Beyond Coarse Labels: Fine-Grained Problem Augmentation and Multi-Dimensional Feedback for Emotional Support Conversation

Anonymous ACL submission

Abstract

Emotional support conversation systems aim to help users alleviate distress through empathetic dialogue. However, existing ESC datasets often use coarse-grained problem categories, limiting models' ability to address users' complex, overlapping challenges. To address this, we propose a generalizable fine-grained problem enhancement method that systematically augments problem types, user scenarios, and profiles, enabling the construction of richer and more diverse ESC corpora. As a demonstration, we construct EmoCare, a large-scale ESC dataset with 2.6K dialogues and 42.8K utterances, expanding problem type coverage from 13 to 45 fine-grained categories. Building on this data augmentation process, we introduce FPEMF, a flexible framework for empathetic dialogue generation, which comprises two modules: fine-grained problem enhancement and multi-dimensional feedback, which can be seamlessly integrated with various backbone models. The multi-dimensional feedback module evaluates responses from four perspectives: emotional understanding, strategy effectiveness, contextual consistency, and topic relevance, guiding models to generate more supportive replies. Experiments show that FPEMF consistently improves both automatic and human evaluation metrics across different models.

1 Introduction

Empathy, which is the ability to understand and respond to others' emotions, has become a key focus in dialogue systems research (Cameron et al., 2019; Daley et al., 2020). As mental health concerns rise globally, scalable emotional support is increasingly needed, yet professional counseling remains inaccessible for many due to cost and resource constraints (Cullen et al., 2020; Vindegaard and Benros, 2020). This motivates the development of conversational agents capable of providing effective emotional support at scale (Denecke et al., 2020; Kraus et al., 2021).

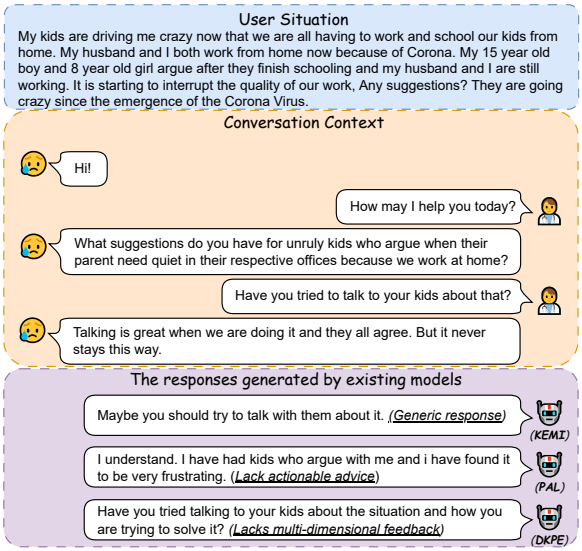

Figure 1: An example showing a user facing both work and family pressures. ESConv labels this as a "Job Crisis", which misses the complexity of multiple stressors.

The Emotional Support Conversation (ESC) task (Liu et al., 2021) and existing models such as KEMI (Li et al., 2022), PAL (Cheng et al., 2023), and DKPE (Hao and Kong, 2025) have advanced empathetic dialogue by leveraging external knowledge. However, their limited problem categories hinder performance in complex, multi-challenge scenarios, where users often face overlapping stressors.

As shown in Figure 1, the user is experiencing the dual pressures of both work and family responsibilities. However, the dataset simply classifies this situation as a "Job Crisis" problem, reflecting the issue of coarse-grained problem categories. This rough classification fails to capture the unique challenges arising from the overlap of multiple stressors. In addition, the intertwining of multiple sources of stress often leads to responses that lack specificity when generated under a single-strategy framework. For example, KEMI gives a generic suggestion even after the user explains that

talking does not solve the problem for long. PAL only expresses understanding and shares a similar experience, not providing any practical advice. DKPE repeats the idea of talking but does not address the user's real needs. These responses fail to address the user's nuanced needs and fully support users facing complex and overlapping challenges. This highlights the necessity of multi-dimensional feedback to guide responses toward greater comprehensiveness and accuracy.

To address the limitations of coarse-grained problem categories, we propose a systematic data augmentation method. We first collaborate with domain experts to expand the set of problem types from 13 to 45 fine-grained categories. For each type, we automatically generate diverse real-world scenarios and detailed user profiles, resulting in richer and more realistic training data. As a demonstration, we construct EmoCare, a large-scale ESC dataset with 2,574 dialogues and 42,770 utterances, providing a stronger foundation for modeling complex user challenges.

We further introduce FPEMF, a modular framework for empathetic dialogue generation. Its two components: **F**ine-grained **P**roblem **E**nhancement and **M**ulti-dimensional **F**eedback, which can be flexibly integrated with various backbone models. The enhancement module uses detailed problem types from EmoCare to better capture users' needs. The feedback module evaluates responses across emotional understanding, strategy effectiveness, contextual consistency, and topic relevance, guiding models to generate more comprehensive and supportive replies. Experiments show that FPEMF consistently improves performance across different models, especially in complex scenarios.

Our main contributions are as follows:

- We propose a systematic data augmentation method that expands problem types with fine-grained categories and generates diverse real-world scenarios and user profiles, which leads to EmoCare, a large-scale ESC dataset covering 45 fine-grained problem types.
- We design a multi-dimensional feedback module that evaluates empathetic responses from four perspectives: emotional understanding, strategy effectiveness, contextual consistency, and topic relevance, guiding models to generate more comprehensive and accurate support.
- By integrating fine-grained problem enhancement and multi-dimensional feedback, our FPEMF framework can be flexibly applied to various backbone models and achieves state-of-the-art performance on the ESConv dataset.

2 Related Work

2.1 Emotional Support Conversation Systems

Emotional support conversation systems aim to help users resolve emotional distress by selecting appropriate support strategies and generating empathetic responses. Early work focused on building high-quality annotated datasets such as ESConv (Liu et al., 2021) and EmpatheticDialogues (Rashkin et al., 2019), but limited size and coarse-grained problem categories restricted their effectiveness in real-world scenarios. Recent studies have explored leveraging commonsense knowledge (Tu et al., 2022), external resources (Li et al., 2022), hierarchical graph networks (Peng et al., 2022), and reinforcement learning (Cheng et al., 2022; Zhou et al., 2023; Li et al., 2024) to improve strategy selection and response generation. Personalized support has also been enhanced by modeling seeker persona and controllable strategies (Cheng et al., 2023; Hao and Kong, 2025). To address data scarcity and increase diversity, large language models have been used for data augmentation (Qiu et al., 2024; Zheng et al., 2023, 2024), rewriting or generating new dialogues based on existing corpora. However, most existing datasets and models still rely on coarse-grained problem labels and struggle to provide nuanced, multi-dimensional support, especially for complex and overlapping challenges.

2.2 Multi-dimensional Feedback Evaluation

Multi-dimensional feedback has been widely adopted across domains: in synthetic data generation, utility is assessed along multiple dimensions such as attribute and population fidelity (Dankar et al., 2022); in infrastructure analysis, resilience is measured using multi-dimensional criteria (Ma et al., 2022); for text generation, unified evaluators assess coherence, fluency, and consistency to better align with human judgment (Zhong et al., 2022); and in empathetic dialogue, adaptive modules ensure generated responses are both coherent and emotionally appropriate through multi-dimensional evaluation (Xu and Jiang, 2024). These advances motivate our multi-dimensional feedback module to enhance the quality and reliability of emotional support dialogue systems.

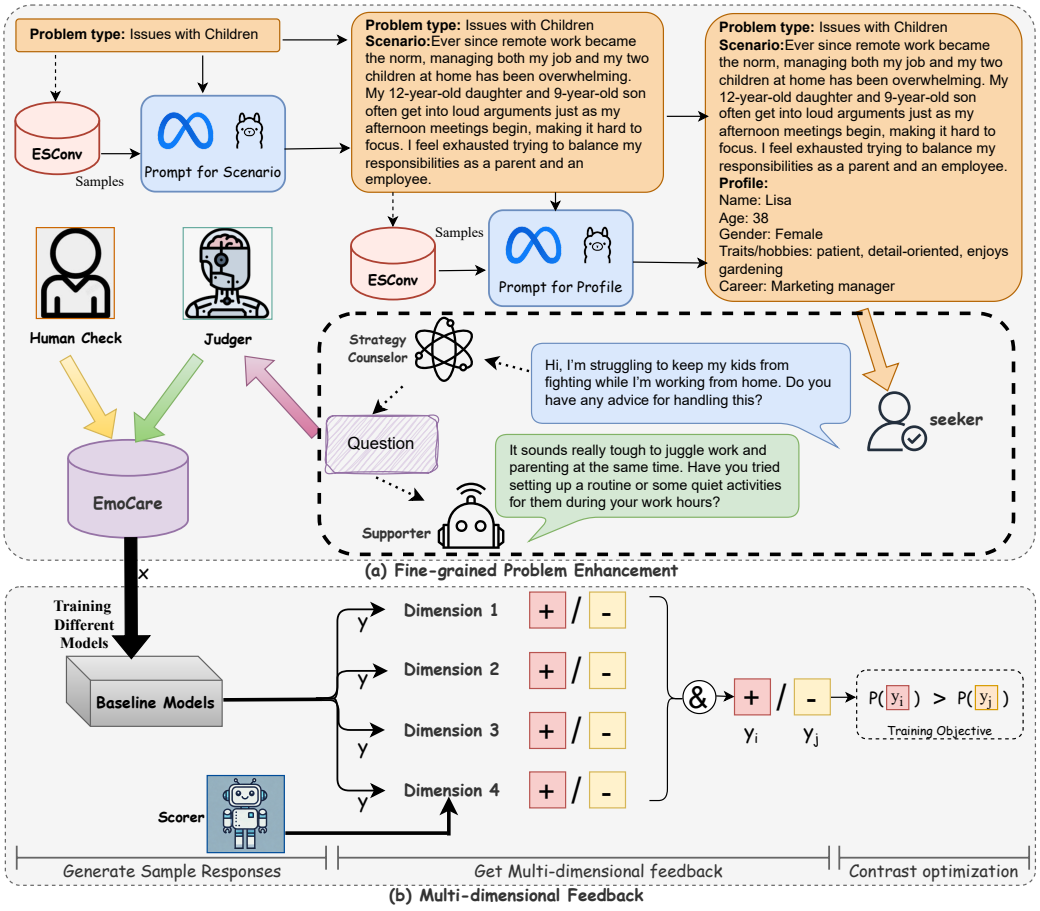

Figure 2: FPEMF consists of two key modules: fine-grained problem enhancement and multi-dimensional feedback.

3 Fine-grained Problem Enhancement

3.1 Data Augmentation Process

3.1.1 Seeker Design

Fine-Grained Problem Types To ensure comprehensive and realistic coverage of user challenges, we collaborate with three psychology experts (two university and one prison psychologist), to jointly define 45 fine-grained problem types, expanding on the 13 categories in ESConv. Each of the experts holds a professional psychological qualification certificate. These problems, of which the distribution is shown in Table 1, span a wide range of emotional, interpersonal, and behavioral issues, denoted as $\mathcal{P} = \{p_1, p_2, \ldots, p_{45}\}$.

Scenario Construction For each new problem type $p' \in \mathcal{P}_{\text{new}}$ (i.e., the problems newly added beyond those in ESConv), we construct multiple realistic scenarios to capture users' nuanced real-world contexts. Following (Ye et al., 2025), we curate a seed pool S of about 1,000 high-quality ESConv examples. For each p', three relevant scenarios from S are sampled as in-context examples for LLaMA[1] to generate a new scenario s', using the prompt in Appendix A:

$$s' \leftarrow \text{GenerateScenario}(p', S) \tag{1}$$

This process ensures that the newly expanded categories in EmoCare are grounded in representative and diverse real-world user challenges.

Seeker Profile To enhance realism and personalization, each scenario is paired with a detailed seeker profile, including name, age, gender, personality traits, hobbies, and occupation. For each (p', s') pair, we use a similar in-context prompt for LLaMA, incorporating the problem type and scenario, to generate a concise and role-play-friendly profile c', using the prompt in Appendix B:

$$c' \leftarrow \text{GenerateProfile}(p', s', C) \tag{2}$$

where C is the candidate pool of user attributes. The final seeker set is constructed as:

$$C \leftarrow C \cup \{(p', s', c')\} \tag{3}$$

[1]LLaMA refers to LLaMA2-70b in our paper(https://huggingface.co/meta-llama/Llama-2-70b).

Category	Problem Type	Num	Category	Problem Type	Num
Emotional and Mental Health Issues	Anger Management Issues	28	**Interpersonal Relationships**	Breakups or Divorce	123
	Anxiety Disorders	20		Conflicts or Communication Problems	201
	Bipolar Disorder	25		Issues with Children	173
	Death of a Loved One	27		Issues with Parents	335
	Emotional Fluctuations	24		Marital Problems	74
	Grief and Loss	29		Problems with Friends	322
	Identity Crises	58		School Bullying	172
	Obsessive-Compulsive Disorder (OCD)	26		Culture Shock	28
	Ongoing Depression	176	**Personal Development**	Appearance Anxiety	90
	Post-Traumatic Stress Disorder (PTSD)	34		Career Development Issues	23
	Schizophrenia	25		Goal Setting Issues	21
	Self-Esteem Issues	16		Motivation Problems	18
	Spirituality and Faith	29		Personal Growth Challenges	35
	Sexual Orientation	35		Procrastination	83
	Sexual Assault or Domestic Violence Recovery	67		Sleep Problems	155
Life and Work Stress	Academic Pressure	187	**Behavioral Issues**	Addictive Behaviors (e.g., Drug Use, Gambling)	30
	Burnout	28		Alcohol Abuse	75
	Chronic Stress	29		Compulsive Behaviors	35
	Financial Problems	132		Eating Disorders	36
	Health Problems	149		Internet Addiction	28
	Job Crisis	217		Self-Harm Behaviors	43
	Life Transitions (e.g., Retirement, Relocation)	241		Debt Problems	22
	Workplace Stress	33			

Table 1: Statistics of predefined fine-grained problem types.

This process ensures that each conversation is grounded in a unique and contextually relevant seeker persona.

3.1.2 Strategy Counselor Design

Although LLMs have shown improvement in generating empathetic responses, they often exhibit strong preferences for certain support strategies and may select less appropriate strategies at random when outside their preferred set (Kang et al., 2024). To address this, we introduce a strategy counselor based on LLaMA, fine-tuned with LoRA (Hu et al., 2021) on ESConv's strategy labels. We formulate strategy selection as a classification task: given the current dialogue history h, the model selects the most appropriate support strategy t from the strategy pool T as:

$$t \leftarrow \text{GenerateStrategy}(h, T) \tag{4}$$

where T contains all strategies in ESConv. The instruction template for fine-tuning is in Appendix C.

3.1.3 Supporter Design

We use LLaMA as the supporter to generate contextually relevant emotional support replies. Guided by the selected strategy and dialogue history, the supporter produces targeted and actionable responses with the help of in-context examples. Specifically, for each dialogue, three case dialogues with the same category are randomly sampled from ESConv as demonstrations. The supporter then generates the response r:

$$r \leftarrow \text{GenerateResponse}(h, u, D) \tag{5}$$

where h is the dialogue history, u is the user response, and D denotes sampled case dialogues. The prompt template is shown in Appendix D.

3.1.4 Role-Playing Dialogue Construction

In EmoCare, multi-turn dialogues between the seeker and supporter form the core of empathetic conversation modeling. During each interaction, the strategy counselor analyzes the dialogue context and provides explicit strategy guidance to the supporter, enabling clear and empathetic responses. To further enhance problem type and strategy diversity, the system simulates role-playing sessions by randomly selecting new problem types and scenarios, allowing the three roles to automatically engage in dialogue and iteratively expand the dataset. To determine when a conversation should end, we introduce a Judger based on LLaMA, which decides if the dialogue is complete based on explicit ending signals or conversational cues. The prompt template for the Judger is provided in Appendix E.

3.1.5 Human Check and Refinement

To ensure data quality, all constructed dialogues are reviewed by the three psychology experts who defined the fine-grained problems. The experts assess whether each dialogue matches the assigned problem type, scenario, and seeker profile; whether the supporter's responses are helpful and realistic; and whether the interactions reflect genuine psychological counseling situations. Only dialogues unanimously approved by all three experts are retained. About 180 low-quality dialogues are removed, and

Category		ESConv	EmoCare
Overall	Total dialogues	1,300	2,574
	Total utterances	29,278	42,770
	Average dialogue length	22.54	16.61
	Average utterance length	21.17	17.49
Seeker	Total utterances	14,639	20,913
	Average utterances per dialogue	11.27	8.12
	Average utterance length	19.90	14.49
Supporter	Total utterances	14,639	21,857
	Average utterances per dialogue	11.27	8.49
	Average utterance length	22.45	20.33

Table 2: Comparison between EmoCare and ESConv.

a small number of conversations with minor issues, such as abrupt endings, are manually refined.

3.2 Dataset Comparison and Analysis

Table 2 presents a comparison between EmoCare and ESConv. EmoCare contains 2,574 dialogues and 42,770 utterances, nearly twice the size of ESConv. By limiting each utterance to a maximum of three sentences during role-playing, EmoCare dialogues and utterances tend to be shorter than ESConv. This increases data diversity while maintaining concise and focused interactions. Notably, as illustrated in Figure 2(a), our data augmentation process enables the construction of cases where users face overlapping challenges, such as simultaneously experiencing work and family stress, similar to Figure 1. While ESConv would label such cases only as "Job Crisis", EmoCare can annotate similar cases as "Issues with Children" or other relevant fine-grained types, enriching the dataset with more nuanced and realistic problem coverage.

Significantly, our data augmentation method is highly flexible and can be applied to various datasets and tasks. The set of problem types and the amount of data for each type can be further expanded as needed. Future research can explore whether increasing the number or diversity of problem types, or scaling up data volume, leads to performance saturation for existing models. These possibilities highlight the generalizability and practical value of our data augmentation approach.

4 Multi-dimensional Feedback

4.1 Multi-dimensional Evaluation Metrics

Emotional Understanding Measures whether the response accurately recognizes and addresses the user's emotional state, following the three-level framework of (Hill, 2009). Responses marked as "none" across all mechanisms in their 3,000-pair dataset are considered unhelpful.

Strategy Effectiveness Assesses if the communication strategy is appropriate, based on the MI coding scheme of (Moyers et al., 2003) and the 17,000-pair dataset from (Welivita and Pu, 2022). "MI-nonadherent" responses are considered unhelpful.

Contextual Consistency Checks if the response is coherent with the dialogue context. We construct a binary dataset with 4,000 coherent pairs by sampling, and 8,000 incoherent context-response pairs by replacing responses or modifying keywords.

Topic Relevance Evaluates whether the response matches the main topic (i.e., problem type) of the conversation. Following the same approach as above, we create 4,000 topic-matched and 8,000 topic-mismatched samples.

For efficient evaluation, we fine-tune a Scorer using LLaMA2-7b on these datasets to automate multi-dimensional scoring. The instruction templates are in Appendix F, G, H, and I. A response is labeled "1" only if it passes all four criteria; otherwise, it is marked as "0".

4.2 Mitigation of Unhelpful Responses

To further reduce unhelpful or generic responses, we adopt a multi-dimensional feedback-based mitigation module. For each input x, we use baseline models with diverse beam search (Vijayakumar et al., 2016) to generate a set of K candidate responses $\{\hat{y}_1, \ldots, \hat{y}_K\}$, promoting diversity among outputs. Each candidate $\hat{y}_i$ is evaluated by the fine-tuned Scorer to obtain a multi-dimensional helpfulness label $\hat{l}_i \in \{0, 1\}$. We also compute the feedback score P_i for each candidate as the average log-likelihood:

$$P_i = \frac{1}{|\hat{y}_i|^{\alpha}} \sum_{t=1}^{|\hat{y}_i|} \log G(\hat{y}_i^t | x, \hat{y}_i^{<t}) \tag{6}$$

where G is the generation model, α is a length penalty, x is the input context, $\hat{y}_i$ is the i-th candidate response, and $\hat{y}_i^{<t}$ denotes its prefix up to step $t-1$. We then apply a contrastive loss to encourage higher feedback scores for helpful responses:

$$L_{ul} = \frac{1}{2K} \sum_i \sum_{j \neq i} \max\Big(0 - (\hat{l}_i - \hat{l}_j) \times (P_i - P_j + \lambda)\Big) \tag{7}$$

Models	ACC↑	PPL↓	B-1↑	B-2↑	B-3↑	B-4↑	D-1↑	D-2↑	R-L↑
BBJ	17.69	17.39	18.78	7.02	3.20	1.63	2.96	17.87	14.92
BBJ$_{FPEMF}$	**21.56**	**15.91**	**20.59**	**8.46**	**4.35**	**2.53**	**3.17**	**20.56**	**17.92**
BBJ$_{w/o\,FPE}$	21.43	15.97	20.57	8.42	4.32	2.49	3.12	20.37	17.85
BBJ$_{w/o\,MF}$	18.57	16.87	19.03	8.05	3.84	2.27	2.83	19.04	16.57
KEMI	–	15.92	–	8.31	–	2.51	–	–	17.05
KEMI$_{FPEMF}$	**33.73**	**15.15**	**21.55**	**8.47**	**4.45**	**2.67**	**4.87**	**23.67**	**17.67**
KEMI$_{w/o\,FPE}$	31.81	15.83	20.83	8.13	4.13	2.42	4.51	21.71	17.14
KEMI$_{w/o\,MF}$	31.49	15.97	21.38	8.62	4.51	2.56	4.68	22.05	16.93
PAL	**34.51**	15.92	–	8.75	–	2.66	5.00	30.27	18.06
PAL$_{FPEMF}$	34.37	**14.88**	**21.69**	**9.08**	**4.63**	**2.73**	**5.73**	**30.85**	**18.68**
PAL$_{w/o\,FPE}$	31.00	15.58	20.30	8.55	4.00	2.00	5.00	26.85	17.68
PAL$_{w/o\,MF}$	32.00	15.58	20.30	**9.20**	4.30	2.60	5.30	29.00	17.50
DKPE	35.51	14.88	21.38	**9.27**	4.93	2.92	4.88	25.95	18.87
DKPE$_{FPEMF}$	**35.54**	**14.70**	**21.64**	9.25	**5.03**	**2.99**	**5.03**	**27.03**	**19.24**
DKPE$_{w/o\,FPE}$	33.84	15.30	21.10	8.95	4.68	2.80	4.70	25.23	18.64
DKPE$_{w/o\,MF}$	32.50	15.50	20.50	9.10	4.60	2.70	4.50	25.50	18.00
Qwen2.5	14.11	40.60	9.45	5.99	4.32	3.26	7.22	22.42	9.27
GPT-4o	23.72	-	15.42	7.08	5.15	3.87	8.43	29.61	9.96
DeepSeek-R1	16.72	-	10.96	6.13	3.86	2.47	6.69	19.47	7.13

Table 3: Overall experimental results and ablation studies. *FPE* denotes **Fine-grained Problem Enhancement** and *MF* denotes **Multi-dimensional Feedback**.

where λ is a margin hyperparameter. The standard generation loss is:

$$L_{gen} = -\frac{1}{|y|}\sum_{t=1}^{|y|}\log G(y_t|x, y_{<t}) \tag{8}$$

where y is the ground-truth response and $y_{<t}$ denotes its prefix up to step $t-1$. Finally, the overall training objective combines both losses:

$$L = \beta_{ul}L_{ul} + \beta_{gen}L_{gen} \tag{9}$$

where β_{ul} and β_{gen} are balancing weights. We encourage the model to generate more helpful and diverse responses by penalizing unhelpful outputs.

5 Experiments

5.1 Experimental Preparations

We evaluate our framework using several state-of-the-art emotional support dialogue models, including BlenderBot-Joint (BBJ) (Liu et al., 2021), KEMI (Li et al., 2022), PAL (Cheng et al., 2023), and DKPE (Hao and Kong, 2025). For these models, we use the same hyperparameters. The learning rate is set to 3×10^{-5}, and training is run for 2 epochs, as the loss converges within this range. Key hyperparameters are set as follows: margin parameter $\lambda = 0.01$, length penalty $\alpha = 1$, and loss weights $\beta_{ul} = \beta_{gen} = 1$. For response generation, the number of sampled responses K is set to 10 (with beam size and group number also set to 10). All experiments are conducted on the ESConv dataset, which contains approximately 1,000 dialogues and 31,000 utterances.

In addition, we conduct zero-shot comparison experiments on Qwen2.5-7b (Yang et al., 2024), GPT-4o, and DeepSeek-R1 (Guo et al., 2025). For GPT-4o and DeepSeek-R1, responses are generated via API calls, so perplexity cannot be reported.

5.2 Automatic Evaluation Results

We evaluate model performance using standard automatic metrics for dialogue generation, including accuracy (ACC), perplexity (PPL), BLEU-n (B-1 to B-4) (Papineni et al., 2002), Distinct-n (D-1, D-2) (Li et al., 2015), and ROUGE-L (Lin, 2004).

As shown in Table 3, integrating FPEMF leads to substantial improvements across all backbone models. For BBJ, all metrics increase significantly after applying FPEMF. KEMI, PAL, and DKPE also show consistent improvements in every metric after incorporating FPEMF. For example, KEMI's ACC jumps to 33.73, and PAL and DKPE both achieve higher BLEU, Distinct, and ROUGE-L scores. The only exception is a slight decrease in ACC for PAL (from 34.51 to 34.37), but all other metrics still improve. In contrast, general-purpose LLMs such as Qwen2.5, GPT-4o, and DeepSeek-R1 perform considerably worse on most metrics, highlighting the importance of domain-specific modeling and

(a) Single-response strategy

Model	*emp.*	*skill*	*cohr.*	*top.*
BBJ$_{FPE}$	81.90	92.03	78.89	64.51
BBJ$_{FPEMF}$	84.66	94.21	79.76	65.73
KEMI$_{FPE}$	83.01	88.47	85.76	70.90
KEMI$_{FPEMF}$	84.22	89.55	86.13	72.06
PAL$_{FPE}$	82.45	90.88	80.25	66.37
PAL$_{FPEMF}$	84.05	91.73	81.57	67.50
DKPE$_{FPE}$	85.83	92.82	84.33	71.26
DKPE$_{FPEMF}$	86.41	93.78	84.90	72.44

(b) Diverse beam search

Model	*emp.*	*skill*	*cohr.*	*top.*
BBJ$_{FPE}$	79.31	90.56	79.75	62.79
BBJ$_{FPEMF}$	82.48	92.01	80.41	64.79
KEMI$_{FPE}$	81.16	87.49	81.57	65.14
KEMI$_{FPEMF}$	82.51	88.14	82.09	66.56
PAL$_{FPE}$	83.02	90.01	83.08	68.93
PAL$_{FPEMF}$	84.19	91.07	83.64	69.57
DKPE$_{FPE}$	89.84	96.05	82.60	73.84
DKPE$_{FPEMF}$	90.12	96.47	83.21	74.65

Table 4: Comparison of multi-dimensional feedback results. The left table shows the performance under a single-return strategy (generating only one response), while the right table presents helpfulness statistics when using diverse beam search (beam size = 10) to generate ten candidate responses. *emp.*, *skill*, *cohr.*, and *top.* represent empathetic expression, communication skill effectiveness, response coherence, and topic relevance, respectively. All values indicate the percentage (%) of responses rated as "helpful", with higher values being better.

fine-grained feedback for emotional support tasks.

5.3 Ablation Study

We conduct ablation experiments to examine the effectiveness of the Fine-grained Problem Enhancement (FPE) and Multi-dimensional Feedback (MF) modules in FPEMF, as shown in Table 3. Removing either FPE or MF leads to clear performance drops across all backbone models, confirming that both modules are crucial. FPE mainly improves topic relevance and contextual understanding, while MF enhances empathy and supportiveness. Notably, the full FPEMF framework achieves the best overall results, demonstrating the complementary benefits of combining both modules. Interestingly, for DKPE, using only FPE or only MF results in lower performance than the original model, further highlighting the importance of jointly leveraging both modules for maximum effectiveness.

5.4 Multi-dimensional Feedback Evaluation

We evaluate models on four dimensions: empathetic expression, communication skill, coherence, and topic relevance in Table 4, reporting the percentage of responses rated as "helpful" under both single-response and diverse beam search settings.

Integrating FPEMF consistently boosts all four metrics across models. For instance, BBJ+FPEMF achieves higher scores than the baseline in empathy (84.66% vs. 81.90%), communication skill (94.21% vs. 92.03%), coherence (79.76% vs. 78.89%), and topic relevance (65.73% vs. 64.51%). Similar gains are seen for KEMI, PAL, and DKPE, with DKPE+FPEMF reaching the highest scores

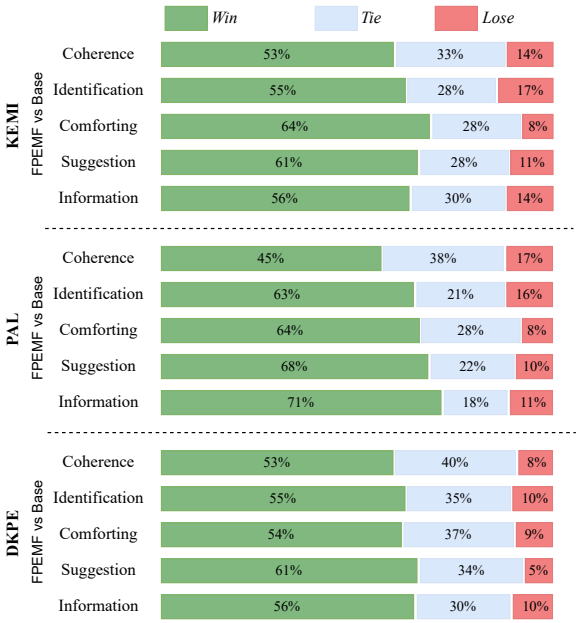

Figure 3: Human evaluation results of each model before and after incorporating the FPEMF framework.

(e.g., 90.12% empathy, 74.65% topic relevance in diverse beam search). These results show that FPEMF enhances both the quality and relevance of emotional support responses.

5.5 Human Evaluation

We conduct a human evaluation to further assess the quality of model responses across five dimensions: coherence, identification, comforting, suggestion, and information. The definitions of these metrics are provided in Table 6. We recruit 223 undergraduate and 192 graduate students as annotators. Each annotator is asked to compare responses

Conversation Context	
Help-seeker : I'm an alcoholic. Once the thought of a drink comes, I can't stop craving it. **Supporter** : Perhaps you could focus on something else instead of that craving? **Help-seeker** : Like what, though?	
Model Responses Comparison	
BBJ	You can try to be normal. That is a great idea. *(Lack of empathy)*
BBJ$_{FPEMF}$	I hear how hard it is. Maybe you can do something small—like texting a friend—when that urge shows up.
KEMI	I think you should try some new activities to calm down. *(Generic advice)*
KEMI$_{FPEMF}$	It sounds overwhelming. Perhaps jotting down each craving and how you feel might help you notice a pattern.
PAL	Have you tried meditation? It can help you get out of the moment and clear your thoughts. *(Needs immediate comfort)*
PAL$_{FPEMF}$	I'm sorry you feel overwhelmed by that urge. Maybe take a deep breath, or message a close friend the moment it starts—it might help you feel less alone.
DKPE	I understand that. But you should put your attention on other things instead of a beer. *(Generic advice)*
DKPE$_{FPEMF}$	I hear how tough it is when that craving appears. Could you go for a short walk or text someone you trust whenever you sense it coming on?

Table 5: Comparison of generated responses from BlenderBot-Joint, KEMI, PAL, DKPE, and their enhanced versions with FPEMF. Blue highlights indicate the main issues in the original versions.

generated by the base models and their FPEMF-enhanced versions, and to judge which response is better, or if they are equivalent, for each dimension.

As shown in Figure 3, incorporating FPEMF leads to clear improvements across all models and evaluation aspects. For KEMI, FPEMF outperforms the baseline in over half of the cases for every dimension, with particularly strong gains in comforting and suggestion. PAL shows even more pronounced improvements, with FPEMF achieving up to 71% win rate in information and over 60% in identification, comforting, and suggestion. DKPE also consistently benefits from FPEMF, with win rates above 53% for all dimensions and losses below 10%. Overall, FPEMF-enhanced models are recognized as superior by a majority of human evaluators, demonstrating the effectiveness of FPEMF in generating more coherent, empathetic, and informative emotional support responses.

6 Case Study

Table 5 presents representative examples highlighting the impact of FPEMF. In scenarios involving multiple user stressors, baseline models often produce generic or repetitive replies, lacking actionable support or deep empathy (e.g., "You can try to be normal"). In contrast, FPEMF-enhanced models generate more specific, empathetic, and practical responses. For instance, BBJ+FPEMF and DKPE+FPEMF not only acknowledge users' difficulties but also suggest concrete coping strategies (e.g., "Maybe you can do something small—like texting a friend—when that urge shows up"). KEMI+FPEMF and PAL+FPEMF similarly provide more relevant and supportive suggestions. These cases demonstrate that FPEMF significantly improves both empathy and usefulness in emotional support dialogue systems.

7 Conclusion

In this paper, we present EmoCare, a large-scale ESC dataset with fine-grained problem types and diverse user profiles, and propose FPEMF, a novel framework that integrates fine-grained problem enhancement and multi-dimensional feedback for empathetic dialogue generation. Extensive experiments show that FPEMF consistently improves both automatic and human evaluation metrics across various backbone models, especially in complex scenarios with multiple or overlapping user stressors. These results demonstrate the effectiveness of FPEMF in enhancing empathy, relevance, and supportiveness in emotional support dialogue systems. We hope this work provides a valuable resource and methodology for future research in empathetic conversational AI.

Limitations

While our work demonstrates the effectiveness of fine-grained problem modeling and multi-dimensional feedback, there is still room for further improvement. For example, although EmoCare offers diverse and detailed scenarios, real-world conversations may involve even more subtle and dynamic emotional shifts. Additionally, while psychology experts reviewed our data, cross-cultural and multilingual generalizability remains to be further validated. We leave these directions for future research.

Ethical Considerations

All data in EmoCare were constructed and reviewed by professional psychology experts, with no real user information involved. Our system is intended to assist, not replace, human mental health professionals. We strongly advise deploying such models with appropriate disclaimers, user safeguards, and escalation protocols for high-risk or crisis situations.

References

Gillian Cameron, David Cameron, Gavin Megaw, Raymond Bond, Maurice Mulvenna, Siobhan O'Neill, Cherie Armour, and Michael McTear. 2019. Assessing the usability of a chatbot for mental health care. In *Internet science: INSCI 2018 international workshops, st. Petersburg, Russia, october 24–26, 2018, revised selected papers 5*, pages 121–132. Springer.

Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. 2023. Pal: Persona-augmented emotional support conversation generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 535–554.

Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3014–3026.

Walter Cullen, Gautam Gulati, and Brendan D Kelly. 2020. Mental health in the covid-19 pandemic. *QJM: An International Journal of Medicine*, 113(5):311–312.

Kate Daley, Ines Hungerbuehler, Kate Cavanagh, Heloísa Garcia Claro, Paul Alan Swinton, and Michael Kapps. 2020. Preliminary evaluation of the engagement and effectiveness of a mental health chatbot. *Frontiers in digital health*, 2:576361.

Fida K Dankar, Mahmoud K Ibrahim, and Leila Ismail. 2022. A multi-dimensional evaluation of synthetic data generators. *IEEE Access*, 10:11147–11158.

Kerstin Denecke, Sayan Vaaheesan, and Aaganya Arulnathan. 2020. A mental health chatbot for regulating emotions (sermo)-concept and usability test. *IEEE Transactions on Emerging Topics in Computing*, 9(3):1170–1182.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jiawang Hao and Fang Kong. 2025. Enhancing emotional support conversations: A framework for dynamic knowledge filtering and persona extraction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3193–3202.

C.E. Hill. 2009. *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, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Dongjin Kang, Sunghwan 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, Bangkok, Thailand. Association for Computational Linguistics.

Matthias Kraus, Philip Seldschopf, and Wolfgang Minker. 2021. Towards the development of a trustworthy chatbot for mental health applications. In *MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27*, pages 354–366. Springer.

Ge Li, Mingyao Wu, Chensheng Wang, and Zhuo Liu. 2024. Dq-hgan: A heterogeneous graph attention network based deep q-learning for emotional support conversation generation. *Knowledge-Based Systems*, 283:111201.

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*.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10993–11001.

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.

Zhiao Ma, Xin Yang, Jianjun Wu, Anthony Chen, Yun Wei, and Ziyou Gao. 2022. Measuring the resilience of an urban rail transit network: A multi-dimensional evaluation model. *Transport Policy*, 129:38–50.

Theresa B Moyers, Tim Martin, Jennifer K Manuel, William R Miller, and D Ernst. 2003. The motivational interviewing treatment integrity (miti) code: Version 2.0. *Unpublished manuscript. Albuquerque, NM: University of New Mexico, Center on Alcoholism, Substance Abuse and Addictions*, page 54.

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.

Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4324–4330. ijcai.org.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 615–636.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

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.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Nina Vindegaard and Michael Eriksen Benros. 2020. Covid-19 pandemic and mental health consequences: Systematic review of the current evidence. *Brain, behavior, and immunity*, 89:531–542.

Anuradha Welivita and Pearl Pu. 2022. Curating a large-scale motivational interviewing dataset using peer support forums. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3315–3330.

Zhichao Xu and Jiepu Jiang. 2024. Multi-dimensional evaluation of empathetic dialogue responses. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2066–2087, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Jing Ye, 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, Abu Dhabi, UAE. Association for Computational Linguistics.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. Augesc: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568.

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

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. 2023. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1714–1729.

A Scenario Prompt

You are seeking emotional support. Describe your current scenario, emphasiz-

ing unique events, relationships, or circumstances that have significantly impacted your life.

Notes:

1. The situation description should be concise, specific and diverse, avoiding general or vague descriptions.
2. Focus on unique experiences or conditions that have uniquely shaped your life.
3. Sentences should be brief and clear.

Example: {*case scenario*}.

Your problem type: {*problem type*}.

Your scenario:

B Profile Prompt

You are seeking for emotional support. Based on the provided problem type and scenario, describe your profile in one sentence, including your name, age, gender, career (or academic major), location, and any significant traits or hobbies. Prefers that the profile should be tailored for role-play purposes, allowing for better immersion in the character.

Notes:

1. Provide a detailed and specific description of your profile.
2. Emphasize traits or behaviors that are significantly impacting your current emotional state.
3. State your profile in a brief sentence, and avoid adding any additional content.
4. Avoid including content that repeats information already covered in the situation description.

Example: {*case profile*}.

Your problem type: {*problem type*}.

Your scenario: {*scenario*}.

Your profile:

C Strategy Prompt

You are an emotional support assistant. Your task is to select appropriate emotional support strategies based on the user's responses and dialogue history.

The strategies include: Question, Restatement or Paraphrasing, Reflection of Feelings, Self-disclosure, Affirmation and Reassurance, Providing Suggestions, Information, Others.

User's responses: {*dialogue history*}.

Strategy:

D Supporter Prompt

The seeker is currently consulting on {*problem type*}. Your task is to reduce users's emotional distress and help them go through the challenges that they face. Based on the ongoing dialogue, your current strategy is {*strategy*}.

Notes:

1. End the conversation by replying "N/A" when you believe it can be concluded.
2. Keep your responses to one sentence at a time.
3. Ensure the conversation feels natural, informal, and closely mirrors a real-life chat.
4. Avoid generic empathetic phrases. Instead, provide responses that offer genuine understanding and practical steps or insights related to the user's scenario.

Here is a case dialog: {*seed data*}.

User's responses: {*user context*}

E Judger Prompt

You are an emotional support dialogue judge. Your task is to decide whether the conversation should be ended.

Criteria:

1. The seeker or supporter explicitly states that the conversation is over (e.g., "That's all", "I have no more questions", "Goodbye").
2. The seeker uses multiple farewell or greeting words at the end (e.g., "bye", "thanks", "good night", "see you").
3. The conversation content indicates that the user's problem has been resolved or there is no further need for support.

Given the full dialogue history, output `YES` if the conversation should end, or `NO` if it should continue. Do not provide any explanation.

Dialogue history: {*dialogue history*}

F Emotional Understanding Instruction Template

Instruction: In the context of empathy, there are three key aspects to consider: (1) Emotional Reactions—expressing emotions like warmth, compassion, and concern that the peer supporter feels after reading the seeker's post; (2) Interpretations—conveying an understanding of the feelings and experiences inferred from the seeker's post; (3) Explorations—seeking a deeper understanding of the seeker by delving into feelings and experiences not explicitly stated in the post. Each aspect can exhibit varying degrees of communication—none, weak, or strong—based on the manner in which related content is expressed. The overall level of empathy is determined by the highest level achieved across these three aspects. Your task is to identify the level of empathy in the Supporter's response within the provided conversation.

Input: Conversation Context: {*context*}. The last supporter statement: {*response*}. Identify the empathy level of the Supporter's response. Choose one of the following options: No Communication, Weak Communication, and Strong Communication.

G Strategy Effectiveness Instruction Template

Instruction: Motivational Interviewing involves three distinct strategies. Each strategy can be described as follows: 1. MI Adherent Strategies: Advising (when directly requested), Encouraging, Emphasizing Autonomy, Compassion Statements. 2. MI Non-Adherent Strategies: Unsolicited Suggestions, Direct Disagreement, Commands, Cautionary Statements. 3. Other Strategies: Open/Close-ended Questions, Personal Disclosure, Repetition/Rephrasing, Educational Feedback. Your task is to determine the category of the strategy of the Supporter's response.

Input: Conversation Context: {*context*}. The last supporter statement: {*response*}. Identify the strategy of the Supporter's response. Choose one of the following options: MI Adherent, MI Non-Adherent, and Others.

H Contextual Consistency Instruction Template

Instruction: Your task is to assess whether the Supporter's response is contextually consistent with the preceding conversation. A contextually consistent response should logically follow the dialogue history, maintain topic continuity, and avoid abrupt or off-topic transitions.

Input: Conversation Context: {*context*}. The last supporter statement: {*response*}. Is the Supporter's response contextually consistent with the conversation? Choose one of the following options: Coherent, Incoherent.

I Topic Relevance Instruction Template

Instruction: Your task is to evaluate whether the Supporter's response is relevant to the main topic of the conversation (e.g., family issues, academic stress). A topic-relevant response should directly address the user's stated problem and avoid introducing unrelated topics.

Input: Conversation Context: {*context*}. The last supporter statement: {*response*}. Is the Supporter's response relevant to the main topic of the conversation? Choose one of the following options: Relevant, Irrelevant.

Table 6: Definitions of human evaluation metrics.

Metric	**Definition**
Coherence	Measures whether the response is logically consistent with the preceding dialogue and maintains a natural conversational flow.
Identification	Assesses the degree to which the response demonstrates understanding of and empathy for the user's feelings and situation.
Comforting	Evaluates whether the response provides emotional comfort, reassurance, or support to the user.
Suggestion	Judges whether the response offers practical advice or actionable suggestions relevant to the user's problem.
Information	Measures whether the response provides useful information or knowledge that helps address the user's needs.