower than that of their counterparts; and 2) when a user's mood is positive, these models struggle to effectively maintain its stability. Therefore, a relative decline in ETV score reflects instability in a model's emotional support capability. Notably, in both English and Chinese contexts, Attention Diversion (AttDep) emerged as the most rapid emotional support strategy for improving user mood, an observation that aligns with real-world patterns.

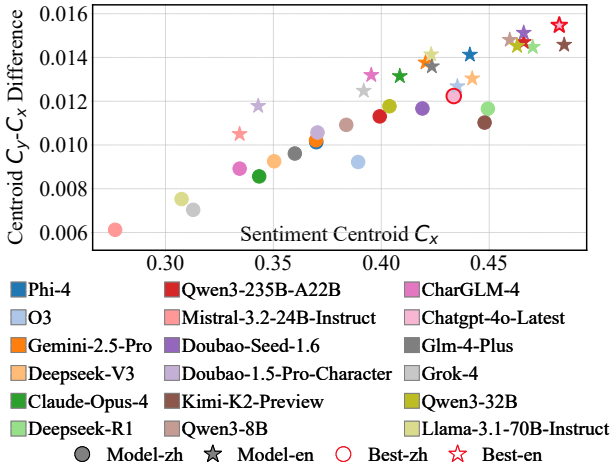

Figure 3: Visualization of the sentiment centroid

Emotional Centroid Position Visualization

We further analyze the emotional guidance capabilities of different LLMs by comparing their sentiment centroids, computed as the expected emotional position under the empirical transition model M. As shown in Fig. 3, the horizontal axis (C_x) represents the overall emotional positivity of the trajectory, while the vertical axis ($C_y - C_x$) captures the emotional concentration or consistency across turns.

The results reveal a clear separation between models: top-performing models, particularly those with high BEL and ETV scores, exhibit both high C_x and $C_y - C_x$ values, in-

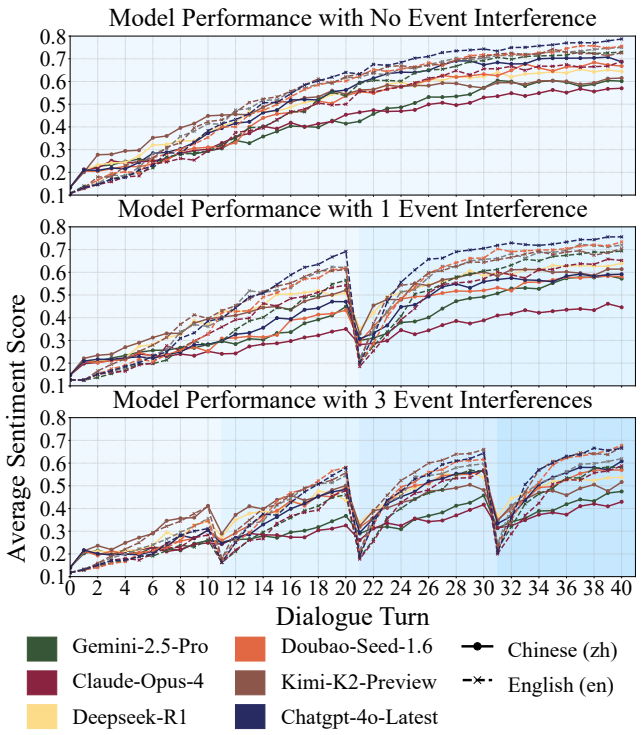

Figure 4: Sentiment dynamics across multi-turn dialogues under different emotional interference conditions.

dicating strong capabilities in steering users toward positive and stable emotional states. In contrast, models with lower centroid values either failed to maintain positive emotional progress or displayed higher volatility in user sentiment trajectories. Notably, several English-instruction-tuned models (e.g., ChatGPT-4o-Latest, kimi-K2-Preview) show superior centroid positioning compared to their Chinese counterparts, suggesting differences in emotional regulation strategies across language-specific pretraining and alignment.

Emotional Trajectory Visualization

To provide a more intuitive understanding of our proposed evaluation framework, we conduct a visual analysis under three levels of emotional disturbance (0, 1, and 3 disturbance events). For this analysis, we selected the multi-turn interaction results from 6 models. As shown in Fig. 4, the user's average emotional trajectory is tracked throughout all dialogues, where the intensity of the background shade indicates the cumulative number of disturbance events the user has experienced. The figure reveals that: (1) Models with higher ETV scores more effectively facilitate the user's emotional recovery from a low state, which corroborates our earlier assertion. (2) In the absence of disturbance events, the models can restore the user's emotion to a neutral level within a relatively short period. (3) Multiple disturbance events impede the rate of emotional recovery; however, models with stronger emotional support capabilities demonstrate greater resilience to such interferences.

Model	ZH		EN	
	Acc(%)	Δ	Acc(%)	Δ
Chatgpt-4o	**76.57**	-	**90.91**	-
Deepseek-V3	73.20	-	89.47	-
RRM-7B	72.96	-	87.44	-
Llama3.1-8B-IT.	74.76	-	88.88	-
Skywork-Reward-Llama-3.1-B-v0.2				
w/o CA	73.82	*ref*	89.11	*ref*
w/ CA	74.61	0.79↑	89.59	0.48↑
Skywork-Reward-V2-Llama-3.1-8B				
w/o CA	72.18	*ref*	89.59	*ref*
w/ CA	<u>74.92</u>	2.74↑	<u>89.83</u>	0.24↑

Table 3: Performance comparison on the human-annotated dataset, with ablation results showing the improvement (Δ) from our causally-adjusted (CA) estimation method. For each case, the model with the highest Acc is marked in **bold**, and the second-highest is <u>underlined</u>.

Causal-Enhanced Estimation of Emotional Dynamics

To evaluate the consistency of our sentiment recognition model with human perception and validate our estimation calibration method, we constructed a human-annotated multi-turn dialogue dataset. This dataset comprises nearly 2,000 Chinese and English multi-turn dialogues selected from the Daily Dialog and CPED corpora. Three expert annotators were employed to re-label the sentiment of each turn in these dialogues with binary annotations, building upon the existing labels. Details can be found in the extended version. As shown in Tab. 3, our findings fall into two aspects. Firstly, comparative experiments demonstrate that our proposed estimation calibration method effectively enhances the model's sentiment recognition capabilities by mitigating the influence of confounding factors. The application of our unbiased estimation method led to improved sentiment recognition performance across different models. Notably, our approach achieves state-of-the-art performance compared to other existing models. Secondly, our evaluation model, combined with the calibration method, shows a high degree of consistency with human judgments, reaching an accuracy of 75% on Chinese dialogues and 90% on English dialogues.

Conclusions

In this paper, we propose a dynamic trajectory analysis framework to evaluate the emotional support capabilities of language models. The core of this framework simulates the real user-model interaction process, guiding the interaction by constructing background contexts, introducing multi-strategy constraints, and incorporating event-driven perturbations. We have designed metrics for dynamic trajectory analysis from three perspectives and utilize causal inference to calibrate the evaluation outcomes. Experiments confirm that our methodology offers a more comprehensive and multifaceted assessment of a model's emotional support capabilities, showing high consistency with human evaluations.

References

Abdin, M.; Aneja, J.; Behl, H.; Bubeck, S.; Eldan, R.; Gunasekar, S.; Harrison, M.; Hewett, R. J.; Javaheripi, M.; Kauffmann, P.; et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Afzoon, S.; Naseem, U.; Beheshti, A.; and Jamali, Z. 2024. Persobench: Benchmarking personalized response generation in large language models. *arXiv preprint arXiv:2410.03198*.

Aldao, A.; Nolen-Hoeksema, S.; and Schweizer, S. 2010. Emotion-regulation strategies across psychopathology: A meta-analytic review. *Clinical psychology review*, 30(2): 217–237.

Bonanno, G. A.; and Burton, C. L. 2013. Regulatory flexibility: An individual differences perspective on coping and emotion regulation. *Perspectives on psychological science*, 8(6): 591–612.

Castillo-Bolado, D.; Davidson, J.; Gray, F.; and Rosa, M. 2024. Beyond prompts: Dynamic conversational benchmarking of large language models. *Advances in Neural Information Processing Systems*, 37: 42528–42565.

Chen, H.; Chen, H.; Yan, M.; Xu, W.; Gao, X.; Shen, W.; Quan, X.; Li, C.; Zhang, J.; Huang, F.; et al. 2024a. Socialbench: Sociality evaluation of role-playing conversational agents. *arXiv preprint arXiv:2403.13679*.

Chen, Y.; Wang, H.; Yan, S.; Liu, S.; Li, Y.; Zhao, Y.; and Xiao, Y. 2024b. Emotionqueen: A benchmark for evaluating empathy of large language models. *arXiv preprint arXiv:2409.13359*.

Chen, Y.; Xing, X.; Lin, J.; Zheng, H.; Wang, Z.; Liu, Q.; and Xu, X. 2023. SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. *arXiv preprint arXiv:2311.00273*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Feng, Q.; Xie, Q.; Wang, X.; Li, Q.; Zhang, Y.; Feng, R.; Zhang, T.; and Gao, S. 2025. EmoCharacter: Evaluating the Emotional Fidelity of Role-Playing Agents in Dialogues. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6218–6240.

Gross, J. J. 2015. Emotion regulation: Current status and future prospects. *Psychological inquiry*, 26(1): 1–26.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hocker, J. L.; and Wilmot, W. W. 2018. *Interpersonal conflict*. McGraw-Hill Education New York, NY.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Kang, D.; Kim, S.; Kwon, T.; Moon, S.; Cho, H.; Yu, Y.; Lee, D.; and Yeo, J. 2024. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. *arXiv preprint arXiv:2402.13211*.

Kassin, S.; Fein, S.; and Markus, H. R. 2024. *Social psychology*. SAGE Publications.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, C. Y.; Zeng, L.; Xiao, Y.; He, J.; Liu, J.; Wang, C.; Yan, R.; Shen, W.; Zhang, F.; Xu, J.; et al. 2025. Skywork-Reward-V2: Scaling Preference Data Curation via Human-AI Synergy. *arXiv preprint arXiv:2507.01352*.

Liu, S.; Zheng, C.; Demasi, O.; Sabour, S.; Li, Y.; Yu, Z.; Jiang, Y.; and Huang, M. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Llanes-Jurado, J.; Gómez-Zaragozá, L.; Minissi, M. E.; Alcañiz, M.; and Marín-Morales, J. 2024. Developing conversational virtual humans for social emotion elicitation based on large language models. *Expert Systems with Applications*, 246: 123261.

Madani, N.; Saha, S.; and Srihari, R. 2024. Steering conversational large language models for long emotional support conversations. *arXiv preprint arXiv:2402.10453*.

Paech, S. J. 2023. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*.

Pearl, J.; Glymour, M.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Sabour, S.; Liu, S.; Zhang, Z.; Liu, J. M.; Zhou, J.; Sunaryo, A. S.; Li, J.; Lee, T.; Mihalcea, R.; and Huang, M. 2024. Emobench: Evaluating the emotional intelligence of large language models. *arXiv preprint arXiv:2402.12071*.

Shanahan, M.; McDonell, K.; and Reynolds, L. 2023. Role play with large language models. *Nature*, 623(7987): 493–498.

Smith, E. M.; Hsu, O.; Qian, R.; Roller, S.; Boureau, Y.-L.; and Weston, J. 2022. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. *arXiv preprint arXiv:2201.04723*.

Sorin, V.; Brin, D.; Barash, Y.; Konen, E.; Charney, A.; Nadkarni, G.; and Klang, E. 2024. Large language models and empathy: systematic review. *Journal of medical Internet research*, 26: e52597.

Tamoyan, H.; Schuff, H.; and Gurevych, I. 2024. Llm roleplay: Simulating human-chatbot interaction. *arXiv preprint arXiv:2407.03974*.

Team, K.; Bai, Y.; Bao, Y.; Chen, G.; Chen, J.; Chen, N.; Chen, R.; Chen, Y.; Chen, Y.; Chen, Y.; et al. 2025.

Kimi K2: Open Agentic Intelligence. *arXiv preprint arXiv:2507.20534*.

Tu, Q.; Fan, S.; Tian, Z.; and Yan, R. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. *arXiv preprint arXiv:2401.01275*.

Wang, P.; Li, L.; Chen, L.; Cai, Z.; Zhu, D.; Lin, B.; Cao, Y.; Liu, Q.; Liu, T.; and Sui, Z. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Welivita, A.; and Pu, P. 2024. Are Large Language Models More Empathetic than Humans? *arXiv preprint arXiv:2406.05063*.

Wu, B.; Sun, K.; Bai, Z.; Li, Y.; and Wang, B. 2025. RAIDEN benchmark: Evaluating role-playing conversational agents with measurement-driven custom dialogues. In *Proceedings of the 31st International Conference on Computational Linguistics*, 11086–11106.

Xu, Z.; and Jiang, J. 2024. Multi-dimensional evaluation of empathetic dialog responses. *arXiv preprint arXiv:2402.11409*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yuan, D.; Chen, Y.; Liu, G.; Li, C.; Tang, C.; Zhang, D.; Wang, Z.; Wang, X.; and Liu, S. 2025. DMT-RoleBench: A Dynamic Multi-Turn Dialogue Based Benchmark for Role-Playing Evaluation of Large Language Model and Agent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25760–25768.

Zhang, H.; Chen, Y.; Wang, M.; and Feng, S. 2024. Feel: A framework for evaluating emotional support capability with large language models. In *International Conference on Intelligent Computing*, 96–107. Springer.

Zhao, H.; Li, L.; Chen, S.; Kong, S.; Wang, J.; Huang, K.; Gu, T.; Wang, Y.; Jian, W.; Liang, D.; et al. 2024. ESC-Eval: Evaluating emotion support conversations in large language models. *arXiv preprint arXiv:2406.14952*.

Zhong, W.; Guo, L.; Gao, Q.; Ye, H.; and Wang, Y. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19724–19731.

Zhou, J.; Huang, Y.; Wen, B.; Bi, G.; Chen, Y.; Ke, P.; Chen, Z.; Xiao, X.; Peng, L.; Tang, K.; et al. 2025. CharacterBench: Benchmarking Character Customization of Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 26101–26110.