MindChat-R0: A Large Language Model for Emotionally Supportive Dialogue through Reinforcement Learning

Dong She
ftdshe@mail.scut.edu.cn
South China University of Technology
Guangzhou, Guangdong, China

Chenxu Zhang
ftcxzhang@mail.scut.edu.cn
South China University of Technology
Guangzhou, Guangdong, China

Xianrong Yao
ftxryao@mail.scut.edu.cn
South China University of Technology
Guangzhou, Guangdong, China

Yang Gao
gaoyang@cs.ecnu.edu.cn
South China University of Technology
Guangzhou, Guangdong, China

Zhanpeng Jin*
zjin@scut.edu.cn
South China University of Technology
Guangzhou, Guangdong, China

Abstract

Emotional Support Conversation (ESC) systems are critical for assisting individuals facing mental health challenges. In this work, we present a reinforcement learning framework to improve ESC systems through structured emotional reasoning. We first collect and clean a dataset of 4,500 real-world support-seeking posts. To guide emotional generation, we introduce Empathetic Chain-of-Thought(ECoT), a structured reasoning format that encourages multi-turn empathy and coherence. Based on this, we train MindChat-R0 (Qwen3-8B as basic model), a Chinese empathetic dialogue agent, using reinforcement learning optimized by ECoT-driven reward signals. LLM-as-a-judge evaluation shows that MindChat achieves the highest average score of 3.863 out of 5.0 across fluency, empathy, and support dimensions (vs. 2.834 for Qwen3-8B-nothink and 2.547 for Qwen3-8B-think). In human preference evaluation, MindChat-R0 also outperforms strong baselines with a win rate of 71.14%, based on pairwise comparisons by human annotators.

CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**.

Keywords

Mental Health, Large Language Model, Emotional Support Conversation, Empathetic Dialogue Systems, Chain-of-Thought Reasoning, Reinforcement Learning

ACM Reference Format:

Dong She, Chenxu Zhang, Xianrong Yao, Yang Gao, and Zhanpeng Jin. 2025. MindChat-R0: A Large Language Model for Emotionally Supportive Dialogue through Reinforcement Learning. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion '25), October 12–16, 2025, Espoo, Finland.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3714394.3756244

*Corresponding author. Email: zjin@scut.edu.cn

UbiComp Companion '25, Espoo, Finland

ACM ISBN 979-8-4007-1477-1/25/10
https://doi.org/10.1145/3714394.3756244

1 Introduction

Mental health has reached critical importance worldwide, with depression recognized as one of the leading causes of disability and emotional distress, affecting approximately 5% of the global adult population [25]. In China, the situation is even more concerning. Over the past three decades, the number of individuals living with depressive disorders increased by 54% [22], and anxiety disorders by 31% [22]. The World Health Organization estimates that among China's 1.4 billion population, approximately 54 million suffer from depression, and 41 million from anxiety disorders [26]. These data underscore the enormous scale of the mental health burden, highlighting an urgent need for scalable, accessible emotional assistance.

One promising intervention is Emotional Support Conversation (ESC)—a dialogue-based approach aimed at providing empathy, validation, and psychological comfort [1]. ESC has proven effective in therapeutic contexts[15] such as counseling, peer support, and online mental health platforms. ESC lies at the intersection of natural language processing (NLP) and affective computing, requiring both linguistic competence and emotional intelligence. With the emergence of large language models (LLMs), researchers have begun applying LLMs to ESC tasks. For example, AugESC [36] uses LLMs to augment ESC data, and ESC-Eval [32] benchmarks LLM-generated responses via role-play evaluation. Other frameworks, such as ExTES [24] and ESCoT [31], guide LLMs using situation-aware strategies or pre-defined support intentions. Despite these advances, current LLM-based ESC systems still face key challenges: they often rely on shallow seeker representations and lack the ability to perform long-horizon emotional reasoning. These limitations highlight the need for more adaptive and cognitively grounded training paradigms.

In parallel, the capabilities of LLMs have been significantly enhanced by the integration of reinforcement learning (RL). OpenAI's o1 models introduced an "exploration–reflection–iteration" loop to extend chain-of-thought(CoT) reasoning in complex tasks [12]. Models such as Marco-o1 [33] and QwQ [21] have further extended this paradigm with structured CoT optimization. More recently,

DeepSeek-R1 [2] departed from supervised fine-tuning entirely, instead relying on thousands of steps of pure RL using Group Relative Policy Optimization (GRPO) [16]. These models exhibit emergent reasoning abilities, suggesting that RL can imbue LLMs with long-term strategy learning and adaptive behavioral alignment. Motivated by these findings, we explore how reinforcement learning can be leveraged for ESC. We introduce MindChat-R0, a reinforcement learning–enhanced large language model for Chinese emotional support conversations, designed to generate responses through an empathetic chain-of-thought (ECoT) framework.

Our ECoT (Empathetic Chain-of-Thought) is a cognitive framework designed to overcome the limitations of shallow emotional reasoning in existing ESC systems by guiding MindChat-R0 to learn and generate responses through structured empathetic reasoning. Specifically, ECoT decomposes empathetic support into three reasoning stages: *situation* (inferring the user's emotional state and context), *motivation* (understanding the user's underlying needs and intent), and *strategy* (formulating a tailored supportive response). By explicitly modeling this three-step cognitive process, ECoT enables MindChat-R0 to generate emotionally resonant and psychologically grounded support, moving beyond surface-level pattern imitation. This design is supported by both psychological theory and recent advances in empathetic dialogue systems, which highlight the value of structured reasoning in generating deeper and more coherent support [8, 14, 27, 37].

Incorporating pure RL into the task of ESC presents two major challenges. First, the model requires a high-quality dataset grounded in authentic Chinese psychological help-seeking scenarios, which are largely absent in existing corpora. Such data must reflect the complexity, emotional nuance, and cultural specificity of real user experiences in mental health contexts. Second, due to the open-ended and emotionally sensitive nature of empathetic dialogue, it remains non-trivial to design an effective and reliable reward function that can accurately guide the learning process in alignment with human-centric communication goals.

To address the first challenge—the lack of domain-specific and realistic emotional support dialogue data—we construct a high-quality Chinese dataset rooted in real-world psychological help-seeking scenarios. Specifically, we curate approximately 4,500 posts from a widely-used online mental health counseling platform. These posts cover a wide range of emotionally charged situations, including academic pressure, interpersonal tensions, romantic difficulties, and clinical mental health concerns. Each entry contains rich psychological cues embedded in natural expressions, making them suitable for modeling nuanced empathic responses. Further details on the dataset construction are provided in Section 2.

The second challenge lies in designing an effective reward signal to supervise ECoT-style response generation during reinforcement learning. In empathetic dialogue, there are no ground-truth answers, and conventional NLP metrics (e.g., BLEU, ROUGE) fall short of evaluating emotional appropriateness, depth, or supportive intent. To overcome this, we adopt the LLM-as-a-judge paradigm [35]. Prior work has explored the use of large language models in various evaluator roles—such as graders [3, 23], assessors [9, 30], critics [5, 13], verifiers [10, 18], and reward/ranking models [11, 20, 29]—demonstrating its effectiveness across summarization, reasoning, and dialogue generation tasks. Section 3 provides a comprehensive description of the reward modeling implementation.

In summary, our contributions are as follows:

- We construct a new dataset of real-world mental health queries by crawling and cleaning online support-seeking posts.
- We propose ECoT, a tree-structured representation to capture emotional coherence and supportive response structure.
- We develop and train MindChat-R0, an RL-based Chinese empathic dialogue model guided by ECoT rewards.

2 Dataset

To facilitate the development of empathetic reasoning in Chinese large language models, we construct a domain-specific dataset grounded in real-world mental health help-seeking scenarios. While existing corpora such as PsyQA [19] provide initial efforts toward modeling psychological dialogues, they are limited in both temporal coverage and linguistic relevance to contemporary Chinese users. In contrast, our objective is to build a high-quality, up-to-date Chinese dataset that reflects current psychological concerns, sociocultural contexts, and natural language expressions observed in real-world help-seeking behavior.

Specifically, we collect approximately 4,500 help-seeking entries from Xinli001[1], one of China's largest online mental health counseling platforms. The data span a recent three-month period (March–May 2025) to ensure temporal relevance. Each entry consists of a self-initiated question or narrative describing the user's psychological distress, encompassing diverse themes such as academic pressure, emotional dysregulation, interpersonal relationships, family conflict, romantic struggles, and mental health symptoms. These posts are characterized by rich emotional expressions and implicitly embedded psychological cues, making them highly suitable for downstream empathetic reasoning and dialogue modeling tasks.

We applied a standardized data cleaning pipeline to ensure linguistic quality and task relevance. Non-user-generated content, duplicates, and extremely short entries (fewer than 20 characters) were removed. Noise such as emojis, URLs, platform artifacts, and non-Chinese characters was filtered using rule-based heuristics. To preserve linguistic coherence and psychological validity, only Mandarin posts containing substantial emotional context were retained. The resulting corpus comprises 4,500 high-quality, user-authored help-seeking queries suitable for downstream empathetic reasoning tasks.

To gain a comprehensive understanding of the distribution and characteristics of the help-seeking data, we conducted a descriptive analysis of the collected corpus. Table 1 presents the distribution of primary psychological categories, along with the corresponding case counts, relative proportions, and average description lengths. The most represented categories include Behavior (13.48%), Interpersonal (11.85%), Family (11.20%), and Emotion (11.15%), reflecting the diversity and complexity of user concerns in real-world mental health discourse. We further analyzed fine-grained subtopics in the help-seeking posts. Commonly occurring themes include relationship, family, emotion, marriage, interpersonal conflict, personal

[1] https://www.xinli001.com/qa?source=pc-home

growth, behavioral issues, and stress management. These topics reflect a wide spectrum of psychological concerns, from emotional regulation and romantic struggles to family dynamics and identity challenges.

Table 1: Distribution of primary categories in the dataset.

Primary Category	Count	Percentage (%)	Mean Length
Interpersonal	474	11.85	288.44
Other	167	4.18	268.31
Marriage	381	9.53	357.37
Family	448	11.20	384.48
Romantic Relationship	399	9.98	326.90
Emotion	446	11.15	280.74
Growth	448	11.20	271.60
Therapy	361	9.03	296.57
Career	330	8.25	286.52
Behavior	539	13.48	226.81

3 MindChat-R0

MindChat-R0 is initialized from the *Qwen3-8B-Base* [28] model, a state-of-the-art open-source language model well-suited for Chinese language understanding and generation. We train MindChat-R0 purely through reinforcement learning to enhance its ability to generate emotionally supportive responses guided by the Empathetic Chain-of-Thought(ECoT) framework. In this section, we present the ECoT framework, followed by the reinforcement learning setup and reward modeling strategy used to optimize MindChat-R0 for emotional support dialogue.

3.1 ECoT Design

To enable emotionally supportive dialogue grounded in human-centered reasoning, we introduce a structured cognitive framework, ECoT, drawing inspiration from psychological theories of empathy and support, as well as recent advances in emotion-aware reasoning. Prior studies have explored multi-stage emotional reasoning in dialogue systems [14, 27, 37] and cognitive modeling for support generation [8]. Building upon these insights, ECoT decomposes empathetic response generation into three levels of reasoning: *situation*, *motivation*, and *strategy*. Each stage reflects a distinct layer of emotional cognition, guiding the model to understand, interpret, and respond with psychologically grounded support. Unlike surface-level generation, ECoT encourages interpretable and adaptive reasoning, aiming to improve depth, relevance, and emotional resonance in open-domain support conversations.

- ***Situation***: The model first identifies and summarizes the user's current situation. Rather than extracting surface-level cues, it interprets the user's emotional and social context—such as stressors, dilemmas, or internal conflicts—through a reflective lens. This mirrors how human therapists often begin with empathetic attunement to the client's narrative [14].
- ***Motivation***: Based on the situation, the model then infers possible underlying emotional needs or motivational states. Inspired by cognitive theories of affect and motivational models in psychotherapy, this step enables the agent to reason about why the user may be feeling a certain way (e.g., desire for control, fear of rejection, need for autonomy), instead of relying solely on predefined emotion categories [37, 37].

Question Title

40多岁，丧偶2年，孩子不同意我谈恋爱，怎么办?
I am in my 40s, widowed for 2 years, and my children don't agree with me having a relationship. What should I do?

Question Description

我中年40多岁丧偶2年多了，有男士主动追求我也没有拒绝一直微信聊天视频.....有一定的经济来源。 要怎么跟女儿沟通她才能同意？不知怎么办?
I am in my 40s and have been widowed for more than two years. A man actively pursued me and I did not refuse. We have been chatting on WeChat and video... I have a certain source of income. How can I communicate with my daughter so that she will agree? I don't know what to do?

Empathetic Chain-of-Thought (ECoT)

<thinking>
<situation>您是一位40多岁的中年女性.....您既希望获得情感上的支持，也希望能够得到女儿的理解和接纳。</situation>
<motivation>您的动机是希望与女儿达成共识.....您希望找到一种既能表达自己情感需求，又能维护家庭关系的方式。</motivation>
<strategy>在这种情况下，建议您采取一种温和、理性的沟通方式.....也可以考虑与女儿一起探讨未来的家庭结构，让她参与决策，增强她的参与感和认同感。</strategy>
</thinking>

<thinking>
<situation>You are a middle-aged woman in your 40s... You want emotional support and your daughter's understanding and acceptance. </situation>
<motivation>Your motivation is to reach a consensus with your daughter... to find a way to express your emotional needs while maintaining family relationships. </motivation>
<strategy>In this case, it is recommended that you adopt a gentle and rational communication method... letting her participate in decision-making, and enhancing her sense of participation and identity. </strategy>
</thinking>

Figure 1: An example of Empathetic Chain-of-Thought (ECoT) reasoning with bilingual annotation, generated by our MindChat-R0.

- ***Strategy***: Finally, the model selects an appropriate support strategy, such as emotional validation, perspective reframing, or gentle encouragement. These strategies are aligned with human support typologies discussed in both psychological counseling and empathic LLM studies [27]. By explicitly modeling this stage, the response moves beyond surface empathy toward actionable support.

This structured decomposition enables the model to provide responses that are emotionally attuned and pragmatically supportive. As illustrated in Figure 1, the user's situation—a middle-aged widow facing intergenerational conflict over romantic relationships—is clearly identified. Her motivation is understood as the desire for

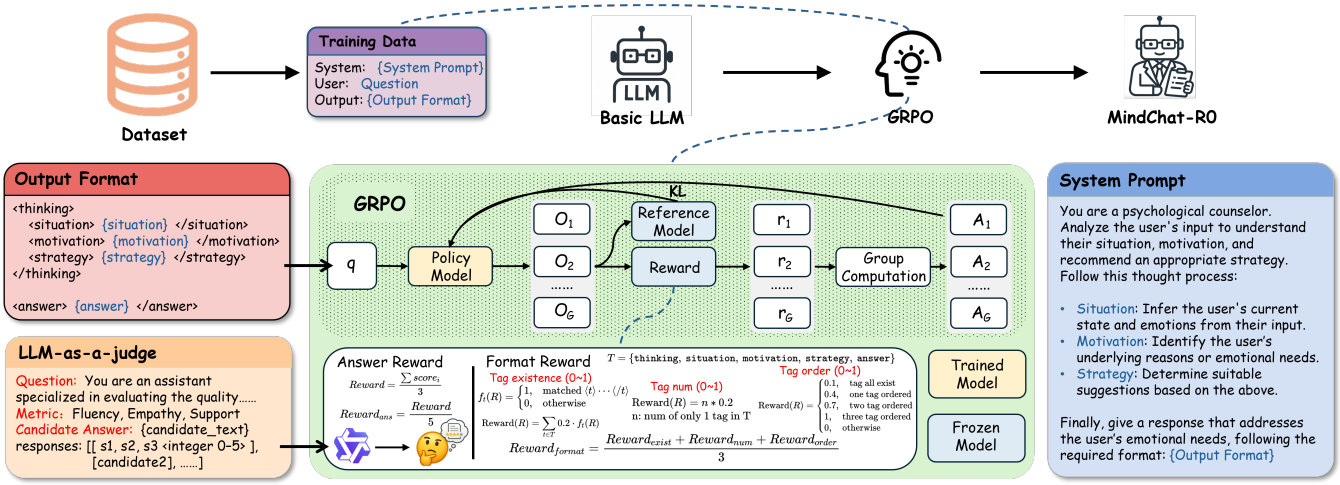

Figure 2: An illustration of the Empathetic Chain-of-Thought (ECoT) reasoning pipeline implemented in our MindChat-R0 framework. Based on our curated bilingual annotated dataset, a basic LLM(Qwen3-8b-Base[28]) is optimized using Group Relative Policy Optimization (GRPO). During training, the model generates G output candidates per query, which are scored by both format and answer rewards. The group-relative advantage A_i is then computed for each output based on its deviation from the batch mean, as defined in Eq. (1), and used to guide policy updates. The system prompt encourages the model to reason through the user's situation, motivation, and strategy before generating an empathetic response in a structured format.

both emotional fulfillment and family harmony. Based on this understanding, the model suggests a gentle and participatory communication strategy to engage her daughter in decision-making. This example showcases how the ECoT framework grounds emotional understanding in real-world context and supports the generation of actionable, empathetic suggestions.

3.2 Reinforcement Learning with ECoT

We build upon the Qwen3-8B-Base [28] model as our foundational language model and apply reinforcement learning with Group Relative Policy Optimization (GRPO) [16] to enhance its capacity for empathetic reasoning. As illustrated in Figure 2, our framework generates multiple ECoT outputs per query and optimizes the model based on their relative performance within the group. Each output is evaluated using reward functions that reflect both response format and content quality, and group-relative advantages are computed to guide policy updates (see Eq. (1)). The training process encourages the model to reason through the user's situation, motivation, and strategy before delivering the final empathetic response. Details of the GRPO method and reward design are presented in Sections 3.2.1 and 3.2.2, respectively.

3.2.1 Group Relative Policy Optimization. To promote the generation of emotionally aligned reasoning chains, we employ *Group Relative Policy Optimization* (GRPO). Unlike methods that evaluate individual rewards independently, GRPO considers the relative quality of sampled outputs within a batch.

At each training step, the current policy π_{old} generates G candidate ECoT responses $\{o_i\}_{i=1}^{G}$, each assigned a reward r_i that reflects both format correctness and answer quality (see details in Sec. 3.2.2).

The group-relative advantage A_i is computed as:

$$A_i = \frac{r_i - \mu_{\{r_1,r_2,\cdots,r_G\}}}{\sigma_{\{r_1,r_2,\cdots,r_G\}}} \tag{1}$$

where $\mu_{\{r_1,r_2,\cdots,r_G\}}$ and $\sigma_{\{r_1,r_2,\cdots,r_G\}}$ denote the mean and standard deviation of rewards within the group. This normalization emphasizes outputs that are not only high-quality in absolute terms but also superior relative to other candidates in the same batch.

Outputs with above-average rewards receive higher advantage values, encouraging the policy to optimize toward relatively better responses. The policy is then updated by maximizing the following objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\Bigg(\min\left(r_i^{ratio}A_i,\ \text{clip}\left(r_i^{ratio},\ 1-\epsilon,\ 1+\epsilon\right)A_i\right)\Bigg) - \beta\mathbb{D}_{\text{KL}}\left(\pi_\theta \,\|\, \pi_{\text{ref}}\right)\Bigg] \tag{2}$$

Where $r_i^{\text{ratio}} = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}$ is the probability ratio between the new and old policies, encouraging higher-quality outputs when A_i is large. The clipping operator $\text{clip}(r_i^{\text{ratio}}, 1-\epsilon, 1+\epsilon)$ stabilizes training by preventing excessively large policy updates. The KL divergence term $\mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})$ constrains the new policy from deviating too far from the reference policy. The coefficient β is a hyperparameter controlling the strength of the KL penalty.

3.2.2 Reward Design. To guide the learning of emotionally coherent and well-structured outputs, we design a composite reward

function consisting of two components: a format reward and an answer reward. The total reward is computed as:

$$r_{\text{total}} = r_{\text{format}} + r_{\text{answer}} \tag{3}$$

where both r_{format} and r_{answer} are normalized to the $[0, 1]$ range.

Format Reward. To encourage outputs that strictly follow the ECoT structure, we design a structured rule-based *Format Reward* composed of three components: tag existence, tag count, and tag order. The tag set is defined as:

$$T = \{\text{thinking}, \text{situation}, \text{motivation}, \text{strategy}, \text{answer}\}$$

The final format reward is calculated as:

$$\mathrm{r}_{\text{format}} = \frac{\mathrm{r}_{\text{exist}} + \mathrm{r}_{\text{num}} + \mathrm{r}_{\text{order}}}{3} \tag{4}$$

where $\mathrm{r}_{\text{format}} \in [0, 1]$ indicates the normalized format quality. Each component also takes values in $[0, 1]$, thereby ensuring the overall score remains within a bounded range.

(1) Tag Existence Reward. This component evaluates whether each tag $t \in T$ appears correctly in the form of both an opening tag $\langle t \rangle$ and a matching closing tag $\langle /t \rangle$. This ensures the output follows XML-like structural conventions.

$$f_t(R) = \begin{cases} 1, & \text{if } \langle t \rangle \cdots \langle /t \rangle \text{ exists} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$\mathrm{r}_{\text{exist}} = \sum_{t \in T} 0.2 \cdot f_t(R) \tag{6}$$

This term rewards the model up to 1.0 if all five tags are properly enclosed. For example, if only three tags are correctly enclosed, $\mathrm{r}_{\text{exist}} = 0.6$.

(2) Tag Count Reward. To discourage repetition or omission, we reward the number of unique tags from T that occur *exactly once*. Let n denote this number:

$$\mathrm{r}_{\text{num}} = n \cdot 0.2 \tag{7}$$

This encourages the model to include all five tags without duplication. For instance, if four tags occur exactly once, the reward is $\mathrm{r}_{\text{num}} = 0.8$.

(3) Tag Order Reward. This component encourages correct sequential ordering among key reasoning tags (e.g., `thinking` → `situation` → `motivation` → `answer`). Rewards are assigned based on how many tag pairs appear in the correct order:

$$\mathrm{r}_{\text{order}} = \begin{cases} 1.0, & \text{if all three key tags are correctly ordered} \\ 0.7, & \text{if two tags are correctly ordered} \\ 0.4, & \text{if only one tag is ordered} \\ 0.1, & \text{if all tags exist but no order is correct} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

This term prioritizes not just tag presence but logical structure. For example, if `thinking`, `situation`, `motivation`, and `answer` appear in sequence, the model gets the full score of 1.0.

Answer Reward. Evaluating the quality of empathetic responses poses a significant challenge, as traditional NLP metrics such as BLEU, ROUGE, or METEOR fail to capture the nuanced human-centric goals of Empathetic Structured Communication (ESC). These metrics focus on lexical overlap or surface similarity and cannot adequately assess whether a response is emotionally supportive, linguistically smooth, or empathetically attuned to the user's feelings.

To address this, we adopt an LLM-as-a-judge approach [35], leveraging a closed-source large language model (Qwen-Plus API) to assign quality scores to model-generated responses. Specifically, following prior work that adopts Likert scale-based evaluation [4], we treat this as a modeling reward problem and assess responses along three key dimensions, each rated on a 5-point scale. The final reward is computed as the average of the three dimension scores and normalized to the $[0, 1]$ range by dividing by 5:

$$r_{\text{answer}} = \frac{1}{3} \sum_{i=1}^{3} s_i \div 5 = \frac{1}{15} \sum_{i=1}^{3} s_i, \tag{9}$$

where s_i denotes the score for the i-th dimension.

- **Fluency and Comfort:** This dimension evaluates whether the response uses natural, smooth, and human-like language. Responses that are rigid, overly templated, or exhibit list-like enumeration are penalized.
- **Empathic Understanding:** This measures how well the response identifies and acknowledges the user's emotional state, reflecting emotional sensitivity and understanding. Responses that feel preachy, detached, or overly factual without emotional resonance receive lower scores.
- **Support and Encouragement:** This assesses whether the response conveys supportive intent and emotional reassurance. Responses that express doubt, challenge the user's perspective, or include rhetorical questioning are penalized for failing to foster a caring tone.

4 Experiment

4.1 Experimental Setup

We conduct all reinforcement learning experiments using the Verl framework [17] on a single node with 4×NVIDIA A800 GPUs (80GB each). Gradient checkpointing is enabled to reduce memory usage, and mixed precision (FP16) training is used throughout. All experiments are conducted using Python 3 with PyTorch and HuggingFace Transformers as underlying toolkits.

We adopt Qwen3-8B-Base [28] as the backbone language model. To enable structured empathetic reasoning learning, we explicitly disable its built-in thinking module during both training and rollout. The model is fine-tuned using full-parameter reinforcement learning on 4,000 real-world help-seeking samples from our self-constructed dataset (see Section 2). An additional 500 samples are used for testing.

Training is conducted for 1000 steps using the GRPO algorithm with a custom composite reward (see Section 3.2.2) that combines structural format and content relevance scores. The learning rate for the actor is set to 1e−6, with micro-batches of size 4 per GPU and mini-batches of size 16. Each input prompt generates $G = 4$ candidate responses for reward comparison. The maximum input

prompt length is 1,024 tokens, and the maximum response length is 1,500 tokens. Rollouts use a temperature of 1.0.

4.2 Evaluation and Results

LLM-as-a-Judge Evaluation. To evaluate the quality of generated responses, we adopt the LLM-as-a-judge framework [35], using the same dimensions as our reward function: Fluency and Comfort, Empathic Understanding, and Support and Encouragement. We report results from two closed-source LLM judges—Qwen-Plus and GPT-4o—each scoring on a 5-point Likert scale with five samples per response. Final scores are averaged.

As shown in Table 2, our model **MindChat** consistently achieves the highest scores across all dimensions under both judges, demonstrating robust and generalizable performance. Despite differences in absolute score levels, the relative ranking remains stable, with MindChat outperforming both baselines—e.g., achieving an average score of **3.863** under Qwen-Plus and **3.588** under GPT-4o.

Table 2: LLM-as-a-Judge evaluation scores (average ± variance) under Qwen-Plus and GPT-4o.

Judge	Model	Fluency	Empathy	Support	Avg.
Qwen-Plus	Qwen3-8B-nothink	1.935 ± 0.00025	3.366 ± 0.00027	3.199 ± 0.00057	2.834 ± 0.00026
Qwen-Plus	Qwen3-8B-think	1.764 ± 0.00027	3.030 ± 0.00029	2.842 ± 0.00058	2.547 ± 0.00022
Qwen-Plus	MindChat (Ours)	**3.516 ± 0.00021**	**4.037 ± 0.00059**	**4.032 ± 0.00086**	**3.863 ± 0.00005**
GPT-4o	Qwen3-8B-nothink	2.433 ± 0.00017	2.537 ± 0.00003	2.426 ± 0.00020	2.465 ± 0.00005
GPT-4o	Qwen3-8B-think	2.414 ± 0.00010	2.401 ± 0.00022	2.263 ± 0.000001	2.359 ± 0.000002
GPT-4o	MindChat (Ours)	**3.468 ± 0.00020**	**3.650 ± 0.00040**	**3.644 ± 0.00006**	**3.588 ± 0.00002**

Human Preference Evaluation. To further assess response quality from a human perspective, we conduct a human evaluation through pairwise preference ranking. Unlike LLM-based evaluations that follow explicit dimensions such as fluency or empathy, human judgments are often more intuitive and holistic. Users tend to evaluate responses based on overall impression rather than scoring along predefined criteria.

We randomly sample 50 user queries from the test set and collect corresponding responses from the three models: Qwen3-8B-nothink, Qwen3-8B-think, and MindChat. Each question-response group is evaluated by 10 human annotators, who are asked to rank the three responses in order of overall preference. Annotators are blind to model identity and encouraged to rely on their instinctive judgment.

As shown in Table 3, our model MindChat achieves a win rate of 71.14%, far surpassing the baselines, and is ranked within the top two positions in 80.60% of cases. It also obtains the lowest mean rank (1.38), indicating both strong first-place preference and consistent ranking across positions. These results further support the effectiveness of our alignment approach in producing responses that are more favored by human users.

Table 3: Human preference evaluation results across 50 questions with 10 annotators.

Model	Win Rate (%)	Top-2 Rate (%)	Mean Rank
Qwen3-8B-nothink	8.57	55.71	2.36
Qwen3-8B-think	20.29	53.72	2.26
MindChat (Ours)	**71.14**	**80.60**	**1.38**

5 Discussion and Conclusion

5.1 Limitations and Future Work

While MindChat-R0 demonstrates strong performance in generating emotionally supportive responses, it remains in an early stage of development. We name this version R0 to reflect its exploratory nature and limited training data of 4,500 samples. This limits the model's ability to generalize to more diverse real-world scenarios. A larger and more diverse dataset is needed to ensure robustness across varied user profiles and emotional scenarios.

Second, the reinforcement learning stage involves a limited number of training steps. The current improvements are visible but potentially shallow; further scaling of RL training is necessary to test the boundaries of reward-driven empathetic reasoning.

Third, the current reward modeling framework relies solely on an Outcome Reward Model (ORM), which evaluates only the final responses. This outcome-level supervision may overlook deficiencies in intermediate reasoning steps, such as emotional inference and supportive intent. Incorporating Process Reward Models (PRMs) [6] could provide more fine-grained feedback aligned with the multi-step reasoning process required by ECoT.

Fourth, our system is built exclusively on Chinese-language dialogues. This limits the applicability of MindChat-R0 in multilingual or cross-cultural settings, where expressions of emotion and support may vary significantly. In future work, integrating multilingual capabilities and cultural adaptation mechanisms will be essential for building more universally empathetic dialogue system.

Finally, the current evaluation setup still lacks human-in-the-loop validation. While both human preference and LLM-judge metrics indicate superior performance, real-time feedback from end users in mental health support contexts would provide more grounded and actionable insights.

5.2 LLM-as-a-judge

We adopt the LLM-as-a-judge paradigm for automatic evaluation. It provides stable and consistent scores, as reflected by the low variance in Table 2. This enables scalable and reproducible assessments of empathetic dialogue systems [7, 34]. It also reduces annotation costs and avoids inter-rater inconsistencies common in subjective tasks. However, this approach also has limitations. Evaluating mental health support is inherently subjective and culturally sensitive. Relying on a single LLM may overlook diverse interpretations of empathy. Future work could explore multi-persona judge ensembles [38], enabling richer and more inclusive evaluations tailored to different emotional and cultural contexts. In addition, combining outcome-level ratings with process-level feedback (e.g., on emotional reasoning steps) may further improve alignment with the goals of empathetic chain-of-thought modeling.

5.3 Conclusion

In summary, this work presents MindChat-R0, a system that leverages Empathetic Chain-of-Thought (ECoT) to enhance emotional support conversations. While still preliminary, it lays the foundation for future advances in affective reasoning and socially aware dialogue systems.

References

[1] Steven A. Beebe, Susan J. Beebe, Mark V. Redmond, and Lynn Salem-Wiseman. 2002. *Interpersonal Communication: Relating to Others.* Allyn and Bacon, Boston.

[2] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL] https://arxiv.org/abs/2501.12948

[3] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment. arXiv:2304.06767 [cs.LG] https://arxiv.org/abs/2304.06767

[4] Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like Summarization Evaluation with ChatGPT. arXiv:2304.02554 [cs.CL] https://arxiv.org/abs/2304.02554

[5] Pei Ke, Bosi Wen, Zhuoer Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. CritiqueLLM: Towards an Informative Critique Generation Model for Evaluation of Large Language Model Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

[6] Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. 2025. Process Reward Models That Think. arXiv:2504.16828 [cs.LG] https://arxiv.org/abs/2504.16828

[7] Philipp Koehn, Zecong Sun, and Benjamin Van Durme. 2024. Evaluating LLM-as-a-Judge: Benchmarking Bias and Reliability. *arXiv preprint arXiv:2402.12349* (2024).

[8] Yoon-Kyung Lee, Sowon Hahn, , Inju Lee, and Minjung Shin. 2024. Enhancing Empathic Reasoning of Large Language Models Based on Psychotherapy Models for AI-assisted Social Support. *Korean Journal of Cognitive Science* 35, 1 (2024), 23–48.

[9] Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. 2024. SALAD-Bench: A Hierarchical and Comprehensive Safety Benchmark for Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 3923–3954. doi:10.18653/v1/2024.findings-acl.235

[10] Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. Deductive verification of chain-of-thought reasoning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 1580, 27 pages.

[11] Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. WizardMath: Empowering Mathematical Reasoning for Large Language Models via Reinforced Evol-Instruct. *arXiv preprint arXiv:2308.09583* (2023).

[12] OpenAI. 2024. Introducing OpenAI o1-preview. https://openai.com/index/introducing-openai-o1-preview/. Accessed: 2025-06-20.

[13] Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. 2024. Agent Q: Advanced Reasoning and Learning for Autonomous AI Agents. arXiv:2408.07199 [cs.AI] https://arxiv.org/abs/2408.07199

[14] Claire Renvoize. 2022. *Conversations on Empathy.* Profile Books.

[15] Giuseppe Riva. 2004. *Cybertherapy: Internet and Virtual Reality as Assessment and Rehabilitation Tools for Clinical Psychology and Neuroscience.* IOS Press.

[16] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300 [cs.CL] https://arxiv.org/abs/2402.03300

[17] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. HybridFlow: A Flexible and Efficient RLHF Framework. *arXiv preprint arXiv: 2409.19256* (2024).

[18] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 377, 19 pages.

[19] Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. PsyQA: A Chinese Dataset for Generating Long Counseling Text for Mental Health Support. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 1489–1503. doi:10.18653/v1/2021.findings-acl.130

[20] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525* (2023).

[21] Qwen Team. 2025. QwQ-32B: Embracing the Power of Reinforcement Learning. https://qwenlm.github.io/blog/qwq-32b/

[22] Wenzhe Tian, Guang Yan, Sheng Xiong, Jie Zhang, Jun Peng, Xinyi Zhang, Yuan Zhou, Ting Liu, Yu Zhang, Peipei Ye, Wei Zhao, and Meng Tian. 2025. Burden of depressive and anxiety disorders in China and its provinces, 1990–2021: findings from the Global Burden of Disease Study 2021. *The British Journal of Psychiatry* (2025), 1–11. doi:10.1192/bjp.2024.267

[23] Luong Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. ReFT: Reasoning with Reinforced Fine-Tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 7601–7614. doi:10.18653/v1/2024.acl-long.410

[24] Yifan Wang, Qiaolin Chen, Ziqing Zhou, and Zheng Liu. 2023. ExTES: Explicit Task Decomposition and Strategy Conditioning for Emotional Support Conversation. *arXiv preprint arXiv:2308.11584* (2023).

[25] World Health Organization. 2023. Depression. https://www.who.int/health-topics/depression. Accessed: 2025-06-20.

[26] World Health Organization. 2023. Mental Health in China. https://www.who.int/china/health-topics/mental-health. Accessed: 2025-06-20.

[27] Jingran Xie, Shun Lei, Yue Yu, Yang Xiang, Hui Wang, Xixin Wu, and Zhiyong Wu. 2025. Leveraging Chain of Thought towards Empathetic Spoken Dialogue without Corresponding Question-Answering Data. arXiv:2501.10937 [cs.CL] https://arxiv.org/abs/2501.10937

[28] 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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388* (2025).

[29] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement. arXiv:2409.12122 [cs.CL] https://arxiv.org/abs/2409.12122

[30] Beichen Zhang, Kun Zhou, Xilin Wei, Wayne Xin Zhao, Jing Sha, Shijin Wang, and Ji-Rong Wen. 2023. Evaluating and Improving Tool-Augmented Computation-Intensive Math Reasoning. *arXiv preprint arXiv:2306.02408* (2023).

[31] Yuanhao Zhang, Xuran Xu, Ziqing Zhou, Jie Fu, and Zheng Liu. 2023. ESCoT: Emotional Support Conversation with Topic-Guided Strategy Planning. *arXiv preprint arXiv:2310.08593* (2023).

[32] Bowen Zhao, Zheng Liu, Danni Wang, Yujia Shen, and Pascale Fung. 2024. ESC-Eval: Evaluating Emotional Support Conversations via Role-play Simulation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language*

Processing (EMNLP).

[33] Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-o1: Towards Open Reasoning Models for Open-Ended Solutions. arXiv:2411.14405 [cs.CL] https://arxiv.org/abs/2411.14405

[34] Dallas Card Zheng, Maarten Sap, Noah A. Smith, and Yejin Choi. 2023. Large Language Models are Human-Level Judges of Text Quality. *arXiv preprint arXiv:2306.05685* (2023).

[35] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 46595–46623. https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf

[36] Shiyi Zheng, Yuxuan Zhang, Mingxuan Wang, Yiwei Zhang, Jinjie Chen, Ruixiang Xu, Qian Liu, Yue Zhang, and Yujie Qian. 2023. AugESC: Augmenting Emotional Support Conversations via Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2023.*

[37] Hao Zhou, Qian Xu, Lei Li, and Minlie Yang. 2022. Exploring Empathy in Multi-turn Open-domain Chatbots. *arXiv preprint arXiv:2203.13127* (2022). https://arxiv.org/abs/2203.13127

[38] Ruichen Zhou, Yujia Zhang, Wenxuan Wang, Yue Zhang, and Yangfeng Ji. 2024. Persona-as-a-Judge: Measuring Empathy with Multi-Persona LLM Evaluation. In *Proceedings of ACL.*