I Don't Need Solution. I Need Emotional Support : Empathetic LLMs based on Emotional Validation

Suhyune Son[1], Jungwoo Lim[1], Myunghoon Kang[1], Seongtae Hong[1], Yuna Hur[2], Evelyn Hayoon Zi[1], Heuiseok Lim[1†],

[1]Department of Computer Science and Engineering, Korea University,
[2]Human-Inspired AI Research

{ssh5131, wjddn803, chaos8527, ghdchlwls123, yj72722, evehy, limhseok}@korea.ac.kr

Abstract

Empathy plays a crucial role in prosocial behavior and supportive human interactions. According to emotional validation theory, effective empathetic conversations require observing and reflecting on the help-seeker's situation before offering emotional support and guidance. While recent advancements in large language models (LLMs) have enabled fluent and coherent dialogue generation, our preliminary study reveals that existing LLMs struggle to generate emotional support response. Instead, they tend to offer repetitive solutions without sufficiently considering the emotional needs of help-seekers. To address this limitation, we propose **EVA**: empathetic LLMs with **E**motional **VA**lidation. EVA enhances empathetic response generation through a two-stage training process: empathy acquisition and emotional validation alignment. For the emotional validation alignment, we introduce the Emotional Validation Aware Dataset (EVAD), which is annotated with levels of emotional validation theory as conversations progress. Additionally, we propose EVAEval, a novel evaluation metric designed to assess whether a model-generated response aligns with emotional validation theory. Experimental results demonstrate that the EVA method significantly improves empathetic response generation, achieving superior performance in both automatic and human evaluations. Furthermore, comprehensive analyses confirm that the EVA method effectively mitigates patterned responses while ensuring adherence to emotional validation principles.

1 Introduction

Empathy plays a crucial role in human psychological processes, facilitating smooth social interactions (Decety, 2010). Responding with empathy in human interactiona through caring and sympathetic concern facilitates prosocial behavior and enhances social bonding (Eisenberg and Eggum, 2009). Moreover, empathy drives the formation of constructive interpersonal and supportive relationships (Reynolds and Scott, 1999), including counseling (Anthony, 1971), psychotherapy (Mitchell and Berenson, 1970), and human relations (Gazda et al., 1977). Among the helping relationships, counseling for mental health care is pivotal since it alleviates emotional distress, anxiety, and psychological concerns at the individuals, groups, organizations, and society (Richards, 2009). Despite the importance of empathy in human dialogue, there are hurdles to an empathetic conversation, such as reluctance to open one's heart or talk freely about one's tough situations at any time (Kang et al., 2024; Burleson, 2003). Therefore, various AI-driven services for empathetic communication, such as virtual therapy or chatbot systems, have emerged (Cameron et al., 2017; Trappey et al., 2022; Qiu and Lan, 2024).

Owing to the advancements in large language models (LLMs) (Han et al., 2021; Bommasani et al., 2021) which have demonstrated exceptional generalization capabilities across various language generation tasks, recent works have applied LLMs to the empathetic dialogue generation (Qian et al., 2023; Chen et al., 2023; Yang et al., 2024b). However, as highlighted in prior studies (Zhao et al., 2024; Kidder et al., 2024; Formanek and Sotolar, 2024), LLMs often exhibit limited empathetic abilities. As shown in Figure 1, the LLM repeatedly suggests possible solutions without adjusting its level of empathy, rather than concerning the help-seeker's emotional status regardless of the progression of the dialogue. Concannon et al. (2023a) highlight that repetitive responses from an empathetic dialogue model can worsen the emotional distress of help-seekers. Therefore, displaying varying levels of empathy is crucial in empathetic dialogues to effectively alleviate the help-seeker's distress (Tao et al., 2024). Achieving this requires

†Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 1–26
July 2-7, 2026

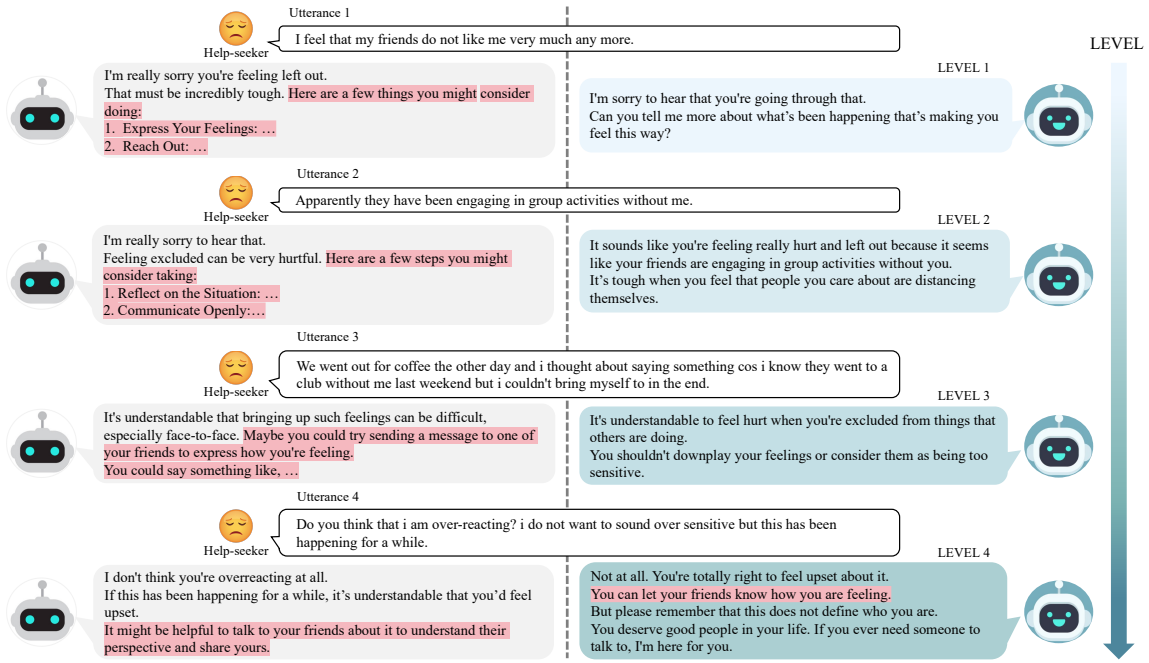


Figure 1: Comparison of responses generated by Mistral (left) and our proposed EVA$_{Mistral}$ (right) to the help-seeker's utterances. Highlighted boxes indicate suggested solutions. The arrows on the right side represent the levels of emotional validation.

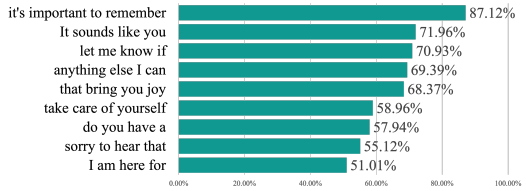
not only recognizing the expressed emotions but also affirming and validating the help-seeker's feelings. Emotional validation strengthens empathy by ensuring that the help-seeker feels genuinely heard, understood, and accepted, which fosters trust and encourages deeper emotional disclosure. By incorporating emotional validation, a dialogue system can dynamically adjust its responses to the evolving emotional context, enabling more supportive interactions (Linehan, 1997).

In this work, we propose an **EVA**: empathetic LLMs with **E**motional **VA**lidation, an effective training method for building an empathetic language model powered by LLM. Inspired by dialectical behavior therapy (DBT) (Linehan, 1997), we introduce emotional validation theory, which defines the principles an empathetic language model should follow. The emotional validation consists of four hierarchical levels: LEVEL 1: Listening and Observing, LEVEL 2: Accurate Reflection, LEVEL 3: Validating, and LEVEL 4: Radical Genuineness. Each level of emotional validation is related to the depth of the dialogue. For instance, as illustrated in Figure 1, EVA-based model tries to perceive the help-seeker's status by asking for a detailed situation (LEVEL 1). Then EVA-based model reflects the help-seeker's emotion by summarizing the emotional transition (LEVEL 2). After that, the model sympathizes with the help-seeker's situation with a comforting response (LEVEL 3), suggesting related action that could relieve the help-seeker's emotional distress (LEVEL 4).

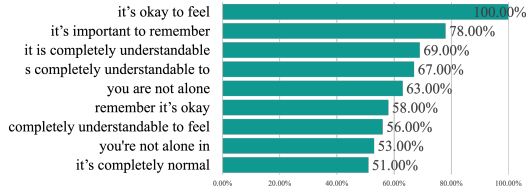
To inject emotional validation theory into the language model, we construct Emotional Validation Aware Dataset (EVAD) and align models with EVAD to generate proper emotional supportive response regarding the progression of the dialogue. We observe that our proposed EVA-based models demonstrate significant improvements compared to the baseline models on general and empathetic dialogue generation metrics. Moreover, we propose a simple yet effective metric, EVAEval, which evaluates whether responses are generated according to emotional validation theory.

The main contributions of this work are summarized as follows: (1) We propose EVA, a novel training framework that enhances empathetic response generation by aligning LLMs with emotional validation theory. (2) We demonstrate that EVA consistently outperforms baseline LLMs across both automatic and human evaluation metrics. (3) We conduct comprehensive qualitative analyses to explain EVA's effectiveness, including studies on response diversity, bias in solution suggestions, and emotional validation behavior. [1].

[1]Our source code is publicly available at



(a) Mistral



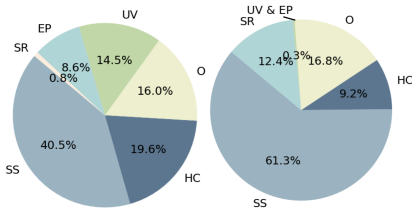
(b) GPT-4o

Figure 2: The proportion of top-9 distinct 4-gram phrases appearing in the responses of Mistral and GPT-4o.

2 Preliminary Studies

In this section, we demonstrate the limitations of the empathetic conversation ability of previous LLMs. First, we prompt Mistral-7B-Instruct (Jiang et al., 2023) and GPT-4o (Achiam et al., 2023) using the instruction shown in Table 3 on 100 dialogues from the ESConv test set (Liu et al., 2021). To evaluate the LLM's empathetic capabilities, we recruited 15 crowd workers without specific psychological expertise to assess whether the responses from Mistral and GPT-4o effectively comforted the help-seeker. As a result, 87.82% of the responses are evaluated as "does not feel comfortable"[2]. This indicates a significant issue with the LLM's empathetic response generation capability. Based on human preference results, we assume that the results are attributed to the limitations of LLMs across two aspects: patterned responses and bias on solution suggestions.

Patterned Responses We first extract the top-9 distinct 4-gram [3] from Mistral and GPT-4o-generated utterances. We then compute the frequency of these phrases across dialogues. As shown in Figure 2, several phrases appear repeatedly across the conversations. This frequent repetition suggests that the model relies on a narrow set of expressions, which may limit its ability to provide emotional support.



(a) Mistral (b) GPT-4o

Figure 3: Proportions of response types in utterances generated by Mistral and GPT-4o on the ESConv dataset.

Biased to Solution Suggestion We further analyze the bias in LLM-generated responses by classifying each utterance according to the empathetic response types defined in Appendix D. As shown in Figure 3, the Suggesting Solution (SS) type constitutes the largest proportion of responses, accounting for 40.5% and 61.3%. This high frequency of solution-suggesting responses indicates a significant bias in the models' empathetic dialogue generation, where the model tends to offer advice or solutions over other types of empathetic responses. Moreover, the Understanding and Validation (UV) type, which conveys emotional validation to the help-seeker, was observed only in a small fraction of the responses. This low percentage highlights a critical shortcoming in the models' ability to recognize and validate the help-seeker's emotions. Overall, the biased distribution of response types identified in this preliminary study explains why human evaluators perceived the models' responses as lacking empathy.

Such repetitive patterns and the tendency to overuse certain response types in LLM-generated utterance can lead to negative feelings among help-seekers. Prior studies (Concannon et al., 2023b; Mohan et al., 2025) have shown that formulaic and repetitive responses from AI systems can induce dissatisfaction, frustration, and even anxiety in users. By introducing emotional validation, we aim to mitigate this issue and provide more adaptive, emotionally supportive interactions.

3 Emotional Validation

To address the observed limitations of existing LLMs in empathetic response generation, we ground our approach in emotional validation theory from Dialectical Behavior Therapy (DBT) (Line-

https://github.com/sonsuhyune/EVA.

[2]The inter-annotator agreement, measured by Fleiss' Kappa, was 0.69, indicating substantial agreement among the evaluators.

[3]We conducted 2- to 6-gram analyses and found that 4-gram phrases most clearly capture characteristic patterns in the model's responses.

LEVEL 1. Listening and Observing – The supporter listens actively without expressing their own thoughts or giving advice.
"I'm sorry to hear that you're going through that."
LEVEL 2. Accurate Reflection – Empathy is expressed by paraphrasing the help-seeker's situation and emotions.
"It sounds like you're feeling really down and discouraged because of the interview results."
LEVEL 3. Validating – The supporter confirms that the help-seeker's feelings are valid and natural to reduce emotional distress.
"Separation can be incredibly difficult, especially when children are involved. It's completely normal to feel disgusted and frustrated with the situation."
LEVEL 4. Radical Genuineness – The supporter acknowledges the help-seeker's inherent worth and shared human experience to facilitate self-validation.
"It's important to prioritize your well-being. If this job is taking a toll on you, consider exploring other career options that align with your values and skills. In the meantime, find ways to decompress, like engaging in hobbies or spending time with loved ones. Remember, it's okay to put yourself first."

Table 1: Descriptions and example utterances for each level of emotional validation.

han, 1997). DBT is a cognitive-behavioural treatment originally developed for individuals with severe emotional dysregulation. A central component of DBT is emotional validation, which involves communicating that another person's feelings, thoughts, and behaviours are understandable and meaningful in light of their situation. Importantly, validation does not necessarily imply agreement; rather, it acknowledges that the person's internal experience makes sense in context.

In DBT, validation is conceptualized as a hierarchical process that progresses from careful attention to more explicit forms of acknowledgment and affirmation. However, the original DBT framework was developed for therapeutic settings that assume richer contextual knowledge, long-term interpersonal rapport, and multimodal cues. These assumptions do not hold in open-domain, text-based dialogue with LLMs.

We therefore adapt DBT's validation principles into four operational levels for supportive multi-turn conversations: (LEVEL 1) Listening and Observing, (LEVEL 2) Accurate Reflection, (LEVEL 3) Validating, and (LEVEL 4) Radical Genuineness. This adaptation enables us to model how emotional support should evolve as dialogue progresses, from understanding the help-seeker's situation to explicitly validating and supportively engaging with their emotions. Table 1 presents the four levels together with their definitions and representative example utterances.

4 Method

In this section, we propose **EVA**: empathetic LLMs with **E**motional **VA**lidation, and its training process. EVA consists of two steps, 1) Empathy Acquisition (ACQ) and 2) Emotional Validation Alignment (ALI). Finally, we evaluate the model's comprehension of emotional validation using EVAEval, a metric that quantifies the distribution of emotional levels within the dialogue.

4.1 EVA

4.1.1 Empathy Acquisition

To let the language model acquire general empathetic generation capability, we fine-tune the model with existing ESC dataset in a supervised manner. We formulate an empathetic dialogue generation task as a multi-turn conversation. The set of dialogues is denoted as $D = \{d_1, ..., d_n\}$, and each dialogue d is converted into $H_m = \{(k_1, s_1), ..., (k_m, s_m)\}$ where H represents the dialogue history, k is the help-seeker's utterance, and s is the supporter's response. Note that m is the number of rounds of dialogue. The final objective is to generate a set of $S_m = \{s_1, ..., s_j, ...s_m\}$. The language modeling loss is computed based on the supporter's responses s given the dialogue history H, and the model is trained to minimize this loss.

4.1.2 Emotional Validation Alignment

Afterward, we advise the model to learn emotional validation by training the human-preference dataset. Despite the explicit supervision on downstream tasks utilizing the empathetic dialogue generation dataset, solely employing the fine-tuned LLM is inadequate (Zheng et al., 2022). Hence, an additional training approach for LLM's empathetic generation is required.

To address this, we adopt direct preference optimization (DPO) (Rafailov et al., 2024), which aligns the model's outputs with human preferences by training on pairs of preferred and rejected responses. To align the model with emotional validation, we use our curated dataset of triplets—(dialogue context, chosen response, rejected response)—annotated according to emotional validation criteria.

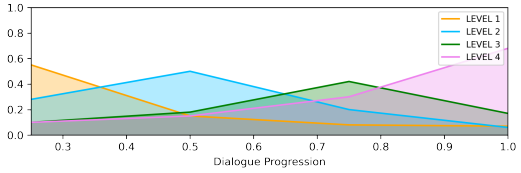


Figure 4: The distribution of chosen responses at each level within each quartile in Emotional Validation Aware Dataset (EVAD).

During training, DPO encourages the model to increase the relative likelihood of preferred responses over rejected ones, compared to the original fine-tuned model. This enables the model to generate emotionally aligned and context-sensitive responses, while avoiding generic or inappropriate outputs—all without relying on reinforcement learning techniques.

EVAD: Emotional Validation Aware Dataset For emotional validation alignment, we construct a human preference dataset, EVAD. Based on the ESConv dataset, we prompt ChatGPT (GPT-3.5-TURBO) (OpenAI-Blog, 2022) to generate utterances at each level of emotional validation given dialogue history and explanation of emotional validation theory[4]. The most preferred responses are labeled as 'chosen', otherwise 'rejected' by human workers[5]. Then, the generated utterances consist of a set of annotated human preferences: chosen and rejected responses. Importantly, all generated responses were evaluated and filtered by human annotators based on emotional validation theory to ensure quality. This procedure allows efficient data augmentation while maintaining alignment with human-centered empathetic principles.

To analyze level distribution of chosen responses over dialogue progression, we illustrate the number of utterances on each level onto the relative position in the dialogue. As in Figure 4, LEVEL 1 appears more frequently at the beginning, while LEVEL 4 increases toward the end. This result indicates that the chosen responses align with the concept of emotional validation as the dialogue progresses.

4.2 EVAEval

To evaluate whether the model follows the emotional validation, we propose a simple yet effective metric, EVAEval. Unlike existing studies that evaluate utterances individually (Sharma et al., 2020), EVAEval assesses the degree of emotional validation across entire multi-turn conversations. We first label the level of each model-generated utterance by using a pre-trained classifier[6].

When the label is assigned, we obtain the EVAEval score for each level l. We calculate the relative position of utterances of level l by normalizing the position with the length of the dialogue. Then, the relative position score is categorized into four quartiles to assess how the levels of emotional validation evolve as the dialogue progresses. To measure the alignment between the model's output distribution $\mathcal{T}_q$ and the distribution of EVAD, as shown in Figure 4, $\mathcal{V}_q$, we compute the squared difference for each quartile. Consequently, EVAEval of level l is calculated as following:

$$\text{EVAEval}^l = \frac{1}{4}\sum_{q=1}^{4}(\mathcal{T}_q^l - \mathcal{V}_q^l)^2 \quad (1)$$

To assess the validity of EVAEval, we measure the Spearman correlation between human evaluation scores and automatic metrics in Appendix H.

5 Experimental Settings

5.1 Datasets

We utilize two empathetic conversation generation datasets for empathy acquisition which are widely used in previous studies: Emotion Support Conversation (ESConv) (Liu et al., 2021), and EmpatheticDialogues (ED) (Rashkin et al., 2019). For emotional validation alignment, we use our Emotional Validation Aware Dataset, EVAD[7].

5.2 Metrics

We use general response generation metrics and empathy-based metrics for balanced evaluation. For general metrics, we use BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) to evaluate the generated response based on the ground-truth utterance. We also employ Dist-N metric (Li et al., 2016) to measure the diversity of the generated response regarding ground-truth utterance. As the ground-truth utterances in emotional support conversations do not always guarantee the best responses due

[4]The prompts are described in Table 6 of Appendix E.

[5]The details of data collection and statistics are described in Appendix E.

[6]The details of the classifier in EVAEval are explained in Appendix F.

[7]The details of dataset, metrics, and, training settings are in Appendix G.

to subjectivity, there exists a limitation of evaluating LLMs with general response metrics. In order to measure the empathetic dialogue generation capability, we employ the Interpretations (IP), Explorations (EX), and Emotional Reactions (ER) of EPITOME (Sharma et al., 2020). Moreover, we evaluate the degree of emotional validation following our EVAEval metrics at each level.

5.3 Models

We evaluate a diverse set of language models to establish baselines and assess the effectiveness of our proposed method. For general-purpose AI assistants, we include GPT-3.5-Turbo and GPT-4o (Achiam et al., 2023). Additionally, we consider ESC-oriented models, including ChatCounselor (Liu et al., 2023), MeChat (Qiu et al., 2023), and EmoLLM[8] (Team, 2024a), which are commonly used in prior studies (Zhao et al., 2024; Cao et al., 2024) as practical baselines for evaluating emotional support capabilities due to the lack of dedicated ESC-specific LLMs. We also include instruction-tuned LLMs as additional baselines: Mistral-7B-Instruct (Jiang et al., 2023), LLaMA-3-8B-Instruct (Meta, 2024), and Qwen2-Instruct (Team, 2024b). The prompt used for all baseline models is provided in Table 3.

6 Experimental Results

In this section, we evaluate the general and empathetic response generation capability. Meanwhile, we also validate our EVA method's effectiveness on empathetic response generation capability by ablating the Empathy Acquisition (ACQ) and Emotional Validation Alignment (ALI).

Furthermore, with our EVAEval metric, we present the emotional validation following capability. Finally, we compare the baseline and EVA-based model with human evaluation. We report the experimental results on the ESConv dataset in Table 2. The experimental results for the ESC-oriented models are provided in Appendix J.

6.1 Empathetic Response Generation

Main Results We observe that our EVA-based models demonstrate comparable performance to the baseline models in general response generation metrics (B-2, R-L, and D-2). Even though the general response generation metrics are limited to measuring empathetic dialogue generation, our EVA-based models outperform the baselines.

We also compare the EVA-based model and baseline with EPITOME (IP, EX, and ER) to evaluate the degree of empathy in generated utterances. In the IP metric, which evaluates the expressions of acknowledgments or understanding of the help-seeker's emotions, the EVA-based model achieves the best performance regardless of baseline. The IP metric aligns with expressing acknowledgments and understanding, which are key to emotional validation. Therefore, the improved performance in the IP metric indicates superior emotional validation capabilities. The EX metric, which considers expressions of active interest, also shows improved performance compared to the baseline. However, the ER metric yields inconsistent results and even shows lower performance than GPT-3.5-Turbo and GPT-4o. This is because the ER metric only measures the degree of emotional expression regardless of the progression of the conversation. Since our objective is not to express emotions in every utterance but rather to demonstrate emotional validation at specific levels, degraded performance is a plausible result.

Ablation When we ablate the model components in the EVA-based models, we find that the removal of ACQ or ALI is critical to the general response generation metrics such as B-2 and R-L. However, the D-2 metric increases when we remove the ALI in the ESConv dataset in $\text{EVA}_{Mistral}$ and EVA_{Qwen}. The increase in the D-2 metric suggests that removing ALI leads to more varied but less contextually aligned responses. This trade-off highlights ALI's contribution to preserving coherence while balancing diversity in empathetic dialogue generation.

Additionally, the IP metric, aligned with our emotional validation theory, declines when either component is removed, regardless of baselines. This result indicates that ACQ and ALI are essential for enhancing empathetic response generation. For the EX metric, w/o ALI results in instability, suggesting its crucial role in maintaining consistency in emotional exploration. Furthermore, ALI removal leads to a decrease in the ER score, implying that ALI facilitates natural emotional expression. In summary, ALI which learns emotional validation with EVAD, is essential for empathetic response generation.

[8] Slipstream-Max/EmoLLM-InternLM2.5-7B-chat-GGUF-fp16

	B-2	R-L	D-2	IP	EX	ER	EVAEval ↓				
							LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	Avg.
GPT-3.5-Turbo	1.27	13.13	59.18	0.0252	0.3971	1.1119	0.0761	0.1274	0.1353	0.1999	0.1347
GPT-4o	1.03	12.45	58.01	0.0374	0.4212	1.1752	0.0641	0.1131	0.2038	0.1476	0.1321
ChatCounselor	0.41	5.41	27.38	0.1181	0.0108	0.0694	0.2963	0.2493	0.1686	0.0336	0.1869
MeChat	1.96	10.45	51.85	0.0595	0.4131	0.5968	0.1022	0.2342	0.0342	0.2246	0.1488
EmoLLM	1.95	10.07	51.19	0.1045	0.3447	0.4431	0.1245	0.1848	0.1528	0.2203	0.1706
Mistral	1.36	8.40	37.53	0.0810	0.4169	0.7857	0.3009	0.2394	0.1356	0.0230	0.1747
EVA$_{Mistral}$	**2.66**	**14.00**	63.55	**0.3050**	**0.4362**	**0.9073**	**0.0033**	**0.0408**	**0.0386**	0.1300	**0.0532**
w/o ACQ	1.47	8.83	40.57	0.1003	0.3861	0.8494	0.3009	0.2434	0.1505	**0.0146**	0.1774
w/o ALI	2.44	13.5	**76.95**	0.2664	0.2934	0.7683	0.0182	0.1349	0.2409	0.3008	0.1737
Qwen	0.83	6.69	30.83	0.0308	0.0849	0.8590	0.2656	0.2434	0.1388	0.0166	0.1661
EVA$_{Qwen}$	**1.94**	**11.85**	71.72	**0.1235**	0.4401	0.6235	**0.0053**	**0.0966**	0.1696	0.2459	**0.1294**
w/o ACQ	0.78	7.01	32.39	0.0193	0.0772	**0.9227**	0.2678	0.2448	**0.1359**	**0.0126**	0.1653
w/o ALI	1.68	11.49	**72.97**	0.1196	**0.5868**	0.5231	0.0095	0.1337	0.2192	0.2634	0.1564
LLaMA	1.43	10.40	52.29	0.0810	0.5637	0.9594	0.1097	0.0167	0.0187	0.0438	0.0472
EVA$_{LLaMA}$	**2.19**	**10.91**	**53.91**	**0.1943**	**0.9286**	0.9646	**0.0743**	**0.0132**	0.0797	**0.0037**	**0.0427**
w/o ACQ	1.48	10.59	51.94	0.0694	0.6236	**0.9835**	0.1653	0.0624	**0.0214**	0.0335	0.0707
w/o ALI	2.07	10.75	51.41	0.1776	0.8803	0.7702	0.0843	0.1006	0.0640	0.0294	0.0696

Table 2: Experimental results of baselines using different training methods with ESConv dataset. B-2, R-L, and D-2 indicate BLEU-2, ROUGE-L, and Dist-2 respectively. We denote empathy acquisition and emotional validation alignment training processes as ACQ and ALI, respectively. Within each model, the highest performance is indicated in **bold**.

6.2 Emotional Validation Following

Main Results We evaluate the degree of emotional validation following based on the EVAEval. In EVAEval, a lower score indicates better performance, as it quantifies the divergence between the predicted level distribution and the gold dataset distribution. Our EVA-based models consistently achieve the lowest average EVAEval scores across all backbone models, outperforming not only general-purpose LLMs but also ESC-oriented models specifically designed for emotional support. This indicates that the EVA-based model effectively follows the concept of emotional validation, generating lower-level (e.g., `LEVEL 1`, `LEVEL 2`) utterances at the early stage and higher-level (e.g., `LEVEL 3`, `LEVEL 4`) utterances toward the end of the conversation.

Ablation In the ablation study, consistent results are observed for all EVA-based models. The average EVAEval score increases when the ALI component is removed from the EVA method. Specifically, the EVAEval score increases by a relatively larger margin in the absence of the ACQ component, which is responsible for acquiring general empathetic response generation capability. This indicates that both components are essential for generating empathetic responses in accordance with emotional validation.

6.3 Human Evaluation

To qualitatively assess the ability of the EVA-based model to generate an empathetic response, we conduct a human evaluation. Given the inherently subjective nature of empathy, we perform a comprehensive human evaluation using a test set of 100 dialogues containing utterances generated by the Mistral and EVA$_{Mistral}$ models[9]. We recruit 15 psychology majors with expertise in emotional validation theory to perform the evaluations. Human workers evaluate each utterance based on four criteria (Comfortness, Comprehensibility, Emotional Validation, and Fluency) [10].

As illustrated in Figure 5, EVA$_{Mistral}$ achieves higher win rates across all evaluation criteria except fluency[11]. Notably, EVA$_{Mistral}$ demonstrates significant improvement in comfortness, with approximately 60% of responses rated as superior to the baseline. This suggests that EVA$_{Mistral}$ generates more effective responses that provide emotional comfort to help-seekers. Furthermore, the EVA$_{Mistral}$ outperforms the Mistral in comprehensibility, indicating an enhanced ability to understand and acknowledge the help-seeker's emotional state. In the emotional validation criteria,

[9]This evaluation includes all possible dialogues from the ESConv test set, excluding only those where the system initiates the dialogue.

[10]The details of the human evaluation and criteria are described in Appendix K.

[11]The inter-annotator agreement, measured by Fleiss' Kappa, was 0.61, which indicates a substantial level of agreement given the inherently subjective nature of the task.

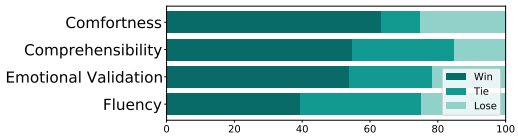


Figure 5: Results of human evaluation on four distinct criteria. The stacked bar chart illustrates the distribution of responses across the 'Win', 'Tie', and 'Lose' categories for each criterion.

EVA$_{Mistral}$ also achieves notable results, demonstrating its successful integration of emotional validation theory into its responses. However, a slight decrease in fluency relative to the baseline may be attributed to the added complexity of incorporating emotional validation components into the response generation process. In summary, these human evaluation results demonstrate that the EVA$_{Mistral}$ model generates responses that enhance emotional support and validation for help-seekers.

7 Analysis

We conduct in-depth analyses to demonstrate the effectiveness of our EVA method, focusing on two research questions: **RQ1**: Does EVA mitigate patterned responses?, **RQ2**: Is EVA still biased to solution suggestions? All analyses are performed on 100 dialogues from the ESConv test set, which includes all possible dialogues except those where the system initiates the conversation. We show the results of EVA$_{Mistral}$ in this section, and comparative analyses of EVA$_{Qwen}$ and EVA$_{LLaMA}$ are described in Appendix M.

7.1 RQ1: Does EVA mitigate patterned responses?

To illustrate how our EVA method mitigates the patterned response within the dialogue compared to the vanilla model, we collect top-4 phrases that appeared frequently in the responses from each model. The phrases are extracted with N-gram rules. As shown in Figure 6, we visualize the number of dialogue where each phrase co-occurs more than 10%, 20%, and 30% within a single dialogue. The results indicate a significant reduction in phrase repetition when the EVA method is applied.

In EVA$_{Mistral}$, although these phrases are among the four most frequently generated across all inferences, they are rarely repeated within a single dialogue. Specifically, in Mistral, each top phrase is repeated more than 30% within a single dialogue, whereas in the EVA-based model, phrases with over 30% repetition barely exist. This suggests that EVA$_{Mistral}$ mitigate the patterned responses.

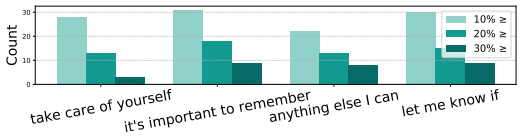
Furthermore, the decrease in repetition ratio indicates that EVA$_{Mistral}$ exhibits a more balanced distribution of phrase usage. While Mistral heavily relies on certain phrases (e.g., "take care of yourself", "let me know if"), EVA$_{Mistral}$ distributes phrase usage more evenly, contributing to more natural and human-like interactions. Previous studies have shown that users are more likely to find responses satisfying and realistic when they are less repetitive and more varied (Concannon et al., 2023a). Therefore, as demonstrated in our human evaluation, EVA$_{Mistral}$'s ability to reduce patterned responses aligns with its improved performance in emotional support and validation. This balanced phrase distribution suggests that EVA$_{Mistral}$ enhances the help-seeker's overall experience by fostering more engaging and natural interactions.

7.2 RQ2: Is EVA still biased to solution suggestions?

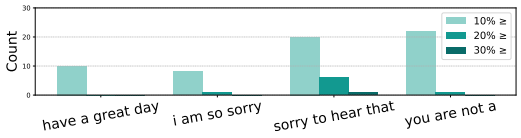
To assess whether the EVA-based model is still biased toward solution suggestions, we analyze the distribution of response types generated by the EVA$_{Mistral}$ as we did in the preliminary study. We assign the types of response which is defined in Appendix D, to each utterance. As illustrated in Figure 7 (a), Mistral primarily relies on Solution Suggestion (SS) responses, comprising 45.0% of results. In contrast, as shown in Figure 7 (b), the EVA$_{Mistral}$ demonstrates a significant decrease in Solution Suggestion (SS) responses type distribution. Furthermore, EVA$_{Mistral}$ generates a wider variety of response types, enhancing its ability to empathize with help-seekers more effectively. Particularly, when applying EVA, the proportion of the Understanding and Validation (UV) type increases significantly from 11.7% to 26.0%. UV type aligns with emotional validation, which is the goal of our study. These results highlight the EVA$_{Mistral}$'s enhanced capacity to engage in empathetic communication by providing responses that acknowledge and validate the help-seeker's emotions rather than merely suggesting solutions.

8 Conclusion

In this work, we proposed the EVA, an effective training method for enhancing LLMs with empathetic response generation capabilities. The EVA is grounded in the emotional validation theory, which

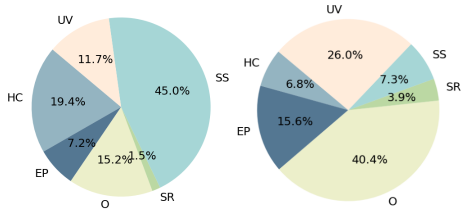


(a) Mistral



(b) EVA$_{Mistral}$

Figure 6: The number of dialogue where each phrase co-occurs more than 10%, 20%, and 30% within a single dialogue at Mistral and EVA$_{Mistral}$. Each bars indicate the ratio of dialogue where each phrase is repeated more than 10%, 20%, and 30% within a dialogue, respectively.



(a) Mistral (b) EVA$_{Mistral}$

Figure 7: Response type proportions in utterances generated with Mistral and EVA$_{Mistral}$.

characterizes four emotional validation levels in accordance with conversational progress. We ensured emotional validation by constructing EVAD and training the model through validation alignment. Additionally, we introduced EVAEval, a novel metric to assess empathetic response generation. Experimental results showed that EVA effectively improves the empathetic response generation capability for LLM in both automatic and human evaluation. In short, emotional validation plays a key role in securing the empathetic response generation capability of LLM. Our work can facilitate future research on utilizing emotional validation theory to develop empathetic conversational agents, as well as designing training methods for aligning human preference to language models.

Limitations

This work has the following limitations: **(1) The Number of Human Workers**: While empathy is inherently subjective and ideally requires a substantial number of annotators, we were constrained by resource limitations and thus employed ten annotators. However, we would like to note that, apart from large-scale studies, most prior work has relied on five or fewer annotators for human evaluation (Yuan et al., 2024; Lee et al., 2022; Fu et al., 2023; Qian et al., 2023; Wang et al., 2025a) in the empathetic response generation task. In contrast, our study involved 15 psychology students, positioning it at the higher end of typical human evaluation settings in empathetic dialogue research. We believe this approach is consistent with established practices and contributes to a more reliable assessment.

(2) Response Type Definition: Due to the lack of theories defining specific response types in the context of emotional support conversations, we collaborated with psychologists to develop our own definitions. Therefore, response types may not capture all possible types comprehensively, which could limit the generalizability of our analysis.

(3) Subjectivity of Empathy: We construct a human preference dataset, EVAD, based on the ESConv dataset, which is widely used for ESC tasks. Given that empathy is inherently subjective, there is a potential for the annotators' inherent biases to influence the data during the annotation process. However, the ESConv dataset is publicly available and carefully curated, with sensitive and private information filtered out during its construction. As a result, we believe that these risks are minimized due to the rigorous filtering procedures and ethical safeguards.

(4) Empathy–Fluency Trade-off. Our human evaluation reveals that EVA-trained models achieve stronger emotional validation but sometimes produce responses that feel less fluent than baseline models. This effect arises from richer empathetic framing and longer, multi-clause utterances. As in Table 14, a highly rated EVA-generated response deeply acknowledged the help-seeker's feelings but was perceived as less smooth than the baseline's shorter and more concise reply. This highlights a trade-off between empathy depth and surface fluency, which we aim to address in future work by exploring generation strategies that balance both aspects.

(5) Dependence on the Classifier for EVAEval. The reliability of our EVAEval metric depends on a RoBERTa-based classifier that labels emotional validation levels. While the classifier achieves a strong F1 score, occasional misclassifications do occur. Most errors happen between adjacent levels (e.g., Level 1 vs. Level 2) rather than distant lev-

els, which minimizes their effect on capturing the overall progression of emotional validation. Nevertheless, this dependence introduces a potential source of noise in our evaluation and represents an area for refinement.

(6) Use of GPT-based Candidates in EVAD. Our EVAD was partially constructed using ChatGPT-generated candidate responses before human preference ranking. Although annotators systematically tagged and removed low-quality generations during the ranking stage, thereby filtering out unsuitable responses, the initial reliance on GPT outputs could still introduce annotation artifacts or subtle biases. Future work should explore dataset construction pipelines that further reduce potential circularity between model-generated data and model evaluation.

Ethical Considerations

We discuss the following potential ethical issues: **(1) Privacy**: In this paper, we employ ESConv and ED dataset, a publicly available and carefully developed benchmark for emotional support conversations. The dataset was collected by crowd-sourced workers, with all sensitive and private information filtered out during the data collection process. **(2) Human Evaluation**: We clarify that we employ three distinct groups of human evaluators tailored to each specific task: 1) For the preliminary study, we recruit 15 general crowd workers without specific psychological expertise to assess whether responses felt comfortable from a general user perspective. 2) For the final human evaluation, we employ 15 psychology majors with expertise in emotional validation theory to ensure rigorous and theory-informed evaluations. 3) For EVAD annotation, 12 annotators with at least a bachelor's degree in psychology were hired to guarantee high-quality preference labels aligned with emotional validation theory. The groups in 1), 2) and 3) were independently hired and did not overlap. We carefully design the recruitment and task instructions to match the purpose of each experiment, thereby minimizing bias and maximizing consistency within each evaluation context. **(3) Potential Risks**: Our approach leverages LLMs, which are known to pose risks such as hallucination and inherent biases. These are recognized challenges in tasks involving LLMs, and we have to care to mitigate potential negative social impacts. **(4) Data Usage and Licensing Compliance**: Our study strictly adheres to the licensing terms of the datasets used. The ESConv dataset is subject to copyright restrictions by CoAI Group, Tsinghua University, and is used solely for academic research. Similarly, ED is released under the CC BY-NC 4.0 license, allowing non-commercial research use. Both datasets were originally developed for empathetic dialogue research, and we use them for training and evaluating empathetic dialogue models, fully aligning with their intended purposes. Additionally, our newly constructed EVAD dataset follows the original datasets' intended research purposes and access conditions, ensuring compliance with all licensing agreements. **(5) Use of AI in Writing**: We acknowledge the use of ChatGPT for minor grammar refinement and rewording in the writing process. However, all research, experimental design, data analysis, and conclusions were conducted solely by the authors without AI assistance.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425). This work was supported by the Commercialization Promotion Agency for R&D Outcomes(COMPA) grant funded by the Korea government(Ministry of Science and ICT)(2710086166). This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

William A Anthony. 1971. A methodological investigation of the" minimally facilitative level of interpersonal functioning". *Journal of Clinical Psychology*, 27(1).

Godfrey T Barrett-Lennard. 1993. The phases and focus of empathy. *British journal of medical psychology*, 66(1):3–14.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx,

Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brant R Burleson. 2003. Emotional support skills. In *Handbook of communication and social interaction skills*, pages 569–612. Routledge.

Gillian Cameron, David Cameron, Gavin Megaw, Raymond Bond, Maurice Mulvenna, Siobhan O'Neill, Cherie Armour, and Michael McTear. 2017. Towards a chatbot for digital counselling. In *Proceedings of the 31st international BCS human computer interaction conference (HCI 2017)*. BCS Learning & Development.

Yaru Cao, Hongzhi Yu, and Fucheng Wan. 2024. Enhancing emotional support conversation system via integrating mental state-strategy reasoning. In *2024 6th International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, pages 259–263. IEEE.

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.

Shauna Concannon, Ian Roberts, and Marcus Tomalin. 2023a. An interactional account of empathy in human-machine communication. *Human-Machine Communication*, 6.

Shauna Concannon, Ian Roberts, and Marcus Tomalin. 2023b. An interactional account of empathy in human-machine communication. *Human-Machine Communication*.

Jean Decety. 2010. The neurodevelopment of empathy in humans. *Developmental neuroscience*, 32(4):257–267.

Nancy Eisenberg and Natalie D Eggum. 2009. Empathic responding: Sympathy and personal distress. *The social neuroscience of empathy*, 6(2009):71–830.

Vojtech Formanek and Ondrej Sotolar. 2024. Quantitative assessment of intersectional empathetic bias and understanding. *arXiv preprint arXiv:2411.05777*.

Fengyi Fu, Lei Zhang, Quan Wang, and Zhendong Mao. 2023. E-core: Emotion correlation enhanced empathetic dialogue generation. *arXiv preprint arXiv:2311.15016*.

George M Gazda et al. 1977. Human relations development: A manual for educators.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

Clara E Hill. 2020. *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. *arXiv preprint arXiv:2106.09685*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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. *arXiv preprint arXiv:2402.13211*.

William Kidder, Jason D'Cruz, and Kush R Varshney. 2024. Empathy and the right to be an exception: What llms can and cannot do. *arXiv preprint arXiv:2401.14523*.

Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2022. Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 669–683.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2019. Empdg: Multiresolution interactive empathetic dialogue generation. *arXiv preprint arXiv:1911.08698*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687*.

Marsha M Linehan. 1997. Validation and psychotherapy.

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Mime: Mimicking emotions for empathetic response generation. *arXiv preprint arXiv:2010.01454*.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date.

Kevin M Mitchell and Bernard G Berenson. 1970. Differential use of confrontation by high and low facilitative therapists. *The Journal of Nervous and Mental Disease*, 151(5):303–309.

Anand Mohan, Dighreandr Singh, Aakanksha Kataria, Hemant Kumar Meena, and Reeta Singh. 2025. The impact of ai on human psychology: A user perspective. *2025 International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI)*, pages 1321–1326.

OpenAI-Blog. 2022. Chatgpt: Optimizing language models for dialogue.

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.

Yushan Qian, Wei-Nan Zhang, and Ting Liu. 2023. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. *arXiv preprint arXiv:2310.05140*.

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*.

Huachuan Qiu and Zhenzhong Lan. 2024. Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions. *arXiv preprint arXiv:2408.15787*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16.

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, Florence, Italy. Association for Computational Linguistics.

William J Reynolds and Brain Scott. 1999. Empathy: a crucial component of the helping relationship. *Journal of psychiatric and mental health nursing*, 6(5):363–370.

Derek Richards. 2009. Features and benefits of online counselling: Trinity college online mental health community. *British Journal of Guidance & Counselling*, 37(3):231–242.

H Rudolph Schaffer. 1996. *Social development*. Blackwell Publishing.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.

Ekaterina Svikhnushina, Iuliana Voinea, Anuradha Welivita, and Pearl Pu. 2022. A taxonomy of empathetic questions in social dialogs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2952–2973.

Dehua Tao, Tan Lee, Harold Chui, and Sarah Luk. 2024. Modeling intrapersonal and interpersonal influences for automatic estimation of therapist empathy in counseling conversation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12692–12696. IEEE.

EmoLLM Team. 2024a. Emollm: Reinventing mental health support with large language models. `https://github.com/SmartFlowAI/EmoLLM`.

Qwen Team. 2024b. Hello qwen2.

Amy JC Trappey, Aislyn PC Lin, Kevin YK Hsu, Charles V Trappey, and Kevin LK Tu. 2022. Development of an empathy-centric counseling chatbot system capable of sentimental dialogue analysis. *Processes*, 10(5):930.

Lanrui Wang, Jiangnan Li, Chenxu Yang, Zheng Lin, Hongyin Tang, Huan Liu, Yanan Cao, Jingang Wang, and Weiping Wang. 2025a. Sibyl: Empowering empathetic dialogue generation in large language models via sensible and visionary commonsense inference. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 123–140.

Xu Wang, Bo Wang, Yihong Tang, Dongming Zhao, Jing Liu, Ruifang He, and Yuexian Hou. 2025b. Ecc: Synergizing emotion, cause and commonsense for empathetic dialogue generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5475–5485.

Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2024. Leveraging large language models to power chatbots for collecting user self-reported data. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–35.

Zhou Yang, Zhaochun Ren, Yufeng Wang, Chao Chen, Haizhou Sun, Xiaofei Zhu, and Xiangwen Liao. 2024a. An iterative associative memory model for empathetic response generation. *arXiv preprint arXiv:2402.17959*.

Zhou Yang, Zhaochun Ren, Wang Yufeng, Haizhou Sun, Chao Chen, Xiaofei Zhu, and Xiangwen Liao. 2024b. An iterative associative memory model for empathetic response generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3081–3092, Bangkok, Thailand. Association for Computational Linguistics.

Jiahao Yuan, Zixiang Di, Zhiqing Cui, Guisong Yang, and Usman Naseem. 2024. Reflectdiffu: Reflect between emotion-intent contagion and mimicry for empathetic response generation via a rl-diffusion framework. *arXiv preprint arXiv:2409.10289*.

Wang Yufeng, Chen Chao, Yang Zhou, Wang Shuhui, and Liao Xiangwen. 2024. Ctsm: combining trait and state emotions for empathetic response model. *arXiv preprint arXiv:2403.15516*.

Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Wang Jian, Dandan Liang, et al. 2024. Esc-eval: Evaluating emotion support conversations in large language models. *arXiv preprint arXiv:2406.14952*.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2022. Augesc: Dialogue augmentation with large language models for emotional support conversation. *arXiv preprint arXiv:2202.13047*.

Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang, and Minlie Huang. 2022. Case: Aligning coarse-to-fine cognition and affection for empathetic response generation. *arXiv preprint arXiv:2208.08845*.

A Comparison of responses

A comparison of responses generated by Mistral-7B-Instruct and our proposed EVA$_{Mistral}$ to the help-seeker's utterances is shown in Figure 1.

B Prompt Template

Prompt
You are a helpful and considerate supporter. Your job is to generate appropriate emotional supporting utterances based on the help-seeker's utterance. Please generate the supporter's utterance based on the given dialogue history. {*dialogue history*} supporter:

Table 3: Prompt templates for LLMs including GPT-3.5-Turbo, GPT-4o, ChatCounselor, MeChat, EmoLLM, Mistral, Qwen, and LLaMA. Note that the experimental results from vanilla LLMs in Table 2 and Table 8 are extracted with this prompt.

C Related Work

C.1 Emotional Support Conversation

Emotional support is a crucial aspect of conversational agents, particularly in social interactions, mental health support, and customer service. Rashkin et al. (2019) explore methods to enhance emotion detection and empathetic responses in conversational agents, emphasizing the importance of contextual understanding and interaction nuances. Building on this foundation, Svikhnushina et al. (2022) highlight the significance of effective questioning in developing empathetic agents. Their empathetic question taxonomy (EQT) ensures that agents can appropriately respond to users' emotional states, thereby fostering meaningful interactions. Liu et al. (2021) proposes the emotional support conversation task along with the ESConv dataset based on helping skills Hill (2020): exploration, comforting, and action.

Recent works have applied LLMs to empathetic dialogue generation, leveraging their strong generalization capability (Wei et al., 2024; Qian et al., 2023). With improved fluency, these models have enabled more natural dialogue. However, studies also report that LLMs often fail to reflect the emotional needs of help-seekers (Zhao et al., 2024; Kang et al., 2024). Earlier models such as MoEL (Lin et al., 2019) and EmpDG (Li et al., 2019) attempted to encode empathy through

mixture-of-experts or multi-resolution representations. MIME (Majumder et al., 2020) proposed mimicking the user's emotions to guide generation. These works paved the way for more controllable and nuanced emotional modeling in dialogue systems. Building on these foundations, more recent studies emphasize controllability and alignment. For instance, CASE (Zhou et al., 2022) models cognition-affect alignment across coarse-to-fine levels, while CTSM (Yufeng et al., 2024) combines trait and state emotions to generate more personalized responses. ECC (Wang et al., 2025b) introduces emotional causes and commonsense to enhance contextual understanding. In terms of learning from preferences, Yang et al. (2024a) and Yuan et al. (2024) show promising results by leveraging memory-based and diffusion-based optimization for emotional alignment.

Overall, these studies reflect a growing trend toward controllable and preference-aligned empathetic response generation. While early approaches focused on emotional expression, recent works shift toward ensuring alignment with the help-seeker's needs and emotional progression. Our work builds upon this line by incorporating emotional validation theory to guide empathetic LLMs in providing context-sensitive and emotionally appropriate responses.

C.2 Levels of Empathy

There are a few works that analyze the progression of emotion towards empathy within the dialogue. Schaffer (1996) introduce four distinct levels on empathy: global empathy, egocentric empathy, empathy for another's feelings, and empathy for another's life condition. These stages elucidate how the depth and complexity of empathy develop, allowing empathy for a nuanced recognition and response to a wide range of emotional expressions. Additionally, the cyclic process highlighted by Barrett-Lennard (1993) focuses on three phases empathetic resonation, communication of understanding, and acknowledgment. This process underscores the importance of feedback loops in refining empathetic exchanges and addressing potential misalignments between internal emotional responses, expressions, and receptions.

On the other hand, Linehan (1997) introduce six levels of emotional validation: 1) listening and observing, 2) accurate reflection, 3) articulating the unverbalized, 4) validating in terms of sufficient, but not necessarily valid, causes , 5) validating as reasonable in the moment, and 6) treating the persona as Valid-Radical Genuineness. Emotional validation plays a pivotal role in counseling by affirming and acknowledging a help-seeker's emotions through deeper understanding of their behavior. While Schaffer (1996) and Barrett-Lennard (1993) present the notion of empathy, it overlook the empirical supportive skills that empathy plays in human conversation. To concentrate conversational situation between help-seeker and supporter, we adopt Linehan (1997)'s emotional validation for our method.

D Response Types

We define the types of empathetic responses and their examples as in Table 4 and Table 5. The types and examples are defined under the supervision of a scholar with a degree in psychology. This list categorizes various supportive phrases into different types of responses that can be used in conversations to convey empathy, support, and encouragement.

The Understanding and Validation (UV) type aims to show that you understand and validate the other person's feelings or situation. Shared Experiences and Relatability (SR) type is about sharing your own experiences or relating to the person's situation, helping them feel understood and less isolated. Encouragement and Positivity (EP) are positive reinforcements that help boost a person's confidence and outlook on the situation. Help and Compassion (HC) type expresses willingness to help and compassion, letting the person know they are not alone and that you care about their well-being with their help. Solution Suggestion (SS) offers suggestions and possible solutions to the person's problem, helping them find a way forward. Finally, the Other Types (O) category includes responses that do not align to any of the previously defined types.

Examples of Empathetic Responses Types
Understanding and Validation (UV)
i can feel how
i can imagine how you feel
i can see that you are
i can see why you feel that way
i can understand
i completely understand
i get it
i hear you
i see what you mean
i see,
i understand
it is understandable
it sounds like
sorry to hear
that makes sense
that must have been hard
you must be feeling
you seem really
your feelings are
Encouragement and Positivity (EP)
everything will be okay
keep going
keep pushing forward
stay strong
that is a great
that's good
things will get better
you are doing great
you are doing your best
you're not alone in this
you are strong
you can do it
you're doing great
you're stronger than you think
you've got this
Help and Compassion (HC)
how can i help?
i am always here
i am here
i am sorry you're going through this
i am with you
i can help
i care about you
i could help
i will do my best to help
if you need any help, let me know
if you need anything else just let me know
you are important to me
you are not alone
you can count on me

Table 4: The types of empathetic responses and their examples for Understanding and Validation (UV), Encouragement and Positivity (EP) and Help and Compassion (HC).

Examples of Empathetic Responses Types
Solution Suggestion (SS)
have you considered
have you thought about
here are few steps
here are few things
i also recommend
i do have some suggestions
i recommend
i think you should
i want you to
i would also recommend
i would suggest
it is important to
it may be a good option
it might be a good option
just try not to
just try to
let's
would you consider
you can try
you could
you have to
you should try
Shared Experiences and Relatability (SR)
i have been
i have felt that way before
i know what that's like
i remember
it is okay to feel this way
we all go through this sometimes

Table 5: The types of empathetic responses and their examples for Solution Suggestion (SS), and Shared Experiences and Relatability (SR).

E EVAD Collection

We construct a human preference dataset, EVAD, for emotional validation alignment. Initially, we obtain an utterance of each level in emotional validation with ChatGPT (GPT-3.5-TURBO). The model inputs a prompt template for each level and generates the utterance based on the dialogue history. As shown in Table 6, we also provide a detailed description of the emotional validation theory in the prompt. Then, we annotate the generated utterances as chosen or rejected by the workers. We employ 12 individuals who majored in psychology and hold at least a bachelor's degree. We provide an explanation of emotional validation theory on the intro page, as shown in Figure 8. Each worker has been remunerated at a rate of $3 per dialogue.

Figure 9 shows that workers rank the utterances from most to least appropriate for each dialogue. We refer to the most and least appropriate utterances as the chosen and rejected, respectively. We adopt a majority voting scheme to decide the final label. To further ensure reliability, we additionally filter out low-agreement cases, defined as instances where more than five annotators disagreed with the majority choice. This procedure allowed us to retain high-quality annotations while minimizing the impact of ambiguous cases. Based on the collected annotations, the Fleiss' Kappa was 0.56, reflecting moderate agreement, which is reasonable given the inherent subjectivity of emotional validation tasks. The statistic of the EVAD is in Table 7. Additionaly, we illustrate the distribution of each level onto the relative position within the dialogue. As in Figure 4, the distribution of chosen responses shows a higher proportion of LEVEL 1 at the beginning of the conversation and a higher proportion of LEVEL 4 towards the end. This result indicates that the chosen responses show deeper empathy as the dialogue progresses, demonstrating consistency with the emotional validation theory.

	Prompt
LEVEL 1	You are a helpful and considerate supporter. Your job is to generate appropriate emotional supporting utterances based on the help-seeker's utterance. Please generate the utterance based on the criteria below LEVEL1: Listening and Observing. LEVEL1 is the listening to and observing of what the help-seeker is saying, feeling, and doing as well as a corresponding active effort to understand what is being said and observed. The supporter should not express their own thoughts or offer advice. {*dialogue history*} supporter:
LEVEL 2	You are a helpful and considerate supporter. Your job is to generate appropriate emotional supporting utterances based on the help-seeker's utterance. Please generate the utterance based on the below criteria LEVEL2: Accurate Reflection LEVEL2 aims to express empathy based on the help-seeker's situation and emotions apprehended in LEVEL 1. The supporter should paraphrase the response of the help-seeker to express empathy {*dialogue history*} supporter:
LEVEL 3	You are a helpful and considerate supporter. Your job is to generate appropriate emotional supporting utterances based on the help-seeker's utterance. Please generate the utterance based on the criteria below LEVEL3: Validating LEVEL3 is delivering the response based on validating the overall help-seeker's situation and feeling. This level aims to alleviate the help-seeker's negative feelings and emotions. {*dialogue history*} supporter:
LEVEL 4	You are a helpful and considerate supporter. Your job is to generate appropriate emotional supporting utterances based on the help-seeker's utterance. Please generate the utterance based on the criteria below LEVEL4. Radical Genuineness LEVEL4 is the task is to recognize the person as he or she is, seeing and responding to the strengths and capacities of the individual while keeping a firm empathic understanding of the client's actual difficulties and incapacities. The purpose of LEVEL 4 is to facilitate self-validation and reduce emotional distress. {*dialogue history*} supporter:

Table 6: Prompt templates for generating emotional validation utterances at different levels using ChatGPT (GPT-3.5-TURBO). Each template provides specific instructions for producing supportive responses based on the help-seeker's dialogue history and the description for four levels of emotional validation.

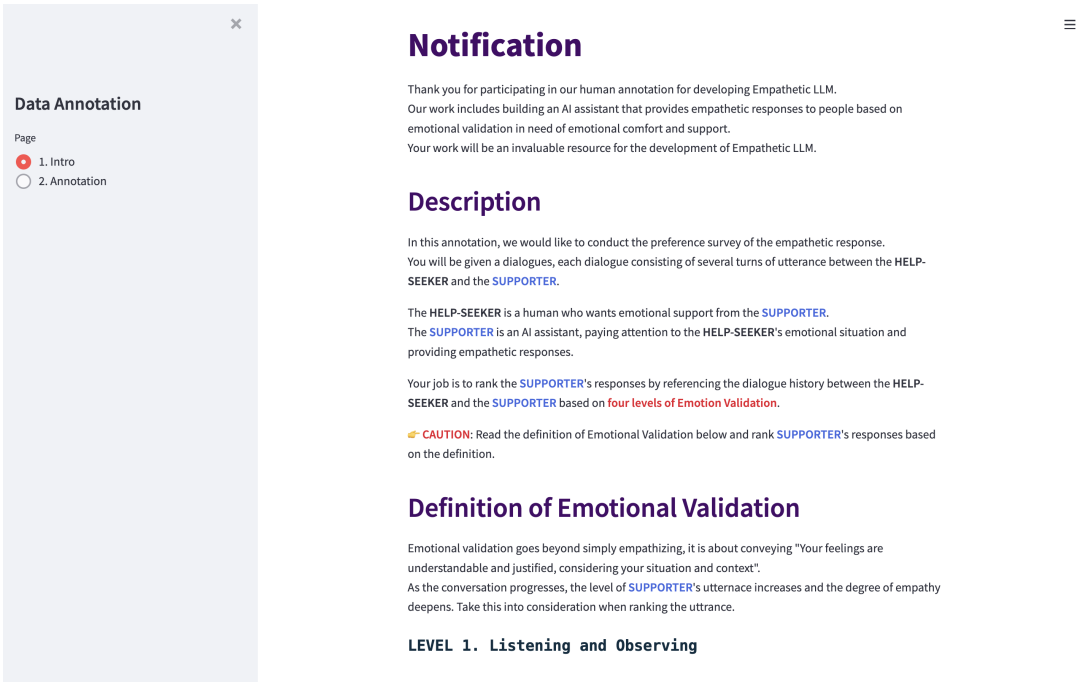


Figure 8: Interface of EVAD collection page - Introduction and Explanation on Emotional Validation

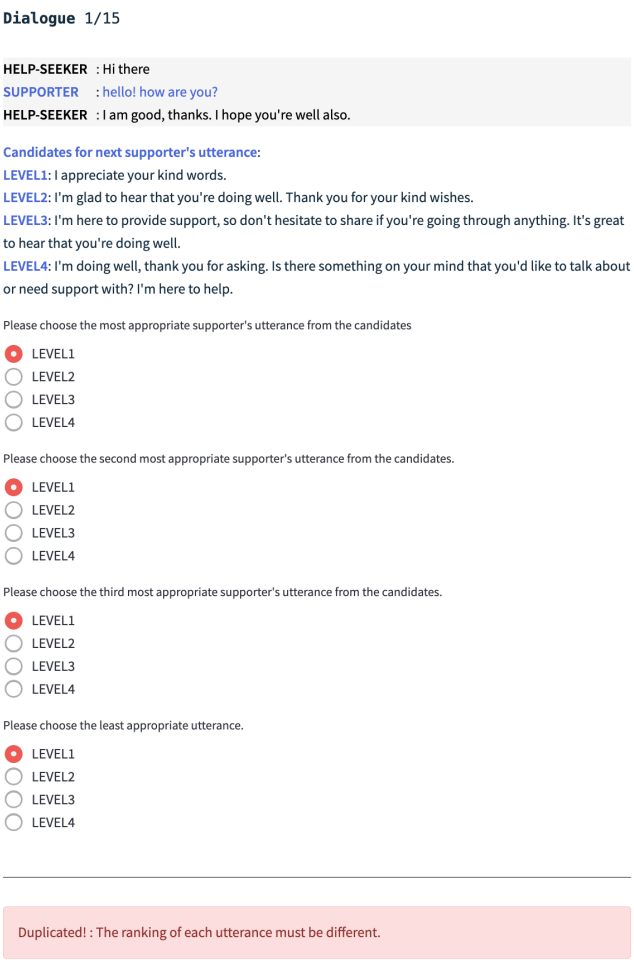


Figure 9: Interface of EVAD collection page - Annotation

	# Examples	Chonsen Utterance			
		LEVEL1	LEVEL2	LEVEL3	LEVEL4
Train	5,157	937	1,209	1,182	1,829
Valid	465	97	126	85	157
Total	5,622	1,034	1,335	1,267	1,986
	# Examples	Rejected Utterance			
		LEVEL1	LEVEL2	LEVEL3	LEVEL4
Train	5,157	1,209	1,114	467	2,367
Valid	465	118	81	48	218
Total	5,622	1,327	1,195	515	2,585

Table 7: Statistics of EVAD dataset: Results of level classification for chosen and rejected utterances.

F The Details of EVAEval

To evaluate whether the generated utterance follows the emotional validation, we train a classifier based on RoBERTa (Liu et al., 2019) with pairs of input utterances and corresponding levels. Utilizing the ESConv dataset, we construct pairs with human workers who majored in psychology and hold at least a bachelor's degree. We provide workers with a detailed explanation of emotional validation theory. Then, the workers annotate each supporter's utterance based on its level of emotional validation. The dataset consists of 4,047 training, 1,012 validation, and 563 test examples. The classifier trained with the constructed dataset achieves an F1 score of 89.56 in the test set.

G The Details of Experimental Setting

Dataset We utilize two English-language empathetic conversation generation dataset for empathy acquisition: EmpatethicDialogues (ED) (Rashkin et al., 2019) and Emotion support conversation (ESConv) (Liu et al., 2021). ED consists of situations involving a wide range of emotions between the help-seeker and the supporter. This dataset is crowdsourced from 810 US-based MTurk workers, meaning it represents a demographic skewed towards US online workers. ED does not explicitly control for age, gender, ethnicity, or socioeconomic diversity, but it does contain dialogues from diverse personal experiences. We use original train/validation/test splits consisting of 19,533 / 2,770 / 2,547 dialogues. ED is publicly available and released under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license, which permits use for academic research purposes only. ESConv includes emotional support conversations where the supporter alleviates emotional distress and aids help-seekers in comprehending and addressing their challenges. ESConv is collected via trained crowdworkers acting as supporters, while help-seekers were either real users experiencing emotional distress or simulated cases based on real experiences. ESConv ensures balanced gender representation, including both male and female help-seekers and supporters. Additionally, 75.2% of the dialogues originate from real personal experiences, ensuring high authenticity in emotional interactions. In the case of ESConv, some exceptions start with the machine's utterance, which is unsuitable for training instruction-tuned models. Therefore, we filter out those exceptions and use subsets of the original datasets consisting of 422/ 94/ 90 dialogues. The dataset is copyrighted by CoAI Group, Tsinghua University (© 2021) and is provided for research purposes only, as specified in its official repository. For emotional validation alignment, we use EVAD which contains 5,622 samples, with balanced coverage of chosen and rejected labels for training and validation.

Metrics We use general response generation metrics and empathy-based metrics for balanced evaluation. For general metrics, we use BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) to evaluate the similarity of the generated response based on the ground-truth utterance. We also employ Dist-N metric (Li et al., 2016) to measure the diversity

of the generated response regarding ground-truth utterance. Here, BLEU scores are computed with the NLTK package (version 3.8.1), while ROUGE scores are computed using the rouge_score package (version 0.1.2). We use default parameter settings for both metrics.

In addition to general response generation metrics, we employ the EPITOME framework (Sharma et al., 2020) to measure the empathetic dialogue generation. The EPITOME framework consists of three independent RoBERTa (Liu et al., 2019) based models that each capture the empathy degree (0, 1, or 2) in the conversation in terms of the three dimensions, Interpretations (IP): the expressions of acknowledgments or understanding of the help-seekers emotion or situation, Explorations (EX): the expressions of active interest in the help-seekers situation, Emotional Reactions (ER): the expressions of emotions such as warmth, compassion, and concern in the help-seekers situation. Unlike prior work that evaluates these mechanisms using accuracy or F1 score with available gold annotations, we do not have gold-standard degree labels for generated responses. Instead, we utilize the EPITOME classifiers to assign a degree score (0, 1, or 2) to each generated response for each mechanism (ER, IP, EX). We then report the average score as a measure of how strongly each communication mechanism is expressed in the model's outputs. Moreover, we evaluate the degree of emotional validation following our EVAEval metrics at each level.

Training Settings For empathy acquisition, we train baseline models with full fine-tuning. We perform 3 epochs with a base learning rate of 1e-6. The batch size is 8, complemented by gradient accumulation steps of 4. We employ DeepSpeed's ZeRO-3 stage optimization (Rajbhandari et al., 2020) for automatic mixed precision training. We choose the best hyperparameters based on the validation set performance with different combinations of batch sizes {2, 4, 8}, learning rates {2e-5, 5e-6, 1e-6}, gradient accumulation steps {1, 2, 4}, and epochs {1, 3, 6, 9}.

For emotional validation alignment, we train the model in PEFT settings with LoRA (Hu et al., 2021). We set PEFT rank, alpha, and dropout to 8, 16, and 0.05, respectively. We train the model for 6 epochs with a base learning rate of 5e-5. The batch size is 8, complemented by gradient accumulation steps of 4. All experiments are conducted on the server configured with Ubuntu 16.04 operating system, Intel(R) Xeon(R) Gold 6230R CPU, 8 NVIDIA RTX A6000, and 128 GB memory. The average training time for empathy acquisition is 2 hours at the ESConv dataset and 10 hours at the ED dataset. For emotional validation alignment, the average training time is 2 hours. Each experiment is conducted three times, and we report the average performance across runs to ensure stability and reproducibility of results.

H Validating EVAEval via Correlation with Human Judgments

To validate the effectiveness of EVAEval, we conducted a correlation analysis between human evaluation scores and automatic metrics. Specifically, we recruited ten annotators with a background in psychology to evaluate 47 dialogues generated by $\text{EVA}_{Mistral}$ on the ESConv dataset. Each annotator assessed whether the supporter's response aligns with the concept of emotional validation ("Does the supporter's utterance align with the concept of emotional validation?"). The inter-annotator agreement, measured by Fleiss' Kappa, was 0.67, indicating substantial agreement among raters.

Since EVAEval measures the deviation from the gold-level distribution (i.e., lower is better), the correlation coefficients are expected to be negative. A higher absolute value of the negative correlation thus indicates better alignment with human judgments. As shown in Table 9, EVAEval exhibits a stronger correlation with human ratings on emotional validation compared to IP, EX, and ER metrics, demonstrating its effectiveness in evaluating empathetic dialogue.

I Experimental Results in ED

I.1 Empathetic Response Generation

In this section, we assess both general and empathetic response generation capabilities with ED dataset. Additionally, we demonstrate the effectiveness of our EVA method for empathetic response generation by conducting ablation studies on Empathy Acquisition (ACQ) and Emotional Validation Alignment (ALI). Furthermore, we utilize our EVAEval metric to evaluate the model's ability to follow emotional validation.

Main Results As depicted in Table 8, we observe that our EVA-based models demonstrate improved performance to the baseline models in general response generation metrics (B-2, R-L, and

	B-2	R-L	D-2	IP	EX	ER	EVAEval ↓				
							LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	Avg.
GPT-3.5-Turbo	1.50	11.77	58.33	0.0336	0.4105	1.1653	0.2192	0.2203	0.1579	0.2144	0.2030
GPT-4o	1.33	11.14	57.54	0.0672	0.4370	1.1644	0.1494	0.2755	0.1937	0.1592	0.1944
ChatCounselor	0.14	3.38	26.44	0.0216	0.0257	0.0300	0.3009	0.2532	0.3721	0.3728	0.3247
MeChat	1.18	12.19	56.70	0.0720	0.4411	0.5931	0.2112	0.1920	0.2021	0.3136	0.2297
EmoLLM	1.22	11.81	74.32	0.1233	0.4510	0.5426	0.1106	0.2280	0.2459	0.2656	0.2125
Mistral	0.60	5.83	32.86	0.0936	0.2881	0.8823	0.2903	0.2532	0.1344	0.0374	0.1788
EVA$_{Mistral}$	**4.38**	**18.57**	81.94	**0.2400**	**0.7731**	0.7635	**0.0625**	**0.1829**	0.2545	0.3332	0.2083
w/o ACQ	0.58	5.51	30.43	0.1296	0.2232	**0.9003**	0.2929	0.1866	**0.1306**	**0.0090**	**0.1548**
w/o ALI	3.67	18.01	**85.67**	0.1776	0.5834	0.7611	0.0682	0.2336	0.2892	0.3332	0.2311
Qwen	0.32	3.88	22.73	0.0768	0.0384	0.8522	0.2918	0.2532	0.1649	**0.0374**	**0.1868**
EVA$_{Qwen}$	**2.76**	**14.9**	**79.01**	**0.1464**	0.4609	**0.8967**	**0.0186**	**0.2175**	0.2596	0.3332	0.2094
w/o ACQ	0.34	3.77	21.73	0.0696	0.0408	0.8943	0.2936	0.2532	**0.1629**	**0.0374**	**0.1868**
w/o ALI	2.38	14.31	77.36	0.1368	**0.5114**	0.8295	0.0275	0.2180	0.2579	0.3332	0.2069
LLaMA	0.88	7.88	41.17	0.1440	0.6338	**1.2376**	0.1770	0.0645	**0.0037**	0.0178	0.0658
EVA$_{LLaMA}$	**2.75**	**16.02**	58.46	**0.2857**	**0.9699**	1.1140	**0.0213**	**0.0529**	0.0252	0.1277	**0.0568**
w/o ACQ	0.85	7.83	40.95	0.1392	0.8931	1.2136	0.1854	0.0835	0.0043	0.0788	0.0880
w/o ALI	2.22	14.70	**69.35**	0.2160	0.6554	0.9807	0.2047	0.2366	0.1308	**0.0085**	0.1451

Table 8: Experimental results of baselines using different training methods with ED dataset. B-2, R-L, and D-2 indicate BLEU2, ROUGE-L, and Dist-2 respectively. We denote empathy acquisition and emotional validation alignment training processes as ACQ and ALI, respectively. Within each model, the highest performance is indicated in **bold**.

Metric	**Emotional Validation (ρ)**
EVAEval (Avg.) vs. Human	**-0.560**
EPITOME-IP vs. Human	-0.219
EPITOME-EX vs. Human	-0.153
EPITOME-ER vs. Human	-0.134

Table 9: Spearman correlation coefficients between automatic metrics (EVAEval and EPITOME) and human evaluation scores. Lower EVAEval scores indicate better alignment with emotional validation theory.

D-2). Even though the general response generation metrics are limited to measure empathetic dialogue generation, our EVA-based models outperform the baseline.

We also conduct a comparative analysis between the EVA-based model and baseline models with EPITOME metrics (IP, EX, and ER) to evaluate the degree of empathy in the generated utterances. Similar to the results observed with ESConv dataset, the EVA-based model consistently outperforms the baseline models at IP and EX metrics. The improved performance in the IP and EX metrics indicates the superior capabilities of EVA-based models in emotional validation and expressing interest, respectively. The ER metric exhibits inconsistent results and even underperforms compared to GPT-3.5-Turbo and GPT-4o because it focuses solely on the degree of emotional expression without considering the progression of the conversation. Given that our goal is to convey emotional validation at specific levels rather than just expressing emotions in every utterance, the decrease in ER performance is understandable.

Ablation Ablation studies on the ED dataset show a pattern consistent with the ESConv dataset. The general response generation metrics, B-2 and R-L, decrease when either ACQ or ALI is omitted. Meanwhile, the D-2 metric increases when the ALI component is removed for both EVA$_{Mistral}$ and EVA$_{LLaMA}$. These results imply a trade-off between diversity and the ability to learn emotional validation. Furthermore, the IP metric, which measures emotional validation, also decreases when the ACQ and ALI is removed.

These findings suggest a trade-off between diversity and the model's ability to learn emotional validation. Furthermore, the IP metric, which measures emotional validation, declines when ACQ or ALI is removed, regardless of the baseline. This result highlights the crucial role of both components in enhancing empathetic response generation. Interestingly, the ER metric shows instability and, in some cases, even improves. This suggests that ER focuses solely on the degree of emotional expression without accounting for the progression of the conversation. Overall, this analysis underscores that both ACQ and ALI are essential for generating empathetic responses.

I.2 Emotional Validation Following

Main Results EVA-based models achieve a higher overall EVAEval score compared to most baselines, except for LLaMA. This discrepancy is primarily attributed to the short dialogue length of the ED dataset, which averages only 2.03 turns per conversation. The limited number of turns constrains the model's ability to produce higher-level utterances later in the conversation.

Despite this limitation, Table 8 shows that, across all foundation models, EVAEval scores for LEVEL 1 and LEVEL 2 are consistently lower than those of the baseline models. This trend suggests that our EVA method successfully generates lower-level utterances in the early stages of the conversation, aligning with the expected progression of emotional validation.

Ablation For the ED dataset, we focus on the LEVEL 1 and LEVEL 2 scores in EVAEval, as its short dialogue turns limit the model's ability to generate higher-level utterances in later stages of the conversation. In the case of LEVEL 1 and LEVEL 2, the score goes higher when we remove ACQ or ALI. This shows that both components contribute to generating empathetic responses following emotional validation.

J Evaluation Results of ESC-oriented Models

Among ESC-oriented models, we exclude EmoLLM from EVA fine-tuning due to its deployment via llama.cpp, which does not support gradient-based training. Therefore, we report results only for ChatCounselor and MeChat.

J.1 Empathetic Response Generation

As shown in Tables 10 and 11, EVA consistently improves performance across both general and empathy-specific metrics. On the ESConv dataset, EVA$_{ChatCounselor}$ significantly improves BLEU-2 and IP, while EVA$_{MeChat}$ achieves strong gains in R-L and D-2, alongside improved empathy scores. On the ED dataset, EVA$_{ChatCounselor}$ and EVA$_{MeChat}$ outperform baselines, with EVA$_{MeChat}$ achieving the highest BLEU-2 and IP across all models.

Ablation results reveal that removing either empathy acquisition (ACQ) or emotional validation alignment (ALI) consistently degrades performance, particularly on empathy-aware metrics (IP, EX). These findings align with the main experimental results (Table 2), supporting the complementary role of both training stages in generating contextually appropriate and emotionally supportive responses.

J.2 Emotional Validation Following

We evaluate emotional progression in generated responses using the EVAEval metric. On both ESConv and ED datasets, EVA-based models achieve the lowest average EVAEval scores, demonstrating strong alignment with the emotional validation framework. For instance, EVA$_{ChatCounselor}$ achieves an EVAEval score of 0.1068 on ESConv and 0.1301 on ED, both significantly lower than their respective baselines.

These findings reinforce the conclusions drawn in the main experiments: EVA effectively guides models to generate low-level (e.g., LEVEL 1, 2) responses early in the dialogue and high-level (e.g., LEVEL 3, 4) responses later, following the theoretical progression of emotional validation. Additionally, the ablation studies show that removing ALI leads to increased EVAEval scores, further confirming its role in preserving emotional structure.

Overall, these ESC-oriented results are consistent with our main findings, highlighting the generalizability of EVA across model architectures and domains. This results demonstrate that EVA not only improves standard and empathy-specific metrics but also promotes emotionally coherent interactions, even in specialized emotional support models.

	B-2	R-L	D-2	IP	EX	ER	EVAEval ↓				
							LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	Avg.
ChatCounselor	0.41	5.41	27.38	0.1181	0.0108	0.0694	0.2963	0.2493	0.1686	**0.0336**	0.1869
EVA$_{ChatCounselor}$	**1.21**	**9.42**	39.14	**0.1729**	**0.3916**	0.1927	**0.0418**	0.1936	**0.1533**	0.0386	**0.1068**
w/o ACQ	0.86	8.41	42.1	0.1148	0.3109	**0.2224**	0.2816	0.2194	0.1539	0.1046	0.1899
w/o ALI	0.78	8.18	**48.46**	0.1518	0.2025	0.006	0.2241	**0.1532**	0.2943	0.1492	0.2052
MeChat	1.96	10.45	51.85	0.0595	0.4131	0.5968	0.1022	0.2342	0.0342	0.2246	0.1488
EVA$_{MeChat}$	**2.04**	11.49	53.79	**0.1142**	**0.4421**	**0.5694**	**0.0642**	**0.2115**	**0.0286**	**0.2148**	**0.1298**
w/o ACQ	1.72	11.94	**55.28**	0.1023	0.4184	0.5913	0.0818	0.2831	0.1274	0.2479	0.1851
w/o ALI	1.29	**12.04**	58.01	0.1045	0.4294	0.4914	0.1012	0.3029	0.1843	0.3857	0.2435

Table 10: Experimental results of ESC-oriented models using different training methods with ESConv dataset. B-2, R-L, and D-2 indicate BLEU-2, ROUGE-L, and Dist-2 respectively. We denote empathy acquisition and emotional validation alignment training processes as ACQ and ALI, respectively. Within each model, the highest performance is indicated in **bold**.

	B-2	R-L	D-2	IP	EX	ER	EVAEval ↓				
							LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	Avg.
ChatCounselor	0.14	3.38	26.44	0.0216	0.0257	0.0300	0.3009	0.2532	0.3721	0.3728	0.3247
EVA$_{ChatCounselor}$	**1.42**	**8.15**	31.94	**0.1843**	**0.0714**	0.1035	**0.0138**	0.2037	**0.1552**	**0.1478**	**0.1301**
w/o ACQ	1.38	7.01	**32.83**	0.0732	0.0493	**0.1395**	0.1043	0.2185	0.1634	0.1528	0.1598
w/o ALI	0.64	6.81	31.38	0.1324	0.0621	0.1149	0.0842	**0.1943**	0.2953	0.2325	0.2016
MeChat	1.18	12.19	**56.70**	0.0720	0.4411	0.5931	0.2112	0.1920	0.2021	0.3136	0.2297
EVA$_{MeChat}$	**2.53**	**15.75**	52.61	**0.2938**	**0.5935**	**0.6821**	**0.1956**	**0.0374**	**0.1643**	0.2843	**0.1704**
w/o ACQ	2.14	13.29	54.87	0.2153	0.5729	0.6184	0.2642	0.0752	0.1739	**0.2143**	0.1819
w/o ALI	1.96	14.6	53.18	0.2319	0.4014	0.5953	0.2153	0.0412	0.1814	0.2703	0.1771

Table 11: Experimental results of ESC-oriented models using different training methods with ED dataset. B-2, R-L, and D-2 indicate BLEU-2, ROUGE-L, and Dist-2 respectively. We denote empathy acquisition and emotional validation alignment training processes as ACQ and ALI, respectively. Within each model, the highest performance is indicated in **bold**.

K The Details of Human Evaluation

We perform a comprehensive human evaluation to evaluate the efficacy of our EVA method in providing emotional validation. The annotators consist of 15 psychology majors with expertise in emotional validation theory. Each example is evaluated by ten different annotators across all criteria, with each annotator compensated $1 per example. We employ four criteria for human evaluation, divided into three main categories: Emotional Support Skills, Emotional Validation, and General Skills. The Emotional Support Skills category includes Comfortness and Comprehensibility. Emotional Validation focuses on how well the responses validate the help-seeker's emotions. Lastly, the General Skills category encompasses Coherence. A detailed description of the evaluation criteria is provided below.

- **Emotional Support**
  - *Comfortness*: Does the help-seeker feel comforted by the supporter's utterance?
  - *Comprehensibility*: Does the supporter understand the help-seeker's feelings?
- **Emotional Validation**
  - *Emotional Validation*: Does the supporter's utterance align with the concept of emotional validation?
- **General**
  - *Fluency*: Does the supporter's utterance flow naturally and smoothly?

L Additional Analyses

In this section, we present two additional analyses to strengthen the empirical evaluation of our proposed method.

L.1 Comparison with Strong In-Context Learning (ICL) Baseline

We additionally evaluated a strong in-context learning (ICL) baseline using chain-of-thought prompting (Mistral$_{CoT}$) to assess the necessity of our multi-stage fine-tuning approach. Table 12 presents the comparative results among Mistral, Mistral$_{CoT}$, and EVA$_{Mistral}$ across general and empathy-specific metrics.

Although Mistral$_{CoT}$ demonstrates improvements over the vanilla Mistral model—particularly in BLEU-2, ROUGE-L, and Interpretations

	B-2	R-L	D-2	IP	EX	ER	EVAEval ↓				
							LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	Avg.
Mistral	1.36	8.40	37.53	0.0810	0.4169	0.7857	0.3009	0.2394	0.1356	0.0230	0.1747
Mistral$_{CoT}$	1.81	10.31	39.38	0.1983	0.3920	0.8164	0.2710	0.1816	0.1031	0.0302	0.1464
EVA$_{Mistral}$	2.66	14.00	63.55	0.3050	0.4362	0.9073	0.0033	0.0408	0.0386	0.1300	0.0532

Table 12: Performance comparison with ICL baseline.

Task	Mistral	EVA$_{Mistral}$
HellaSwag	76.00	75.66
CoQA	60.00	65.28
MMLU	61.60	59.55
WikiText	9.86	9.02
HumanEval	28.63	26.21

Table 13: Zero-shot performance on general capabilities.

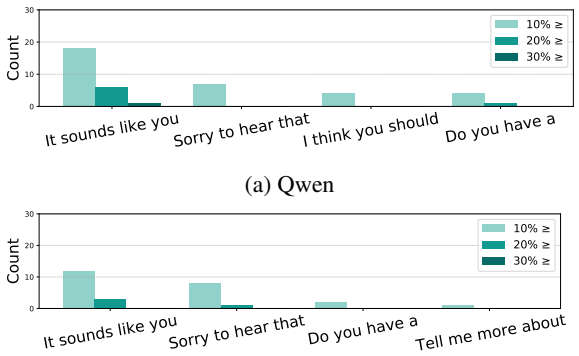
(IP)—its overall performance remains notably lower than that of EVA$_{Mistral}$. In particular, the EVAEval scores indicate that Mistral$_{CoT}$ still fails to produce emotionally aligned responses, showing inconsistent progression across emotional validation levels. In contrast, EVA$_{Mistral}$ consistently follows the intended level distribution (Levels 1–4), as guided by emotional validation theory.

These findings confirm that while strong ICL settings offer moderate gains, they are insufficient for achieving stable, theory-aligned empathetic responses. The proposed multi-stage fine-tuning approach remains essential for modeling emotional validation effectively.

L.2 Evaluation on General Language and Reasoning Tasks

To examine whether the empathy-aligned fine-tuning degrades the model's general capabilities, we evaluated EVA$_{Mistral}$ on a diverse set of standard benchmarks: HellaSwag, CoQA, MMLU, WikiText, and HumanEval. Table 13 presents a comparison between the baseline Mistral and EVA$_{Mistral}$ under 0-shot settings.

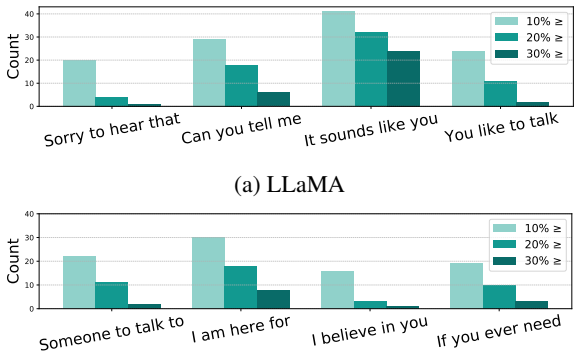
Despite the empathy-focused alignment, EVA$_{Mistral}$ demonstrates comparable performance across all tasks. Notably, it achieves higher accuracy on CoQA, suggesting potential gains in contextual understanding. These results indicate that the proposed empathy-oriented fine-tuning does not significantly compromise the model's general reasoning or coding abilities. Instead, EVA$_{Mistral}$ retains the broad utility of the underlying language model while gaining empathetic capabilities.



(a) Qwen

(b) EVA$_{Qwen}$

Figure 10: The number of dialogue where each phrase co-occurs more than 10%, 20%, and 30% within a single dialogue at Qwen and EVA$_{Qwen}$. Blue, orange, and green bars indicate the ratio of dialogue where each phrase is repeated more than 10%, 20%, and 30% within a conversation, respectively.



(a) LLaMA

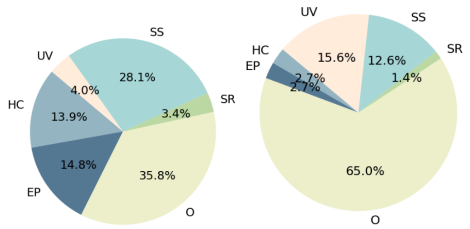
(b) EVA$_{LLaMA}$

Figure 11: The number of dialogue where each phrase co-occurs more than 10%, 20%, and 30% within a single dialogue at LLaMA and EVA$_{LLaMA}$. Blue, orange, and green bars indicate the ratio of dialogue where each phrase is repeated more than 10%, 20%, and 30% within a conversation, respectively.

M Analyses on Qwen and LLaMA

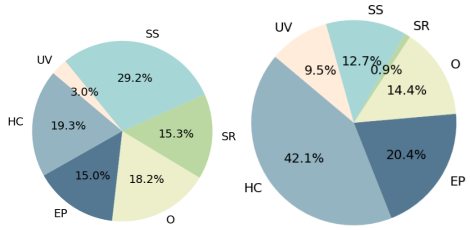
M.1 RQ1: Does EVA mitigate patterned responses?

The comparison results of the repetition ratio of EVA$_{Qwen}$ and EVA$_{LLaMA}$ with their respective baselines are presented in Figure 10 and Figure 11, respectively. The EVA-based model demonstrates a significant reduction in the repetition of phrases in both the Qwen and LLaMA models. Notably, EVA$_{Qwen}$ does not repeat any phrases more than 30%. This indicates that the EVA method enables the model to learn to generate responses with diverse phrases, improving its ability to produce emotionally validated responses. Additionally, by using a broader range of phrases, the EVA-based model enhances the richness and variety of dialogue generation.



(a) Qwen (b) EVA$_{Qwen}$

Figure 12: Response type proportions in utterances generated with Qwen and EVA$_{Qwen}$.



(a) LLaMA (b) EVA$_{LLaMA}$

Figure 13: Response type proportions in utterances generated with LLaMA and EVA$_{LLaMA}$.

M.2 RQ2: Is EVA still biased to solution suggestions?

The comparison results of the response types generated by EVA$_{Qwen}$ and EVA$_{LLaMA}$ are presented in Figure 12 and Figure 13, respectively. For the EVA$_{Qwen}$ and EVA$_{LLaMA}$, UV response type's proportion slightly increases, and SS decreases compared to vanilla. This result indicates that the EVA method not only affects the SS type of answers but also gives positive impacts by expressing compassion and encouragement.

Model	Generated Response
$\text{EVA}_{Mistral}$	*"You're absolutely right to feel upset about this. It shows how much you care about your friends, and it must really hurt to feel excluded. You don't deserve to be treated that way, and it's okay to tell them how you feel if you're ready. ... "*
Mistral	*"I understand you feel left out. Maybe you can try telling your friends how you feel."*

Table 14: Qualitative comparison showing the empathy–fluency trade-off. The EVA-generated response was rated higher in Emotional Validation but lower in Fluency due to its longer structure.