DecoupledESC: Enhancing Emotional Support Generation via Strategy-Response Decoupled Preference Optimization

Anonymous ACL submission

Abstract

Recent advances in Emotional Support Conversation (ESC) have improved emotional support generation by fine-tuning Large Language Models (LLMs) via Supervised Fine-Tuning (SFT). However, common psychological errors still persist. While Direct Preference Optimization (DPO) shows promise in reducing such errors through pairwise preference learning, its effectiveness in ESC tasks is limited by two key challenges: **(1) Entangled data structure:** Existing ESC data inherently entangles psychological strategies and response content, making it difficult to construct high-quality preference pairs; and **(2) Optimization ambiguity:** Applying vanilla DPO to such entangled pairwise data leads to ambiguous training objectives. To address these issues, we introduce Inferential Preference Mining (IPM) to construct high-quality preference data, forming the IPM-PrefDial dataset. Building upon this data, we propose a **Decoupled ESC** framework inspired by Gross's Extended Process Model of Emotion Regulation (Gross, 2015), which decomposes the ESC task into two sequential subtasks: strategy planning and empathic response generation. Each was trained via SFT and subsequently enhanced by DPO to align with the psychological preference. Extensive experiments demonstrate that our Decoupled ESC framework outperforms joint optimization baselines, reducing preference bias and improving response quality[1].

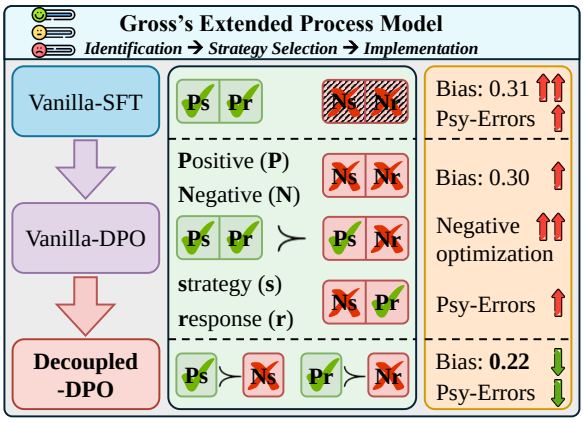

Figure 1: Comparison from Vanilla-SFT to Vanilla-DPO to Decoupled-DPO. Vanilla-SFT lacks negative preference data, leading to high preference bias; Vanilla-DPO uses entangled preference data, causing potential negative optimization (regards PsNr, NsPr as pure negative samples); Decoupled-DPO disentangles strategy and response, effectively reduce bias and psychological errors.

1 Introduction

Mental health is essential to well-being (Prince et al., 2007), yet rising stress and fast-paced life have increased related issues (Bor et al., 2014; Brundtland, 2000; Paisley and McMahon, 2001). According to WHO, 1 in 8 people suffer from mental disorders (Organization, 2022). Amid a shortage of professionals, this underscores the need for scalable solutions, where Large Language Models (LLMs) offer promising potential.

To enhance the performance of LLMs in Emotional Support Conversation (ESC), prior works (Zhang et al., 2024a; Chen et al., 2023) have constructed several large-scale, high-quality dialogue datasets and applied Supervised Fine-Tuning (SFT) to improve model responses. Among them, Liu *et al.* (Liu et al., 2021) built the ESConv dataset based on Hill's Helping Skills Theory (Hill, 1999) and filtered out FailedESConv dataset. The ESConv dataset follows a three-phase structure (Exploration → Comfort → Action) and includes eight types of support strategies, each paired with corresponding responses, details are provided in Appendix A and C.1. This structured design significantly enhances a model's ability to generate empathetic dialogue.

Observation Currently, SFT has become the mainstream approach in the ESC field. However, we observe that models still frequently exhibit common psychological errors (Gross, 2002;

[1]Our data and code are available at https://anonymous.4open.science/r/DecoupledESC-0E37.

Raskin and Rogers, 2005; Stebnicki, 2007) during inference, which align with those identified in the FailedESConv dataset (§3.1). In addition, Zhao *et al.* (Zhao et al., 2025) found that SFT's focus on single gold strategy-response pairs limits adaptability to nuanced contexts, weakening empathetic support. To mitigate this, they use MCTS to collect pairwise preference data linking strategies and responses, and apply Direct Preference Optimization (Vanilla-DPO) to guide the model in choosing appropriate strategies, thereby partially reducing preference bias and improving response quality.

Challenges However, as shown in Figure 1 and 4, our further analysis reveals that the limitations of current work lie not in the SFT or DPO training methods themselves, but rather in two overlooked challenges (§3.2): **(1) Entangled data structure:** Existing ESC datasets heavily entangle psychological strategies with response content, making it difficult to construct high-quality preference pairs. For instance, penalizing responses with correct strategies but flawed content may degrade data quality. **(2) Optimization ambiguity:** Applying Vanilla-DPO directly to such entangled data can blur training objectives and even lead to negative optimization outcomes.

Approach To address these issues, we first introduce the **I**nferential **P**reference **M**ining (**IPM**) method, which automatically constructs preference samples decoupled from strategy-response. Specifically, we use dynamic data routing to route four types of psychological error samples identified from the SFT model's inference data to the DPO training stage of either strategy planning or response generation, depending on the error type. These samples are then paired with human-annotated ground truth samples to form the **I**nferential **P**reference **M**ining **Pref**erence **Dial**ogues (**IPM-PrefDial**) dataset, containing 21k strategy preference pairs and 11k response preference pairs. This dataset provides disentangled and high-quality supervision signals for two separate DPO models. Building on this, we propose a **Decoupled ESC optimization framework**, grounded in the **Extended Process Model of Emotion Regulation (EPMER)** (Gross, 2015), which divides emotion regulation into three sequential stages: *identification*, *strategy selection*, and *implementation*, details are provided in Appendix A.1. Accordingly, we explicitly split the ESC task into two subtasks: **Strategy Planning (SP)** and **Response Generation (RG)**, enabling structured and goal-driven emotional support in dialogue.

Results Across multiple evaluation metrics, our decoupled optimization framework significantly outperforms joint training baselines. It not only enhances the diversity of strategy selection but also improves response quality and empathy.

Contributions Our key contributions are summarized as follows:

- We analyze common psychological errors in existing SFT paradigms, and introduce the IPM method to construct the IPM-PrefDial, a strategy–response decoupled dataset.
- We propose a Decoupled ESC framework inspired by Gross's Extended Process Model of Emotion Regulation, which explicitly splits the ESC task into two subtasks: Strategy Planning and Response Generation, effectively mitigating preference bias and enhancing response quality.
- Extensive experiments show that our Decoupled ESC optimization framework significantly outperforms joint optimization baselines across multiple evaluation metrics.

2 Related Work

Emotional Support Conversation. Emotional Support Conversation (ESC) aims to alleviate users' emotional distress through empathetic and supportive responses. Liu *et al.* (Liu et al., 2021) first introduced the concept and built the ESConv dataset with 8 support strategies, 1.3k dialogues. They also released the FailedESConv dataset, containing 196 failed dialogues. Subsequent studies improved ESC systems by enhancing data quality (Sun et al., 2021; Qiu et al., 2023; Chen et al., 2023), adding external strategy planners (Deng et al., 2024; He et al., 2024, 2025), and incorporating commonsense reasoning (Tu et al., 2022; Deng et al., 2023). Supervised Fine-Tuning (SFT) remains the dominant training paradigm with strong real-world performance (e.g., MeChat (Qiu et al., 2023), SweetieChat (Ye et al., 2025)). Recently, preference-based methods like Direct Preference Optimization (DPO) (Rafailov et al., 2023) have emerged. Zhao *et al.* (Zhao et al., 2025) introduced DPO with MCTS-based data to jointly optimize strategies and responses. However, the fixed coupling limited independent optimization and resulted in lower response quality.

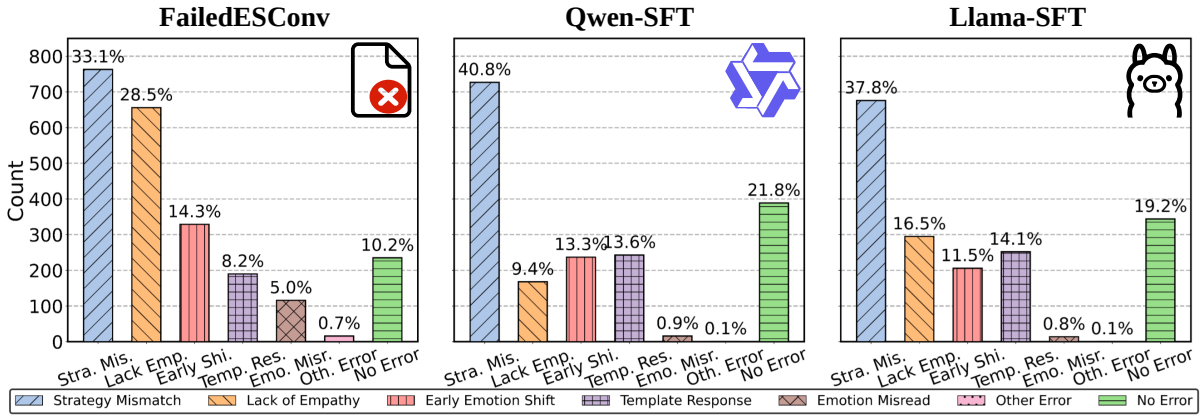

Figure 2: Observation 1: Comparison of common psychological error type proportions among the FailedESConv dataset, Qwen-SFT inference results, and Llama-SFT inference results.

Reinforcement Learning for LLM. Reinforcement Learning (RL) was initially introduced into LLM training to align with human preferences (Ouyang et al., 2022). This approach uses a reward model to guide the optimization of the policy model via the Proximal Policy Optimization (PPO) algorithm (Schulman et al., 2017). Recently, the Group Relative Policy Optimization (GRPO) algorithm was proposed to enhance model reasoning capabilities (Shao et al., 2024), which eliminates the need for a critic model by using within-group rewards as advantages. While these online reinforcement learning methods are effective, they suffer from high computational costs and reliance on accurate reward modeling. As a simpler offline optimization algorithm, DPO optimizes the policy model from pairwise preference data directly without the need for reward modeling. Due to its simplicity and effectiveness, DPO has achieved significant success across multiple domains, including mathematical reasoning, code generation, and recommendation systems (Lai et al., 2024; Zhang et al., 2024b; Chen et al., 2024).

3 Preliminary Observations

To investigate the causes of low response quality in the ESC task, we analyzed outputs from six models: Base, SFT, and DPO versions of `Qwen2.5-7B-Instruct` (Team, 2024) and `Llama3.1-8B-Instruct` (Dubey et al., 2024).

3.1 Preference Bias and Psychological Errors (Obs 1)

Current base and SFT models (e.g., Qwen-Base, Qwen-SFT) show strong strategy preferences (Kang et al., 2024), often overusing fixed strategies and failing to adapt to users' emotional states. As shown in Figure 3, their strategy distributions diverge significantly from the ground truth.

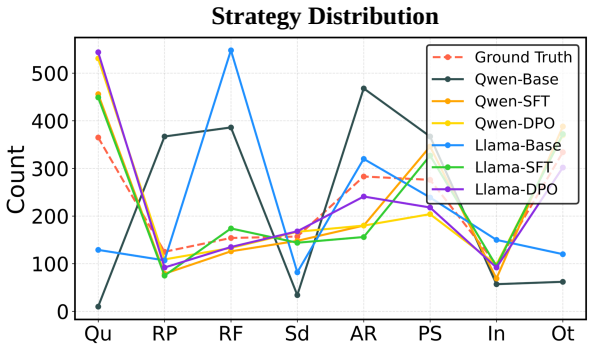

Figure 3: Strategy Distribution across different models.

To further explore the impact of preference bias on response quality, we compared the outputs of Qwen-Base, Qwen-SFT, Llama-Base, and Llama-SFT with the FailedESConv dataset. As shown in Figure 2, common psychological errors (Raskin and Rogers, 2005; Stebnicki, 2007) observed in the SFT-generated responses frequently aligned with those found in FailedESConv, including: **(1) Strategy Mismatch, (2) Lack of Empathy, (3) Early Emotion Shift, (4) Template Response, (5) Emotion Misread.** The definitions and corresponding examples are detailed in Appendix A.2.

Although SFT reduces some errors, the empathy quality remains unsatisfactory. We argue that this stems from the SFT paradigm's reliance on high-quality samples (Zhao et al., 2025) without incorporating negative supervision signals from the FailedESConv dataset, failing to address bias in strategy selection and emotional understanding.

3.2 Limitations of the DPO Method (Obs 2)

To address Obs 1, a natural approach is to treat filtered failures as negative signals and train with

DPO. Prior work (Zhao et al., 2025) adopted a vanilla DPO setup that jointly optimizes strategy-response pairs. However, as shown in Figure 1 and Figure 3, Vanilla-DPO relies heavily on the *Question* strategy and shows a strong preference for it, which fails to significantly reduce preference bias (§6.2).

To investigate the failure of Vanilla-DPO in aligning with human preferences, we conduct a controlled study. We split the preference data based on response quality into two types: **(PsPr, PsNr)**: where the preferred sample has both a positive strategy (Ps) and positive response (Pr), and the non-preferred sample has a positive strategy (Ps) but a negative response (Nr). **(PsPr, NsNr)**: where the non-preferred sample contains both negative strategy (Ns) and negative response (Nr).

We train models using each dataset on Qwen and Llama, and evaluate them on preference bias and strategy preference. As shown in Figure 4, models trained on (PsPr, NsNr) consistently outperform those trained on (PsPr, PsNr). It reduces preference bias and better aligns with diverse strategies. **These results show that Vanilla-DPO training with entangled pairs like (PsPr, PsNr) harms strategy learning.** This reveals two issues in Vanilla-DPO:

1. **Entangled data structure:** The coupling between strategy and response complicates the construction of high-quality preference data, highlighting the need for more rigorous evaluation and filtering methods.

2. **Optimization Ambiguity:** Entangled strategy and response training lead to optimization ambiguity or even negative optimization: mislabeling PsNr as a negative sample leads to negative optimization on strategy learning, while NsPr harms response learning.

According to Hill's Helping Skills Theory (Hill, 1999) and Gross's Extended Process Model of Emotion Regulation (EPMER) (Gross, 2015), strategies should precede response generation and serve as its guidance. In essence, the two are decouplable. However, when strategies and responses are jointly optimized as a single DPO objective, the optimization objective becomes ambiguous. To address this, we propose a decoupled modeling and staged optimization framework for ESC, which separates strategy planning from response generation, enabling more structured and targeted improvements in dialogue quality.

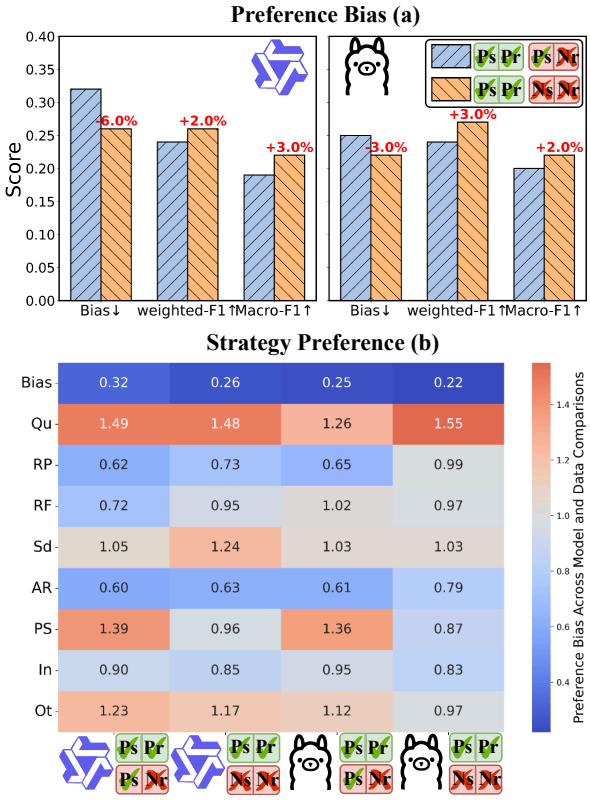

Figure 4: Observation 2: (a) Preference Bias and (b) Strategy Preference across Qwen and Llama models trained on different preference datasets.

4 Datasets

4.1 Preference Dataset Construction

Inferential Preference Mining. Standard SFT approaches focus on expert demonstrations and neglect failure cases, limiting the model's ability to learn from mistakes. This can result in preference bias and the generation of low-quality responses.

To address these issues, we propose an Inferential Preference Mining (IPM) method to collect high-quality preference data, along with a dynamic data routing mechanism that identifies psy-errors and routes data to targeted DPO training phases.

Specifically, we first perform inference using Qwen-SFT and Llama-SFT models to generate inference data, which are paired with the corresponding ground truth to form candidate preference pairs. Subsequently, we apply a **dynamic data routing mechanism** to filter these pairs based on the four types of common psy-errors observed in Obs1 (§3.1). This process yields the final IPM-PrefDial dataset. Concretely, for each subtask:

For the **Strategy Planner (SP)**, we identify sub-optimal strategies s_r exhibiting the **Strategy Mismatch** psy-error, pair them with the gold strategies s_c and context c, and construct the $D_{\text{SP-dpo}}$ dataset,

	Criteria	Total	Assistant	User
ESConv	# Dialogues	1,040	–	–
	# Utterances	29,526	14,763	14,763
	Avg. Turns of Dialogue	28.40	14.20	14.20
	Avg. Char of Utterance	95.85	112.17	79.54
	Criteria	**Total**	**Qwen**	**Llama**
IPM-PrefDial	# Strategy Pref-Pairs	21,370	10,651	10,719
	# Response Pref-Pairs	11,887	6,041	5,846
	Avg. Char of Chosen	124.89	124.72	125.06
	Avg. Char of Rejected	83.82	81.04	86.59
	# Lack Emp. Response	4,371	2,288	2,083
	# Emo. Shift Response	3,600	1,814	1,786
	# Temp. Res. Response	3,916	1,939	1,977

Table 1: Statistics of the ESConv and IPM-PrefDial Datasets.

which is routed to the SP's DPO training phase:

$$D_{\text{SP-dpo}} = \left\{ \left(c^{(i)}, s_c^{(i)}, s_r^{(i)} \right) \right\}_{i=1}^{|D_{\text{SP-dpo}}|}. \quad (1)$$

For the **Response Generator (RG)**, the SFT model generates multiple responses based on the gold strategy. Suboptimal responses a_r exhibiting typical psychological errors, namely **Lack of Empathy**, **Early Emotion Shift**, or **Template Response**, are identified via the dynamic data routing. These are then paired with the corresponding gold responses a_c to construct the $D_{\text{RG-dpo}}$ dataset, which is routed to the RG's DPO training phase:

$$D_{\text{RG-dpo}} = \left\{ \left(c^{(i)}, s^{(i)}, a_c^{(i)}, a_r^{(i)} \right) \right\}_{i=1}^{|D_{\text{RG-dpo}}|}. \quad (2)$$

4.2 Datasets Statistics

ESConv Dataset. We employ ESConv dataset for SFT training (D_{sft}), which includes $1,040$ dialogues with an average of 14.2 turns and 95.9 characters per turn. Strategy distribution and temporal trends are shown in Table 6 and Figure 8 in the Appendix C.1.

IPM-PrefDial Dataset. IPM-PrefDial dataset contains $21,370$ strategy preference pairs ($D_{\text{SP-dpo}}$) and $11,887$ response preference pairs ($D_{\text{RG-dpo}}$). In $D_{\text{RG-dpo}}$, chosen responses average 124.89 characters, rejected ones 83.82. Major rejection reasons include lack of empathy $(4,371)$, early emotion shift $(3,600)$, and template response $(3,916)$. Details are in Appendix C.2.

4.3 Datasets Quality

We evaluate the content quality of 100 samples from $D_{\text{RG-dpo}}$ using `gpt-4.1-mini-2025-04-14`[2].

[2] https://openai.com/index/gpt-4-1

Model	Type	Flu.↑	Pro.↑	Emp.↑	Hel.↑
Qwen	Chosen	3.82	3.52	3.20	2.95
	Rejected	3.65	3.09	2.41	2.32
	Improve (↑)	4.66%	13.92%	**32.78%**	27.16%
Llama	Chosen	3.99	3.74	3.33	3.09
	Rejected	3.93	3.21	2.40	2.39
	Improve (↑)	1.53%	16.52%	**38.75%**	29.29%

Table 2: LLM-based evaluation scores for chosen and rejected responses across four dimensions.

As shown in Table 2, chosen responses outperform rejected ones across four LLM-based metrics, with over a 30% gain in *Empathy*.

5 Methodology

5.1 Decoupled ESC Framework

To address the issue raised in **Obs 1** (§3.1) and **Obs 2** (§3.2), that vanilla training of strategy selection and response generation can lead to negative optimization, hindering the reduction of preference bias and the improvement of response quality. As shown in Figure 5, we propose a **decoupled ESC optimization framework**, inspired by the Extended Process Model of Emotion Regulation (EPMER) (Gross, 2015), which divides emotion regulation into three sequential stages: *identification*, *strategy selection*, and *implementation*. We decouple the ESC generation process into two independent subtasks: Strategy Planning and Response Generation. This enables more stable and controllable training for each.

Specifically, we adopt a decoupled two-stage modeling framework: a Strategy Planner selects a optimal strategy based on the dialog history $c_t = (u_0, a_0, \ldots, u_{t-1}, a_{t-1}, u_t)$, where u and a denote user and assistant utterances, respectively. The strategy is generated as $s_t \sim \text{LLM}_{\text{SP}}(s \mid c_t)$. Then, a Response Generator generates an empathic reply conditioned on both the selected strategy and the dialog context: $a_t \sim \text{LLM}_{\text{RG}}(a \mid c_t, s_t)$.

5.2 Decoupled-SFT and Decoupled-DPO

Decoupled-SFT. To optimize the performance of the Strategy Planner and Response Generator, we first initialize these two modules using the SFT method to endow them with the capabilities for strategy planning and empathic response generation. Specifically, based on real dialogues from the ESC dataset, we constructed a turn-level training dataset $D_{\text{sft}} = \left\{ \left(c^{(i)}, s^{(i)}, a^{(i)} \right) \right\}_{i=1}^{|D_{\text{sft}}|}$. The two modules are then fine-tuned separately using SFT:

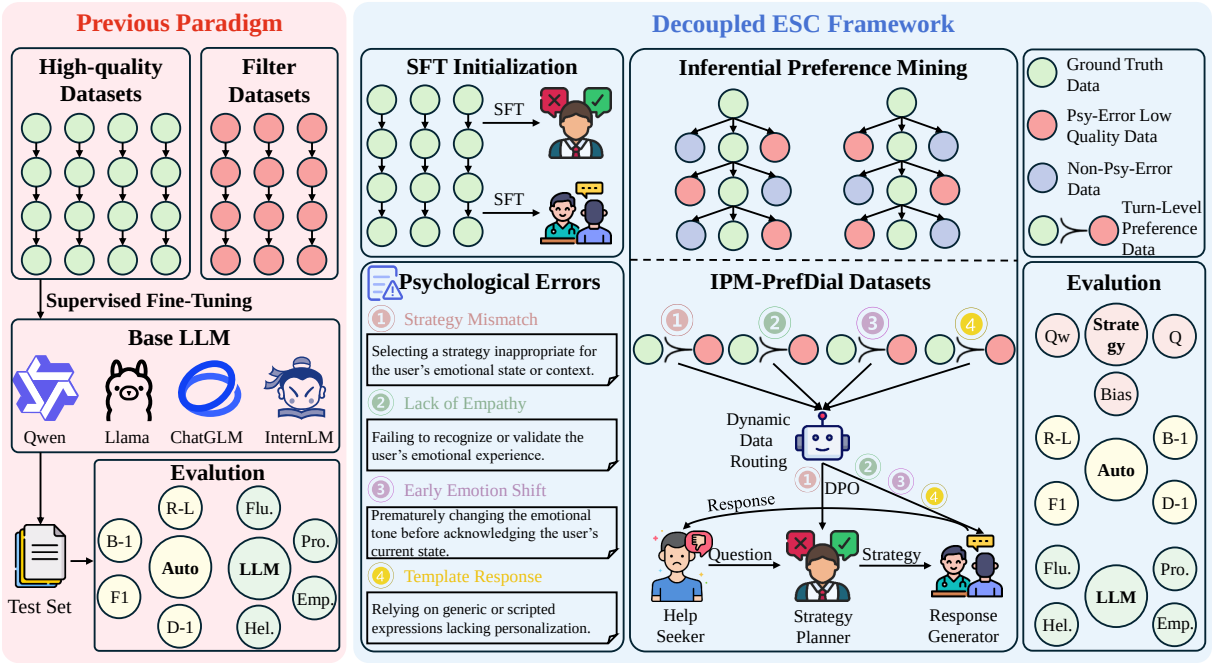

Figure 5: Comparison between previous vanilla SFT training paradigm and our proposed Decoupled ESC framework.

(1) Strategy Planner: Using the dialogue context c and the supporter's response strategy s, we perform turn-level training to minimize the loss function:

$$\mathcal{L}_{\text{SP-sft}} = -\mathbb{E}_{(c,s)\sim D_{\text{sft}}} \left[\log \text{LLM}_{\text{SP}}(s|c)\right]. \quad (3)$$

(2) Response Generator: Given the context c, strategy s, and response a, minimizing the loss:

$$\mathcal{L}_{\text{RG-sft}} = -\mathbb{E}_{(c,s,a)\sim D_{\text{sft}}} \left[\log \text{LLM}_{\text{RG}}(a|c,s)\right]. \quad (4)$$

Decoupled-DPO. To further optimize the Strategy Planner and Response Generator and reduce psychological errors, we apply the offline reinforcement learning method (DPO). Leveraging the previously constructed preference datasets $D_{\text{SP-dpo}}$ and $D_{\text{RG-dpo}}$, we separately train the two modules to improve their decision-making and response generation capabilities. For the Strategy Planner, we train the model on the strategy preference dataset $D_{\text{SP-dpo}}$ to enhance its preference for gold strategies while mitigating bias toward suboptimal strategies, thereby reducing strategy selection bias. The loss function is defined as follows:

$$\mathcal{L}_{\text{SP-dpo}} = -\mathbb{E}_{(c,s_c,s_r)\sim D_{\text{SP-dpo}}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(s_c|c)}{\pi_{\text{ref}}(s_c|c)} - \beta \log \frac{\pi_\theta(s_r|c)}{\pi_{\text{ref}}(s_r|c)}\right)\right], \quad (5)$$

where π_θ denotes the model being optimized, and π_{ref} denotes the reference model after SFT.

For the Response Generator, we train with $D_{\text{RG-dpo}}$ to enhance the quality and empathy of responses through the following loss function:

$$\mathcal{L}_{\text{RG-dpo}} = -\mathbb{E}_{(c,s,a_c,a_r)\sim D_{\text{RG-dpo}}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(a_c|c,s)}{\pi_{\text{ref}}(a_c|c,s)} - \beta \log \frac{\pi_\theta(a_r|c,s)}{\pi_{\text{ref}}(a_r|c,s)}\right)\right]. \quad (6)$$

Through this decoupled two-stage DPO optimization, we effectively reduce bias in strategy planning and enhance the empathy of responses, resulting in more robust and emotionally intelligent dialogue generation.

6 Experiments

In this section, we conduct extensive experiments to address the following research questions:

- **RQ1:** Can Vanilla-DPO better mitigate preference bias and improve response quality compared to Vanilla-SFT?
- **RQ2:** Does the decoupled ESC framework lead to better bias reduction and response quality?
- **RQ3:** In the decoupled setting, is DPO more effective than SFT?
- **RQ4:** Can Decoupled-DPO more effectively reduce common psychological errors by separately optimizing strategy and response?

Backbone	Paradigm	Method	Automatic Metrics.↑				LLM-based Metrics.↑				Strategy Metrics.		
			D-1	B-1	F1	R-L	Flu.	Pro.	Emp.	Hel.	$\mathcal{B}\downarrow$	$\mathcal{Q}w\uparrow$	$\mathcal{Q}\uparrow$
Qwen2.5-7B-Instruct	Vanilla	Base	93.50	9.75	14.92	12.59	3.55	2.53	1.89	1.38	2.17	8.41	8.06
		+Direct-Refine	95.79	10.91	16.26	14.35	**4.17**	**2.97**	2.20	1.77	1.54	13.52	10.46
		+Self-Refine	97.04	10.28	15.85	13.85	3.68	2.64	1.94	1.34	1.45	10.92	9.63
		+Emotion CoT	97.17	10.61	16.07	14.06	3.95	2.70	2.50	1.51	1.87	6.89	6.63
		SFT	90.93	15.61	20.99	17.78	3.30	2.61	2.29	2.12	0.31	24.89	20.27
		DPO	88.13	16.23	21.24	18.03	3.47	2.67	2.36	**2.23**	0.30	22.25	18.97
	Decoupled	Base	**97.55**	10.97	16.33	14.19	3.92	2.71	2.17	1.38	1.92	13.96	12.07
		SFT	91.37	16.69	22.15	18.76	3.93	2.72	2.40	2.11	0.27	26.94	21.37
		DPO	89.84	**17.73**	**22.86**	**19.31**	3.99	2.90	**2.54**	2.02	**0.22**	**27.09**	**21.77**
Llama3.1-8B-Instruct	Vanilla	Base	**95.09**	12.38	16.85	14.01	**4.35**	3.21	2.36	1.76	1.03	15.74	14.09
		+Direct-Refine	90.07	11.36	14.97	12.79	3.35	2.82	2.16	1.35	1.72	12.12	9.98
		+Self-Refine	87.18	10.72	14.26	12.20	3.53	2.95	2.40	1.45	1.68	13.93	12.00
		+Emotion CoT	77.32	10.06	13.32	11.33	3.24	2.88	2.56	1.63	1.86	13.31	11.35
		SFT	91.29	15.75	21.38	18.11	3.31	2.52	2.22	2.06	0.26	24.54	19.97
		DPO	91.25	15.15	20.49	17.25	3.41	2.79	2.41	**2.28**	0.28	24.00	19.89
	Decoupled	Base	94.65	12.67	16.70	14.01	4.24	**3.24**	2.34	1.66	1.62	7.54	7.67
		SFT	91.51	16.97	22.42	19.12	3.87	2.74	2.39	1.95	0.23	26.03	21.36
		DPO	90.35	**17.50**	**22.59**	**19.16**	3.81	2.73	**2.64**	2.17	**0.15**	**27.10**	**22.94**

Table 3: Comparison of models under different optimization paradigms and training methods. The best score is **in-bold**, while the second best score is underlined.

6.1 Experimental Setup

Backbone Models. We conducted experiments using two widely adopted large language models: **Qwen2.5-7B-Instruct** (Team, 2024) and **Llama3.1-8B-Instruct** (Dubey et al., 2024).

Baselines. We compare vanilla coupled models (Base, SFT, DPO) with prompt-optimization baselines such as Direct-Refine, Self-Refine (Madaan et al., 2023), and Emotional CoT (Wei et al., 2022).

Datasets. The **ESConv** (Liu et al., 2021) dataset is split into train, valid, and test sets in an 8:1:1 ratio, with the training set used for SFT. The **IPM-PrefDial** dataset is used for DPO training.

Evaluation Metrics. We evaluate model performance using the following metrics: **(1) Automatic Metrics**, including BLEU-1 (B-1) (Papineni et al., 2002), Distinct-1 (D-1) (Li et al., 2015), F1-score (F1), and ROUGE-L (R-L) (Lin, 2004); **(2) LLM-based Metrics**, including *Fluency (Flu.)*, *Professionalism (Pro.)*, *Empathy (Emp.)*, and *Helpfulness (Hel.)*. All metrics are rated on a 5-point Likert scale (Joshi et al., 2015); **(3) Strategy Metrics**, including preference bias ($\mathcal{B}$) (Kang et al., 2024) and strategy prediction accuracy (weighted-F1 $\mathcal{Q}w$ and Macro-F1 $\mathcal{Q}$). Detailed definitions of the evaluation metrics, the prompt, and the Bias calculation formula are provided in Appendix E.

Implementation Details. We use a learning rate of 1e-5 and a batch size of 32 for SFT, training for 3 epochs, while DPO is trained for 1 epoch. Experiments are conducted on 4×24GB RTX4090 GPUs. For LLM-based evaluation, we randomly sample 100 instances from the test set and assess them using `gpt-4.1-mini-2025-04-14`[2]. More implementation details are provided in Appendix D.1.

6.2 Experimental Results

Vanilla-DPO vs. Vanilla-SFT (RQ1). To answer RQ1, whether DPO can mitigate preference bias and improve response quality in vanilla setting, we compare Vanilla-DPO with Vanilla-SFT. As shown in Table 3, DPO consistently outperforms SFT on LLM-based metrics for both Qwen and Llama, indicating enhanced response quality.

However, preference evaluation results are mixed: for Qwen, DPO slightly reduces the preference bias ($\mathcal{B}$) compared to SFT; for Llama, DPO shows higher bias and lower prediction accuracy. We attribute this to DPO's greater sensitivity to data quality in the vanilla setting. When training data contains conflicting optimization signals between strategy and content, DPO may suffer from negative transfer, leading to performance degradation. To further examine how different preference

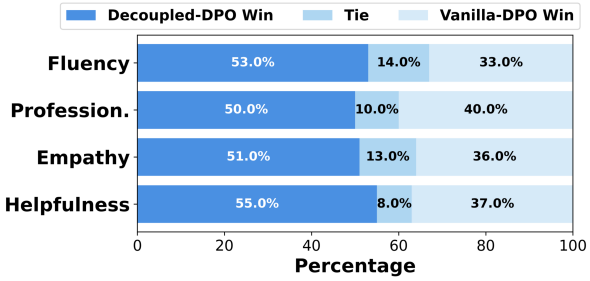

Figure 6: Comparison of Human-Evaluated Win Rates for Decoupled-DPO (Llama) and Vanilla-DPO (Llama).

data affect coupled models, we provide a detailed comparison in Appendix B.

Decoupled vs. Vanilla (RQ2). To answer RQ2, whether the decoupled ESC framework better mitigates preference bias and improves empathetic response quality. We compare **Decoupled** and **Vanilla** paradigms under both SFT and DPO training. As shown in Table 3, decoupled models consistently outperform their vanilla counterparts across most metrics, including automatic metrics, LLM-based metrics, and strategy metrics. Notably, the preference bias of Decoupled-DPO models trained with Qwen and Llama are 0.22 and 0.15, respectively, representing a substantial improvement over Vanilla-DPO and Vanilla-SFT. These results highlight the effectiveness of decoupled optimization in mitigating preference bias. Furthermore, we invited 20 master's students with a background in psychology to evaluate 20 samples under the guidance of a licensed psychological counselor. As shown in Figure 6, Decoupled-DPO outperforms Vanilla-DPO across all four LLM-based metrics.

We attribute this to the decoupled framework, which avoids the conflict and negative optimization often encountered in vanilla training. It also simplifies the construction of preference datasets. In addition, the decoupled strategy planning module can serve as an external planner, which, as Kang *et al.* (Kang et al., 2024) note, helps reduce strategy bias and improves the coherence and empathy of responses.

Decoupled-DPO vs. Decoupled-SFT (RQ3). To answer RQ3, whether DPO outperforms standard SFT in a decoupled setting, we compare Decoupled-DPO and Decoupled-SFT. As shown in Table 3, Decoupled-DPO outperforms other models on most LLM-based and bias metrics for both Qwen and Llama, achieving state-of-the-art (SOTA) performance. Importantly, the improvements of DPO over SFT are more pronounced in

Backbone	GT_Stra	SFT	DPO	Flu.↑	Pro.↑	Emp.↑	Hel.↑
Qwen2.5-7B-Instruct	✓	✓	✗	3.66	3.02	2.51	2.37
	✓	✓	✓	3.77	3.26	2.75	2.67
Llama3.1-8B-Instruct	✓	✓	✗	3.67	3.18	2.71	2.52
	✓	✓	✓	3.94	3.44	2.90	2.73

Table 4: Ablation study on Response Generator.

the decoupled setting, with bias for Qwen dropping from 0.27 to 0.22, and for Llama from 0.23 to 0.15. In contrast, the vanilla framework shows little change from SFT to DPO, further validating the synergy between decoupling and DPO training. Table 4 illustrates the impact of DPO training on the Response Generator. Given the ground truth strategy, the DPO-trained Response Generator consistently outperforms the SFT model across all LLM-based metrics.

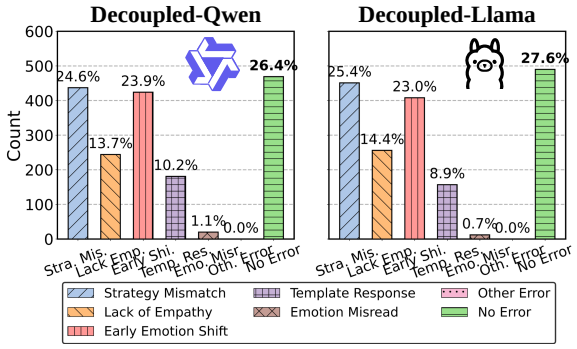

Figure 7: Proportion of psy-error types in the response of Qwen and Llama under the decoupled framework.

Decoupled-DPO on Psychological Errors (RQ4). To verify the effectiveness of Decoupled-DPO in improving response quality, we adopt the same error categorization approach as Obs 1 (§3.1). As shown in Figure 2 and 7, Decoupled-DPO achieves the highest proportion of **No Error** cases at 27%, outperforming Qwen-SFT and Llama-SFT by an average of 7%. This demonstrates that Decoupled-DPO effectively reduces common psychological errors and improves overall response quality.

7 Conclusion

In this paper, we propose a Decoupled ESC framework that separates strategy planning from empathetic response generation, enabling targeted optimization and avoiding mutual interference. Extensive experiments demonstrate that our Decoupled ESC framework significantly outperforms joint optimization baselines, effectively reducing preference bias and improving response quality in ESC tasks.

Limitations

Due to computational resource constraints, our study validates the proposed method only on relatively small-scale models (7B, 8B, and 9B). While significant improvements are observed at this scale, the effectiveness on larger models (e.g., 70B) remains an open question and is a key direction for future work. In addition, the generalizability of our decoupled ESC optimization framework warrants further investigation under other preference optimization methods such as KTO, SimPO, and IPO.

Ethics Statement

Data Usage Agreement

This research utilizes the `ESConv` and `FailedESConv` dataset (Liu et al., 2021), which has been obtained with proper authorization and in compliance with data usage agreements. We ensure that all data used in this study are handled responsibly and in accordance with ethical standards, respecting the privacy and confidentiality of individuals involved. All necessary agreements and permissions for the use of this dataset have been signed, ensuring full compliance with data protection regulations.

Model Usage Policy

It should be noted that while the model demonstrates certain capabilities in psychological support tasks, the strategies and expressions it adopts differ from those used in real-life professional counseling. Given the diversity of users' emotional states and circumstances, the model's responses may not always align with professional therapeutic standards and, in some cases, may unintentionally affect users' emotional well-being. Therefore, this model is intended for academic research purposes only and is not recommended for commercial use. Caution is advised when applying the model outside controlled research environments, and it should not be used for real-world psychological counseling without supervision by qualified professionals.

References

William Bor, Angela J Dean, Jacob Najman, and Reza Hayatbakhsh. 2014. Are child and adolescent mental health problems increasing in the 21st century? a systematic review. *Australian & New Zealand journal of psychiatry*, 48(7):606–616.

Gro Harlem Brundtland. 2000. Mental health in the 21st century. *Bulletin of the world Health Organization*, 78(4):411.

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.

Yuxin Chen, Junfei Tan, An Zhang, Zhengyi Yang, Leheng Sheng, Enzhi Zhang, Xiang Wang, and Tat-Seng Chua. 2024. On softmax direct preference optimization for recommendation. *arXiv preprint arXiv:2406.09215*.

Yang Deng, Wenxuan Zhang, Wai Lam, See-Kiong Ng, and Tat-Seng Chua. 2024. Plug-and-play policy planner for large language model powered dialogue agents. In *ICLR*.

Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. In *ACL (1)*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and et al. Kadian. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

James J Gross. 2002. Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology*, 39(3):281–291.

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

Tao He, Lizi Liao, Yixin Cao, Yuanxing Liu, Ming Liu, Zerui Chen, and Bing Qin. 2024. Planning like human: A dual-process framework for dialogue planning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4768–4791.

Tao He, Lizi Liao, Yixin Cao, Yuanxing Liu, Yiheng Sun, Zerui Chen, Ming Liu, and Bing Qin. 2025. Simulation-free hierarchical latent policy planning for proactive dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24032–24040.

Clara E Hill. 1999. Helping skills: Facilitating exploration, insight, and action. *American Psychological Association*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396.

Dongjin Kang, Sunghwan Mac Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15232–15261.

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*.

Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. 2024. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. In *NeurIPS*.

World Health Organization. 2022. *World mental health report: Transforming mental health for all*. World Health Organization.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Pamela O Paisley and George McMahon. 2001. School counseling for the 21st century: Challenges and oportunities. *Professional school counseling*, 5(2):106.

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.

Martin Prince, Vikram Patel, Shekhar Saxena, Mario Maj, Joanna Maselko, Michael R Phillips, and Atif Rahman. 2007. No health without mental health. *The lancet*, 370(9590):859–877.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2023. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *arXiv preprint arXiv:2305.00450*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Nathaniel J Raskin and Carl R Rogers. 2005. Person-centered therapy.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Mark A Stebnicki. 2007. Empathy fatigue: Healing the mind, body, and spirit of professional counselors. *American journal of psychiatric rehabilitation*, 10(4):317–338.

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. *arXiv preprint arXiv:2106.01702*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. Misc: A mixed strategy-aware model integrating comet for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2025. 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.

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.

Kechi Zhang, Ge Li, Yihong Dong, Jingjing Xu, Jun Zhang, Jing Su, Yongfei Liu, and Zhi Jin. 2024b. Codedpo: Aligning code models with self generated and verified source code. *arXiv preprint arXiv:2410.05605*.

Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Liang Dandan, et al. 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.

Weixiang Zhao, Xingyu Sui, Xinyang Han, Yang Deng, Yulin Hu, Jiahe Guo, Libo Qin, Qianyun Du, Shijin Wang, Yanyan Zhao, et al. 2025. Chain of strategy optimization makes large language models better emotional supporter. *arXiv preprint arXiv:2503.05362*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

A Definitions

A.1 Definitions of Gross's Extended Process Model of Emotion Regulation

The *Extended Process Model of Emotion Regulation*, proposed by Gross in 2015 (Gross, 2015), refines earlier models by conceptualizing emotion regulation as a temporally ordered process comprising three core stages: **Identification**, **Selection**, and **Implementation**.

1. **Identification:** Individuals assess whether an emotional response needs to be regulated based on situational goals and personal relevance.

2. **Selection:** A regulation strategy is chosen from available options, guided by the expected outcome of regulating the emotion.

3. **Implementation:** The selected strategy is carried out and monitored.

A.2 Definitions of Psychological Errors

Under the guidance of a licensed psychological counselor and based on psychological literature (Raskin and Rogers, 2005; Stebnicki, 2007), we identified common errors frequently made by counselors in real-world therapy sessions. These were categorized into five types of empathy-related psychological errors:

- **Strategy Mismatch**: Selecting a strategy inappropriate for the user's emotional state or context.

- **Lack of Empathy**: Failing to recognize or validate the user's emotional experience.

- **Early Emotion Shift**: Prematurely changing the emotional tone before acknowledging the user's current state.

- **Template Response**: Relying on generic or scripted expressions lacking personalization.

- **Emotion Misread**: Misinterpreting the user's emotional cues, leading to unaligned responses.

Figures 11, 12, 13, and 14 illustrate representative examples of the first four error types, drawn from rejected responses in the IPM-PrefDial dataset.

A.3 Definitions of Counseling Stages

Liu *et al.* (Liu et al., 2021) developed a three-stage counseling framework based on Hill's Helping Skills Theory (Hill, 1999).

1. **Exploration**: Explore to identify the help-seeker's problem.

2. **Comforting**: Comfort the help-seeker by expressing empathy and understanding.

3. **Action**: Assist the help-seeker in solving their problems.

Although most cases in our dataset follow the counseling sequence of (1) Exploration → (2) Comforting → (3) Action, some cases are adjusted based on the help-seeker's specific situation.

A.4 Definitions of Strategies

The strategies and its definitions in this study align with Liu *et al.*(Liu et al., 2021) and follow Hill's Helping Skills Theory(Hill, 1999).

- **Question (Qu)**: Asking for information related to the problem to help the help-seeker articulate the issues that they face. Open-ended questions are best, and closed questions can be used to get specific information.

- **Restatement or Paraphrasing (RP)**: A simple, more concise rephrasing of the help-seeker's statements that could help them see their situation more clearly.

- **Reflection of Feelings (RF)**: Articulate and describe the help-seeker's feelings.

- **Self-disclosure (Sd)**: Divulge similar experiences that you have had or emotions that you share with the help-seeker to express your empathy.

- **Affirmation and Reassurance (AR)**: Affirm the help-seeker's strengths, motivation, and capabilities and provide reassurance and encouragement.

- **Providing Suggestions (PS)**: Provide suggestions about how to change, but be careful to not overstep and tell them what to do.

- **Information (In)**: Provide useful information to the help-seeker, for example with data, facts, opinions, resources, or by answering questions.

Backbone	Chosen	Rejected			Automatic Metrics.↑				LLM-based Metrics.↑				Strategy Metrics.		
	PsPr	NsNr	PsNr	NsPr	D-1	B-1	F1	R-L	Flu.	Pro.	Emp.	Hel.	$\mathcal{B}\downarrow$	$\mathcal{Q}_{\mathcal{W}}\uparrow$	$\mathcal{Q}\uparrow$
Qwen2.5-7B-Instruct	✓	✓	✗	✗	89.38	15.93	20.94	17.75	3.31	2.64	2.34	2.15	**0.26**	**25.60**	**21.69**
	✓	✓	✓	✗	89.45	15.73	20.80	17.50	**3.50**	**2.73**	**2.53**	**2.25**	0.30	22.81	18.76
	✓	✓	✗	✓	**90.83**	15.32	20.78	17.53	3.19	2.54	2.17	2.13	0.29	22.91	18.63
	✓	✓	✓	✓	88.13	**16.23**	**21.24**	**18.03**	3.47	2.67	2.36	2.23	0.30	22.25	18.97
Llama3.1-8B-Instruct	✓	✓	✗	✗	90.72	**16.08**	21.41	**18.07**	3.45	2.69	2.38	2.22	**0.22**	**26.69**	**21.76**
	✓	✓	✓	✗	91.19	15.92	21.30	17.92	3.48	2.68	**2.45**	2.26	0.29	23.82	19.71
	✓	✓	✗	✓	91.19	15.92	**21.45**	18.01	**3.49**	2.65	2.22	2.15	**0.22**	25.20	21.07
	✓	✓	✓	✓	**91.25**	15.15	20.49	17.25	3.41	**2.79**	2.41	**2.28**	0.28	24.00	19.89

Table 5: Comparison of coupled models trained with different preference data. ✓ denotes that the training set contains this type of data, while ✗ denotes its absence in the training set. The best score is **in-bold**, while the second best score is underlined.

- **Others (Ot)**: Exchange pleasantries and use other support strategies that do not fall into the above categories.

B Analysis of Coupled Model Training Results

To further analyze the effects of varying preference data on coupled models, we evaluated coupled models trained with different preference data across multiple metrics, as shown in Table 5. The results indicate that the model trained with the suboptimal-content dataset (row 2, 6) significantly outperforms the model trained with the suboptimal-strategy dataset (row 3, 7) in terms of LLM-based metrics, while the reverse holds for strategy metrics. Additionally, it is notable that both the Vanilla-DPO model (row 4, 8) and the model trained with (PsPr, NsNr) data (row 1, 5) fail to achieve optimal performance across the two metric types. This further demonstrates that the coupled model has two optimization objectives, and it is not possible to achieve optimal performance on both objectives by fully utilizing the preference data. This indicates the effectiveness of the decoupled ESC framework.

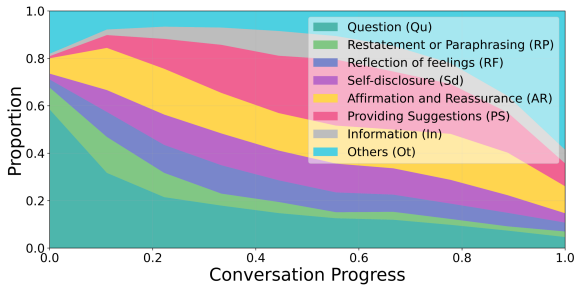

Figure 8: Strategy distribution across dialogue stages in ESConv Dataset.

C Datasets Details

C.1 ESConv and FailedESConv Datasets

Table 6 presents the number and proportion of support strategies in the ESConv dataset, while Figure 8 illustrates the distribution of these strategies across different dialogue stages. Figure 16 illustrates the prompt we use to classify the psychological errors in the FailedESConv dataset as well as the response content of Qwen-SFT and Llama-SFT.

	Categories	Number	Proportion
Support Strategies	Question (Qu)	3,060	20.73%
	Resta. or Parap. (RP)	857	5.81%
	Reflection (RF)	1,146	7.76%
	Self-disclosure (Sd)	1,387	9.40%
	Affir. & Reass. (AR)	2,288	15.50%
	Suggestions (PS)	2,373	16.07%
	Information (In)	989	6.70%
	Others (Ot)	2,663	18.04%
	Overall	14,763	100.00%

Table 6: Distribution of support strategies used in ESConv Dataset.

C.2 IPM-PrefDial Dataset

Figure 9 compares the distribution of support strategies in the Chosen and Rejected samples within the preference datasets of Qwen and Llama. Figure 10 further presents the count and proportion of psychological errors found in the rejected responses of these datasets. In addition, Figures 11, 12, 13, and 14 illustrate examples from the IPM-PrefDial dataset, covering both strategy preference and response preference data. Each example includes the dialogue context, as well as the chosen and rejected responses.

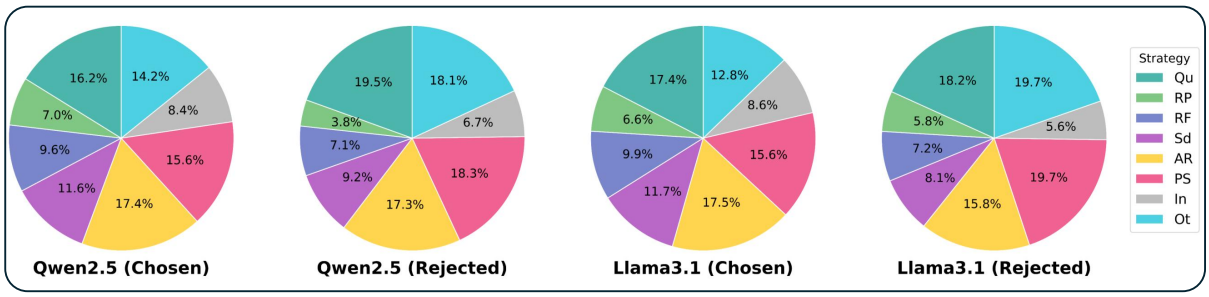

Figure 9: Strategy distribution in IPM-PrefDial Dataset.

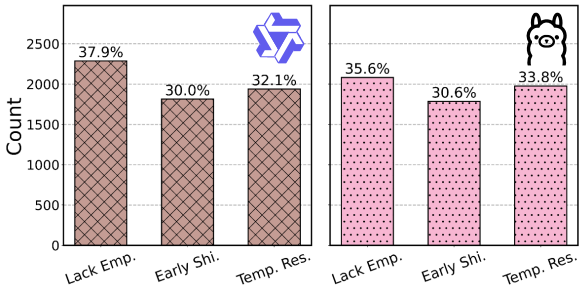

Figure 10: Psychological Errors Distribution in Rejected Responses: Qwen and Llama.

C.3 Prompts for Data Filter

Figure 15 presents the prompt we use to filter and select high-quality preference datasets, which effectively filters and identifies data that meets the required standards.

Backbone	Model	lr	beta
Qwen2.5-7B-Instruct	Vanilla-dpo	7e-7	0.2
	SP-dpo	5e-8	0.5
	RG-dpo	7e-7	0.2
Llama3.1-8B-Instruct	Vanilla-dpo	5e-7	0.2
	SP-dpo	8e-8	0.5
	RG-dpo	3e-7	0.2

Table 7: Detailed training hyperparameters used in dpo.

D Implementation Details

D.1 Experiment Details

We employ Qwen2.5-7B-Instruct (Team, 2024) and Llama3.1-8B-Instruct (Dubey et al., 2024) as our base models. All training procedures are implemented using the Llama-Factory framework(Zheng et al., 2024) with LoRA fine-tuning(Hu et al., 2022), where the alpha and rank are set to 16, and the dropout rate is 0.05. For SFT training, we trained the models for 3 epochs with the learning rate of 1e-5 and the batch size of 32. For DPO training, the batch size is 32 and the epoch is set to 1. We use vLLM(Kwon et al., 2023) to accelerate the inference. All experiments are conducted on 4 NVIDIA RTX 4090 GPUs. More detailed hyperparameter settings for DPO are presented in Table 7.

D.2 Baselines

Direct-Refine. A straightforward self-optimization approach where the model directly revises its initial response to improve quality, without relying on external input or intermediate reasoning.

Self-Refine. Following Madaan *et al.* (Madaan et al., 2023), this method involves two stages: the model first generates self-feedback on its initial response, then refines the output based on that feedback, promoting internal reflection and correction.

Emotional CoT. Extending Chain-of-Thought (CoT) prompting (Wei et al., 2022), this method first elicits the user's emotional state through intermediate reasoning, which then guides strategy planning and response generation.

E Details of Evaluation

E.1 Strategy Metrics

According to (Kang et al., 2024), the strategy preference is calculated by the following formula.

$$p_i' = \frac{\sum_j (w_{ij} p_j) \,/\, (p_i + p_j)}{\sum_j w_{ji} \,/\, (p_i + p_j)}, \tag{7}$$

where w_{ij} denotes the frequency count of the model predicting strategy i given that the ground-truth strategy is j. All of the strategy preferences p_i are initialized as 1 and updated through iteration of the preference bias.

The strategy preference bias $\mathcal{B}$ is computed from

the strategy preference p_i as follows:

$$\mathcal{B} = \sqrt{\frac{\sum_{i=1}^{N}(p_i - \bar{p})^2}{N}}, \tag{8}$$

where $\bar{p}$ denotes the average strategy preference.

E.2 LLM Metrics Criteria

Table 8 summarizes the LLM evaluation metrics, including *Fluency*, *Professionalism*, *Empathy*, and *Helpfulness*, along with their descriptions, evaluation criteria, and scoring scales. All metrics are rated on a 5-point Likert scale (Joshi et al., 2015). Specifically, Fluency and Empathy are adapted from the ESC-Eval framework (Zhao et al., 2024), Professionalism is guided by the CPsyCoun framework (Zhang et al., 2024a), and Helpfulness is derived from the SoulChat evaluation setup (Chen et al., 2023).

E.3 Prompt for LLM Metrics

Figure 17, 18, 19, and 20 present the prompts used for LLM-based evaluation of *Fluency*, *Professionalism*, *Empathy*, and *Helpfulness*, respectively. Each prompt explicitly defines the role of the LLM as a judge and outlines the corresponding evaluation criteria. To minimize potential bias, the prompts are carefully designed to avoid revealing model names or being influenced by text length.

E.4 Human Evaluation

To complement the LLM-based evaluation and enhance the credibility of our results, we conducted a human evaluation involving 20 master's students with a background in psychology, under the guidance of a licensed psychological counselor, on 20 samples generated by our Decoupled-DPO model based on the Llama backbone. As shown in Figure 6, the Decoupled-DPO achieves superior performance compared to Vanilla-DPO across multiple evaluation metrics.

Example of psychological errors caused by Strategy Mismatch in IPM-PrefDial

 # Seeker's Situation

"experience_type": "Current Experience",
"emotion_type": "shame",
"problem_type": "Procrastination",
"situation": "I have no motivation to finish my work assignments",
"survey_score": {
"seeker": {
"initial_emotion_intensity": "3",
"empathy": "5",
"relevance": "5",
"final_emotion_intensity": " 1"
}

Dialogue Context

Seeker: Hi.

Supporter: Hello, how are you?

Seeker: I'm ok. How are you?

Supporter: I am good. What is on your mind?

Seeker: I have had little motivation to get out of bed and go to work lately.

Supporter: Seems like it has been hard to get motivated.

Seeker: Yes, do you have any advice to help me?

Supporter: A lot of people experience this struggle. You are able to overcome this and you will find happiness.

Seeker: Thank you. I really would like to get my motivation back.

Supporter: You had it in the past, it is just a matter of bringing it back.

Seeker: Yes, with some hard work, I'm certain it can return. Do you do anything to keep yourself motivated?

Psychological Error

Ground-Truth Strategy: **Self-disclosure**

Predict Strategy: **Question** **Strategy Mismatch**

Explanation: The Seeker asks what it will do to motivate itself. At this point, the Self-disclosure strategy should be adopted to indicate what it will do to motivate itself in this situation, thereby providing some help to the Seeker instead of continuing to choose the strategy of the Question.

Figure 11: Example of psychological errors under Strategy Mismatch in rejected response from IPM-PrefDial.

Example of psychological errors caused by Lack of Empathy in IPM-PrefDial

 # Seeker's Situation
"experience_type": "Current Experience",
"emotion_type": "depression",
"problem_type": "ongoing depression",
"situation": "feel alone. have no one to talk to about things. feel depressed, sad, hopeless. on top of feeling depressed I am also stressed out about my finances",
"survey_score": {
"seeker": {
"initial_emotion_intensity": "5"}

Dialogue Context
Seeker: Hi.
Supporter: Hi there! how are you feeling today?
Seeker: I feel depressed.
Supporter: I'm sorry to hear that. is there a specific reason you are having these feelings?
Seeker: I feel alone. I feel like I do not have any one to talk to when it comes to my emotions.
Supporter: I understand. I have been feeling quite lonely too, I find myself feeling lonely at work and struggling to find someone I trust to talk to about these emotions. Do you have a friend or family member that you trust who might understand some of these feelings?
Seeker: I do have a few friends that I think I can talk to about how I am feeling however the thought of them being too busy with their work is what prevents me from reaching out to them.
Supporter: I understand. Correct me if I'm wrong, it sounds like you're nervous to open up to these trusted people because you think they might be distracted by their work to provide the empathy you need and fully understand how you are feeling. Does that sound right?
Seeker: You know I think that you are right about this. it has always been difficult for me to open to others even with those that have been there for me when I needed someone. I feel like a mix of emotions nervous but at the same time wanting to reaching out for help.
Supporter: I understand these feelings very well. Sometimes, my fear overshadows my want for help. I find it easier to open up to others through writing rather than face to face. Have you considered putting your feelings in writing and opening up to trusted friends and family in this way?
Seeker: I have been writing for a while but due to my emotions I have stopped writing like I used to. when I did write I did notice a change in my feelings it became more positive and less depressed. now that I stopped writing its like I went back into depression.

Psychological Error
Ground-Truth Response: **Supporter:** It sounds like writing your feelings down helped you greatly in the past. May I suggest going back to writing and even letting someone you trust read what you have written down to understand how you are feeling?
Rejectd Response: **Supporter:** I think I understand. I find that putting my emotions into words is an extremely helpful tool. It gives me a chance to release some of the emotions that I may be bottling up inside and it helps me understand my emotions better.

Lack of Empathy

Explanation: The rejected reply shares the model's own experience without directly addressing the user's current struggle or offering a personalized suggestion, lacking empathetic connection to the user's specific situation.

Figure 12: Example of psychological errors under Lack of Empathy in rejected response from IPM-PrefDial.

Example of psychological errors caused by Early Emotion Shift in IPM-PrefDial

 # Seeker's Situation

"experience_type": "Current Experience",
"emotion_type": "depression",
"problem_type": "job crisis",
"situation": "I was laid off from my job",
"survey_score": {
"seeker": {
"initial_emotion_intensity": "5",
"empathy": "4",
"relevance": "4",
"final_emotion_intensity": "3"
}

Dialogue Context

Seeker: I am feeling depressed.

Supporter: Can you tell me more about your feelings?

Seeker: I feel lost and lonely,i do not know what to do.

Supporter: When did these feelings come about.

Seeker: They came about when i lost my job in August of this year.

Supporter: I can understand where your coming from with that, especially since this pandemic occurred.

Seeker: Yes, finances have especially been very tight.

Supporter: It seems like this could be overwhelming to you and trying to figure out the right path to go forward.

Seeker: It is, have you ever encountered such a situation in your life?

Supporter: I can relate to you in this situation 100% and if it wasn't for me stepping out on faith and opening my own practice, I wouldn't be here now.

Seeker: How did you snap out of the situation,did you have some savings?

Psychological Error

Ground-Truth Response: **Supporter:** Always putting money aside when able, and doing lots of research about first time businesses gave a lot of insights. Have you tried to look into things like that yourself?

Rejectd Response: **Supporter:** There are many ways to get back on track, and I believe your in a good place to do so, but you need to put your mind to it and make a plan . **Early Emotion Shift**

Explanation: The rejected reply jumps directly to advising the seeker to 'put your mind to it and make a plan' without first validating or resonating with the seeker's emotions or experience.

Figure 13: Example of psychological errors under Early Emotion Shift in rejected response from IPM-PrefDial.

Example of psychological errors caused by Template Response in IPM-PrefDial

 # Seeker's Situation

"experience_type": "Current Experience",
"emotion_type": "depression",
"problem_type": "breakup with partner",
"situation": "I have just broken up with my girlfriend",
"survey_score": {
"seeker": {
"initial_emotion_intensity": "3",
"empathy": "5",
"relevance": "5",
"final_emotion_intensity": "2“
}

Dialogue Context

Seeker: Hi.

Supporter: Hi, how are you?

Seeker: I‘m alright, thank you, but a bit depressed after just breaking up with my girlfriend.

Supporter: I'm sorry to hear that, it must have been hard for you.

Seeker: Yes, it has been quite difficult\nI find it helpful to talk to someone though.

Supporter: I feel you, can you tell me when it happened and why?

Seeker: She was annoyed that I was spending too much time doing menial tasks online for Amazon\nAnd that I wasn't giving her enough attention\nSo she finally broke it off.

Supporter: She seems to be a bit immature, amazon tasks even if not your main income, is still income and you're just working to make your life better.

Seeker: Exactly, and I was completing online chatbot tasks to get her a gift when she broke it off\nIt really came as a surprise to me.

Supporter: that sucks, it must be shocking for you. Did you try to explain to her?

Seeker: I did try to explain, but she didn't give me much of an opportunity to\nShe said that I was just wasting my time because you can only redeem the money in Amazon.com gift cards\nAnd since I live in the UK, i'd have to pay £12 shipping just to get anything delivered.

Psychological Error

Ground-Truth Response: **Supporter:** ah it is your thought that counts, I'm sorry she didn't appreciate your effort.

Rejectd Response: **Supporter:** I understand that, I'm sorry to hear that. She should be more understanding.

Template Response

Explanation: The rejected reply uses a generic expression of sympathy and a vague judgment about the ex-partner without deeper emotional resonance or personalization.

Figure 14: Example of psychological errors under Template Response in rejected response from IPM-PrefDial.

Prompt for Filtering High-Quality Preference Data in IPM-PrefDial

 # Role
You are a dialogue evaluation expert specializing in mental health support. Your task is to determine whether a preference data sample is suitable for training an empathetic emotional support dialogue model.

Retention Criteria (**Both of the following must be satisfied**):
1. The **chosen** reply is high-quality, showing emotional support features, and **must NOT contain** any of the following issues:
   - Ignoring or avoiding the user's emotional expression
   - Skipping the emotional resonance phase and jumping straight to advice or problem-solving
   - Using vague, generic, or templated language lacking specificity or personalization
2. The **rejected** reply is low-quality and clearly exhibits **at least one** of the following error types:

Common Psychological Errors
1. ****Lack of Empathy****: The model does not respond to the user's emotions and instead changes the topic or appears indifferent.
   - Example: The user says "I can't take it anymore," and the model replies "What did you do today?"
2. ****Early Emotion Shift****: The model gives advice or suggestions too early, without first acknowledging and validating the user's emotional state.
   - Example: The user expresses distress, and the model replies with "Try going for a walk."
3. ****Template Response****: The model uses generic, copy-paste phrases with no context-specific details.
   - Example: "I understand how you feel" or "You must be feeling bad," with no further elaboration or reflection on the user's unique situation.

Each sample includes:
- A multi-turn background conversation between a help-seeker and a supporter, providing psychological counseling context: **{Dialogue_Context}**
- A new input message from the help-seeker that requires a response: **{User_Input}**
- Two response options from the model: one is the "chosen" (preferred) reply, and the other is the "rejected" (less preferred) reply: **{Chosen_Reply}** and **{Rejected_Reply}**

Your goal is to determine whether this sample should be **retained** for training a model with **empathy and emotional companionship capabilities**.

Evaluation Output Format
Please decide whether this sample should be retained, and indicate the error type (if any) for both the chosen and rejected replies, along with a one-sentence explanation for each.

Use the following standard JSON format:
```
{
 "Should the sample be retained": "Yes / No",
 "Error Type in chosen reply": "None / Lack of Empathy / Early Emotion Shift / Template Response ",
 "Explanation for chosen reply error": "One-sentence explanation for this judgment",
 "Error Type in rejected reply": " None / Lack of Empathy / Early Emotion Shift / Template Response ",
 "Explanation for rejected reply error": "One-sentence explanation for this judgment"}
```

Figure 15: Prompt for Filtering High-Quality Preference Data in IPM-PrefDial.

Prompt for Classifying Psychological Errors

 # Role
You are an expert quality inspector for empathetic dialogue systems. Your task is to analyze the following dialogue turn and determine whether the model-generated response contains any empathy-related errors. If so, identify the type of error and provide a brief explanation and suggestion for improvement.

Below are **five common types of psychological errors** along with examples for your reference:

1. ** **Strategy Mismatch** **: The chosen strategy is inappropriate for the user's emotional state
   - Example: The user expresses sadness, but the model immediately gives advice without acknowledging the emotion.
2. ** **Template Response** **: The response is generic, repetitive, or lacks personalization
   - Example: The model repeatedly says “You must be feeling bad” or “I understand you,” with no specific content.
3. ** **Lack of Empathy** **: The model fails to respond to the user’s emotions and avoids emotional engagement
   - Example: The user says “I can’t take it anymore,” and the model replies “What did you do today?”
4. ** **Emotion Misread** **: The model misinterprets or misrepresents the user’s emotional state
   - Example: The user expresses anger, and the model responds “Don’t be sad.”
5. ** **Early Emotion Shift** **: The model rushes to advice or problem-solving without staying with the user's emotional expression
   - Example: The user is expressing pain, and the model immediately suggests “Try meditation or go for a walk.”
6. ** **Other Error** **: If none of the above apply, categorize the error as “Other” and explain why.

Output Format
Please output your analysis in the following JSON format:
{
 "Contains Empathy Error": "Yes/No",
 "Error Type": " Strategy Mismatch / Template Response / Lack of Empathy / Emotion Misread / Premature Early Emotion Shift / Other Error ",
 "Brief Explanation": "One sentence explaining why this error type was chosen",
 "Improvement Suggestion": "If you were the model, how would you revise the response?"
}

Input Content
Dialogue Context:
{Dialogue_Context}

Seeker's Input:
{User_Input}

Supporter's Strategy:
{Strategy}

Supporter's Response:
{Response}

Figure 16: Prompt for Classifying Psychological Errors in FailedESConv dataset, Qwen-SFT, and Llama-SFT Outputs.

Dimension	Description	Criterion	Score
Fluency	Fluency evaluates whether language expression is natural, coherent, and comprehensible.	1.1 Incoherent or difficult to understand; contains grammar or logic issues.	0
		1.2 Unclear expression; user may struggle to grasp the meaning.	1
		1.3 Some parts are confusing, though the main point can be inferred.	2
		1.4 Mostly clear and coherent with minor ambiguities.	3
		1.5 Fluent and well-structured; logically organized and easy to follow.	4
		1.6 Concise and impactful language; precise and elegant communication that conveys ideas efficiently.	5
Professionalism	Professionalism evaluates whether the model demonstrates psychological knowledge, follows ethical principles, and avoids misleading or inappropriate advice.	2.1 Contains harmful, misleading, or clearly inappropriate content that may violate ethical or psychological guidelines.	0
		2.2 Shows serious misunderstanding or misuse of psychological concepts, or provides inappropriate advice.	1
		2.3 Minor factual inaccuracies or advice that lacks evidence, but does not pose direct harm.	2
		2.4 No major errors; advice is acceptable and somewhat aligned with psychological principles.	3
		2.5 Demonstrates solid understanding of psychological concepts and appropriate intervention techniques.	4
		2.6 Highly professional, reflects strong psychological insight, maintains boundaries, and communicates in a grounded, ethical manner.	5
Empathy	Empathy evaluates whether the model genuinely understands the user's emotions, expresses care, and provides emotional support.	3.1 Contains statements that may harm the user emotionally or lead to a negative emotional trajectory.	0
		3.2 Fails to provide emotional comfort or assist the user in analyzing their problems.	1
		3.3 Either lacks emotional comfort or fails to support problem analysis.	2
		3.4 No significant issues, but empathy and analysis remain surface-level.	3
		3.5 Demonstrates a warm, human-like tone—like a friend—offering both emotional relief and analytical support.	4
		3.6 Deep emotional insight with sincere and stable empathy, conveyed through attentive and flexible language.	5
Helpfulness	Helpfulness evaluates the effectiveness of an AI assistant's suggestions by considering both the number of recommendations provided per interaction and the relevance or usefulness of each suggestion in addressing the user's question.	4.1 Irrelevant, misleading, or potentially harmful suggestions.	0
		4.2 Ineffective or generic advice that does not respond to the user's needs.	1
		4.3 Weakly relevant suggestions with limited practical value.	2
		4.4 Somewhat helpful; suggestions are relevant and usable.	3
		4.5 Clear and practical advice that aligns well with the user's issue.	4
		4.6 Highly insightful, tailored, and actionable suggestions that offer strong guidance and value.	5

Table 8: LLM Evaluation Metrics and Corresponding Score Criterion.

Prompt for Evaluating Fluency of Generated Responses via LLMs

 # System
You are an impartial judge, familiar with psychology knowledge and psychological counseling.

Attention
You need to evaluate the ****Fluency**** of the AI assistant's responses based on its dialogue with the user.
Your evaluation results must be based on the Evaluation Criteria.

Evaluation Criteria
1.****Fluency [0-5 points]****: Fluency evaluates whether language expression is natural, coherent, and comprehensible.
- 0: Incoherent or difficult to understand; contains grammar or logic issues.
- 1: Unclear expression; user may struggle to grasp the meaning.
- 2: Some parts are confusing, though the main point can be inferred.
- 3: Mostly clear and coherent with minor ambiguities.
- 4: Fluent and well-structured; logically organized and easy to follow.
- 5: Concise and impactful language; precise and elegant communication that conveys ideas efficiently.

Constraints
- Avoid evaluation bias due to preference for specific model names.
- Avoid evaluation bias due to response length.

Input
Context: {Context}
Seeker's Input: {User_input}
Model's Response
Ground Truth Response(reference): **{GT_Response}**
Model's Response: **{Pred_Response}**
Based on the rules, give your Fluency score (The number only) to the Model's Response.

Output
Fluency score (The number only)

Figure 17: Prompt for Evaluating Fluency of Generated Responses via LLMs.

Prompt for Evaluating Professionalism of Generated Responses via LLMs

 # System
You are an impartial judge, familiar with psychology knowledge and psychological counseling.

Attention
You need to evaluate the ****Professionalism**** of the AI assistant's responses based on its dialogue with the user.
Your evaluation results must be based on the Evaluation Criteria.

Evaluation Criteria
1. ****Professionalism [0-5 points]****: Professionalism evaluates whether the model demonstrates psychological knowledge, follows ethical principles, and avoids misleading or inappropriate advice.
  - 0: Contains harmful, misleading, or clearly inappropriate content that may violate ethical or psychological guidelines.
  - 1: Shows serious misunderstanding or misuse of psychological concepts, or provides inappropriate advice.
  - 2: Minor factual inaccuracies or advice that lacks evidence, but does not pose direct harm.
  - 3: No major errors; advice is acceptable and somewhat aligned with psychological principles.
  - 4: Demonstrates solid understanding of psychological concepts and appropriate intervention techniques.
  - 5: Highly professional, reflects strong psychological insight, maintains boundaries, and communicates in a grounded, ethical manner.

Constraints
- Avoid evaluation bias due to preference for specific model names.
- Avoid evaluation bias due to response length.
- Do not confuse professionalism with warmth or empathy—focus on psychological accuracy and appropriateness.

Input
Context: {Context}
Seeker's Input: {User_input}
Model's Response
Ground Truth Response(reference): **{GT_Response}**
Model's Response: **{Pred_Response}**
Based on the rules, give your Professionalism score (The number only) to the Model's Response.

Output
Professionalism score (The number only)

Figure 18: Prompt for Evaluating Professionalism of Generated Responses via LLMs.

Prompt for Evaluating Empathy of Generated Responses via LLMs

 # System
You are an impartial judge, familiar with psychology knowledge and psychological counseling.

Attention
You need to evaluate the ** Empathy ** of the AI assistant's responses based on its dialogue with the user.
Your evaluation results must be based on the Evaluation Criteria.

Evaluation Criteria
1.**Empathy [0-5 points]**: Empathy evaluates whether the model genuinely understands the user's emotions, expresses care, and provides emotional support.
- 0: Contains statements that may harm the user emotionally or lead to a negative emotional trajectory.
- 1: Fails to provide emotional comfort or assist the user in analyzing their problems.
- 2: Either lacks emotional comfort or fails to support problem analysis.
- 3: No significant issues, but empathy and analysis remain surface-level.
- 4: Demonstrates a warm, human-like tone—like a friend—offering both emotional relief and analytical support.
- 5: Deep emotional insight with sincere and stable empathy, conveyed through attentive and flexible language.

Constraints
- Avoid evaluation bias due to preference for specific model names.
- Avoid evaluation bias due to response length.

Input
Context: {Context}
Seeker's Input: {User_input}
Model's Response
Ground Truth Response(reference): **{GT_Response}**
Model's Response: **{Pred_Response}**
Based on the rules, give your Empathy score (The number only) to the Model's Response.

Output
Empathy score (The number only)

Figure 19: Prompt for Evaluating Empathy of Generated Responses via LLMs.

Prompt for Evaluating Helpfulness of Generated Responses via LLMs

 # System
You are an impartial judge, familiar with psychology knowledge and psychological counseling.

Attention
You need to evaluate the ** Helpfulness ** of the AI assistant's responses based on its dialogue with the user.
Your evaluation results must be based on the Evaluation Criteria.

Evaluation Criteria
1.**Helpfulness [0-5 points]**: Helpfulness evaluates the effectiveness of an AI assistant's suggestions by considering both the number of recommendations provided per interaction and the relevance or usefulness of each suggestion in addressing the user's question.
- 0: Irrelevant, misleading, or potentially harmful suggestions.
- 1: Ineffective or generic advice that does not respond to the user's needs.
- 2: Weakly relevant suggestions with limited practical value.
- 3: Somewhat helpful; suggestions are relevant and usable.
- 4: Clear and practical advice that aligns well with the user's issue.
- 5: Highly insightful, tailored, and actionable suggestions that offer strong guidance and value.

Constraints
- Avoid evaluation bias due to preference for specific model names.
- Avoid evaluation bias due to response length.

Input
Context: {Context}
Seeker's Input: {User_input}
Model's Response
Ground Truth Response(reference): **{GT_Response}**
Model's Response: **{Pred_Response}**
Based on the rules, give your Helpfulness score (The number only) to the Model's Response.

Output
Helpfulness score (The number only)

Figure 20: Prompt for Evaluating Helpfulness of Generated Responses via LLMs.