EmoReflect: A Reflective Emotion Reasoning Framework for Empathetic Support Conversations

Kuan-Lin Yu[1] **Kuan-Shun Su**[1] **Bing-Wei Lin**[1]
Wen-Ni Liu[2] **Lung-Hao Lee**[2,*]

[1] Institute of Manufacturing Information and Systems,
National Cheng Kung University, Tainan, Taiwan
[2] Institute of Artificial Intelligence Innovation,
National Yang Ming Chiao Tung University, Hsinchu, Taiwan
[*] lhlee@nycu.edu.tw

Abstract

Large language models (LLMs) have shown strong potential for emotional support conversations, yet directly generated responses may remain emotionally misaligned, overly generic, or insufficiently actionable. We propose EmoReflect, a psychology-inspired multi-agent framework that formulates emotional support generation as a draft-reflect-refine process. EmoReflect first predicts the user's valence-arousal (VA) state as auxiliary affective grounding and generates an initial supportive draft. Four specialized agents then review the draft from complementary perspectives: cognitive reflection, emotional regulation, user needs, and reality-oriented guidance. Their critiques are selectively integrated to produce the final response. Experiments on ESConv with three LLM backbones show that EmoReflect consistently improves overall response quality over standalone generation under both GPT-based and human evaluations, particularly in emotional understanding, helpfulness, practical actionability, and perspective richness. Ablation results further demonstrate the complementary contributions of the reflective agents and VA grounding, with cognitive reflection and emotional regulation showing particularly strong effects. These findings highlight the effectiveness of affectively grounded, multi-perspective reflection for improving LLM-based emotional support.

Keywords: Emotional support conversation, valence-arousal prediction, multi-agent reflection

1 Introduction

Emotional support is important for helping individuals cope with stress, frustration, and other challenging situations, yet timely interpersonal support may not always be available. Recent advances in large language models (LLMs) have created new opportunities for conversational systems that provide accessible emotional support (Yu and McGuinness, 2024; Bn et al., 2025). Research on empathetic dialogue has established benchmarks for recognizing and responding to users' emotional experiences (Rashkin et al., 2019), while Emotional Support Conversation (ESC) further emphasizes helping users cope through appropriate support strategies. ESConv (Liu et al., 2021) provides help-seeker–supporter conversations annotated with explicit support strategies, and subsequent studies have explored controlled empathetic generation, psychotherapy-inspired reasoning, and multi-agent collaboration (Hsu et al., 2023; Lee et al., 2023; Xu et al., 2025).

Despite these advances, directly generated LLM responses may still be emotionally misaligned, overly generic, or insufficiently useful. As illustrated in Fig. 1, a response may appear relevant to the latest utterance while failing to account for the user's broader emotional and conversational context. Effective support requires not only emotional acknowledgment but also appropriate consideration of the user's needs, cognitive perspective, and practical situation. Moreover, multidimensional evaluation suggests that empathetic dialogue quality extends beyond surface-level fluency (Xu and Jiang, 2024). These limitations motivate moving beyond one-step generation toward a process that explicitly reviews and improves an initial response before presenting it to the user.

To this end, we propose **EmoReflect**, a psychology-inspired multi-agent reflective

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 198–206
November 6-7, 2026 ©2026 Association for Computational Linguistics

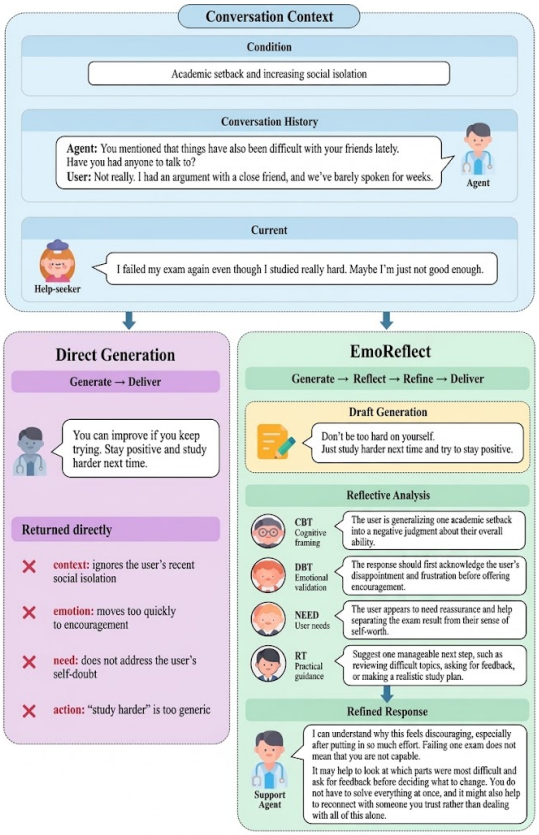

Figure 1: Illustrative comparison between direct LLM response generation and the proposed EmoReflect framework.

framework that formulates emotional support generation as a *draft–reflect–refine* process. EmoReflect first predicts valence–arousal (VA) information as auxiliary affective grounding and generates an initial supportive draft. Four specialized agents then review the same draft from complementary perspectives: *cognitive reflection*, *emotional regulation*, *user needs*, and *reality-oriented guidance*. Their critiques are integrated through a refinement module to produce the final response. Building on iterative reflection (Shinn et al., 2023; Madaan et al., 2023), multi-agent reasoning (Wu et al., 2024), and psychotherapy-inspired empathetic reasoning (Lee et al., 2023), EmoReflect distinguishes itself by using specialized agents as reflective critics of a shared initial response rather than relying on a single reasoning trajectory. Experiments on ESConv with three LLM backbones and both GPT-based and human evaluations demonstrate consistent overall improvements over standalone generation.

2 Related Work

2.1 Emotional Support Conversation

Recent studies on emotional support conversations have moved beyond empathy recognition toward explicit modeling of affective signals, support strategies, and conversational objectives. Sentiment-enhanced models incorporate emotion-related signals into empathetic dialogue modeling (Lin et al., 2023), while controlled generation has been explored to guide empathetic responses (Hsu et al., 2023). Other studies consider longer-term support objectives (Yang et al., 2025) and multi-agent collaboration for coordinating support strategies (Xu et al., 2025). While these approaches improve supportive response generation, less attention has been paid to systematically reviewing an already generated response from multiple perspectives. EmoReflect addresses this gap by combining affective grounding with multi-perspective reflection before final response generation.

2.2 Multi-Agent and Psychology-Inspired Reflection

Reflection improves LLM outputs through intermediate feedback and iterative revision. Reflexion uses verbal feedback for behavioral refinement (Shinn et al., 2023), while Self-Refine iteratively revises outputs using self-generated feedback (Madaan et al., 2023). Multi-agent frameworks such as AutoGen further enable role-specialized collaboration (Wu et al., 2024). In empathetic dialogue, Chain of Empathy introduces psychotherapy-inspired reasoning (Lee et al., 2023), while MultiAgentESC applies multi-agent collaboration to emotional support generation (Xu et al., 2025).

Building on these directions, EmoReflect employs multi-agent reasoning as a *post-generation reflective mechanism*. Specialized agents review the same initial draft from complementary psychological and support-oriented perspectives, and their critiques guide subsequent refinement. By separating generation, reflection, and refinement, EmoReflect differs from approaches that primarily use multiple agents during response generation and provides a structured mechanism for reconsidering an initial supportive response before final output.

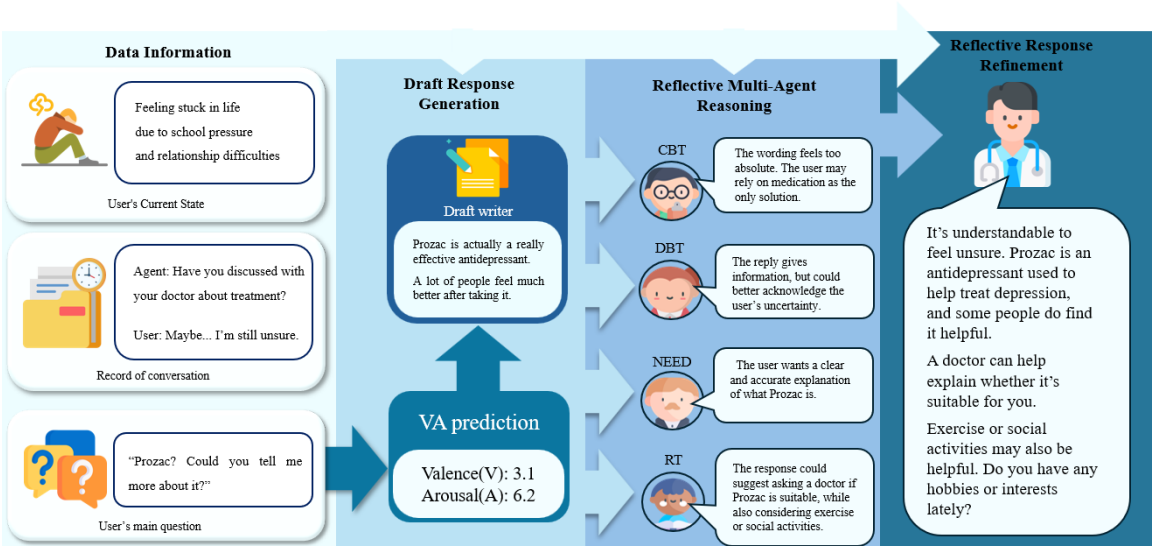

Figure 2: Overall framework of the proposed EmoReflect system.

3 Methodology

We propose **EmoReflect**, a psychology-inspired multi-agent reflective framework for emotional support dialogue. As illustrated in Fig. 2, EmoReflect adopts a *draft–reflect–refine* process that explicitly separates initial response generation from multi-perspective evaluation and subsequent refinement. Let $H_t = (s_t, C_{1:t-1})$ denote the user context at dialogue turn t, where s_t represents the user's current situational state and $C_{1:t-1}$ denotes the preceding conversation history. Given H_t and the current user utterance u_t, EmoReflect first estimates the user's affective state in the valence–arousal (VA) space, which serves as auxiliary emotional grounding for subsequent processing. Based on the dialogue context, current utterance, and predicted VA representation, the framework generates an initial supportive draft. Rather than presenting this draft directly to the user, four specialized reflective agents independently evaluate it from complementary psychological perspectives. Their critiques are then jointly incorporated by a refinement module to revise the initial draft and generate the final supportive response. This formulation enables EmoReflect to explicitly incorporate affective grounding and multi-perspective reflection into the response generation process, allowing potential deficiencies in the initial draft to be identified and addressed before producing the final response.

3.1 Draft Response Generation

Before generating a supportive response, EmoReflect explicitly grounds the generation process in the user's current affective state. Specifically, given the latest user utterance u_t, we estimate its affective characteristics using the two-dimensional valence–arousal (VA) space (Russell, 1980). The resulting representation,

$$e_t = (v_t, a_t) = f_{\mathrm{VA}}(u_t), \tag{1}$$

consists of the predicted valence v_t, which characterizes the degree of emotional positivity or negativity, and arousal a_t, which reflects the level of emotional activation or intensity. Unlike discrete emotion categories, the continuous VA representation provides a compact yet fine-grained characterization of the user's affective state, allowing emotionally different situations to be distinguished even when they share similar surface expressions or sentiment polarity (Yu et al., 2016; Lee et al., 2022).

The predicted VA representation is subsequently incorporated as *affective grounding* for response generation. Given the user context H_t, the current utterance u_t, and the affective representation e_t, the backbone LLM generates an initial supportive draft as

$$d_t = f_{\mathrm{draft}}(H_t, u_t, e_t). \tag{2}$$

By explicitly conditioning generation on VA information, EmoReflect provides the LLM with an additional affective signal beyond the emotional cues implicitly encoded in the dialogue

text. This grounding encourages the generated response to account for both the emotional direction and intensity of the user's current state. For example, user utterances with similarly negative valence may require different forms of support depending on whether they exhibit low arousal, such as sadness or discouragement, or high arousal, such as anxiety or frustration.

Accordingly, the initial draft is grounded jointly in the broader conversational context and the user's current affective state before multi-agent reflection is performed. The VA representation is not intended to prescribe a specific response strategy; rather, it serves as an auxiliary affective reference that helps align the initial response with the user's emotional condition and provides a consistent emotional signal for the subsequent reflective agents.

3.2 Reflective Multi-Agent Reasoning

The core of EmoReflect is a *reflective multi-agent reasoning* stage in which four specialized agents independently review the same initial draft d_t. Given the user context H_t, current utterance u_t, predicted VA representation e_t, and draft d_t, each agent produces a perspective-specific critique:

$$c_{i,t} = f_{\text{agent}_i}(H_t, u_t, e_t, d_t), \quad i = 1, \ldots, 4. \quad (3)$$

The agents serve as *reflective critics* rather than response generators, identifying potential weaknesses for subsequent refinement. The VA representation is retained as a shared affective reference throughout the reflective process.

Inspired by psychotherapy-oriented reasoning (Lee et al., 2023), EmoReflect decomposes reflection into four complementary perspectives:

Cognitive Reflection Agent (CBT). Inspired by Cognitive Behavioral Therapy (Rush, 1979), this agent examines whether the draft reinforces self-blame, rigid assumptions, or one-sided interpretations, and encourages more balanced cognitive framing.

Emotional Regulation Agent (DBT). Inspired by Dialectical Behavior Therapy (Linehan, 1993, 2025), this agent evaluates emotional validation, supportive tone, and alignment with the user's affective state, using VA information as additional emotional grounding.

User Needs Agent (NEED). This agent assesses whether the draft addresses the user's immediate needs, such as comfort, reassurance, explanation, or practical advice, and identifies responses that are overly generic or irrelevant.

Reality-Oriented Guidance Agent (RT). This agent evaluates whether suggested actions are realistic, feasible, and contextually appropriate while avoiding overly simplistic or prescriptive guidance.

Together, the four agents provide complementary critiques of cognitive framing, emotional alignment, user needs, and practical guidance. Their feedback enables EmoReflect to identify different weaknesses in the initial draft before final response refinement.

3.3 Final Response Refinement

The final stage of EmoReflect performs *selective response refinement* based on the multi-perspective feedback produced by the reflective agents. Given the user context H_t, current utterance u_t, predicted VA representation e_t, initial draft d_t, and the set of agent critiques $c_{i,t}i = 1^4$, the refinement module generates the final response as

$$r_t = f\text{refine}\left(H_t, u_t, e_t, d_t, c_{i,t}{}_{i=1}^{4}\right). \quad (4)$$

Rather than mechanically aggregating or incorporating every critique, the refinement module jointly reconsiders the original dialogue context and selectively integrates feedback that is relevant and compatible with the user's situation. This process allows potentially conflicting or unnecessary suggestions to be filtered while preserving the contextual coherence, emotional consistency, and linguistic naturalness of the initial response.

The predicted VA representation continues to serve as an affective reference during refinement, helping the module maintain alignment between the final response and the user's estimated emotional direction and intensity. Meanwhile, the four critiques provide complementary evidence regarding cognitive framing, emotional validation and regulation, immediate support needs, and realistic guidance. By jointly considering these signals, the refinement module can revise specific weaknesses

identified in the initial draft without unnecessarily altering appropriate content.

Consequently, the final response r_t represents a context-aware synthesis of the initial draft and multi-perspective reflective feedback rather than a simple combination of agent outputs. Through this selective refinement process, EmoReflect aims to produce responses that are more emotionally aligned with the user, responsive to their immediate support needs, contextually coherent, and practically helpful.

4 Evaluation

4.1 Experimental Setup

Datasets. We evaluate EmoReflect on ESConv (Liu et al., 2021), a multi-turn emotional support dataset containing help-seeker–supporter conversations. Each instance includes the preceding dialogue history, current situational state, and latest help-seeker utterance for response generation.

For affective grounding, the VA predictor is trained on the English subset of SemEval-2026 Task 3 Dimensional Aspect-Based Sentiment Analysis (Lee et al., 2026; Yu et al., 2026). The aspect-level valence and arousal annotations, ranging from 1 to 9, are averaged within each sentence to construct sentence-level targets. The resulting predictor estimates $e_t = (v_t, a_t)$ as auxiliary affective grounding for EmoReflect.

Implementation Details. All experiments are implemented using PyTorch and Hugging Face. The VA predictor is based on DeBERTa-v3-large (He et al., 2020) and jointly predicts valence and arousal using MSE loss. It is fine-tuned with AdamW for 5 epochs using a learning rate of 1×10^{-5}, batch size of 16, and maximum input length of 128.

We evaluate three instruction-tuned LLM backbones: Qwen2.5-7B[1], Llama-3.1-8B[2], and Gemma-2-9B[3], all under zero-shot settings. *Standalone* directly generates responses without VA grounding or reflection, whereas *EmoReflect* applies the complete draft–reflect–refine process. For reflective agents, the maximum input and output lengths are 16,384 and 256 tokens, respectively. We use temperature = 0.3, top-p = 0.9, repetition penalty = 1.1, and FP16 inference.

Evaluation Metrics. Generated responses are evaluated along six dimensions: *Emotional Understanding* (EU), *Fluency and Naturalness* (FN), *Helpfulness of Suggestion* (HS), *Practical Actionability* (PA), *Perspective Richness* (PR), and *Safety and Responsibility* (SR). Each dimension is rated on a five-point scale, and their sum forms the overall *Total* score.

4.2 GPT-Based Evaluation Results

For automatic evaluation, we employ GPT-5.4-mini as an LLM-based evaluator. For each evaluation instance, generated responses are anonymized and randomly ordered. The evaluator receives the dialogue context together with each candidate response and applies the same six-dimensional evaluation rubric described above. A score ranging from 1 to 5 is independently assigned to each dimension based solely on the provided conversational context and generated response. This evaluation protocol is applied consistently across all systems to reduce potential bias associated with model identity or generation method.

Table 1 presents the GPT-based evaluation results on the full ESConv evaluation set. Overall, incorporating EmoReflect improves the Total score across all three LLM backbones, although the magnitude of improvement varies across models. Qwen2.5-7B exhibits the largest gain, with its Total score increasing from 20.34 to 22.42. Llama-3.1-8B also improves from 20.32 to 21.14, while Gemma-2-9B shows a smaller increase from 20.02 to 20.24. These results suggest that the proposed draft–reflect–refine framework can improve emotional support generation across different LLM families, while its effectiveness may depend on the underlying backbone.

Among the six evaluation dimensions, the improvements are particularly evident for *Helpfulness of Suggestion* (HS) and *Practical Actionability* (PA). For Qwen2.5-7B, EmoReflect increases HS from 3.04 to 3.48 and PA from 2.48 to 3.10, representing the largest dimension-level gains observed for this backbone. Similar improvements are found for Llama-3.1-8B, where HS increases from 2.78

[1]https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[2]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[3]https://huggingface.co/google/gemma-2-9b-it

Table 1: GPT-based evaluation results on the full ESConv evaluation set.

Model	EU	FN	HS	PA	PR	SR	Total
Qwen2.5-7B	3.64	4.40	3.04	2.48	2.24	4.54	20.34
+ **EmoReflect**	**3.82**	**4.58**	**3.48**	**3.10**	**2.72**	**4.72**	**22.42**
Llama-3.1-8B	3.56	**4.54**	2.78	2.24	2.52	4.68	20.32
+ **EmoReflect**	**3.66**	4.50	**2.94**	**2.52**	**2.82**	**4.70**	**21.14**
Gemma-2-9B	3.60	4.50	2.84	2.32	**2.20**	**4.56**	20.02
+ **EmoReflect**	**3.64**	**4.52**	**2.98**	**2.40**	2.16	4.54	**20.24**

to 2.94 and PA from 2.24 to 2.52. Gemma-2-9B also achieves improvements in both dimensions, although with smaller margins. These findings indicate that multi-perspective reflection is particularly useful for identifying limitations in an initial response and refining it toward more helpful and practically applicable support.

EmoReflect also generally improves *Emotional Understanding* (EU), with scores increasing from 3.64 to 3.82 for Qwen2.5-7B, from 3.56 to 3.66 for Llama-3.1-8B, and from 3.60 to 3.64 for Gemma-2-9B. This trend suggests that incorporating explicit VA grounding together with reflective evaluation can help maintain better alignment with the user's emotional state. Improvements in *Perspective Richness* (PR) are also observed for Qwen2.5-7B and Llama-3.1-8B, increasing from 2.24 to 2.72 and from 2.52 to 2.82, respectively.

In contrast, the effects on *Fluency and Naturalness* (FN) and *Safety and Responsibility* (SR) are relatively small. Qwen2.5-7B improves on both dimensions, whereas Llama-3.1-8B shows a slight decrease in FN from 4.54 to 4.50, and Gemma-2-9B exhibits minor decreases in PR and SR. These small variations suggest that reflective refinement primarily contributes to the substantive quality of emotional support—particularly helpfulness, actionability, and emotional understanding—rather than uniformly improving every response dimension. Overall, the results demonstrate that EmoReflect can enhance direct LLM generation, with the strongest benefits observed when the underlying backbone can effectively incorporate and synthesize multi-perspective reflective feedback.

4.3 Human Evaluation Results

We additionally conduct human evaluation on 50 randomly sampled ESConv instances using the same six-dimensional rubric. Three graduate-student annotators independently evaluate the generated responses. None of the annotators has formal training in psychology or counseling; therefore, the evaluation focuses on the perceived quality of emotional support rather than clinical or therapeutic validity. Before annotation, all evaluators are provided with detailed definitions and scoring criteria for the six dimensions. As in the GPT-based evaluation, responses are anonymized and randomly ordered to conceal the corresponding backbone model and generation setting. The reported results are averaged across annotators and evaluation instances.

To assess annotation reliability, we calculate ordinal Krippendorff's α for each evaluation dimension. The resulting agreement scores are 0.686 for EU, 0.648 for FN, 0.708 for HS, 0.686 for PA, 0.684 for PR, and 0.661 for SR. Overall, these values indicate moderate to substantial agreement among the annotators across the six evaluation dimensions.

Table 2 presents the human evaluation results on 50 randomly sampled ESConv instances. Consistent with the GPT-based evaluation, EmoReflect improves the overall response quality across all three LLM backbones. The Total score increases from 24.40 to 24.82 for Qwen2.5-7B, from 22.75 to 24.00 for Llama-3.1-8B, and from 19.77 to 21.57 for Gemma-2-9B. These results provide additional evidence that the proposed draft–reflect–refine framework can enhance emotional support responses across different LLM families.

At the dimension level, EmoReflect generally produces consistent improvements across

Table 2: Human evaluation results on a randomly sampled subset of 50 ESConv instances.

Model	EU	FN	HS	PA	PR	SR	Total
Qwen2.5-7B	4.26	**4.19**	4.11	3.81	3.96	4.07	24.40
+ **EmoReflect**	**4.37**	4.15	**4.22**	**3.93**	**4.00**	**4.15**	**24.82**
Llama-3.1-8B	4.04	3.89	3.63	3.56	3.70	3.93	22.75
+ **EmoReflect**	**4.33**	**4.11**	**3.96**	**3.67**	**3.93**	**4.00**	**24.00**
Gemma-2-9B	3.37	3.59	3.26	3.11	2.96	3.48	19.77
+ **EmoReflect**	**3.89**	**3.93**	**3.52**	**3.15**	**3.15**	**3.93**	**21.57**

the six evaluation criteria. For Qwen2.5-7B, improvements are observed in five of the six dimensions, including *Emotional Understanding* (EU), *Helpfulness of Suggestion* (HS), *Practical Actionability* (PA), *Perspective Richness* (PR), and *Safety and Responsibility* (SR), while *Fluency and Naturalness* (FN) shows only a minor decrease from 4.19 to 4.15. For Llama-3.1-8B, EmoReflect improves all six dimensions, with particularly noticeable gains in EU (4.04→4.33), HS (3.63→3.96), and PR (3.70→3.93). These improvements suggest that reflective reasoning can strengthen both the emotional responsiveness and substantive helpfulness of generated support.

The largest overall improvement is observed for Gemma-2-9B, whose Total score increases by 1.80 points, from 19.77 to 21.57. EmoReflect improves all six dimensions for this backbone, with especially pronounced gains in EU (3.37→3.89), FN (3.59→3.93), and SR (3.48→3.93). This result indicates that reflective refinement can provide substantial benefits even when the initial standalone responses receive comparatively lower human-evaluation scores. At the same time, the relatively small improvement in PA (3.11→3.15) suggests that the magnitude of the benefit varies across evaluation dimensions.

Although the magnitude of improvement differs between GPT-based and human evaluation, both evaluation protocols consistently show higher Total scores for EmoReflect across all three backbones. Notably, GPT-based evaluation shows the largest improvement for Qwen2.5-7B, whereas human evaluators observe the largest gain for Gemma-2-9B. This difference suggests that LLM-based and human evaluators may emphasize different characteristics of emotional support responses. Nevertheless, the consistent overall improvements across both evaluation settings support the effectiveness of multi-perspective reflection for enhancing emotional support generation.

4.4 Component Ablation Study

To examine the contribution of individual components in EmoReflect, we conduct an ablation study using Qwen2.5-7B as the backbone. As shown in Table 3, removing any component results in a lower Total score than the full EmoReflect model.

Among the reflective agents, removing the *Cognitive Reflection Agent* (CBT) leads to the largest performance degradation, reducing the Total score from 22.42 to 21.24 ($\Delta = -1.18$). Removing the *Emotional Regulation Agent* (DBT) results in the second-largest decrease to 21.43 ($\Delta = -0.99$). These results suggest that examining cognitive framing and emotional appropriateness constitutes an important part of the reflective process. In comparison, removing the *User Needs Agent* (NEED) decreases the score by 0.61, while removing the *Reality-Oriented Guidance Agent* (RT) produces a smaller but still observable reduction of 0.28. Together, these results indicate that the four agents provide complementary perspectives, with cognitive and emotional reflection contributing particularly strongly to the final response quality.

Removing VA grounding also reduces the Total score, from 22.42 to 22.26 ($\Delta = -0.16$). Although this decrease is smaller than those caused by removing individual reflective agents, the result indicates that explicit affective information provides an additional signal beyond the textual dialogue context. This finding supports the role of VA as *auxiliary affective grounding* rather than as the primary

Table 3: Ablation results using Qwen2.5-7B. Δ denotes the difference from the full model.

Setting	Total	Δ
Full EmoReflect	**22.42**	–
w/o VA	22.26	−0.16
w/o RT	22.14	−0.28
w/o NEED	21.81	−0.61
w/o DBT	21.43	−0.99
w/o CBT	21.24	−1.18

driver of response quality: the predicted valence and arousal representation complements the multi-agent reflective process by providing an explicit reference to the user's estimated emotional direction and intensity.

Overall, the ablation results demonstrate that EmoReflect benefits from the combination of multi-perspective reflection and VA-based affective grounding. In particular, the larger performance drops observed when CBT and DBT are removed highlight the importance of jointly reconsidering cognitive framing and emotional alignment, while NEED, RT, and VA provide additional complementary information for refining the initial supportive response.

5 Conclusions

In this work, we presented **EmoReflect**, a psychology-inspired multi-agent reflective framework for emotional support dialogue that reformulates response generation as a *draft–reflect–refine* process. Rather than directly returning an LLM-generated response, EmoReflect incorporates valence–arousal information as auxiliary affective grounding and employs four specialized agents to review the initial draft from complementary perspectives, including cognitive reflection, emotional regulation, user needs, and reality-oriented guidance. Their critiques are subsequently integrated through selective refinement to produce the final supportive response.

Experiments on ESConv across three LLM backbones demonstrate that EmoReflect consistently improves overall response quality over standalone generation under both GPT-based and human evaluations. The improvements are particularly evident in dimensions related to emotional understanding, helpfulness, practical actionability, and perspective richness. Component ablation further confirms the complementary contributions of the proposed framework, with cognitive reflection and emotional regulation showing particularly strong effects, while VA grounding provides an additional affective signal beyond the textual dialogue context.

Overall, our findings suggest that emotional support generation can benefit from explicitly reconsidering an initial response from multiple perspectives before it is presented to the user. By combining affective grounding with structured reflective reasoning, EmoReflect provides a general framework for moving beyond one-step LLM generation toward more contextually appropriate, emotionally aligned, and practically helpful emotional support responses. Future work will investigate adaptive selection and coordination of reflective agents, more robust affective grounding from conversational history, and broader evaluation across emotional support scenarios and LLM backbones.

Acknowledgments

This study was partially supported by the National Science and Technology Council, Taiwan, under the grant NSTC 114-2221-E-A49-059-MY3. This work was also financially supported by the Co-creation Platform of the Industry Academia Innovation School, NYCU.

References

Suhas Bn, Yash Mahajan, Dominik O Mattioli, Andrew M Sherrill, Rosa I Arriaga, Christopher Wiese, and Saeed Abdullah. 2025. The pursuit of empathy: evaluating small language models for PTSD dialogue support. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30888–30910.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Jia-Hao Hsu, Jeremy Chang, Min-Hsueh Kuo, and Chung-Hsien Wu. 2023. Empathetic response generation based on plug-and-play mechanism with empathy perturbation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2032–2042.

Lung-Hao Lee, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese EmoBank: building valence-arousal resources for dimensional sentiment analysis. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(4).

Lung-Hao Lee, Liang-Chih Yu, Natalia V Loukachevitch, Ilseyar Alimova, Alexander Panchenko, Tzu-Mi Lin, Zhe-Yu Xu, Jian-Yu Zhou, Guangmin Zheng, Jin Wang, Sharanya Awasthi, Jonas Becker, Jan Philip Wahle, Terry Ruas, Shamsuddeen Hassan Muhammad, and Saif M. Mohammad. 2026. DimABSA: building multilingual and multidomain datasets for dimensional aspect-based sentiment analysis. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 40492–40518.

Yoon Kyung Lee, Inju Lee, Minjung Shin, Seoyeon Bae, and Sowon Hahn. 2023. Chain of empathy: enhancing empathetic response of large language models based on psychotherapy models. *arXiv preprint arXiv:2311.04915*.

Tzu-Mi Lin, Jung-Ying Chang, and Lung-Hao Lee. 2023. NCUEE-NLP at WASSA 2023 Shared Task 1: empathy and emotion prediction using sentiment-enhanced RoBERTa transformers. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 548–552.

Marsha Linehan. 1993. *Cognitive-behavioral treatment of borderline personality disorder*. Guilford press.

Marsha M Linehan. 2025. *DBT skills training manual*. Guilford Publications.

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.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

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.

A John Rush. 1979. Cognitive therapy for depression. *Australian and New Zealand Journal of Psychiatry*, 13(1):13–16.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2024. AutoGen: enabling next-gen LLM applications via multi-agent conversation. In *Proceedings of the First Conference on Language Modeling*.

Yangyang Xu, Jinpeng Hu, Zhuoer Zhao, Zhangling Duan, Xiao Sun, and Xun Yang. 2025. Multiagentesc: a LLM-based multi-agent collaboration framework for emotional support conversation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4665–4681.

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

Ting Yang, Li Chen, and Huimin Wang. 2025. Towards open-ended emotional support conversations in LLMs via reinforcement learning with future-oriented rewards. *arXiv preprint arXiv:2508.12935*.

Hong Qing Yu and Stephen McGuinness. 2024. An experimental study of integrating fine-tuned large language models and prompts for enhancing mental health support chatbot system. *Journal of Medical Artificial Intelligence*, 7:16.

Liang-Chih Yu, Jonas Becker, Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Lung-Hao Lee, Ying-Lung Lin, Jin Wang, Jan Philip Wahle, Terry Ruas, Natalia Loukachevitch, Alexander Panchenko, Ilseyar Alimova, Lilian Diana Awuor Wanzare, Nelson Odhiambo, Bela Gipp, Kai-Wei Chang, and Saif Mohammad. 2026. SemEval-2026 Task 3: Dimensional Aspect-Based Sentiment analysis (DimABSA). In *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, pages 3753–3778.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016. Building Chinese affective resources in valence-arousal dimensions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 540–545.