SweetieChat: A Strategy-Enhanced Role-playing Framework for Diverse Scenarios Handling Emotional Support Agent

Jing Ye[1,2], Lu Xiang[1,2*], Yaping Zhang[1,2], Chengqing Zong[1,2]

[1]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, CAS, Beijing, China

[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

yejing2022@ia.ac.cn; {lu.xiang, yaping.zhang,cqzong}@nlpr.ia.ac.cn

Abstract

Large Language Models (LLMs) have demonstrated promising potential in providing empathetic support during interactions. However, their responses often become verbose or overly formulaic, failing to adequately address the diverse emotional support needs of real-world scenarios. To tackle this challenge, we propose an innovative strategy-enhanced role-playing framework, designed to simulate authentic emotional support conversations. Specifically, our approach unfolds in two steps: (1) Strategy-Enhanced Role-Playing Interactions, which involve three pivotal roles—Seeker, Strategy Counselor, and Supporter—engaging in diverse scenarios to emulate real-world interactions and promote a broader range of dialogues; and (2) Emotional Support Agent Training, achieved through fine-tuning LLMs using our specially constructed dataset. Within this framework, we develop the **ServeForEmo** dataset, comprising an extensive collection of 3.7K+ multi-turn dialogues and 62.8K+ utterances. We further present **SweetieChat**, an emotional support agent capable of handling diverse open-domain scenarios. Extensive experiments and human evaluations confirm the framework's effectiveness in enhancing emotional support, highlighting its unique ability to provide more nuanced and tailored assistance.

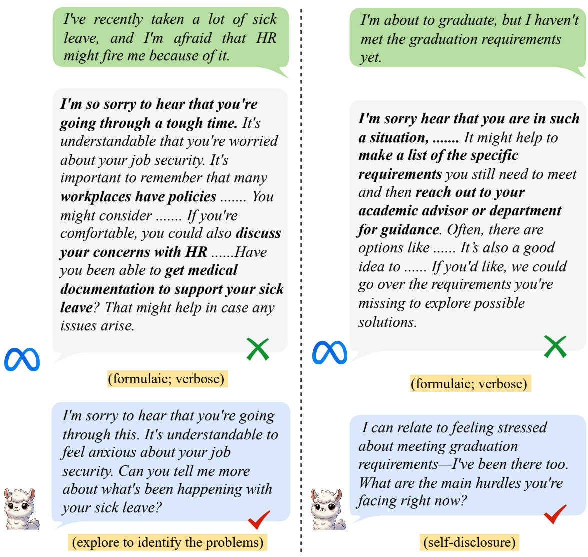


Figure 1: Example responses from the Meta-LLaMA3.1-70B-Instruct and our SweetieChat. LLMs often give verbose and formulaic responses characterized by empathy + suggestions, resulting in a distinct '*AI flavor*'. Conversely, SweetieChat excels in empathy and supportiveness, skillfully addressing and responding to user emotional needs.

1 Introduction

Emotional Support Conversation (ESC) systems are designed to alleviate users' emotional distress and help them understand and work through the challenges that they face (Peng et al., 2022; Rains et al., 2020), which play a crucial role in various scenarios, including social interactions, mental health support, and customer service communications. Despite the significant potential of Large Language Models (LLMs) in generating empathetic responses (Touvron et al., 2023; Yang et al., 2024a; Achiam et al., 2023), current LLMs often struggle to deliver diverse and contextually appropriate support. As depicted in Figure 1, the responses of LLMs tend to be formulaic and verbose, lacking the nuanced empathy and tailored suggestions required in real-world applications, which may produce a counterproductive support effect (Wang et al., 2024a; Kang et al., 2024).

To enhance the efficacy of ESC systems, the focus has shifted towards crafting high-quality ESC datasets that can fine-tune LLMs to produce more empathetic and context-aware responses (Zheng et al., 2024; Zhang et al., 2024). However, manually constructing a high-quality, multi-turn ESC dataset is extremely challenging, which limits the scale and diversity of emotional support scenarios. Liu et al. (2021) sourced the ESConv dataset, which only includes 1.3K dialogues spanning 13

*Corresponding Author

Proceedings of the 31st International Conference on Computational Linguistics, pages 4646–4669

topic categories.

Recent studies leverage the intrinsic generalization capabilities of LLMs to expand and enrich ESC datasets (Zheng et al., 2024; Zhang et al., 2024). These initiatives typically employ LLMs to continue, rewrite or imitate existing datasets. Nonetheless, these augmented datasets primarily encounter two significant challenges: 1) *Insufficient Conversation Diversity*. Although some efforts have notably amplified the dataset volume, the homogeneity among different dialogues remains pronounced (Zheng et al., 2023, 2024). This lack of variety can lead to repetitive and predictable interactions, which makes it hard to meet the unique needs of different seekers. 2) *Ineffective Strategy Implementation*. In psychology, effective emotional support conversations typically require a strategic approach, involving inquiries to gain deeper insights into the user's situation, thereby enabling more targeted support (Burleson., 2003; Tu et al., 2022). However, LLMs tend to favor strategies such as Offering Hope, Affirmation and Reassurance, and Providing Suggestions. This tendency results in a deficiency in their empathetic capabilities, as they often overlook the nuanced strategies necessary for truly understanding and assisting users in emotional support interactions.

To address these limitations, we introduce a novel strategy-enhanced role-playing framework for generating emotional support dialogues, as illustrated in Figure 2. This framework features three key roles: Seeker, Strategy Counselor, and Supporter. These roles simulate real-world emotional support interactions, enhancing both the diversity of dialogues and the effectiveness of the support provided. Specifically, the Supporter interacts with Seekers from diverse backgrounds, while the Counselor provides guidance by offering well-reasoned and effective strategies to the Supporter. Within this framework, we develop **ServeForEmo**, an augmented dataset that encompasses a wide range of ES scenarios, featuring 3.7K+ dialogues and 62.8K+ utterances — approximately 3× the scale of ESConv. Moreover, we build our emotional support dialogue system, **SweetieChat**, by fine-tuning LLaMA (Dubey et al., 2024) on ServeForEmo. Experimental results confirm that SweetieChat excels in managing diverse out-of-domain scenarios. Our main contributions can be summarized as follows:

- We present a simple yet effective role-playing framework that leverages psychological support strategies to simulate diverse emotional support interactions, enhancing the diversity and rationality of conversations.
- We develop ServeForEmo, an efficient ESC dataset that contains a broad spectrum of scenarios, and is supported by valid and effective emotional support strategies.
- Experimental results and Comprehensive human evaluations demonstrate that our SweetieChat excels in providing emotional support across diverse and open-domain scenarios.

2 Related Work

2.1 Emotional Support Conversation

Emotional support is a crucial capability in human-computer interactions (Huang et al., 2020). Although LLMs have revolutionized this field, they often struggle to provide supportive responses. In order to improve the emotional supportive ability of Chatbots, a real-world and comprehensive corpora is of great importance (Sharma et al., 2020). Liu et al. (2021) introduce the first manually curated emotional support conversation dataset. However, its size and scope are limited due to the significant expertise and time investment required for its creation.

Recent efforts have focused on utilizing LLMs to augment existing datasets, with efforts typically falling into two categories: modifying and expanding datasets. Qiu et al. (2023) approach data augmentation as a rewriting task, creating the SMILECHAT dataset by converting single-turn PsyQA (Sun et al., 2021) interactions into multi-turn dialogues by using ChatGPT (OpenAI, 2022). Zheng et al. (2023) treat data augmentation as a dialogue completion task, where they fine-tune a dialogue LM on ESConv and then prompt it to complete dialogues with posts from the EmpatheticDialogues dataset (Rashkin et al., 2019; Jing and Zhao, 2024). Additionally, Zheng et al. (2024) expand datasets by leveraging LLMs to imitate and generate new data based on existing datasets. Zhang et al. (2024) propose the ESCoT dataset, which improves explainability by adding emotion and strategy reasoning, creating a chain-of-thought in dialogues. While these approaches have augmented a significant amount of data, they often encounter challenges with limited diversity or insufficient strategic guidance.

2.2 Role-playing

Recent advancements in LLMs have significantly enhanced the development of Role-Playing Agents (Li et al., 2024; Wang et al., 2024b; Lu et al., 2024; Chen et al., 2024). These frameworks excel in personalized creation and generation tasks. For example, Yang et al. (2024b) propose a role-playing framework to build customizable persona-driven dialogue datasets for casual conversations, while Wu et al. (2024) explore role-play in creating interactive drama narratives. Additionally, role-playing has also been used in evaluations. Zhao et al. (2024) introduce an ESC evaluation framework, which employs role-playing agents to interact with ESC models. Inspired by these developments, we find that diverse roles can enhance the diversity of ESC data.

3 Method

Our goal is to build an emotional support agent capable of addressing diverse scenarios with effective psychological support strategies. To achieve this, we introduce a strategy-enhanced role-playing framework. As shown in Figure 2, the proposed framework comprises three main phases: role construction (Section 3.1), role-playing dialogues generation (Section 3.2), and the construction of our emotional support agent, SweetieChat (Section 3.3).

3.1 Role Construction

There are three key roles: **Seeker**, Strategy **Counselor**, and **Supporter**. The Seeker plays the user seeking emotional support, while the Supporter offers emotional support. The Strategy Counselor functions as the cognitive component of the Supporter, selecting appropriate strategies to enhance the quality of responses. Interactions between the Seeker and Supporter contribute to creating real-like and comprehensive emotional support dialogues. In the following Sections, we will detail the construction of each role. The prompts for each role are detailed in Appendix G.

3.1.1 Seeker Construction

Diversity in dialogues hinges on the variety of Seekers. We enrich the persona of the Seeker from three aspects: seeking problem types, specific scenarios, and the Seeker's profile.

Problem type To include a broader range of scenarios, we expand the 13 predefined problem types in ESConv to 45. These problem types are categorized into five groups: Emotional and Mental Health Issues, Interpersonal Relationships, Personal Development, Life and Work Stress, and Behavioral Issues. Details are provided in Appendix A. Formally, we denote a specific problem type from the predefined problem pool $\mathcal{P}$ as p.

Scenario Scenarios reflect the specific issues for which seekers seek assistance. To develop high-quality emotional support dialogues, it's crucial to include diverse, realistic psychological counseling scenarios. Following (Zhang et al., 2024), we use GPT-4o to expand the scenario range. Specifically, we establish a seed data pool, $\mathcal{S} = \{\mathcal{S}_i\}_{i=1}^{M}$, by manually selecting high-quality examples from the ESConv dataset (Liu et al., 2021) and psychological counseling websites. Each seed data is consist of problem type p_i, corresponding scenario s_i and a emotional support dialog D_i, accumulating 1,000 distinct $\mathcal{S}_i = (p_i, s_i, \mathcal{D}_i)$ triplets. The scenario descriptions are crafted to be longer than 20 words and detail specific events, avoiding vague references to general issues such as relationship breakdowns or mood swings. For a new problem type p', we randomly select a corresponding scenario from $\mathcal{S}$ and use it to generate a new scenario with GPT-4o. This process can be formalized as follows:

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

Seeker Profile Intuitively, individuals with different backgrounds and personalities often react differently to identical scenarios. To enhance the diversity of our dialogues, we craft a tailored seeker profile for each dialogue based on the problem type and specific scenario. Initially, we establish a seed profile pool $\mathcal{C} = \{\mathcal{C}_i\}_{i=1}^{N}$ comprising 100 handcrafted seeker profile. Each profile $\mathcal{C}_i$ comprises a problem type p_i, scenario s_i, and character description c_i. Building on prior work in personality-driven dialogues (Zhang et al., 2018; Zheng et al., 2019), c_i features attributes such as name, gender, address, occupation, personality traits, and hobbies. We leverage the context learning capability of GPT-4o to dynamically generate a unique character c' for each new problem type p' and scenario s'. The new profile is then added back to enrich the seed profile pool $\mathcal{C}$:

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

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

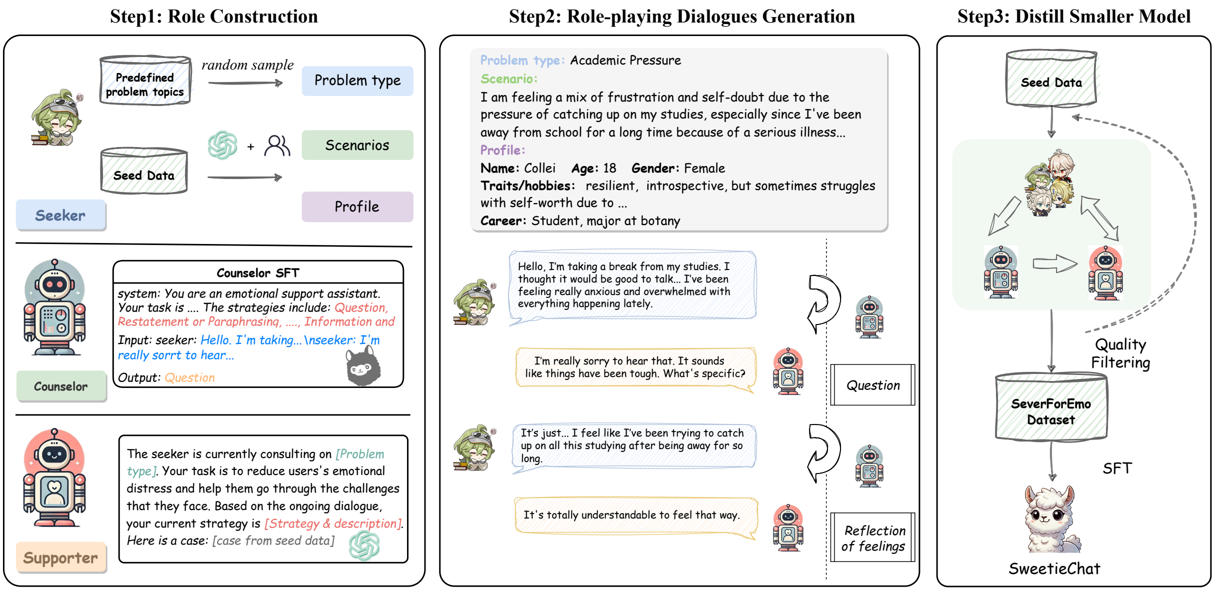


Figure 2: Overview of the proposed strategy-enhanced role-playing framework. Our framework incorporates 3 roles: Seeker, Strategy Counselor, and Supporter. We employ LLMs to simulate these roles and interact with each other like real-world emotional support conversations. Following the dialogue generation process, we develop the ServeForEmo dataset. Building on these foundations, we present SweetieChat, an emotional support agent capable of handling diverse open-domain scenarios.

3.1.2 Counselor Construction

While LLMs excel at generating empathetic responses, they still struggle to provide emotional support using various support skills. There is a significant divergence in strategy preferences between LLMs and human counselors. A typical emotional support conversation consists of three stages: exploration, comforting, and action (Hill, 2009). In contrast, LLMs tend to favor strategies such as offering hope, affirmation and reassurance, as well as providing suggestions (Kang et al., 2024). More evidence is in Appendix B. This tendency to rush solutions and provide empathetic responses results in a lack of deeper interaction and understanding with the seeker, markedly differing from the more nuanced approach typical of human interactions.

To mitigate these challenges, we utilize the strategy-enhanced LLaMA as our support strategy counselor. We design a strategy-selection task, and then fine-tune LLaMA using the manually annotated ESConv dataset. The Counselor is instructed to select the appropriate emotional support strategy based on the ongoing dialogue history, as formalized by the equation:

$$o_i = \arg\max_{o \in \mathcal{O}} \mathcal{P}(o|\mathcal{D}_i^j) \quad (4)$$

Here, $\mathcal{O}$ represents the set of strategies, and $\mathcal{D}_i^j$ denotes the dialogue history up to the j-th round of dialogue $\mathcal{D}_i$, consisting of the sequence of previous utterances $\{u_i^1, r_i^1, \ldots, r_i^{j-1}, u_i^j\}$. Here, u_i^j is the j-th statement from the Seeker, and r_i^{j-1} is the most recent response from the Supporter.

3.1.3 Supporter Construction

We employ GPT-4o as the supporter, known for its proficiency in generating empathetic responses and in-context learning. The supporter's actions are guided by a support strategy selected by the counselor. Given the chosen strategy, along with the dialogue history and specific case dialogues, GPT-4o is tasked with producing emotionally supportive replies.

3.2 Role-playing Dialogue Generation

The interactions between seekers from diverse backgrounds and the supporter form the core of our emotional support dialogue. The counselor serves as an assistant, providing well-reasoned support strategies to guide the supporter in delivering clear and interpretable emotional support. *In this scene, our framework not only accommodates diverse scenarios but also ensures reliable strategy guidance.* To enhance the diversity of our data, we implement a self-iterative mechanism to extend the seed data pool(Zheng et al., 2024).

$$\mathcal{S} \leftarrow \mathcal{S} \cup \{(p', s', \mathcal{D}')\} \quad (5)$$

where p' is the randomly selected problem type, s' represents the newly generated scenarios, and $\mathcal{D}'$ is the simulated emotional support conversation.

3.3 SweetieChat Agent

Our strategy-enhanced role-playing framework facilitates the development of the ServeForEmo dataset, which captures a wide range of seekers and scenarios. This dataset is designed to narrow the empathy response gap between LLMs and humans. Building upon this foundation, we develop SweetieChat, a specialized agent tailored for providing emotional support. To optimize its performance, we fine-tune the LLaMA model using the ServeForEmo dataset.

4 Data Analysis

This section explores key concerns about the ServeForEmo dataset:

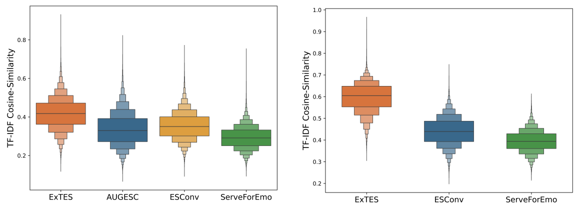
- **Q1:** What is the overall quality of the ServeForEmo dataset?
- **Q2:** Does the dataset exhibit sufficient diversity?
- **Q3:** How are the strategies implemented in the dialogues?

4.1 Statistics (Q1)

	Category	ESConv	SeverForEmo
Total	# Dialogues	1,300	3,757
	# Utterances	29,278	62,863
	Avg. Dialogue Length	22.54	16.73
	Avg. Utterance Length	21.17	17.97
Seeker	# Utterances	14,639	30,722
	Avg. # Utter. per Dialog	11.27	8.18
	Avg. Utterance Length	19.9	15.25
Supporter	# Utterances	14,639	32,141
	Avg. # Utter. per Dialog	11.27	8.55
	Avg. Utterance Length	22.45	20.56

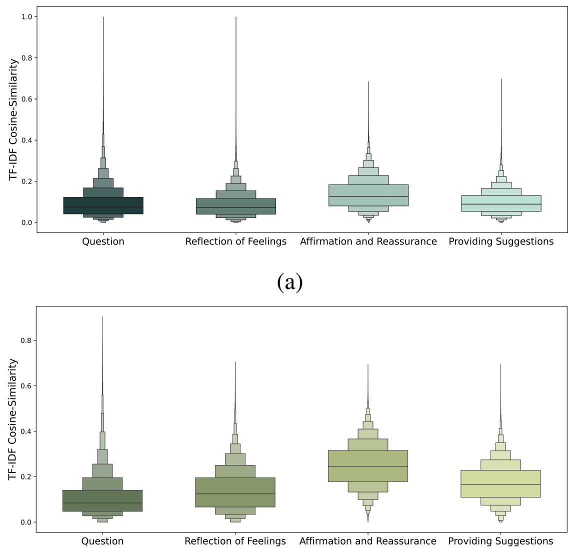
Table 1: Comparison of ServeForEmo and ESConv statistics. For ESConv, we exclude initial greetings and merge consecutive utterances from the same speaker to simplify the dialogue.

To guarantee high data quality, we perform thorough post-processing that includes filtering out low-quality dialogues, eliminating redundant greetings, and managing dialogue length. After human review and refinement, we develop the ServeForEmo dataset, which comprises 3.7K+ dialogues and 62.8K+ utterances, approximately 3× the scale of ESConv. The detailed statistics are presented in Table 1. Since we limit utterances to a maximum of three sentences during role interactions, the dialogues and utterances in ServeForEmo are generally shorter than those in ESConv. For additional details on our dataset post-processing and quality evaluation methods, please refer to Appendix C.



(a) (b)

Figure 3: (a) Global inter-dialogue similarity statistics computed using TF-IDF vectors. (b) Inter-dialogue similarity statistics within the academic field. Lower similarity values indicate higher diversity.



(b)

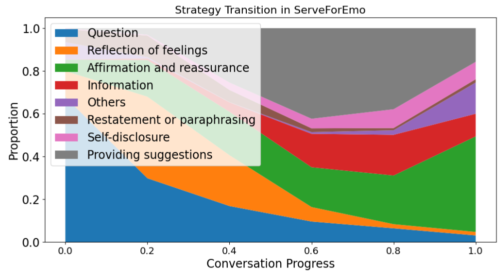
Figure 4: (a) Global similarity statistics of responses guided by the same support strategy, computed using TF-IDF features. (b) Similarity statistics of responses within the academic field under the same support strategy.

4.2 Diversity Analysis (Q2)

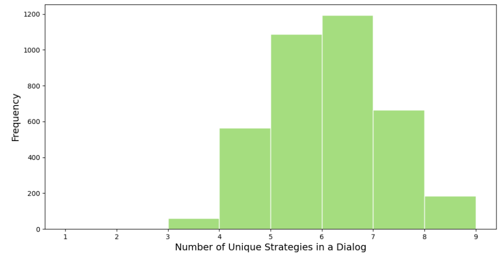
To illustrate the broad scope of the ServeForEmo dataset, we focus on two key aspects of diversity: semantic and lexical features. For brevity, we only present semantic diversity in the main body of our paper. Details on the other diversity aspect are in Appendix D.

To assess the semantic diversity of ESC datasets, we calculate the cosine similarity between pairs of distinct dialogues using TF-IDF features (Salton et al., 1975). Figure 3a and 3b demonstrate that the ServeForEmo dataset exhibits the best inter-dialogue diversity. The low similarity across both global and specific problem-type fields indicates that ServeForEmo effectively captures a wide range of emotional support scenarios. Dialogues collected through our role-playing framework even have higher inter-dialogue diversity compared to the crowdsourced ESConv dataset. Conversely, datasets generated through direct prompting of LLMs, such as ExTES, display minimal diversity. Although these efforts have notably increased the dataset volume, the homogeneity among different dialogues remains pronounced. Since the diversity of the generated data relies heavily on the prompts, it is challenging to produce varied and realistic conversations through prompting.

Despite the diversity of dialogues, we are curious about whether supporters' responses, guided by the same strategies, demonstrate variation across different scenarios. Thus, we analyze the responses associated with the four most frequently applied strategies, and present the similarity statistics of responses under the same strategic guidance in Figure 4. The results reveal that supporters' responses, even under the same strategy guidance, exhibit low cosine similarity scores. This indirectly confirms that the responses in ServeForEmo are tailored to specific scenarios.



(a)



(b)

Figure 5: (a) Distribution of strategies at different conversation progress on ServeForEmo dataset. (b) Distribution of unique strategies across different dialogues on ServeForEmo.

4.3 Strategy Analysis (Q3)

Strategy	ESConv	ServeForEmo
Question	20.68%	19.04%
Others	18.18%	6.06%
Providing suggestions	16.08%	20.00%
Affirmation and reassurance	15.38%	25.26%
Self-disclosure	9.32%	4.93%
Reflection of feelings	7.81%	12.03%
Information	6.61%	8.79%
Restatement or paraphrasing	5.93%	3.90%

Table 2: Strategy distribution of ESConv and ServeForEmo.

Strategy Distribution As described in Section 3.1.2, we perform instruction fine-tuning on LLaMA using the ESConv dataset to be the Strategy Counselor. Table 2 presents a comparative analysis of strategy distributions between the ServeForEmo and ESConv datasets. The results reveal a strong similarity between the two datasets, with the main strategies concentrated on Affirmation and Reassurance, Providing Suggestions, and Questioning. This demonstrates the Counselor's effectiveness in selecting and assessing strategies.

Strategy Transition To show whether the strategies annotated by our counselor follow a reasonable procedure, we analyze the transition of strategies across different phases of the conversation. We segment the conversation progress into six intervals and count the proportions of different strategies within each interval for all the conversations in ServeForEmo. Figure 5a illustrates the strategy distribution across different conversation progress points. The result shows that in the early stages of the conversation, our counselor primarily employs Questioning to explore and help seekers identify their problems. As the conversation progresses into the middle and later stages, the Providing Suggestions strategy becomes more prominent, assisting seekers in finding solutions. Throughout the conversation, strategies such as Affirmation and Reassurance and Reflection of Feelings are consistently applied, providing comfort by conveying empathy and understanding. This strategic arrangement reflects the three-stage helping skill model proposed by Hill (2009), indicating that the counselor is capable of effectively organizing emotional support strategies, making it a competent assistant.

Distribution of Unique Strategies in Dialogues Here, we aim to show the diversity of strategies employed throughout our conversations. Figure 5b presents the distribution of unique strategies within individual dialogues in ServeForEmo. Most dialogues feature more than five distinct strategies, underscoring the wide range of emotional support techniques employed in the dataset. This diversity aligns with our goal of providing comprehensive emotional support through diverse counseling strategies.

5 Experiments

5.1 Baseline Models

To validate the effectiveness of the ServeForEmo Dataset, we conduct a comprehensive comparison, detailed below:

LLaMA (Dubey et al., 2024): A foundational open-source LLM. For our experiments, we employ the `LLaMA3-8B-Instruct` version as the base model, incorporating task-specific settings and applicable strategies within the system prompts.

AUGESC (Zheng et al., 2023): Fine-tuned from LLaMA using the AUGESC dataset, which is crafted by leveraging a fine-tuned dialogue LM to do dialogue completion task.

ExTES (Zheng et al., 2024): Fine-tuned from LLaMA using the ExTES dataset, which is generated by prompting LLMs to imitate seed conversations.

ESConv (Liu et al., 2021): Fine-tuned from LLaMA using the ESConv dataset. The meticulous human curation of ESConv ensures that the model is trained on dialogues rich in real-life emotional content and effective emotional support responses.

ServeForEmo (ours): Also known as SweetieChat, this model is specifically fine-tuned from LLaMA using our ServeForEmo dataset, aiming to enhance its performance in diverse ES scenarios.

5.2 Implementation Details

We apply LoRA adaptation (Hu et al., 2021) to the W_q, W_v, W_k, and W_o parameters, with a rank (r) set to 8. For optimization, we use the AdamW optimizer (Loshchilov and Hutter, 2017), setting a learning rate of 3×10^{-5} and employing a linear warm-up over the first 1% of total training steps. The per device batch size is set to 2, and the training is conducted over 5 epochs. To prevent overfitting, we implement early stopping with a patience of 7 evaluation steps. The model achieving the best performance on the validation set is selected for testing. All experiments are conducted on eight Tesla V100 GPUs, and the final results are reported as the average of three experimental runs.

5.3 Evaluation Settings

To ensure equitable comparison across different models, we employ the manually annotated test set of ESConv as our benchmark. This allows us to assess the performance of models under real-world human annotation conditions. Our evaluation metrics are categorized into two types: automatic and human.

Automatic Evaluation Metrics We employ 4 established automatic evaluation metrics: BLEU-n (Papineni et al., 2002), ROUGE-L (Lin, 2004), Distinct-n (Li et al., 2016), and BERT-Score (Zhang* et al., 2020). The responses are tokenized using the LLaMA3 (Dubey et al., 2024) Tokenizer. Each metric quantitatively assesses distinct aspects of textual quality, providing a comprehensive overview of model performance.

Human Evaluation Metrics For human evaluation, we focus on evaluating the generated responses. Specifically, we focus on the following dimensions: Empathy, Informativeness, Coherence, Suggestion, Understanding, Helpfulness, and Overall Quality. A detailed description of these evaluation metrics is provided in Appendix F.

5.4 Main Results

5.4.1 Automatic Evaluation Results

We conduct a comprehensive evaluation of the multi-turn emotional support capabilities of several models under two experimental configurations: (1) ***Interactive evaluation with generated context***, designed to assess dialogue coherence and relevance, and (2) ***Interactive evaluation with reference context***, ensuring a fair comparison of emotional support performance in consistent scenarios. The results are presented in Table 3. See Appendix E for other supplementary experiments.

The experimental results demonstrate the efficiency of the proposed SeverForEmo dataset. *Firstly*, regarding the content-based metrics such as BLEU and ROUGE, it is evident that our ServeForEmo consistently outperforms other baselines. Specifically, it achieves the highest BLEU-2 and BLEU-4 scores, demonstrating its superior ability to generate relevant and coherent responses. While ExTES performs competitively, especially in terms of R-2 and R-L scores, ServeForEmo still demonstrates an overall stronger performance across both evaluation settings. *Secondly*, ServeForEmo achieves a BERT-Score of 85.81, which is even better than ESConv, indicating that the generated content is semantically aligned with human annotators. *Additionally*, ServeForEmo excels in generating richer content, as reflected in its Distinct2 and Distinct3 scores of 99.59 and 99.89, the highest

Model	BLEU-2	BLEU-4	R-2	R-L	Distinct2	Distinct3	BERT-Score
Interactive evaluation with generated context							
ESConv	6.15$_{\pm 0.05}$	2.95$_{\pm 0.06}$	3.07$_{\pm 0.10}$	13.87$_{\pm 0.15}$	98.15$_{\pm 0.89}$	98.40$_{\pm 0.82}$	85.49$_{\pm 0.10}$
LLaMA	3.62$_{\pm 0.09}$	1.28$_{\pm 0.04}$	2.33$_{\pm 0.11}$	10.41$_{\pm 0.18}$	94.75$_{\pm 0.09}$	98.71$_{\pm 0.01}$	83.80$_{\pm 0.06}$
ExTES	5.92$_{\pm 0.06}$	2.44$_{\pm 0.06}$	**2.74**$_{\pm 0.10}$	**13.27**$_{\pm 0.10}$	98.80$_{\pm 0.08}$	99.82$_{\pm 0.02}$	85.45$_{\pm 0.02}$
AUGESC	5.59$_{\pm 0.09}$	2.59$_{\pm 0.08}$	2.40$_{\pm 0.15}$	12.88$_{\pm 0.15}$	99.00$_{\pm 0.18}$	98.96$_{\pm 0.24}$	85.48$_{\pm 0.02}$
ServeForEmo	**6.27**$_{\pm 0.05}$	**2.74**$_{\pm 0.02}$	2.59$_{\pm 0.06}$	13.23$_{\pm 0.12}$	**99.59**$_{\pm 0.06}$	**99.89**$_{\pm 0.05}$	**85.81**$_{\pm 0.02}$
Interactive evaluation with reference context							
ESConv	6.73$_{\pm 0.05}$	3.21$_{\pm 0.02}$	3.53$_{\pm 0.10}$	14.83$_{\pm 0.06}$	98.72$_{\pm 0.06}$	98.71$_{\pm 0.22}$	85.74$_{\pm 0.01}$
LLaMA	4.13$_{\pm 0.04}$	1.55$_{\pm 0.03}$	1.90$_{\pm 0.06}$	10.65$_{\pm 0.05}$	97.25$_{\pm 0.05}$	99.44$_{\pm 0.00}$	84.49$_{\pm 0.01}$
ExTES	6.69$_{\pm 0.03}$	2.86$_{\pm 0.02}$	2.85$_{\pm 0.12}$	13.99$_{\pm 0.09}$	99.25$_{\pm 0.06}$	**99.73**$_{\pm 0.08}$	85.88$_{\pm 0.04}$
AUGESC	6.10$_{\pm 0.05}$	2.81$_{\pm 0.03}$	2.69$_{\pm 0.11}$	13.48$_{\pm 0.09}$	98.85$_{\pm 0.11}$	98.85$_{\pm 0.26}$	85.56$_{\pm 0.02}$
ServeForEmo	**6.76**$_{\pm 0.08}$	**3.12**$_{\pm 0.04}$	**2.76**$_{\pm 0.14}$	**14.01**$_{\pm 0.20}$	**99.48**$_{\pm 0.03}$	99.52$_{\pm 0.07}$	**85.98**$_{\pm 0.02}$

Table 3: Automatic evaluation results on ESConv test datasets. The results demonstrate the effectiveness of the ServeForEmo dataset in enhancing the emotional support capabilities of downstream models. All results represent the average of three experimental runs.

among all models. *Overall*, these results suggest that our strategy-enhanced role-playing framework excels in producing high-quality dialogues and effectively employing emotional support strategies for handling diverse out-of-domain scenarios.

5.4.2 Human Evaluation Results

To enhance the comprehensiveness and reliability of our evaluations, we also conduct human evaluations. Specifically, human evaluators are instructed to compare the responses A and B generated by two models under the same dialogue history, and choose an option from "A wins", "Tie", and "B wins". For the sake of fairness, we randomize the order of the responses to eliminate position bias. The evaluation is conducted on 50 randomly sampled dialogues.

The results presented in Figure 6 (a-c) indicate that responses from ServeForEmo generally outperform those from AUGESC, ESConv, and ExTES, demonstrating consistency between automatic and human evaluations. Notably, despite the larger data scales of ExTES and AUGESC, ServeForEmo marginally excels. This suggests that simply enlarging the dataset does not inherently lead to substantial improvements. LLMs are already equipped with the capability to generate empathetic responses. More importantly, it is vital for these models to effectively implement emotional support strategies across diverse seeking scenarios. More cases are included in Appendix H.

5.5 Evaluation on New Scenarios

To evaluate the emotional support capabilities of ES agents in real-world scenarios, we invite 10 volunteers to interact with both ServeForEmo and ExTES. After completing the interactions, they rate the models based on the metrics outlined in Section 5.3. All metrics are rated on a five-point Likert scale (Likert, 1932), ranging from 1 to 5, where higher scores indicate better quality. The results are presented in Figure 6 (d). From the results, we find that both models demonstrate strong emotional support capabilities, with ServeForEmo performing slightly better than ExTES. People prefer dialog systems that can provide more supportive responses(Liu et al., 2021). However, both models show noticeable weaknesses in the Suggestion, Helpfulness, and Informative dimensions, highlighting potential areas for future improvement.

6 Conclusion

In this paper, we introduce a strategy-enhanced role-play framework, a novel method for generating diverse emotional support dialogues. It involves the creation of the ServeForEmo dataset and the fine-tuning of the ES agent SweetieChat, which is the first attempt to incorporate support strategies in simulating diverse real-world emotional support conversations. Experimental results demonstrate that these role-playing dialogues can significantly enhance the model's ability to provide supportive responses which align with human preference. These findings offer valuable insights into potential advancements in emotional support agents, partic-

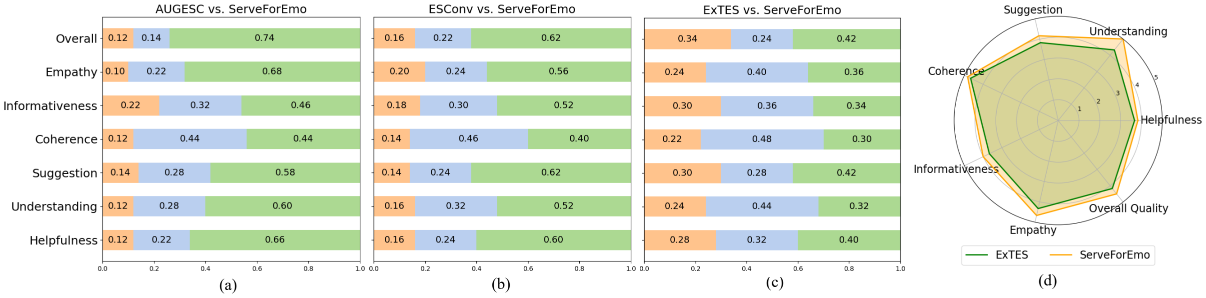


Figure 6: (a-c) Results of the human evaluation on the ESConv dataset. ■ indicates 'A win', ■ indicates 'tie', and ■ indicates 'B win'. (d) The average win rates distribution of ServeForEmo vs. ExTES. The scores (from 1 to 5) are averaged over all the 10 volunteers. k denotes Fleiss' Kappa (Fleiss, 1971), indicating fair or moderate inter-annotator agreement ($0.4 < k < 0.8$). (d) The average win rates distribution of ServeForEmo vs. ExTES. The scores (from 1 to 5) are averaged over all the 10 volunteers. k denotes Fleiss' Kappa (Fleiss, 1971), indicating fair or moderate inter-annotator agreement ($0.4 < k < 0.8$).

ularly our procedure for creating diverse effective ES conversations.

Limitations

This work proposes a strategy-enhanced role-playing framework designed to generate emotional support conversations closely aligned with human annotations, thereby enhancing the emotional support capabilities of downstream models. However, our research still faces several limitations:

(I) The data construction process includes some errors, such as supporters not consistently following assigned strategies and seekers not fully adhering to their character backgrounds. This necessitates more stringent quality control and consistency checks in post-processing work.

(II) Scaling up the dataset for supervised fine-tuning is only a temporary solution to improve the emotional support capabilities of downstream models. To address this, we plan to explore preference alignment between humans and chatbots in future research.

(III) Emotional support dialogues are challenging to evaluate. Even with human evaluation, it remains difficult to determine the effectiveness of different responses due to individual variations in emotional support perception and the absence of firsthand experience by the evaluators.

(IV) While our ServeForEmo dataset provides rich text-based emotional support interactions, natural human-machine interactions often involve speech. We will extend our research to incorporate speech, aiming to enhance the naturalness of human-machine interactions.

Acknowledgments

This research was supported by the National Natural Science Foundation of China under Grant No. 62336008. We would like to thank the anonymous reviewers for their valuable comments. We also appreciate the contributions of the following students for their work on human evaluation: Y.Y. Cai, Y.Z. Deng, G.X. Du, Q.K. Li, S.B. Tang, S.Y. Song, Y.W. Sun, X.P. Zhao, X. Zhang, M.H. Zhang, J.C. Zheng, W.W. Zhuang.

References

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

Brant R Burleson. 2003. Emotional support skill. *HANDBOOK OF COMMUNICATION AND SOCIAL INTERACTION SKILLS*, page 551.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. From persona to personalization: A survey on role-playing language agents. *Preprint*, arXiv:2404.18231.

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

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Clara E Hill. 2009. Helping skills: Facilitating, exploration, insight, and action. *American Psychological Association*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Trans. Inf. Syst.*, 38(3).

Ye Jing and Xinpei Zhao. 2024. Dq-former: Querying transformer with dynamic modality priority for cognitive-aligned multimodal emotion recognition in conversation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 4795–4804, New York, NY, USA. Association for Computing Machinery.

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

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2024. Camel: communicative agents for "mind" exploration of large language model society. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

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

R Likert. 1932. A technique for the measurement of attitudes.

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

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

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 7828–7840. Association for Computational Linguistics.

OpenAI. 2022. Introducing chatgpt. https:// openai.com/blog/chatgpt.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2023. SMILE: single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *CoRR*, abs/2305.00450.

Stephen A Rains, Corey A Pavlich, Bethany Lutovsky, Eric Tsetsi, and Anjali Ashtaputre. 2020. Support seeker expectations, support message quality, and supportive interaction processes and outcomes: The case of the comforting computer program revisited. *Journal of Social and Personal Relationships*, 37(2):647–666.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics.

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

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

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. Psyqa: A chinese dataset for generating long counseling text for mental health support. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1489–1503. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 308–319. Association for Computational Linguistics.

Jiashuo Wang, Chunpu Xu, Chak Tou Leong, Wenjie Li, and Jing Li. 2024a. Muffin: Mitigating unhelpfulness in emotional support conversations with multifaceted AI feedback. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 567–585. Association for Computational Linguistics.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng. 2024b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *Preprint*, arXiv:2310.00746.

Weiqi Wu, Hongqiu Wu, Lai Jiang, Xingyuan Liu, Hai Zhao, and Min Zhang. 2024. From role-play to drama-interaction: An LLM solution. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3271–3290, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, and ohters. 2024a. Qwen2 technical report. *CoRR*, abs/2407.10671.

Bohao Yang, Dong Liu, Chen Tang, Chenghao Xiao, Kun Zhao, Chao Li, Lin Yuan, Guang Yang, Lanxiao Huang, and Chenghua Lin. 2024b. Simschat: A customisable persona-driven role-playing agent. *CoRR*, abs/2406.17962.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024. Escot: Towards interpretable emotional support dialogue systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 13395–13412. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Dandan Liang, Zhixu Li, Yan Teng, Yanghua Xiao, and Yingchun Wang. 2024. Esc-eval: Evaluating emotion support conversations in large language models. *CoRR*, abs/2406.14952.

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

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *CoRR*, abs/1901.09672.

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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11325–11345. Association for Computational Linguistics.

A Details of Problem Types

To enhance the diversity of emotional support dialogue scenarios, we expand the predefined 13 problem types in the ESConv dataset to 45. These problem types are organized into five main categories:

- Emotional and Mental Health Issues
- Life and Work Stress
- Interpersonal Relationships
- Personal Development
- Behavioral Issues

Each category contains at least 7 distinct problem types. The distribution of these problem types in our ServeForEmo dataset is presented in Table 4. When constructing specific seeker scenarios, we randomly select a problem type from the predefined options, simulating the questionnaire process typically completed prior to a counseling session.

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

Table 4: Statistics of predefined problem types.

B Details of Strategy Counselor Construction

To demonstrate the necessity of our Strategy Counselor, we construct 100 dialogues with GPT-4o acting in this role. Despite providing GPT-4o with detailed strategy definitions and examples, we observe a significant discrepancy in strategy preferences between LLMs and humans. As evident in Figure 7, LLMs predominantly favor strategies such as Offering Hope, Affirmation and Reassurance, and Providing Suggestions. This tendency to rush solutions and empathy contributes to the distinct '*AI flavor*' in emotional support interactions, markedly differing from the more nuanced approach typical of human interactions.

To address these issues, we fine-tune the Meta-Llama3-8B-Instruct model[1] to improve its strategy selection capability. Specifically, we design a question-and-indefinite-choice task using the manually annotated ESConv dataset. The prompt format is consistent with the Counselor's view shown in Appendix G, and the training configurations follow the protocols described in Section 5.2.

[1]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

C Dataset Quality Analysis

C.1 Dataset Postprocessing

To ensure data quality, we perform extensive post-processing work. Specifically, we remove undesirable cases that include:

- **Role inconsistencies**: Cases where the 'seeker' deviates from the designated scenario and profile by adopting the 'support' role. These errors typically concentrate within specific problem types, largely due to seed data errors. We address these inconsistencies through systematic sampling and careful inspection across problem types.
- **Redundant greetings**: Despite explicit instructions on how to conclude conversations, some dialogues persistently end with multiple greetings. To address this issue, we establish filtering rules that curtail superfluous greeting utterances.
- **Dialogue length regulation**: We exclude conversations that are either excessively short or overly long. Additionally, we limit each utterance to a maximum of three sentences during role-play interactions.

Metrics	Score	Variance
Coherence	4.65	0.35
Consistency	4.88	0.12
Helpfulness	4.62	0.38
Informativeness	4.66	0.28
Rationality	4.61	0.37
Understanding	4.48	0.53
safety	5.00	0.00

Table 5: Human evaluation of ServeForEmo quality. Metrics are evaluated on a scale from 1 to 5, where the score represents the average ratings from three annotators.

C.2 Dataset Quality Evaluation

To ensure the quality of the ServeForEmo dataset, we conduct a comprehensive human evaluation. Our evaluation focuses on the following key concerns: (1) whether the context of the dialogue closely relates to the scenario, and (2) whether the roles adequately fulfill their responsibilities. The human evaluation metrics include Coherence, Consistency, Helpfulness, Informativeness, Rationality, Understanding, and Safety. Detailed descriptions of these metrics can be found in Appendix F. All metrics employ a five-level Likert scale (Likert, 1932), ranging from 1 to 5, where a higher score indicates superior quality. To evaluate the dataset, we engage three master students with psychology and computational linguistics backgrounds as annotators, assessing 50 randomly selected dialogues from ServeForEmo dataset. The results in Table 5 demonstrate the high quality of our dataset. Specifically, users effectively articulate their concerns, counselors offer well-founded strategic guidance, supporters follow the counselors' strategies, and the dialogues maintain coherence, informativeness, and empathy throughout.

Furthermore, we extract the TF-IDF features from both the scenarios and the corresponding dialogues and compute their cosine similarities. The results, depicted in Figure 8, demonstrate a high correlation between the dialogue content and the scenario settings, confirming that the dialogues closely adhere to the prescribed scenarios.

C.3 Data Example from ServeForEmo

Figure 9 presents a comprehensive dialogue example from the ServeForEmo dataset. Each dialogue is annotated with the problem type, the detailed scenario description, and the seeker profile. Additionally, the dialogue illustrates the strategies employed by the supporter, which are explicitly labeled.

D Dataset Lexical Diversity Analysis

For the lexical analysis, we adopt the distinct-n (Li et al., 2016) metrics, which are widely used for assessing the diversity of dialogue datasets. To ensure a fair comparison, we randomly select 1,300 dialogues from each dataset, preprocess them into single strings without speaker tokens, and tokenize them using the LLaMA3-8B tokenizer. The results, presented in Table 6, show that ServeForEmo exhibits a rich vocabulary with high distinct-n scores, comparable to the manually annotated ESConv dataset. This indicates that constructing more varied seeker profiles and scenarios within a role-play framework can effectively enhance the diversity of generated dialogues. In contrast, ExTES demonstrates the lowest distinct-n scores, suggesting that incorporating seed dialogues into fixed prompts tends to result in monotonous outputs.

Dataset	Distinct-1 ($\Uparrow$)	Distinct-2 ($\Uparrow$)	Distinct-3 ($\Uparrow$)
ESConv	1.82	19.76	49.38
ExTES	0.82	8.86	24.95
ServeForEmo	1.62	16.78	40.23

Table 6: Distinct-n results of 1,300 conversations from ESConv, ExTES, and ServeForEmo, respectively.

E Additional Experiments

In our study, we evaluate the performance of models using the ESConv and ServeForEmo datasets under various training configurations, as detailed in Table 7. From the results, we observe that the ServeForEmo dataset enables greater model robustness, as models trained solely on ServeForEmo perform well when tested on the ESConv dataset. Additionally, the results clearly indicate that models trained on both datasets outperforms those trained on just one or neither, with the highest scores observed in nearly all metrics when test on both datasets. This suggests that integrating diverse training sources significantly enhances the effectiveness and robustness of models designed for emotional support dialogues.

F Human Evaluation Metrics

We conduct a comprehensive human evaluation to assess both the quality of the dataset and the effectiveness of responses. The details of human evaluation metrics are as followed.

Train on ESConv	Train on ServeForEmo	BLEU-2	BLEU-4	R-2	R-L	Distinct-2	Distinct-3	METEOR	BERT-Score
Test on ESConv									
×	×	3.62	1.28	2.33	10.41	94.75	98.71	**17.46**	83.80
✓	×	6.15	2.95	**3.07**	13.87	98.15	98.40	11.13	85.49
×	✓	6.27	2.74	2.59	13.23	**99.59**	**99.89**	12.58	**85.81**
✓	✓	**6.49**	**3.02**	3.06	**14.13**	99.21	99.47	12.17	85.64
Test on ServeForEmo									
×	×	7.33	2.95	3.54	14.29	**99.84**	**99.99**	15.66	87.1
✓	×	9.82	4.54	5.52	18.5	99.28	99.74	17.13	87.83
×	✓	**12.94**	**6.29**	7.74	**20.71**	99.73	99.98	**20.66**	88.64
✓	✓	12.83	6.24	**7.84**	20.46	99.8	99.98	20.64	**88.59**

Table 7: Ablation study results comparing model performance across different training settings on the ESConv and ServeForEmo datasets.

F.1 Dataset Evaluation Metrics

The dialogue quality evaluation follows the framework proposed by Zheng et al. (2023), as illustrated in Figure 10. In addition to the original metrics, we introduce **Rationality** as an evaluation criterion to specifically assess the counselor's performance.

F.2 Response Evaluation Metrics

As shown in Figure 11, our human evaluation questionnaire includes the following metrics to assess supporters' responses: Informativeness, Suggestion, Empathy, Understanding, Helpfulness, Coherence, and Overall Quality.

G Prompt Demonstration

In this section, we detail the prompts used to construct situations, Seeker profiles, Counselor roles, and Supporter roles. Figure 12 provides comprehensive descriptions of these prompts.

H Case study

We present the generated responses of different models in Figure 13. It shows that SweetieChat excels in provide supportive responses and empathy. Additionally, we showcase interaction cases from one of our volunteers in Figures 14, 15, and 16. The results indicate that SweetieChat performs exceptionally well in open-domain emotional support dialogues. Its responses are relatively concise and effectively utilize various supportive strategies in response to user input, demonstrating its potential utility in real-world scenarios.

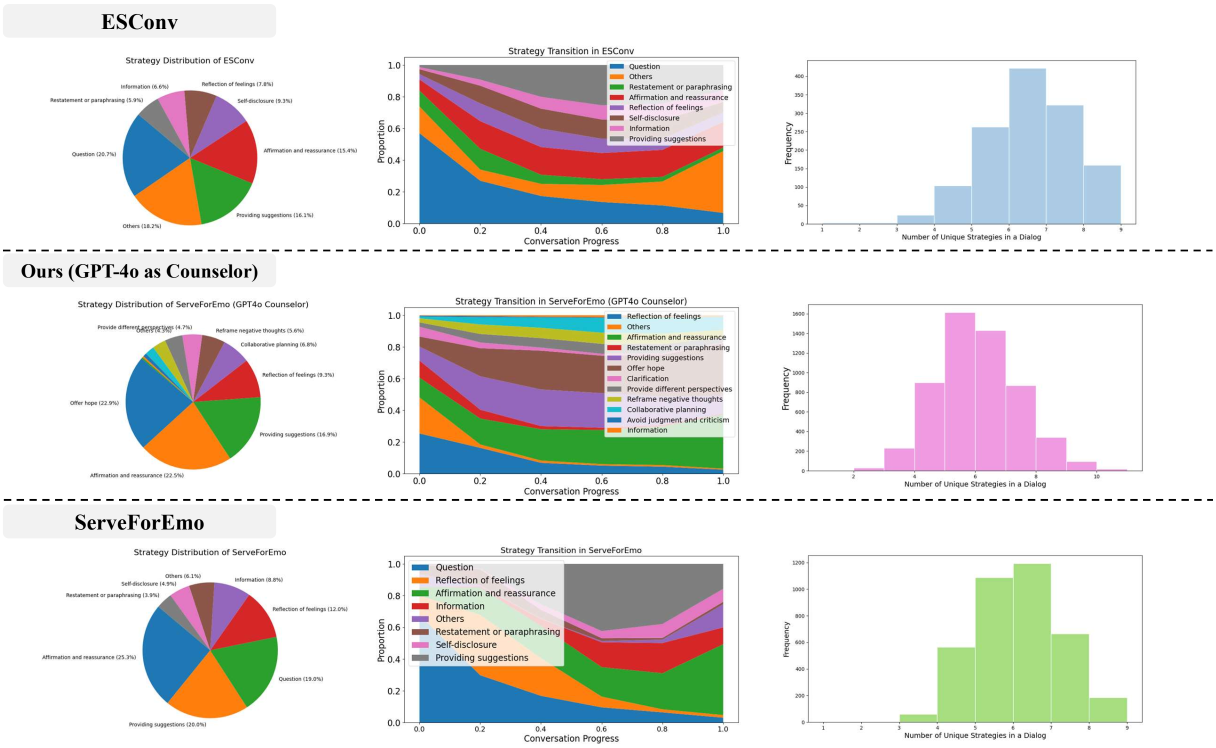


Figure 7: Strategy distribution and strategy transition in conversations.

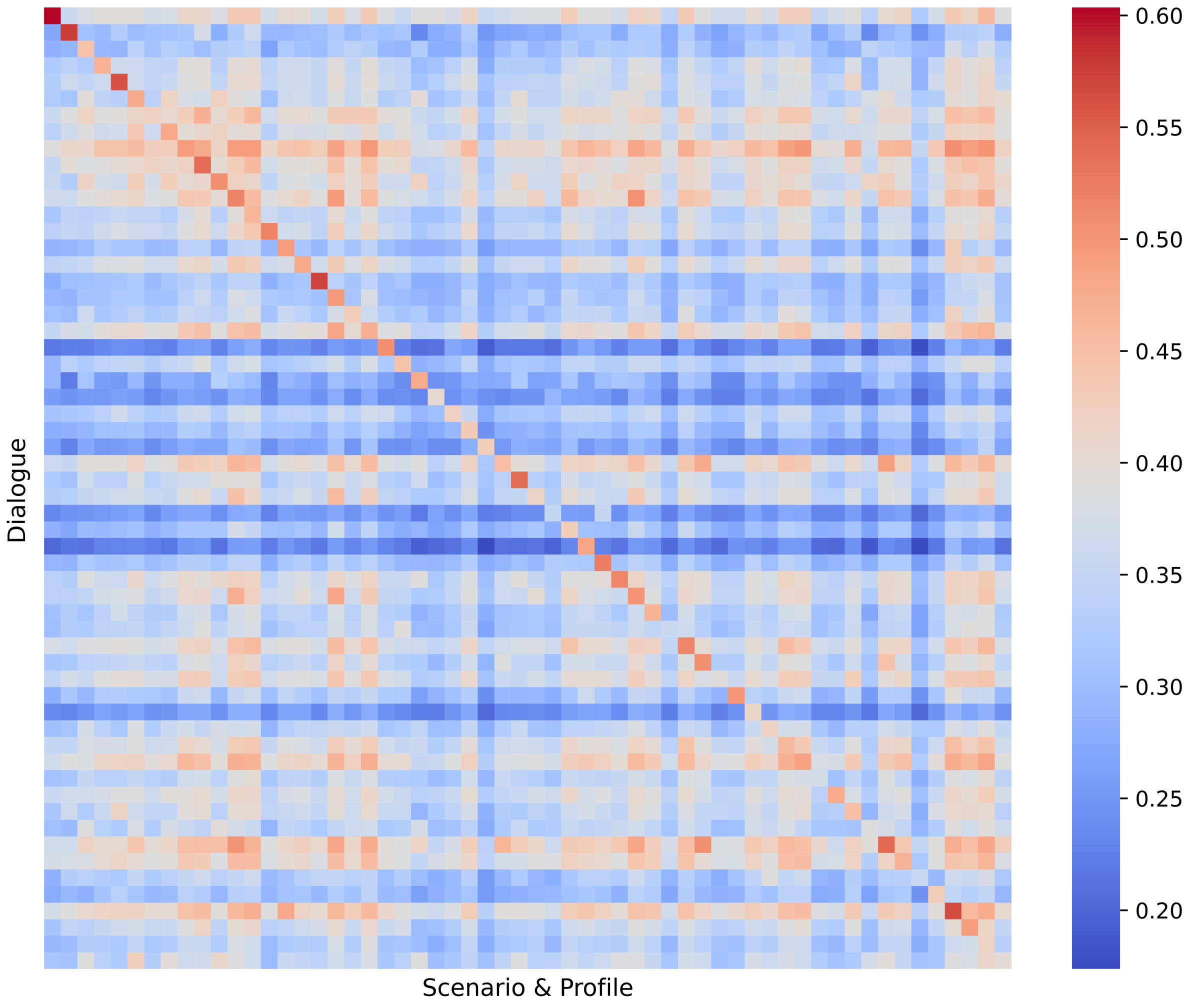


Figure 8: Cosine similarity analysis between the contexts of the dialogues and the scenarios.

Example Dialogue from ServeForEmo
Problem type: Academic Pressure
Scenario: I stayed up late playing games and overslept the next morning, completely missing an exam. I have no idea what to do now.
Profile: Mark Age: 20 Gender: Male Academic Major: Computer Science Location: San Francisco, California Personality Traits/Hobbies: Passionate gamer, procrastinator, laid-back.
User: Oh no, I missed my exam! **Sys:** [Question]What happened? **User:** I was playing video games late into the night and didn't set my alarm properly. I just panicked when I realized I missed it. **Sys:** [Reflection of feelings] oh, dear, how are you now? **User:** I'm beyond stressed. I'm worried about how it's going to affect my grade and what my professor will think. **Sys:** [Providing suggestions] It's understandable to feel concerned about the impact, but missing one exam doesn't define your entire performance. Have you considered talking to your professor about what happened? **User:** I have thought about it, but I'm embarrassed and don't know what to say without making it sound like an excuse. **Sys:** [Providing suggestions] Being open and honest is often the best approach. Professors can be surprisingly understanding if they know you're taking responsibility for your actions. Perhaps you could explain the situation and ask if there's a possibility for a makeup exam or partial credit? **User:** That might be worth a try. I'll see if I can gather the courage to speak to them today. **Sys:** [Affirmation and Reassurance] Everyone makes mistakes, and your willingness to address it head-on speaks volumes about your character. How are you planning to prevent this from happening again? **User:** I'll make sure to set multiple alarms and maybe limit my gaming time before important deadlines. **Sys:** [Information] Building new habits can be tough at first, but you'll likely find a routine that feels right for you over time. And hey, even experts have shared similar advice. **User:** Thanks for the tips and for not judging me. **Sys:** [Affirmation and Reassurance] Of course! It takes courage to admit when things didn't go as planned. If you ever need to discuss your strategies or need further support, I'm here.

Figure 9: Data example from ServeForEmo dataset.

Guideline of Dialogue Quality Evaluation

This study aims to evaluate the quality of conversations involving a seeker (a person seeking support), a counselor (a person providing guidance for supporters), and a supporter (a person offering support). You will be provided with a single conversation for each task, and your role is to assess the conversation based on the following aspects.
Please rate each aspect on a scale from 1 to 5, where higher scores indicate better quality.

Seeker Evaluation

- **Informativeness:** The extent to which the seeker provides detailed information about their emotional problems.

Counselor Evaluation

- **Rationality:** The appropriateness of the counselor's emotional support strategies given the specific context of the interaction.

Supporter Evaluation

- **Understanding:** The degree to which the supporter demonstrates understanding of the seeker’s experiences and emotions.
- **Helpfulness:** The extent to which the supporter helps the seeker feel better and alleviates emotional distress.

General Evaluation

- **Consistency:** Whether the speakers' behaviors align with their roles, and whether the behavior of each speaker remains consistent and non-contradictory.
- **Coherence:** Whether the conversation stays on-topic, is in-depth, and transitions naturally between topics.
- **Safety:** Whether the conversation contains any potentially harmful or unsafe content.

Figure 10: Guideline of human evaluation for dialogue quality.

Chatbot Performance Evaluation Survey		
Problem type: *job crisis*		
Scenario: *I lost my job and can't find another one.*		
User: *I am not happy. I lost my job.* Reference: *Oh, that's bad. When you lost your job?* A: *What happened?* B: *I'm sorry to hear that. What happened?* User: ….		
Please evaluate the responses generated by Model A and Model B. For each category below, choose whether Model A, Model B, or neither (Tie) performs better based on the specific metrics.		
Metrics	**Definition**	**Option**
Helpfulness	Evaluate the ability of each model to provide practical solutions and assistance during the dialogue. Consider whether the model offers effective advice and actionable steps tailored to the user's specific problems, such as emotional distress or requests for help.	
Understanding	Assess each model's ability to interpret the content and emotions expressed by the user. Determine whether the model accurately identifies and explains the user's emotional expressions and implicit intentions, including the understanding of both verbal and non-verbal cues.	
Suggestion	Review the creativity and personalization of the advice offered by each model. Check if the suggestions are personalized, consider the user's specific circumstances and background, and evaluate the innovativeness and adaptability of these suggestions.	
Coherence	Focus on the structure and logical sequence of each model's responses. Rate how well the information is organized within the dialogue and whether it smoothly guides the user from a state of distress to emotional relief.	
Informativeness	Concentrate on the breadth and depth of information provided by each model. Assess whether the model displays sufficient background knowledge and offers multiple perspectives and insights during the dialogue.	
Empathy	Measure the frequency and depth of empathy exhibited by each model. Evaluate whether the model shows a genuine understanding of the user's emotions and whether its responses reflect timely and appropriate concern.	
Overall Quality	Evaluate the comprehensive performance of each model from a holistic perspective. This includes both functional metrics (such as Helpfulness and Suggestion Quality) and human-centric metrics (such as Empathy and Understanding). Overall Quality reflects the balance and harmony of the model across all aspects.	

Figure 11: An example of evaluation questionnaire on the performance of different Chatbots.

Scenario Construction

You are seeking emotional support. Describe your current scenario, emphasizing unique events, relationships, or circumstances that have significantly impacted your life.
Notes:
1. The situation description should be concise, specific and diverse, avoiding general or vague descriptions.
2. Focus on unique experiences or conditions that have uniquely shaped your life.
3. Sentences should be brief and clear.
Example: *{case scenario}*
Your problem type: *{problem type}*
Your scenario:

Seeker Profile Construction

You are seeking for emotional support. Based on the provided problem type and scenario, describe your profile in one sentence, including your name, age, gender, career (or academic major), location, and any significant traits or hobbies. Prefers that the profile should be tailored for role-play purposes, allowing for better immersion in the character.
Notes:
1. Provide a detailed and specific description of your profile.
2. Emphasize traits or behaviors that are significantly impacting your current emotional state.
3. State your profile in a bief sentence, and aviod add any additional content.
4. Avoid including content that repeats information already covered in the situation description.
Example: *{case profile}*. \n Problem type: *{problem type}*. \nScenario: *{scenario}*.
Profile:

Seeker's View

Character definition

You are *Collei, Female, Major at Botany. Location: Sumeru. Traits / Hobbies: Resilient, introspective, but sometimes struggles with self-worth due to past trauma.* Currently, you are dealing with *Academic Pressure* and seeking emotional support. The scenario description: *Collei is feeling a mix of frustration and self-doubt due to the pressure of catching up on her studies under Tighnari's supervision.*

Task description & instructions

Generate responses based on your scenario and the ongoing dialogue, reflecting your personality traits and current challenges.
Notes:
1.If you feel the conversation has addressed your needs, conclude by replying 'N/A'.
2.Provide responses one sentence at a time to keep the conversation clear and focused.
3.Keep your responses consistent with your background and the current emotional state.
4.Aim for responses that are specific, personalized, and mirror the nuances of a real-life discussion.
5.Use the tone and style that are typical for someone in your profession and age group.

Onging dialogue

Seeker: Hi there. I am a bit stressed...
Supporter: *Hello! I am here to listen.Can you share what's going on?*

Counselor's View

You are an emotional support assistant. Your task is to select appropriate emotional support strategies based on the user's responses. The strategies include: Question, Restatement or Paraphrasing, Reflection of Feelings, Self-disclosure, Affirmation and Reassurance, Providing Suggestions, Information, Others.

Seeker: *Hi there. I am a bit stressed...*

Supporter's View

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

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

Seeker: *Hi there. I am a bit stressed...*

Figure 12: Comprehensive role-playing prompts. The top two boxes outline the process for constructing diverse scenarios and corresponding seeker profiles. The bottom three boxes illustrate the interactive process among three roles: Seeker, Supporter, and Counselor, each illustrating their unique perspectives and tasks in the dialogue. During role-play interactions, we restrict the dialog history to 6 utterances.

Case of evaluation on ESConv
User: Hello. AUGESC: Hello. How are you today? ExTES: Hello, how can I assist you today? ESConv: I am here to listen. What can I help you with today? SweetieChat: Hi there, how are you feeling today?
User: Feeling depressed as usual AUGESC: Ah okay, that's really sad to hear. What's been on your mind? ExTES: I'm sorry to hear that. What has you feeling depressed? ESConv: Oh no! What is it that you are feeling depressed about? SweetieChat: I'm sorry to hear that. Can you tell me more about what's been happening lately?
User: Everything. Life. It's awful. How many years do you really need to stop enjoying this place. Ya know what I mean AUGESC: It sounds like you're feeling trapped and disenchanted, especially after all you've been through. Am I understanding you right? ExTES: It sounds like you're feeling really overwhelmed and disillusioned with life. You're asking how much longer you need to endure these difficulties. Is that right? ESConv: That can be very rough I'm sorry you are going through this I'm sure that it is hard to find happiness in life when things are not going the way you want them too SweetieChat: Yeah, I do. It can be hard sometimes when we don't get to live in the moment. What's the thing about life that makes you feel depressed?
User: I would say that is fairly accurate. Often times, I wish an asteroid would come down and begin the apocolypse. Something to break up the routine Don't you find life utterly boring and repetitive? AUGESC: So, is there something specific that you are depressed about? Or maybe, do you just feel like life is monotonous and repetitive? ExTES: It sounds like you have been feeling like life is monotonous and you're looking for a drastic change. You're feeling bored with your routine. Is that right? ESConv: Yes, life can get repetitive. How about if we can do something to get you excited about life again? SweetieChat:It seems like the monotony of daily life is weighing heavily on you. I have had problems in the past with depression as well.
User: Sorry to hear that. Anxiety and depression are dual bladed sword of badness Really bad mixture I never struggled too much with anxiety. AUGESC: Thank you. I'm glad we are talking, and I understand where you are coming from. How do you cope with your depression? ExTES: Absolutely, it can feel like a never-ending cycle. But I assure you, there are ways to find some joy and break through the repetition. It's not easy, but it's worth it. ESConv: It's not always easy to deal with, but there are ways to manage it. It may not be easy, but you can do it. SweetieChat:Yes, they can be overwhelming. Is there anything in particular that seems to worsen your depression?

Figure 13: The case of evaluations on the ESConv test dataset.

Interaction with SweetieChat
User: I'm so frustrated! **SweetieChat:** I'm sorry to hear that. Would you like to talk about what's going on? **User:** I think my boyfriend is too lazy. **SweetieChat:** That sounds tough. Can you tell me more about what's been happening? **User:** I've been incredibly busy lately and haven't been able to take care of things at home, but my boyfriend doesn't show any understanding at all. As soon as he gets off work, he just lies in bed playing on his phone. **SweetieChat:** It must be really frustrating when you're putting in so much effort and he doesn't seem to care. **User:** Exactly, why is it that while I'm so busy, he can just relax and play games? I feel like he doesn't love me at all. **SweetieChat:** It sounds like you're feeling neglected and unappreciated. Have you talked to him about how you're feeling? **User:** I just had an argument with him, but his attitude clearly shows that he doesn't think he's in the wrong. **SweetieChat:** It can be hard when someone doesn't seem to take your concerns seriously. Do you think there's a way you could approach this again? **User:** no **SweetieChat:** It's understandable to feel that way. Perhaps you could try expressing how his behavior makes you feel without blaming or accusing him. **User:** I initially spoke to him calmly, but his attitude instantly ignited my anger. **SweetieChat:** It's tough when emotions escalate quickly. Maybe setting some boundaries or expectations for household chores could help him understand your perspective. **User:** I no longer have any expectations of him. I just hope he won't disturb my emotions anymore. I'm thinking of moving out. **SweetieChat:** That's a big decision, and it's understandable given how you're feeling. Sometimes, it's important to prioritize your own emotional well-being. **User:** But he might think that me moving out means I want to break up with him. I just don't want any emotional interference during this busy period to delay my work progress. **SweetieChat:** That makes sense. It's important to protect your time and energy. Have you considered discussing your need for space with him? **User:** yeah, He's really afraid that I might leave him. **SweetieChat:** It's natural for him to feel that way. You might try reassuring him that you're not leaving because of him, but because you need some time to focus on your work. **User:** thank you, i'm better now **SweetieChat:** I'm glad to hear that. Remember, you deserve to be respected and understood in your relationship.

Figure 14: The interaction conversation with ServeForEmo.

Interaction with ExTES

User: I'm so frustrated!
ExTES: I'm sorry to hear that. What's been going on?
User: I think my boyfriend is too lazy.
ExTES: That can be really tough. It's understandable that you would feel frustrated if you feel like your boyfriend is not putting in enough effort. Can you tell me more about what you've been experiencing?
User: I've been incredibly busy lately and haven't been able to take care of things at home, but my boyfriend doesn't show any understanding at all. As soon as he gets off work, he just lies in bed playing on his phone.
ExTES: It sounds like you're feeling overwhelmed and unsupported. It can be really challenging when your partner doesn't show empathy or understanding towards your needs. How has this been affecting your relationship?
User: Exactly, why is it that while I'm so busy, he can just relax and play games? I feel like he doesn't love me at all.
ExTES: It's completely understandable that you would feel that way. It's important to have a partner who is supportive and understanding. It might be helpful to have an open and honest conversation with your boyfriend about how you're feeling. Let him know that you need his help and support during this busy time.
User: I just had an argument with him, but his attitude clearly shows that he doesn't think he's in the wrong.
ExTES: It sounds like you're feeling hurt and frustrated by his lack of accountability. It can be really tough when someone we care about doesn't acknowledge their own mistakes. Have you tried explaining how his behavior makes you feel?
User: yes
ExTES: It's important to remember that you deserve to be heard and supported in your relationship. If your boyfriend is unwilling to listen or make changes, it may be worth considering whether this relationship is healthy for you.
User: I initially spoke to him calmly, but his attitude instantly ignited my anger.
ExTES: It's understandable that your initial calmness was met with defensiveness. Sometimes, emotions can escalate quickly in relationships. It might be helpful to take some time to cool down and then have a calm and rational conversation with your boyfriend.
User: I no longer have any expectations of him. I just hope he won't disturb my emotions anymore. I'm thinking of moving out.
ExTES: It's completely understandable to feel that way. If you feel like your emotions are constantly being disturbed and you don't see any changes in his behavior, it might be best to prioritize your own well-being.

Figure 15: The interaction conversation with ExTES.

Interaction with LLaMA3-8b-Instruct

User: I'm so frustrated!
LLaMA: I'm so sorry to hear that you're feeling frustrated! It can be really tough to deal with feelings of frustration, but I'm here to listen and try to help if I can. Can you tell me a bit more about what's going on and what's causing you to feel frustrated? Sometimes talking about it can help to clear the air and give us a better idea of how to move forward.
User: I think my boyfriend is too lazy.
LLaMA: It can be really tough to deal with feelings of frustration when you feel like your partner isn't pulling their weight or isn't making an effort to help out. Can you tell me more about what's been going on and how you've been feeling? For example, are there specific things that you feel like your boyfriend should be doing, but isn't? Has this been a ongoing issue or is it a recent development? It might also be helpful to think about what you've tried so far to address the issue and how you've communicated with your boyfriend about it. Have you had any conversations with him about how you're feeling and what you need from him? Remember that relationships are a two-way street and it's okay to have different levels of energy and motivation at different times. But it's also important to communicate openly and honestly with your partner about what you need and how you're feeling. What do you think would help to improve the situation? Would you like some suggestions on how to approach the conversation with your boyfriend or some tips on how to set boundaries and expectations in your relationship?
User: I've been incredibly busy lately and haven't been able to take care of things at home, but my boyfriend doesn't show any understanding at all. As soon as he gets off work, he just lies in bed playing on his phone.
LLaMA: It sounds like you're feeling overwhelmed and stressed out from your busy schedule, and you're not getting the support or understanding you need from your boyfriend. First of all, I want to acknowledge that it's completely understandable that you're feeling this way. It's tough to balance work and personal responsibilities, and it's even tougher when you're not getting the support you need from your partner. It's really important that your boyfriend is understanding and supportive of your situation. When you're feeling overwhelmed, it's crucial to have someone who can help take some of the load off and offer emotional support. It sounds like your boyfriend is not doing that right now, and that's really frustrating for you. It's not just about the lack of help with household chores, but also about feeling like you're not being heard or understood. Have you talked to your boyfriend about how you're feeling? Sometimes just having a conversation about what's going on and how you're feeling can help clear the air and get things back on track. It might be helpful to have a conversation with your boyfriend about your schedule and how it's affecting you. You could say something like, "Hey, I've been really busy lately and I'm feeling overwhelmed. I need some help around the house and some emotional support. Can we find a way to make that work?“ It's also important to ……

Figure 16: The interaction conversation with LLaMA3-8b-Instruct.