MultiAgentESC: A LLM-based Multi-Agent Collaboration Framework for Emotional Support Conversation

Yangyang Xu[1,2], Jinpeng Hu[3,†], Zhuoer Zhao[2], Zhangling Duan[2], Xiao Sun[3], Xun Yang[1,†]
[1]University of Science and Technology of China,
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center,
[3]Hefei University of Technology
yangyangxu@mail.ustc.edu.cn

Abstract

The development of Emotional Support Conversation (ESC) systems is critical for delivering mental health support tailored to the needs of help-seekers. Recent advances in large language models (LLMs) have contributed to progress in this domain, while most existing studies focus on generating responses directly and overlook the integration of domain-specific reasoning and expert interaction. Therefore, in this paper, we propose a training-free Multi-Agent collaboration framework for ESC (MultiAgentESC). The framework is designed to emulate the human-like process of providing emotional support through three stages: dialogue analysis, strategy deliberation, and response generation. At each stage, a multi-agent system is employed to iteratively enhance information understanding and reasoning, simulating real-world decision-making processes by incorporating diverse interactions among these expert agents. Additionally, we introduce a novel response-centered approach to handle the one-to-many problem on strategy selection, where multiple valid strategies are initially employed to generate diverse responses, followed by the selection of the optimal response through multi-agent collaboration. Experiments on the ESConv dataset reveal that our proposed framework excels at providing emotional support as well as diversifying support strategy selection[1].

1 Introduction

In recent years, escalating pressures from personal and occupational demands have significantly impacted individuals' mental health, while the shortage of psychologists has limited access to adequate support and care. Emotional Support Conversation (ESC) (Liu et al., 2021) is designed to comprehend the emotional issues of seekers, alleviate their psychological stress, and assist them in overcoming challenges through automatic natural language processing (NLP) techniques. Recently, many methods have been proposed for this task. For example, Tu et al. (2022) employs COMET (Bosselut et al., 2019), a commonsense reasoning model, to extract the user's fine-grained emotional states from the dialogue history, and utilizes a mixture of strategies to generate an emotional support response. However, these methods need to be trained on specific datasets, as well as suffer from limited generalization and extension, which restricts their application in real-world scenarios. Therefore, LLMs have been employed in this domain for their exceptional in-context learning and generalization abilities (Lee et al., 2023; Friedman et al., 2023). For example,

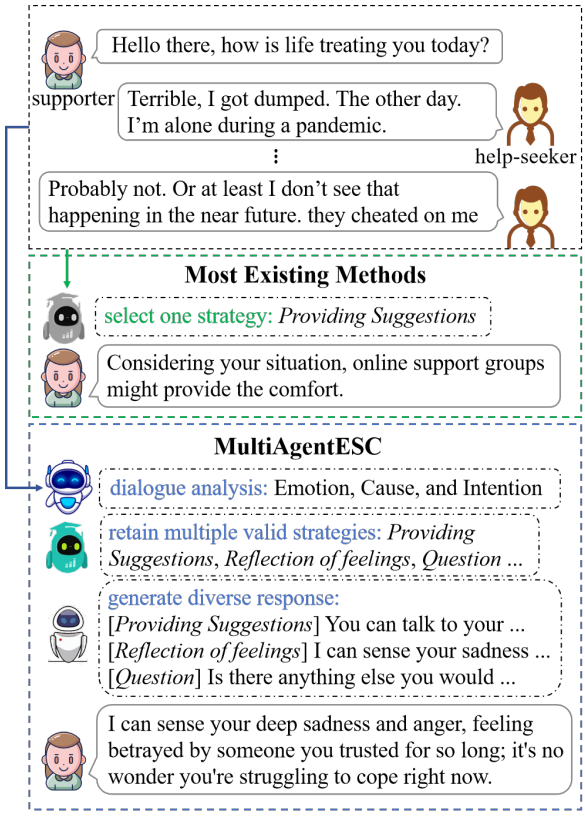

Figure 1: Our proposed MultiAgentESC framework and other methods to provide emotional support.

[†]Corresponding author.
[1]Our code is released at https://github.com/MindIntLab-HFUT/MultiAgentESC.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 4665–4681
November 4-9, 2025

Chen et al. (2023a) conceptualizes the ESC task as a mixed-initiative dialogue generation process, wherein it employs well-designed prompts to steer the LLM towards producing high-quality responses. Zheng et al. (2024) leverages LLMs as a "counseling teacher" to enhance the emotional support response capabilities of smaller models.

Despite the progress achieved by these efforts, they still face several noteworthy challenges. **First**, most existing methods tend to generate responses directly, with limited consideration with respect to the social interdependence theory (Johnson, 2003) and the Helping Skills Theory (Hill and O'Brien, 1999), which emphasize domain-specific reasoning and expert cooperation during the process of providing emotional support. **Second**, existing approaches mainly focus on *strategy-centered* frameworks, selecting appropriate strategies before generating the responses. However, these methods overlook the one-to-many characteristic in support strategy selection (Xu et al., 2022), where multiple strategies may be effective for a given patient context, posing challenges in directly identifying the optimal strategy. **Third**, the inherent characteristics of LLMs, such as the preference bias toward specific strategies, hinder their effectiveness in providing emotional support, as demonstrated in Kang et al. (2024); Zhao et al. (2023a).

Therefore, in this paper, we propose a multi-agent collaboration framework that emulates the process of psychological experts providing emotional support without relying on supervised training. A comparison of our approach with existing methods is presented in Figure 1. The framework facilitates seamless collaboration among specialized agents, with each agent focusing on a specific aspect so that agent cooperation mechanisms can further refine the LLM's ability to generate effective emotional support responses. In addition, on the one hand, we integrate similar cases into the prompt to provide additional prior knowledge. On the other hand, we introduce a novel response-centered solution that retains multiple valid support strategies. Afterward, these strategies are applied to generate distinct responses, which are subsequently evaluated and chosen through multi-agent debate and cooperation. In doing so, our proposed method can alleviate LLM preference bias and the one-to-many problem in the support strategy selection. Experimental results on the standard ESConv dataset demonstrate that our proposed MultiAgentESC framework not only delivers effective emotional support but also enhances the diversity of support strategy utilization.

2 Related Work

2.1 Emotional Dialogue

NLP has received sustained research attention over the past decades, with applications spanning tasks from named entity recognition to natural language generation (Hu et al., 2022a; Lample et al., 2016; Liu and Lapata, 2019; Hu et al., 2022b; Wang et al., 2025a). Within this landscape, Emotional Dialogue (Zhou et al., 2018) is a promising field dedicated to enhancing the ability of dialogue systems to respond to and engage with various emotions. Recent research (Zhao et al., 2023a; Welivita and Pu, 2024) suggests that LLMs hold significant potential in generating emotional responses. These approaches are broadly classified into two categories: training-based methods (Zheng et al., 2023; Chen et al., 2023b; Hu et al., 2024; Zhang et al., 2024c; Dai et al., 2025) and training-free methods (Su et al., 2023; Li et al., 2024; Abbasian et al., 2024). For example, SoulChat (Chen et al., 2023b) substantially improves its empathetic abilities by fine-tuning on a multi-turn empathetic dialogue dataset constructed with GPT-3.5-turbo. Zhang et al. (2024a) fine-tunes CPsyCounX on a high-quality multi-turn consultation dialogue dataset, achieving superior performance compared to other methods. However, these methods suffer from the need for a large amount of specialized data and computational resources, which necessitates the development of train-free approaches. For example, ECoT (Li et al., 2024) introduces a plug-and-play prompting technique that aligns with human emotional intelligence, enhancing LLMs' performance across various emotional generation tasks. Therefore, in this paper, we also propose a training-free framework to provide effective emotional support.

2.2 Emotional Support Conversation

Emotional Support Conversation (ESC), proposed by Liu et al. (2021), is designed to offer users professional emotional support, helping them effectively address their emotional difficulties (Xu et al., 2025). In recent years, the prevailing approach in ESC involves modifying and fine-tuning Pre-trained Language Models (PLMs) like BlenderBot (Roller et al., 2021) and BART (Lewis et al., 2020). For instance, MISC (Tu et al., 2022) and TransESC (Zhao et al., 2023b) utilize Blenderbot-small as

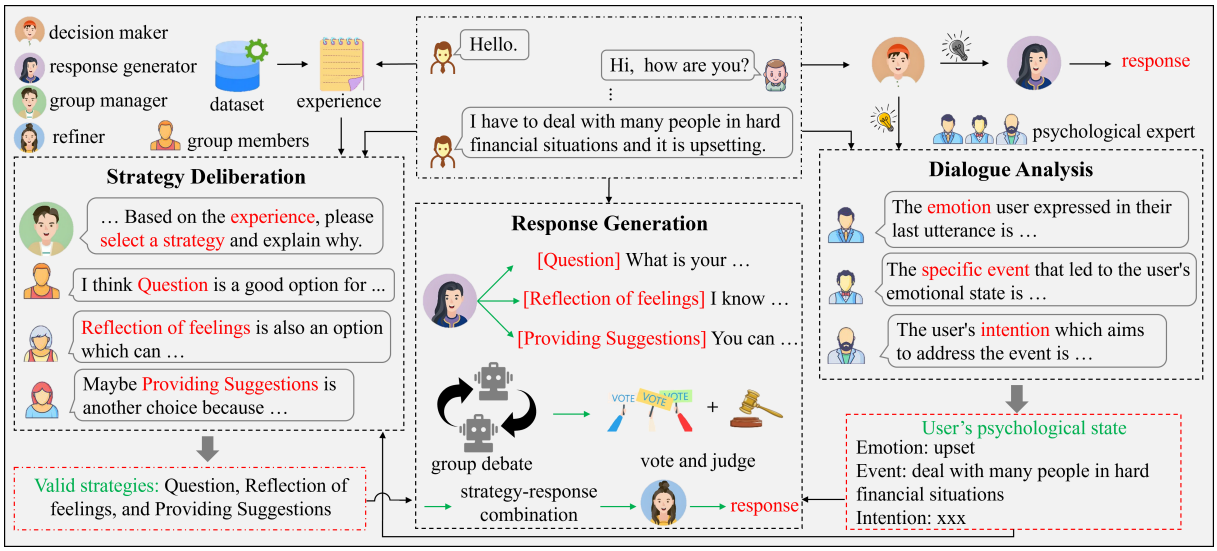

Figure 2: The overview of our proposed MultiAgentESC, which consists of three key stages: Dialogue Analysis, Strategy Deliberation, and Response Generation.

their backbone and incorporate additional modules that enable the model to perceive fine-grained emotional information. Meanwhile, the remarkable in-context learning and generalization capabilities of LLMs present new opportunities to advance this domain. Most existing studies (Ye et al., 2025; Zhang et al., 2024b; Zheng et al., 2024) focus on the construction of specialized datasets, which serve as a fundamental resource for emotional support. In addition, some researchers conducted an in-depth analysis on whether LLM is a good emotional supporter. For example, Kang et al. (2024) suggests that the inherent preference bias in LLMs (Pan et al., 2024, 2025) significantly impedes their ability to deliver effective emotional support. Inspired by this, we propose a novel framework named MultiAgentESC to mitigate these issues in this work.

2.3 LLM-based Multi-Agent System

The application of LLM-based multi-agent systems (MAS) has expanded across multiple domains, showcasing their robust planning and reasoning abilities in various complex scenarios (Qian et al., 2024; Zeng et al., 2024a; Liu et al., 2025; Hu et al., 2025a,b). To facilitate the development of LLM applications, a series of multi-agent frameworks has been proposed (Wu et al., 2024; Hong et al., 2024; Chen et al., 2024). Among them, AutoGen (Wu et al., 2024) is a prominent open-source framework designed for developing AI agents and enabling multi-agent collaboration to solve complex tasks, which facilitates the efficient construction of agentic workflows. In this paper, we propose a multi-agent system to provide effective emotional support to help-seekers, implemented based on AutoGen.

3 Methodology

In this section, we present the details of the MultiAgentESC framework. As demonstrated in Figure 2, MultiAgentESC consists of three stages. Firstly, during **Dialogue Analysis** stage, multiple agents play different roles to extract the user's psychological state from the dialogue context. Then in the **Strategy Deliberation** stage, we retrieve similar cases from the dataset and integrate them into the following deliberation process to alleviate the preference bias of LLMs. Following this, multiple valid strategies are retained for further utilization. In the **Response Generation** stage, these strategies are used to generate diverse responses, from which the optimal response is selected through multi-agent debate and collaboration. All the prompts can be found in the Appendix A.

3.1 Stage 1: Dialogue Analysis

As described in Algorithm 1, during the dialogue analysis stage, decision maker A_d firstly determines whether the dialogue context necessitates multi-agent collaborative analysis. We propose that a dialogue context can be considered sufficiently simple to forgo multi-agent collaboration for further analysis under either of the following conditions: (1) there are fewer than 5 exchanges in the conversation, or (2) the dialogue context

Algorithm 1 Dialogue Analysis Stage

Input: Dialogue history H, decision maker A_d, emotion agent A_e, cause agent A_c, intention agent A_i, response generator A_r
Output: User's psychological state S or emotional support response R

$F \leftarrow A_d(H)$
if $F \neq$ True **then**
    $R \leftarrow A_r(H)$
    Return R
else
    $E \leftarrow A_e(H)$
    $C \leftarrow A_c(H, E)$
    $I \leftarrow A_i(H, E, C)$
    $S \leftarrow (E, C, I)$
    Return S
end if

does not reflect the user's psychological state[2].In this case, the response generator A_r directly provides emotional support through zero-shot learning mechanisms.

When multi-agent collaborative analysis is required, the framework deploys three specialized agents: A_e for emotional state extraction, A_c for causal event identification, and A_i for intention recognition. These agents operate sequentially to extract the user's psychological state S, which is subsequently utilized in the following strategy deliberation and response generation stage.

3.2 Stage 2: Strategy Deliberation

The ESC task presents a one-to-many challenge in strategy selection (Xu et al., 2022), where multiple valid strategies can be appropriately aligned with a single dialogue context. Therefore, it is difficult to select the most appropriate strategy solely based on the contextual information. Additionally, the inherent biases of LLMs present a substantial obstacle to providing effective emotional support, as these models demonstrate a marked tendency to favor specific support strategies while overlooking potentially more appropriate alternatives.

To address these limitations, we propose an enhanced approach that integrates prior knowledge into the strategy deliberation process and maintains a diverse set of potentially effective support strategies through multi-agent discussion. Specifically, we retrieve semantically similar cases from the dataset by employing an off-the-shelf pre-trained model SBERT[3] to calculate the cosine similarity between the user's last utterance and all candidate entries. Subsequently, we identify the top-k samples with the highest similarity scores as experience and integrate them into the following deliberation process through the group manager[4]. Additionally, group members are encouraged to prioritize the selection of diverse strategies during the deliberation process, thereby ensuring that multiple distinct strategies are retained as valid candidates S_s for final response generation.

Algorithm 2 Response Generation Stage

Input: Dialogue history H, User's psychological state S, Strategy set S_s, response generator A_r, group members S_a, Judge A_j, Refiner A_{re}
Output: Response R

$S_r \leftarrow [\,]$
for s in S_s **do**
    $r \leftarrow A_r(H, S, s)$
    Add r to S_r
end for
$H_d \leftarrow \text{Debate}(H, S, S_r, S_a)$
$H_r \leftarrow \text{Reflect}(H, S, S_r, S_a, H_d)$
$S_v \leftarrow \text{Vote}(H_r)$
if $\text{len}(S_v) > 1$ **then**
    $R \leftarrow A_j(H, S, S_v)$
else
    $R \leftarrow S_v[0]$
end if
$R \leftarrow A_{re}(H, S, R)$
Return R

3.3 Stage 3: Response Generation

As illustrated in Algorithm 2, during the response generation stage, the response generator A_r produces emotional support responses for each strategy within S_s derived from the last stage. To select the optimal response, we employ a multi-expert cooperation mechanism where psychological counseling specialists S_a, each holding different initial opinions to support different strategy-response combinations, engage in a group debate to articulate their perspectives. Following initial position presentations, the experts participate in reflective discourse, critically revising their stances through

[2]We utilize a LLM to make the judgment by crafting an appropriate prompt.

[3]https://huggingface.co/sentence-transformers/all-roberta-large-v1

[4]In AutoGen (Wu et al., 2024), the group manager is a special agent used to initialize and manage the group discussion.

mutual exchange. This process culminates in a voting mechanism to identify the strategy-response pairing with maximal consensus. In cases of tied votes, judge A_j will determine the final selection. This innovative approach effectively transforms the complex challenge of strategy selection into a more tractable response evaluation task, thereby offering a new solution for the one-to-many problem. Ultimately, the refiner A_{re} evaluates the response based on the following criteria: (1) whether this response is consistent with the ongoing conversation, (2) whether it aligns with the strategy, and (3) whether it effectively helps alleviate the user's emotional stress, and provide a refined version.

4 Experiments

4.1 ESConv Dataset

Our experiments are conducted on the ESConv (Liu et al., 2021) dataset, a benchmark for emotional support conversations, with approximately 1K conversations and 31K utterances. This dataset is collected in a help-seeker and supporter mode with crowdworkers. In each conversation, the supporter provides emotional support to the seeker with a bad emotional state through professional conversational skills, which can be categorized into eight distinct types (e.g., *Question*, *Reflection of feelings*, and *Providing Suggestions*)[5]. In light of the substantial time and computational costs associated with inference using LLMs, we randomly select 100 conversations as the test set to validate the effectiveness of our proposed MultiAgentESC. The remainder will be utilized for extracting cases analogous to the target dialogue, serving as valuable experiential references.

4.2 Baselines

Our proposed MultiAgentESC is a training-free framework based on LLMs and prompt engineering techniques. For a fair comparison, we compare it to the following training-free baselines: **Zero-shot** (Brown et al., 2020), **Few-shot** (Brown et al., 2020), **Zero-shot CoT** (Kojima et al., 2022), **Few-shot CoT** (Wei et al., 2022), **Self-consistency** (Wang et al., 2023b), **Self-Refine** (Madaan et al., 2023), **Mixed-Initiative** (Chen et al., 2023a), **ESCoT** (Zhang et al., 2024b), **CogChain** (Cao et al., 2024), and **Cooper** (Cheng et al., 2024). Moreover, a comparison is conducted between two multi-agent system (MAS) approaches utilizing different topological structures: **MAS(Chain)** and **MAS(Debate)**. In MAS(Chain), three agents are assigned to dialogue analysis, strategy selection, and response generation, respectively, and operate within a sequential chain topology where information is passed unidirectionally between agents. In MAS(Debate), three agents independently generate responses using distinct strategies, thereby exploring a diverse range of potential solutions. The final response is determined through a judge agent, which synthesizes the outputs to form an optimized decision. More details about them are described in Appendix D.

4.3 Implementation Details

In this study, we employ LLaMA3-70b (Grattafiori et al., 2024) and Qwen2.5-32b (Yang et al., 2024a), both derived from Ollama[6], an open-source tool for managing and deploying LLMs, as the foundational models[7]. We search 10 similar cases from the dataset as experience and inject them into the strategy deliberation process. The strategy deliberation process is designed with three participants, and in the subsequent response generation stage, the size of the debate group matches the number of strategies under consideration, guaranteeing that each member initially represents a distinct viewpoint. This study is implemented based on the AutoGen (Wu et al., 2024), an open-source programming framework for building AI agents and facilitating cooperation among multiple agents to solve tasks. We set the temperature to 0 during the inference process, and all experiments are conducted on an A800 GPU.

4.4 Evaluation Metrics

Automatic Evaluation. For the evaluation of response generation, we employ the following metrics: (1) Distinct-1 (**D-1**) and Distinct-2 (**D-2**) (Li et al., 2016) to measure response diversity, and (2) BLEU-1 (**B-1**), BLEU-2 (**B-2**), BLEU-3 (**B-3**) (Papineni et al., 2002), along with **F1** and ROUGE-L (**R-L**) (Lin, 2004), to evaluate response quality. Additionally, (3) we analyze the **strategy distribution** across different methods to validate the effectiveness of our approach in mitigating the preference bias of LLMs.

[5]The details of strategies can be found in Appendix B.

[6]https://ollama.com/

[7]All results are based on Qwen2.5-32b unless specified.

Method	LLaMA3-70b							Qwen2.5-32b						
	D-1↑	D-2↑	B-1↑	B-2↑	B-3↑	F1↑	R-L↑	D-1↑	D-2↑	B-1↑	B-2↑	B-3↑	F1↑	R-L↑
Zero-shot (Brown et al., 2020)	6.65	31.25	17.52	5.22	2.17	18.11	14.45	6.37	32.02	**17.76**	5.31	2.27	18.20	14.59
Few-shot (Brown et al., 2020)	6.53	30.78	17.78	**5.58**	2.23	18.25	14.57	6.40	32.22	17.59	5.29	2.27	18.23	14.52
Zero-shot CoT (Kojima et al., 2022)	5.06	21.12	17.28	4.73	1.63	17.75	13.78	5.37	24.96	17.41	5.28	2.16	17.12	13.51
Few-shot CoT (Wei et al., 2022)	5.01	21.60	16.66	4.64	1.77	16.18	12.71	5.17	25.91	16.43	4.68	1.97	16.67	12.92
Self-consistency (Wang et al., 2023b)	5.43	24.11	15.64	4.07	1.54	15.49	12.25	5.66	30.45	16.16	4.42	1.74	16.53	12.80
Self-Refine (Madaan et al., 2023)	6.37	32.62	14.97	4.69	2.14	17.11	13.35	6.13	33.46	16.24	5.08	2.15	17.96	14.31
Mixed-Initiative (Chen et al., 2023a)	6.03	28.12	15.42	4.35	1.88	16.01	12.59	6.15	30.39	15.26	4.10	1.61	15.11	11.94
ESCoT (Zhang et al., 2024b)	5.08	21.54	16.63	4.93	2.03	17.43	13.84	6.58	34.48	15.32	4.53	1.98	16.18	13.02
CogChain (Cao et al., 2024)	4.92	20.89	15.02	4.06	1.59	16.31	12.97	5.79	34.56	16.13	4.54	1.79	16.34	12.99
Cooper (Cheng et al., 2024)	5.38	31.29	17.18	5.13	2.11	17.54	13.99	5.71	33.41	16.22	4.89	1.86	16.55	13.34
MAS(Chain)	5.11	28.96	16.87	4.67	2.06	17.52	13.62	5.07	29.71	16.59	4.97	2.17	17.26	13.78
MAS(Debate)	5.43	32.15	17.05	5.33	2.21	17.68	14.07	5.42	34.14	16.95	5.15	2.16	17.48	13.94
MultiAgentESC (Ours)	**6.84**	**33.88**	**17.83**	5.56	**2.41**	**18.60**	**14.69**	**6.78**	**35.15**	17.66	**5.38**	**2.35**	**18.30**	**14.66**

Table 1: Comparison of our MultiAgentESC against state-of-the-art baselines in terms of the automatic evaluation. The best results among all methods are highlighted in **bold**.

5 Results and Analysis

5.1 Automatic Evaluation

Response Generation. As demonstrated in Table 1, our proposed MultiAgentESC framework delivers strong performance across most automatic evaluation metrics, implemented with both LLaMA3-70b and Qwen2.5-32b. Specifically, MultiAgentESC achieves superior results on the D-1 and D-2 metrics, highlighting its exceptional capability in generating diverse emotional support responses. This observation can be attributed to its innovative approach of incorporating experience and collaborative interactions among multiple agents, which mitigates the preference bias of LLM in strategy selection, thereby enhancing the diversity of strategy choices and response generation. Furthermore, MultiAgentESC achieves promising results on reference-based metrics like B-n, F1, and R-L, showcasing its proficiency in generating high-quality responses. This may be attributed to the following two factors. On the one hand, MultiAgentESC performs an in-depth analysis of the dialogue context from multiple perspectives, facilitating the extraction of user-related information. On the other hand, the effective multi-agent interaction mechanism facilitates optimal response selection among multiple candidates, which may further enhance the quality of the generated responses.

Strategy Selection. Figure 3 illustrates the strategy distributions obtained by various approaches. The results indicate that strategy-centered methods, such as Zero-shot CoT, Self-consistency, CogChain, and ESCoT, exhibit a strong propensity to favor a limited subset of strategies. Among the eight support strategies, only four (*Reflection of feelings*, *Affirmation and Reassurance*, *Others*, and *Question*) show utilization rates above 2%. Moreover, the distribution among these strategies is extremely unbalanced: *Reflection of feelings* dominates with over 70% usage, while both *Others* and *Question* each account for less than 5%. In contrast, our proposed MultiAgentESC demonstrates a more balanced and diversified distribution of strategy selection, effectively mitigating the risk of over-reliance on a limited set of strategies. All strategies within MultiAgentESC maintain utilization rates exceeding 3%, with none surpassing one-third of the total usage. This observed phenomenon is likely primarily driven by two key contributing factors. First, experience provides a valuable reference in the process of strategy deliberation, effectively mitigating the inherent preference bias of LLMs in strategy selection. Second, the response-centered approach we adopt prioritizes the selection of optimal responses over predetermined strategies, thereby potentially mitigating the preference of LLMs toward specific strategies.

Further Analysis. As shown in Figure 3, although our proposed MultiAgentESC framework exhibits promising capabilities in selecting diverse strategies, slight discrepancies persist compared with the standard strategy distribution observed in the ESConv dataset. The following are some potential explanations we have identified: (1) Some strategies serve similar functions. For example, *Question* and *Restatement or Paraphrasing* can be used to inquire about the user's current situation. In such cases, the more commonly used strategy, *Question*, is more likely to be selected by the LLM. (2) During the initial phase of the dialogue, we employ Zero-shot to generate responses directly, so

MultiAgentESC vs.	Zero-shot			Few-shot CoT			Self-consistency			Self-Refine		
	Win	Lose	Tie	Win	Lose	Tie	Win	Lose	Tie	Win	Lose	Tie
Fluency	**48***	22	30	**52***	18	30	**50***	19	31	**44***	19	37
Identification	**46***	19	35	**55***	13	32	**53***	15	32	**46***	22	32
Comforting	**41**	24	35	**48**	21	31	**49***	20	31	**42**	20	38
Suggestion	**52***	16	32	**54***	18	28	**54***	19	27	**40**	22	38
Overall	**49**	21	30	**53***	19	28	**52***	18	30	**44***	21	35

Table 2: The results of human evaluations(%). Our proposed MultiAgentESC performs better than all other methods (sign test, * represents p-value < 0.05).

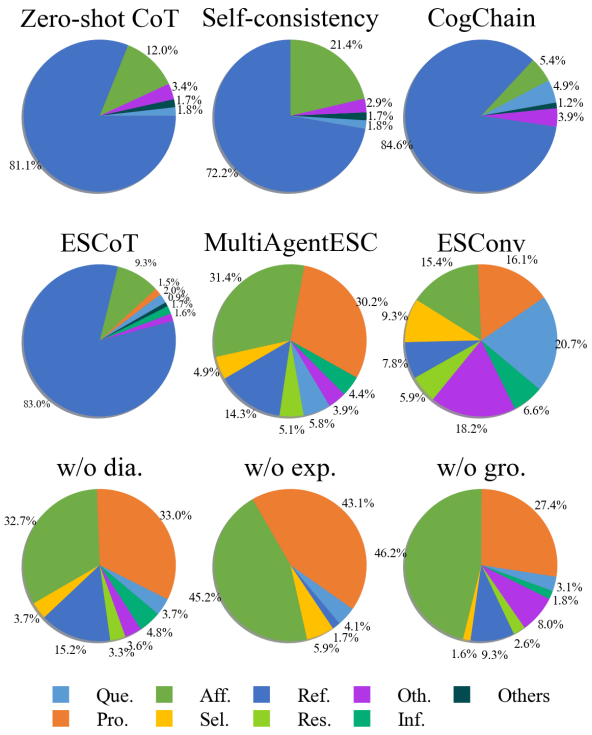

Figure 3: Strategy distributions of MultiAgentESC and compared methods. For simplicity, we abbreviate each strategy using its first three letters (e.g., Que. represents *Question*). Note that Oth. is the abbreviation of *Others* strategy, while "Others" represents the collection of strategies with utilization rates below 1%.

there are no strategies considered in the statistical process. As a result, strategies that are typically employed at this stage, such as *Question*, *Restatement or Paraphrasing*, and *Others*, show low utilization rates.

5.2 Human Evaluation

Following Liu et al. (2021); Zhao et al. (2023b), we recruit three postgraduate students with psychology backgrounds as annotators to evaluate the performance of our proposed MultiAgentESC framework and other methods. Specifically, we randomly sample 100 dialogues from the test set of the ESConv dataset and instruct the annotators to assume the role of help-seekers under these dialogue scenarios. Each annotator is tasked with comparing all responses generated by our method against those produced by other methods across the 100 dialogues. To ensure fairness, the annotators are blinded to the source model of each response. Given MultiAgentESC and a compared baseline method, the annotators are required to choose which model performed better (or tie) across the following dimensions: (1) **Fluency**: which model generates more coherent and smooth responses; (2) **Identification**: which model is more effective at identifying your problems; (3) **Comforting**: which model is better at comforting you; (4) **Suggestion**: which model provides more useful suggestions; (5) **Overall**: which model provides more effective emotional support.

As illustrated in Table 2, MultiAgentESC exhibits superior performance across all evaluation metrics compared to the baseline methods. Specifically, it generates more fluent and contextually coherent responses, which may be attributed to the in-depth analysis of dialogue context. Moreover, MultiAgentESC outperforms other methods in *Identification*, *Comforting*, and *Suggestion*, demonstrating its ability to effectively employ diverse support strategies. This success validates the advantages of integrating experience into the process of strategy deliberation and the efficacy of our response-centered approach.

5.3 GPT-4o Judgements

In this section, we employ GPT-4o (Hurst et al., 2024), a large language model with robust reasoning capabilities, to evaluate the performance of our proposed framework and other baseline methods. Similar to human evaluation, we assess the performance of different methods based on the following five aspects: **Fluency**, **Identification**, **Comforting**, **Suggestion**, and **Overall**. The difference lies

Methods	Flu.	Ide.	Com.	Sug.	Ove.
Zero-shot	4.702	3.323	4.132	3.115	3.823
Few-shot CoT	4.614	3.109	4.068	2.738	3.764
Self-consistency	4.602	3.258	3.875	2.727	3.685
Self-Refine	4.715	3.427	3.972	3.104	3.792
MultiAgentESC (Ours)	**4.786**	**3.729***	**4.323***	**3.462***	**4.028***

Table 3: The results of GPT-4o judgements.

Variants	D-1	D-2	B-1	B-2	R-L
w/o dialogue analysis	6.70	33.98	17.55	5.28	14.38
w/o experience	6.50	33.31	17.64	5.32	14.54
w/o group discussion	6.57	34.44	17.50	5.31	14.58
MultiAgentESC (Ours)	**6.78***	**35.15***	**17.66**	**5.38**	**14.66***

Table 4: Results of the ablation study.

in the fact that we simultaneously score all the methods, including our proposed MultiAgentESC framework, rather than comparing other baselines with MultiAgentESC one by one. The prompt utilized for this is illustrated in Appendix C.

As shown in Table 3, the evaluation results of GPT-4o closely align with those of human evaluation. Our proposed MultiAgentESC framework achieves optimal results across all the evaluation aspects. Notably, MultiAgentESC demonstrates promising advancements in both *Identification* and *Suggestion* aspects. This may be because MultiAgentESC can effectively employ strategies such as *Question* and *Providing Suggestions*, thereby enhancing their ability to identify users' emotional problems and provide suggestions.

5.4 Ablation Study

To explore the impact of individual components within our proposed MultiAgentESC framework, we conduct an ablation study by designing five distinct variants, as detailed in Table 4 and Figure 3. These variants include: (1) **w/o dialogue analysis** (w/o dia.), which eliminates the dialogue analysis module; (2) **w/o experience** (w/o exp.), where experience is removed from the strategy deliberation phase; and (3) **w/o group discussion** (w/o gro.), where both strategy selection and response generation are handled by a single agent rather than through collaboration of multiple agents.

As shown in Table 4, the ablation of each component can lead to a drop in the automatic evaluation results, demonstrating the essential role of these components in delivering effective emotional support. Furthermore, as demonstrated in Figure 3, the strategy distribution of *w/o dialogue analysis* closely aligns with that of MultiAgentESC, indicating that this component has limited impact on the strategy selection process. In contrast, both the *w/o group discussion* and *w/o experience* exhibit substantial deviations from MultiAgentESC in their strategy distributions, with the latter showing more pronounced differences. These findings suggest two key insights: (1) the incorporation of experience plays a crucial role in reducing the preference bias of LLM during strategy selection, and (2) our response-centered approach contributes to mitigating this bias, albeit to a lesser extent.

5.5 Case Study

In Table 5, we present a case with responses generated by MultiAgentESC and other baseline methods. This case involves a help-seeker facing a dilemma between continuing work, which may endanger their son's health, and quitting, which would lead to financial instability. The baseline methods, including Zero-shot, Few-shot, and Self-consistency, predominantly employ the *Reflection of feeling* strategy. While these approaches acknowledge the emotional state of the help-seeker, they fail to provide actionable suggestions, leading to ineffective emotional support. Self-Refine, through its self-optimization mechanism, attempts to address this limitation by incorporating the *Providing Suggestions* strategy. However, its responses lack sufficient emotional empathy, resulting in a rigid and less engaging interaction style. In contrast, MultiAgentESC demonstrates superior performance by conducting a comprehensive analysis of the dialogue context. It accurately identifies the user's desperation emotion and the event that led to the user's emotional state. By leveraging the collaborative capabilities of multiple agents and incorporating experience as contextual references, our approach effectively combines the *Providing Suggestion* strategy with empathetic engagement. This dual focus not only offers practical solutions but also ensures emotional resonance, which shows a remarkable alignment with the ground truth.

6 Conclusion

In this paper, we propose MultiAgentESC, an innovative training-free framework that simulates human-like emotional support processes through three key stages: dialogue analysis, strategy deliberation, and response generation. Similar cases are incorporated to mitigate the preference bias of LLMs. Furthermore, we introduce a novel

Dialogue history	**Assistant**: I understand how you feel. Can you tell me a little more about how it makes you feel when they don't honor your requests? **User**: It makes me feel trapped. Frustrated. ... **User**: I feel trapped because, what are my options? Continue to work.....my son gets sick and dies. His pediatrician has already warned me how deadly this could be for me. **User**: And if I quit? No money? I have a 5 year old I have to take care of too.
Zero-shot	It sounds like you're feeling overwhelmed by the impossible choices you're facing, and it's understandable that you'd feel trapped and frustrated.
Few-shot CoT	**[Reflection of feelings]** I sense your desperation and fear, feeling stuck between protecting your son's health and providing for your family's well-being.
Self-consistency	**[Reflection of feelings]** I sense your desperation and anxiety about protecting your son's health while being stuck between your job requirements and financial responsibilities.
Self-Refine	Have you considered speaking with HR or management about implementing stricter safety protocols or exploring remote work options?
MultiAgentESC (Ours)	**[Providing suggestions]** It's heartbreaking to feel trapped between your job and your son's health. Have you considered exploring remote work options or discussing possible accommodations with your employer to minimize the risk?
Ground-Truth	**[Providing suggestions]** I can see exactly how frustrated you are in this situation. Before I share some ideas do you have any ideas of ways to combat this issue?

Table 5: Case study of the generated responses by our proposed MultiAgentESC and the baselines.

response-centered approach to effectively resolve the one-to-many challenge in strategy selection. Results on the ESConv dataset demonstrate that our proposed framework exhibits strong capabilities in both delivering effective emotional support and enhancing the diversity of support strategies. In the future, we will try to build a multimodal emotional support dialogue system that empowers MultiAgentESC with the ability to understand multimodal emotions (Han et al., 2025; Song et al., 2023, 2024) based on Multimodal LLMs (Zeng et al., 2024b; Song et al., 2025). In addition, the ability to generalize across domains (Yang et al., 2024b; Wang et al., 2023a, 2025b) will also be seriously considered in our multimodal ESC system.

7 Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) grant U22A2094 and grant 62402158, and also by the advanced computing resources provided by the Supercomputing Center of the USTC.

8 Limitations

While our method outperforms the baselines, there are several challenges that call for further exploration. First, a primary limitation of MultiAgentESC lies in its difficulty in distinguishing between analogous strategies, thereby compromising its effectiveness in providing emotional support. This underscores the crucial need for specialized LLMs designed to provide emotional support. Second, MultiAgentESC still faces the persistent challenge of biased strategy utilization, highlighting the necessity to investigate more efficient methods for strategy selection.

9 Ethics Statement

Although our proposed method demonstrates efficacy in diversifying strategy selection and generating emotional support responses, several ethical considerations must be addressed to ensure its responsible deployment. The application of AI techniques, especially LLMs, in real-world settings carries inherent risks. Without appropriate protective measures, they may generate harmful content. Consequently, our objective is to deliver emotional support within the context of daily dialogue, without replacing professional psychological therapy. Ethical considerations also encompass privacy and data protection. In this work, we utilize the ESConv dataset, which is a well-established open-access benchmark for the ESC task, without personal information involved.

References

Mahyar Abbasian, Iman Azimi, Mohammad Feli, Amir M Rahmani, and Ramesh Jain. 2024. Empathy through multimodality in conversational interfaces. *arXiv preprint arXiv:2405.04777*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for knowl-

edge graph construction. In *Association for Computational Linguistics (ACL)*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Yaru Cao, Zhuang Chen, Guanqun Bi, Yulin Feng, Min Chen, Fucheng Wan, Minlie Huang, and Hongzhi Yu. 2024. Enhancing emotional support conversation with cognitive chain-of-thought reasoning. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 175–187. Springer.

Maximillian Chen, Xiao Yu, Weiyan Shi, Urvi Awasthi, and Zhou Yu. 2023a. Controllable mixed-initiative dialogue generation through prompting. *ACL 2023*.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, et al. 2024. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *ICLR*.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023b. 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.

Yi Cheng, Wenge Liu, Jian Wang, Chak Tou Leong, Yi Ouyang, Wenjie Li, Xian Wu, and Yefeng Zheng. 2024. Cooper: coordinating specialized agents towards a complex dialogue goal. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, pages 17853–17861.

Chongyuan Dai, Jinpeng Hu, Hongchang Shi, Zhuo Li, Xun Yang, and Meng Wang. 2025. Psyche-r1: Towards reliable psychological llms through unified empathy, expertise, and reasoning. *arXiv preprint arXiv:2508.10848*.

Luke Friedman, Sameer Ahuja, David Allen, Zhenning Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, Harsh Lara, et al. 2023. Leveraging large language models in conversational recommender systems. *arXiv preprint arXiv:2305.07961*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Zhiyuan Han, Beier Zhu, Yanlong Xu, Peipei Song, and Xun Yang. 2025. Benchmarking and bridging emotion conflicts for multimodal emotion reasoning. In *ACM Multimedia*.

Clara E Hill and Karen M O'Brien. 1999. Helping skills: Facilitating exploration, insight, and action.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. In *12th International Conference on Learning Representations, ICLR 2024*.

Jinpeng Hu, Tengteng Dong, Luo Gang, Hui Ma, Peng Zou, Xiao Sun, Dan Guo, Xun Yang, and Meng Wang. 2024. Psycollm: Enhancing llm for psychological understanding and evaluation. *IEEE Transactions on Computational Social Systems*.

Jinpeng Hu, Zhuo Li, Zhihong Chen, Zhen Li, Xiang Wan, and Tsung-Hui Chang. 2022a. Graph enhanced contrastive learning for radiology findings summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4677–4688.

Jinpeng Hu, Yaling Shen, Yang Liu, Xiang Wan, and Tsung-Hui Chang. 2022b. Hero-gang neural model for named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1924–1936.

Jinpeng Hu, Hongchang Shi, Chongyuan Dai, Zhuo Li, Peipei Song, and Meng Wang. 2025a. Beyond emotion recognition: A multi-turn multimodal emotion understanding and reasoning benchmark. *arXiv preprint arXiv:2508.16859*.

Jinpeng Hu, Ao Wang, Qianqian Xie, Hui Ma, Zhuo Li, and Dan Guo. 2025b. Agentmental: An interactive multi-agent framework for explainable and adaptive mental health assessment. *arXiv preprint arXiv:2508.11567*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

David W Johnson. 2003. Social interdependence: interrelationships among theory, research, and practice. *American psychologist*, 58(11):934.

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*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 22199–22213.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. 2023. Prompted llms as chatbot modules for long open-domain conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4536–4554.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, pages 110–119.

Zaijing Li, Gongwei Chen, Rui Shao, Dongmei Jiang, and Liqiang Nie. 2024. Enhancing the emotional generation capability of large language models via emotional chain-of-thought. *arXiv preprint arXiv:2401.06836*.

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

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740.

Ziyan Liu, Chunxiao Fan, Haoran Lou, Yuexin Wu, and Kaiwei Deng. 2025. Mind: A multi-agent framework for zero-shot harmful meme detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 923–947.

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

Haowen Pan, Yixin Cao, Xiaozhi Wang, Xun Yang, and Meng Wang. 2024. Finding and editing multimodal neurons in pre-trained transformers. In *ACL (Findings)*.

Haowen Pan, Xiaozhi Wang, Yixin Cao, Zenglin Shi, Xun Yang, Juanzi Li, and Meng Wang. 2025. Precise localization of memories: A fine-grained neuron-level knowledge editing technique for llms. In *ICLR*.

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.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. 2024. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*.

Peipei Song, Dan Guo, Xun Yang, Shengeng Tang, and Meng Wang. 2024. Emotional video captioning with vision-based emotion interpretation network. *IEEE Transactions on Image Processing*, 33:1122–1135.

Peipei Song, Dan Guo, Xun Yang, Shengeng Tang, Erkun Yang, and Meng Wang. 2023. Emotion-prior awareness network for emotional video captioning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 589–600.

Shezheng Song, Xiaopeng Li, Shasha Li, Shan Zhao, Jie Yu, Jun Ma, Xiaoguang Mao, Weimin Zhang, and Meng Wang. 2025. How to bridge the gap between modalities: Survey on multimodal large language model. *IEEE Transactions on Knowledge and Data Engineering*.

Guinan Su, Yanwu Yang, and Jie Guo. 2023. Prompt your mind: Refine personalized text prompts within your mind. *arXiv preprint arXiv:2311.05114*.

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

Chengyu Wang, Shan Zhao, Tianwei Yan, Shezheng Song, Wentao Ma, Kuien Liu, and Meng Wang. 2025a. Hierarchical label-enhanced contrastive learning for chinese ner. *IEEE Transactions on Neural Networks and Learning Systems*.

Shanshan Wang, Yiyang Chen, Zhenwei He, Xun Yang, Mengzhu Wang, Quanzeng You, and Xingyi Zhang. 2023a. Disentangled representation learning with causality for unsupervised domain adaptation. In *Proceedings of the 31st ACM international conference on multimedia*, pages 2918–2926.

Shanshan Wang, Houmeng He, Xun Yang, Zhipu Liu, Yuanhong Zhong, Xingyi Zhang, and Meng Wang. 2025b. Exploring invariance matters for domain generalization. *IEEE Transactions on Image Processing*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

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

Anuradha Welivita and Pearl Pu. 2024. Is chatgpt more empathetic than humans? *arXiv preprint arXiv:2403.05572*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversation. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Xiaohan Xu, Xuying Meng, and Yequan Wang. 2022. Poke: Prior knowledge enhanced emotional support conversation with latent variable. *arXiv preprint arXiv:2210.12640*.

Yangyang Xu, Zhuoer Zhao, Xiao Sun, and Xun Yang. 2025. Prompt learning with multiperspective cues for emotional support conversation systems. *IEEE Transactions on Computational Social Systems*.

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

Xun Yang, Tianyu Chang, Tianzhu Zhang, Shanshan Wang, Richang Hong, and Meng Wang. 2024b. Learning hierarchical visual transformation for domain generalizable visual matching and recognition. *International Journal of Computer Vision*, 132(11):4823–4849.

Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2025. Sweetiechat: A strategy-enhanced role-playing framework for diverse scenarios handling emotional support agent. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4646–4669.

Yifan Zeng, Yiran Wu, Xiao Zhang, Huazheng Wang, and Qingyun Wu. 2024a. Autodefense: Multi-agent llm defense against jailbreak attacks. *arXiv preprint arXiv:2403.04783*.

Zhen Zeng, Leijiang Gu, Xun Yang, Zhangling Duan, Zenglin Shi, and Meng Wang. 2024b. Visual-oriented fine-grained knowledge editing for multimodal large language models. *arXiv preprint arXiv:2411.12790*.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024a. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13947–13966.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024b. Escot: Towards interpretable emotional support dialogue systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13395–13412.

Yiqun Zhang, Fanheng Kong, Peidong Wang, Shuang Sun, Lingshuai Wang, Shi Feng, Daling Wang, Yifei Zhang, and Kaisong Song. 2024c. Stickerconv: Generating multimodal empathetic responses from scratch. *arXiv preprint arXiv:2402.01679*.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023a. Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*.

Weixiang Zhao, Yanyan Zhao, Shilong Wang, and Bing Qin. 2023b. Transesc: Smoothing emotional support conversation via turn-level state transition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6725–6739.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584*.

Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. 2024. Self-chats from large language models make small emotional support chatbot better. In *Proceedings of the 62nd Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 11325–11345.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: emotional conversation generation with internal and external memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, pages 730–738.

A Prompt Template

In this work, we have meticulously designed prompts to implement the MultiAgentESC framework and compared baselines as shown in Table 6, 7, and 8.

B Emotional Support Strategies

The definitions of strategies are the same as (Liu et al., 2021).

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

Restatement or Paraphrasing A simple, more concise rephrasing of the user's statements that could help them see their situation more clearly.

Reflection of feelings Articulate and describe the user's feelings.

Self-disclosure Divulge similar experiences that you have had or emotions that you share with the user to express your empathy.

Affirmation and Reassurance Affirm the user's strengths, motivation, and capabilities and provide reassurance and encouragement.

Providing Suggestions Provide suggestions about how to change, but be careful to not overstep and tell them what to do.

Information Provide useful information to the user, for example with data, facts, opinions, resources, or by answering questions.

Others Exchange pleasantries and use other support strategies that do not fall into the above categories.

C Prompt Used for GPT-4o Judgements

We use GPT-4o to score our method and the compared baselines. The prompts are shown in Figure 4.

D Baselines

- **Zero-shot** (Brown et al., 2020), **Few-shot** (Brown et al., 2020), **Zero-shot CoT** (Kojima et al., 2022), **Few-shot CoT** (Wei et al., 2022): They are widely used prompting techniques in LLMs, differing in example usage and reasoning steps.
- **Self-consistency** (Wang et al., 2023b): It boosts the performance of CoT by selecting the most consistent answer from multiple reasoning paths.
- **Self-Refine** (Madaan et al., 2023): Self-Refine generates an initial response and its feedback, then uses this feedback to iteratively refine the response.
- **ESCoT** (Zhang et al., 2024b) and **CogChain** (Cao et al., 2024): They are designed to address ESC-related tasks. Based on their approaches, we design specific prompts for each and compare them with our proposed method.
- **Mixed-Initiative** (Chen et al., 2023a): Mixed-Initiative utilizes a well designed prompt with user's background information, such as emotion type, problem type and user's situation to generate high-quality responses.
- **Cooper** (Cheng et al., 2024): It highlights the multifaceted nature of complex dialogue goals by employing multiple specialized agents, each dedicated to addressing a specific aspect of the dialogue objectives.

Prompt Template

 # Role
You are a judge with a background in psychology and linguistics.

Task
You are provided with a dialogue history between the Assistant and the User, along with 5 responses provided by 'A', 'B', 'C', 'D', 'E'.
Please score these 5 responses from the following aspects and provide convincing reasons.

Evaluation Aspects
Fluency (1-5 points): Please evaluate the fluency of the response. Does the response flow fluently within the dialogue and coherent with the context?

Identification (1-5 points): Please evaluate the effectiveness of the response in identifying issues. Does the response thoroughly explore the user's situation and successfully pinpoint the problems?

Comforting (1-5 points): Please evaluate the ability of the response to provide comfort. Does the response demonstrate skill in offering reassurance and empathy, making you feel more at ease and supported during the interaction?

Suggestion (1-5 points): Please evaluate the quality of the suggestions provided by the response. Does the response offer useful and practical recommendations to address your problems?

Overall (1-5 points): Please evaluate the overall emotional support provided by the response. Generally, how much do you favor this response?

Constraints
- Avoid any position biases and ensure that the order in which the answers were presented does not influence your decision.
- Do not allow the length of the answers to influence your evaluation.
- Do not favor certain names of the assistants. Be as objective as possible.

Workflow
For the following dialogue history and 5 response,
dialogue history: {context}
A: {response1}
B: {response2}
C: {response3}
D: {response4}
E: {response5}

Output your final verdict by strictly following this format:
Fluency: [A_rating], [B_rating], [C_rating], [D_rating], [E_rating]; The reasons for the scores are as follows: [reasons],
Identification: [A_rating], [B_rating], [C_rating], [D_rating], [E_rating]; The reasons for the scores are as follows: [reasons],
Comforting: [A_rating], [B_rating], [C_rating], [D_rating], [E_rating]; The reasons for the scores are as follows: [reasons],
Suggestion: [A_rating], [B_rating], [C_rating], [D_rating], [E_rating]; The reasons for the scores are as follows: [reasons],
Overall: [A_rating], [B_rating], [C_rating], [D_rating], [E_rating]; The reasons for the scores are as follows: [reasons].

Figure 4: GPT-4o evaluation prompt.

Prompt 1: Dialogue Analysis (Emotion)

Instruction
You are a psychological counseling expert. You will be provided with a dialogue context between an 'Assistant' and a 'User'.
Please infer the emotional state expressed in the user's last utterance.
Dialogue context
{context}
Your answer must include the following elements:
Emotion: the emotion user expressed in their last utterance.
Reasoning: the reasoning behind your answer.
Your answer must follow this format:
Emotion: [emotion]
Reasoning: [reasoning]

Prompt 2: Dialogue Analysis (Specific Event)

Instruction
You are a psychological counseling expert. You will be provided with a dialogue context between an 'Assistant' and a 'User'.
Another agent analyzes the conversation and infers the emotional state expressed by the user in their last utterance.
Dialogue context
{context}
Emotional state
{emotion}
Please infer the specific event that led to the user's emotional state based on the dialogue context.
Your answer must include the following elements:
Event: the specific event that led to the user's emotional state.
Reasoning: the reasoning behind your answer.
Your answer must follow this format:
Event: [event]
Reasoning: [reasoning]

Prompt 3: Dialogue Analysis (Intention)

Instruction You are a psychological counseling expert. You will be provided with a dialogue context between an 'Assistant' and a 'User'.
Other agents have analyzed the conversation, inferring the emotional state expressed by the user in their last utterance and the specific event that led to the user's emotional state.
Dialogue context
{context}
Emotional state
{emotion}
Event
{cause}
Please reasonably infer the user's intention based on the dialogue context, with the goal of addressing the event that lead to their emotional state.
Your answer must include the following elements:
Intention: user's intention which aims to address the event that lead to their emotional state.
Reasoning: the reasoning behind your answer.
Your answer must follow this format:
Intention: [intention]
Reasoning: [reasoning]

Prompt 4: Strategy Deliberation (Group Discussion Initiator)

You will be provided with a dialogue context between an 'Assistant' and a 'User'.
Psychologists have analyzed the conversation, inferring the emotional state expressed by the user in their last utterance, the specific event that led to the user's emotional state and user's intention aiming to address the event that lead to their emotional state.
Dialogue context
{context}
Emotional state
{emotion}
Event
{cause}
Intention
{intention}
Based on the provided information and dialogue context, please select a strategy for the 'Assistant' to generate an appropriate response, and explain why. The following are examples of different strategies, all presented in the format of <post \n[strategy] response>.
Examples
{examples}
Your answer must include the following elements:
Strategy: Strategy for generating a response. The strategy must appear in the examples. Please choose different strategies as much as possible.
Reasoning: the reasoning behind your answer.
Your answer must follow this format:
Strategy: [strategy]
Reasoning: [reasoning]

Table 6: Prompts in MultiAgentESC and baselines.

Prompt 5: Response Generation (Response Generation with Strategy)

You will be provided with a dialogue context between an 'Assistant' and a 'User'.
Psychologists have analyzed the conversation, inferring the emotional state expressed by the user in their last utterance, the specific event that led to the user's emotional state and user's intention aiming to address the event that lead to their emotional state.
Dialogue context
{context}
Emotional state
{emotion}
Event
{cause}
Intention
{intention}
Please generate a response from the Assistant's perspective using the {strategy} strategy.
The following are examples of this strategy, all presented in the format of.
Examples
{examples}
Your answer must be fewer than 30 words and must follow this format:
Response: [strategy] [response]

Prompt 6: Response Generation (Group Debate Initiator)

You will be provided with a dialogue context between an 'Assistant' and a 'User'.
Psychologists have analyzed the conversation, inferring the emotional state expressed by the user in their last utterance, the specific event that led to the user's emotional state and user's intention aiming to address the event that lead to their emotional state.
Dialogue context
{context}
Emotional state
{emotion}
Event
{cause}
Intention
{intention}
Based on the provided information and dialogue context, please select the most appropriate response from the following options and explain why.
Response
{responses}
Your answer must include the following elements:
Response: the most appropriate response and the strategy used in this response.
Reasoning: the reasoning behind your answer.
Your answer must follow this format:
Response: [strategy] [response]
Reasoning: [reasoning]

Prompt 7: Response Generation (Group Reflection Initiator)

You will be provided with a dialogue context between an 'Assistant' and a 'User'.
Psychologists have analyzed the conversation, inferring the emotional state expressed by the user in their last utterance, the specific event that led to the user's emotional state and user's intention aiming to address the event that lead to their emotional state.
Dialogue context
{context}
Emotional state
{emotion}
Event
{cause}
Intention
{intention}
Based on the provided information and the context of the dialogue, a group discussion is taking place to determine which response is the most appropriate.
Discussion content
{discussion content}
You should carefully analyze the various different viewpoints above, reflect on your own thoughts, and ultimately arrive at a convincing result.
Your thought can be changed if you believe the viewpoints of others are more reasonable.
Your answer must include the following elements:
Response: the most appropriate response and the strategy used in this response.
Reasoning: the reasoning behind your answer.
Your answer must follow this format:
Response: [strategy] [response]
Reasoning: [reasoning]

Table 7: Prompts in MultiAgentESC and baselines.

Prompt 8: Baselines (Zero-shot)

Instruction
You are a psychological counseling expert. You will be provided with a dialogue context between an 'Assistant' and a 'User'.
Your task is to play a role as 'Assistant' and generate a response based on the given dialogue context.
Dialogue context
{context}
Your answer must be fewer than 30 words and must follow this format:
Response: [response]

Prompt 9: Baselines (Few-shot)

You are a psychological counseling expert. You will be provided with a dialogue context between an 'Assistant' and a 'User'.
Your task is to play the role of 'Assistant' and generate a response based on the given dialogue context.
The following are some examples, all presented in the format of <context\n response>.
Examples
{examples}
Dialogue context
{context}
Your answer must be fewer than 30 words and must follow this format:
Response: [response]

Prompt 10: Baselines (Zero-shot CoT)

Instruction
You are a psychological counseling expert. You will be provided with a dialogue context between an 'Assistant' and a 'User'.
Dialogue context
{context}
You should select an appropriate emotional support strategy first and then generate a strategy-constrained response.
{strategy definition}
Please ensure that you are absolutely fair and do not overly favor any particular strategy.
Your answer must include the following elements:
Strategy: the most appropriate strategy.
Reasoning: the reason why you choose this strategy.
Response: strategy-constrained response. Response must be fewer than 30 words.
Your answer must follow this format:
Strategy: [strategy]
Reasoning: [reasoning]
Response: [response]
Let's think step by step!

Prompt 11: Baselines (Few-shot CoT)

Instruction
You are a psychological counseling expert. You will be provided with a dialogue context between an 'Assistant' and a 'User'.
Dialogue context
{context}
You should select an appropriate emotional support strategy first and then generate a strategy-constrained response.
{strategy definition}
Please ensure that you are absolutely fair and do not overly favor any particular strategy.
The following are some examples, all presented in the format of <context\n strategy\n reasoning\n response>.
Examples
{examples}
Your answer must include the following elements:
Strategy: the most appropriate strategy.
Reasoning: the reason why you choose this strategy.
Response: strategy-constrained response. Response must be fewer than 30 words.
Your answer must follow this format:
Strategy: [strategy]
Reasoning: [reasoning]
Response: [response]

Table 8: Prompts in MultiAgentESC and baselines.