Is ChatGPT Equipped with Emotional Dialogue Capabilities?

Weixiang Zhao, Yanyan Zhao*, **Xin Lu, Shilong Wang, Yanpeng Tong, Bing Qin**
Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
{wxzhao, yyzhao}@ir.hit.edu.cn

Abstract

This report presents a study on the emotional dialogue capability of ChatGPT, an advanced language model developed by OpenAI. The study evaluates the performance of ChatGPT on emotional dialogue understanding and generation through a series of experiments on several downstream tasks. Our findings indicate that while ChatGPT's performance on emotional dialogue understanding may still lag behind that of supervised models, it exhibits promising results in generating emotional responses. Furthermore, the study suggests potential avenues for future research directions.

1 Introduction

Emotional dialogue technology is a promising research area that aims to equip chatbots with human-like emotions, enabling them to recognize, understand, and express emotions in their interactions with users, thus generating more engaging and diverse responses. Consequently, emotional dialogue robots have gained significant academic and industrial attention. In recent years, it has emerged as a core technology for enhancing the performance of various application products, including open-domain chatbots (Zhou et al., 2020), intelligent customer service (Chen et al., 2019), and voice assistants (Kepuska and Bohouta, 2018). By integrating emotional dialogue technology into these products, chatbots can better understand the needs and emotions of users, providing services that align with user intent. Overall, emotional dialogue technology represents a promising research area with the potential to enhance the performance of various AI applications by enabling robots to interact with users in a more empathetic and human-like manner.

With the advent of ChatGPT[1], the field of conversational robots has undergone a revolution. As an advanced large-scale language model, ChatGPT has brought about unprecedented semantic understanding and response generation capabilities for conversational robots, greatly improving the interaction experience with human users. Considering the significant breakthrough of ChatGPT in basic conversational technology, as well as recent research analyzing its performance in various traditional natural language processing tasks (Qin et al., 2023; Bang et al., 2023; Kocoń et al., 2023; Wang et al., 2023), whether it exhibits emotional intelligence in dialogue has not yet been explored. Therefore, we are interested in the impact of ChatGPT on the development of emotional dialogue technology. In this report, we will explore the performance of ChatGPT on multiple tasks in the field of emotional dialogue, analyze its strengths and weaknesses, and consider future research directions.

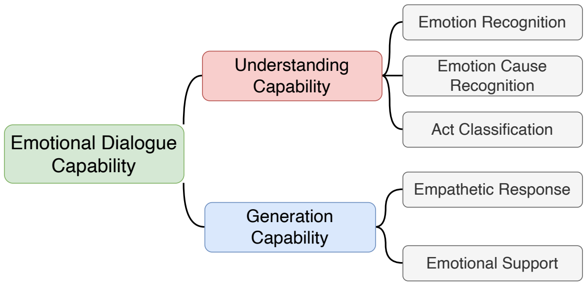

Figure 1: The emotional dialogue capability of a chatbot can be divided into two aspects: understanding and generation capability, with several downstream tasks.

2 Task Settings

As shown in Figure 1, we will compare and analyze the performance of ChatGPT in various downstream tasks based on two dimensions:

- **Understanding Capability**: Is ChatGPT capable of accurately understanding and interpreting the user's emotions?
- **Generation Capability**: Is ChatGPT capable of eliciting empathy or support towards such

*Corresponding author
[1]https://chat.openai.com

emotional states of the user?

Detailed definitions and evaluations of each task will be elaborated in the following sections.

3 Evaluation Methods

We directly refer to the experimental results of the original papers for state-of-the-art (SOTA) models in each task. For ChatGPT's performance testing, we use the "gpt-3.5-turbo" model of the OpenAI public API (version up to March 8). We evaluated ChatGPT's performance in both zero-shot and few-shot prompting settings for above tasks.

4 Understanding Capability

4.1 Emotion Recognition

Emotion Recognition in Conversations (ERC) is a classification task aimed at categorizing emotions within conversational utterances. The input for this task consists of a continuous dialogue, while the output entails the emotional classification of all utterances present. Figure 2 illustrates a straightforward example. Emotion recognition in conversational utterances cannot be reduced to simple single-sentence emotion recognition; it necessitates a holistic examination of the conversation's background, context, and speaker information.

Emotion Recognition in Conversations (ERC) has extensive applications in a variety of conversational settings, including sentiment analysis of comments on social media platforms and emotional assessment of clients in artificial customer service environments. Furthermore, dialogue emotion recognition can be implemented in chatbots to real-time assess users' emotional states, facilitating the generation of emotionally-driven responses.

4.1.1 Task Definition

Given a conversation with several utterances, the goal of Emotion Recognition in Conversations (ERC) is to detect the emotions of all utterances. Formally, given a conversation $C = \{u_1, u_2, ..., u_N\}$ consisting of a sequence of N utterances, the task is to map the utterance sequence to corresponding emotion label sequence $Y = \{y_1, y_2, ..., y_N\}$.

4.1.2 Dataset and Evaluation Metrics

We evaluate the performance of ChatGPT and all baseline models on four publicly available datasets, IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019a), EmoryNLP (Zahiri and Choi, 2017) and DailyDialog (Li et al., 2017).

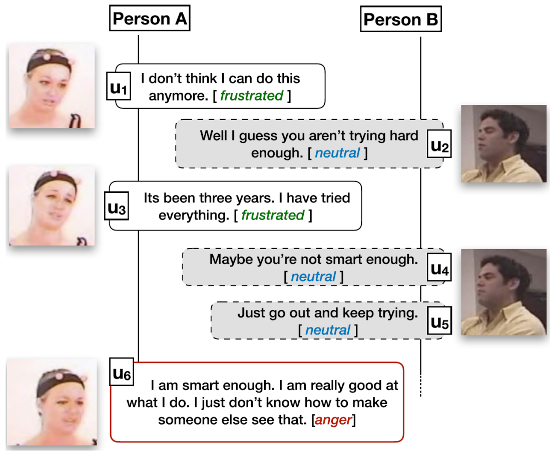

Figure 2: An example (Poria et al., 2019b) of Emotion Recognition in Conversations (ERC).

IEMOCAP The IEMOCAP dataset[2] was collected by SAIL lab at USC. It consists of approximately 12 hours of multimodal conversation data, we only use the text modality in this report. The dataset contains 152 dialogues with a total of 7,433 utterances, and it comes with six emotion categories: happy, sad, neutral, anger, excited and frustrated.

MELD The MELD dataset[3] contains the conversations from Friends TV show transcripts, which is a multimodal extension of the EmotionLines dataset (Hsu et al., 2018). In this report, we only use the text modality. The dataset contains 1,433 dialogues with a total of 13,708 utterances, and it comes with seven emotion categories: neutral, surprise, fear, sadness, joy, disgust and anger.

EmoryNLP The EmoryNLP dataset[4] also contains the conversations from Friends TV show transcripts. The dataset contains 897 dialogues with a total of 12,606 utterances, and it comes with seven emotion categories: sad, mad, scared, powerful, peaceful, joyful and neutral.

DailyDialog The DailyDialog dataset[5] is a high-quality multi-turn dialogue dataset, with conversations reflecting various topics in daily life. The dataset contains 13,118 dialogues with a total of 102,979 utterances, and it comes with seven emotion categories: neutral, happiness, surprise,

[2]https://sail.usc.edu/iemocap/
[3]https://affective-meld.github.io/
[4]https://github.com/emorynlp/emotion-detection
[5]http://yanran.li/dailydialog

Model	IEMOCAP	MELD	EmoryNLP	DailyDialog
DialogueRNN	64.76	63.61	37.44	57.32
IEIN	64.37	60.72	-	-
COSMIC	65.28	65.21	38.11	58.48
DialogXL	65.94	62.41	34.73	54.93
DAG-ERC	68.03	63.56	39.02	59.33
DialogueCRN	66.20	58.39	-	-
CauAIN	67.61	65.46	-	58.21
CoMPM	69.46	66.52	38.93	**60.34**
MuCDN	-	65.37	40.09	-
SPCL	**69.74**	**67.25**	**40.94**	-
ChatGPT 0-shot	44.97	57.30	37.47	40.66
ChatGPT 1-shot	47.46	58.63	35.60	42.00
ChatGPT 3-shot	48.58	58.35	35.92	42.39

Table 1: Comparison of ChatGPT and other baselines on ERC task.

sadness, anger, disgust and fear.

For the IEMOCAP, MELD and EmoryNLP datasets, most papers currently use the Weighted-F1 metric for evaluation; for the DailyDialog dataset, due to the extremely high proportion of neutral utterances, most papers currently use the Micro-F1 metric that excludes the neutral category for evaluation.

4.1.3 Main Results

We selected the current state-of-the-art (SOTA) models: DialogueRNN (Majumder et al., 2019), IEIN (Lu et al., 2020), COSMIC (Ghosal et al., 2020), DialogXL (Shen et al., 2021a), DAG-ERC (Shen et al., 2021b), DialogueCRN (Hu et al., 2021), CauAIN (Zhao et al., 2022b), CoMPM (Lee and Lee, 2022), MuCDN (Zhao et al., 2022d) and SPCL (Song et al., 2022) as baseline models and tested the performance of the baseline models and ChatGPT on ERC task.

The experimental results are shown in Table 1, from which it can be seen that ChatGPT generally has a performance gap of 3-18 percentage points compared to the most advanced fine-tuned models.

4.1.4 Case Study

We show a dialog from the DailyDialogue dataset, which simulates a conversation scenario between a doctor and a patient, as shown in Table 2.

4.1.5 Analysis and Discussion

In the case study section, we display potential annotation errors in the dataset using red font and instances where ChatGPT rectifies annotation errors in green font. Furthermore, ChatGPT's prediction results include labels in yellow font, signifying an additional issue we identified: an inconsistency between ChatGPT and the dataset's guidelines. Examining these actual prediction samples reveals that the primary challenge for ChatGPT is the deviation between its criteria and those of the dataset. While the dataset annotation likely adhered to specific guidelines to determine the corresponding emotion for a given situation, ChatGPT operates under its own interpretation and standards. For instance, in the dialogue between the physician and patient, when the patient describes their headache symptoms, the dataset annotation classifies the emotion as neutral, while ChatGPT deems it sadness. This discrepancy cannot be attributed to one being correct and the other incorrect but rather highlights the differing standards employed.

Upon further discussion, this misalignment of standards may not stem from ChatGPT's capabilities but could be attributed to the few-shot prompting setting. As annotation guidelines become increasingly intricate and involved, it becomes implausible to encompass them with merely a handful of examples, which is an inherent constraint of few-shot prompting.

This insight allows for conjecture on possible future directions: if the objective is not to strictly conform to specific guidelines, then enhancements based on few-shot prompting settings, such as ChatGPT, are viable. However, utilizing dataset labels for evaluation may be unsuitable, potentially necessitating extensive human evaluations. Conversely,

Speaker	Dialogue Content	Annotation	Prediction
A	Good morning. What's the matter with you?	Neutral	Neutral
B	Good morning, doctor. I have a terrible headache.	Neutral	Sadness
A	All right, young man. Tell me how it got started.	Neutral	Neutral
B	Yesterday I had a runny nose. Now my nose is stuffed up. I have a sore throat. And I'm afraid I've got a temperature. I feel terrible.	Neutral	Sadness
A	Don't worry, young man. Lat me give you an examination. First let me take a look at your throat. Open your mouth and say 'ah'.	Neutral	Neutral
B	Ah.	Neutral	Neutral
A	Your throat is inflamed. And your tongue is heavily coated. You have all the symptoms of influenza.	Neutral	Fear
B	What am I supposed to do then?	Neutral	Fear
A	A good rest is all you need, and drink more water. I'll write you a prescription.	Neutral	Happiness
B	Thank you very much.	Neutral	Happiness

Table 2: An example of ChatGPT's prediction on ERC task.

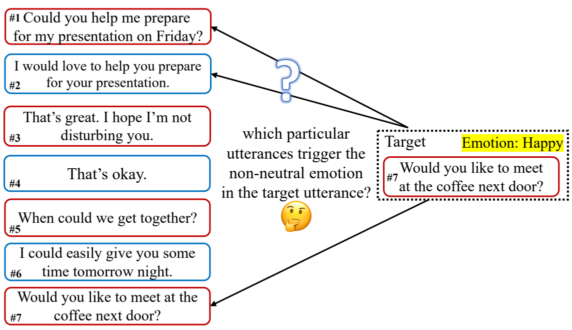

Figure 3: An example from RECCON-DD dataset (Poria et al., 2021) for identifying the causal utterances.

if the goal is to strictly adhere to specific guidelines, few-shot prompting settings may not be the optimal choice, with supervised fine-tuned models remaining the superior alternative.

4.2 Emotion Cause Recognition

Step further on ERC, recognizing emotion cause in conversations serves to fully understand the emotional situations of the user and is beneficial to improve interpretability and performance in affect-based models. And Poria et al. (2021) introduce a new task named RECCON with annotated emotion causes in conversations. It includes two different sub-tasks: Causal Span Extraction (CSE) and Causal Emotion Entailment (CEE). We focus on the CEE sub-task in this report and its goal is to predict which particular utterances in the conversation history trigger the non-neutral emotion in the target utterance, as illustrated in Figure 3.

4.2.1 Task Definition

First, we define the problem of the CEE task. Given a conversation that consists of t consecutive utterances $\{u_1, u_2, \cdots, u_t\}$ with the corresponding emotion label $\{e_1, e_2, \cdots, e_t\}$ between two speakers, the goal of this task is to predict which particular utterances u_i $(i \leq t)$ in the conversational history are responsible for the non-neutral emotion e_t in the target utterance u_t. And u_i is a positive example if it contains the cause of non-neutral emotion in the target utterance and a negative example otherwise.

4.2.2 Dataset and Evaluation Metrics

In this study, we evaluate the performance of ChatGPT and all baseline models on the widely-used benchmark dataset RECCON-DD[6], which features utterance-level emotion labels and emotion cause labels annotated by Poria et al. (2021), based on the popular dialogue dataset DailyDialog (Li et al., 2017). Specifically, we focus on the causal pairs extracted from the dialogue history, filtering out any repetitive instances.

To be consistent with all baseline methods, we report the F1 scores of both negative and positive causal pairs and the macro F1 scores of them.

[6]https://github.com/declare-lab/RECCON.

Model	Neg. F1	Pos. F1	macro F1
RoBERTa-Base	88.74	64.28	76.51
RoBERTa-Large	87.89	66.23	77.06
KEC	88.85	66.55	77.70
KBCIN	89.65	68.59	79.12
TSAM	**90.48**	**70.00**	**80.24**
ChatGPT 0-shot	85.25	51.33	68.29
ChatGPT 1-shot	82.10	52.84	67.47

Table 3: Comparison of ChatGPT and other baselines on CEE task.

4.2.3 Main Results

We compare the performance of ChatGPT 0-shot and 1-shot ability with that of supervised baseline models. To be more specific, we select the pretrained-model based methods (Poria et al., 2021; Zhang et al., 2022) and commonsense-based methods (Li et al., 2022; Zhao et al., 2022a).

Results are presented in Table 3. On one hand, ChatGPT achieve comparable performance on the recognition of negative causal pairs. On the other hand, it still holds a 11.95% gap in terms of macro F1 score compared to the SOTA supervised method.

4.2.4 Analysis and Discussion

Through an analysis of error cases in ChatGPT, it has been identified that the primary cause of the performance gap between ChatGPT's Pos. F1 score and the state-of-the-art (SOTA) lies in the presence of numerous emotionally charged samples within the target sentence itself. ChatGPT has shown a tendency to overlook these types of examples and instead focuses on identifying causal statements from the conversation context. This observation aligns with the previously discussed analysis of emotion recognition in dialogues. One of the primary reasons for ChatGPT's poor performance is the considerable discrepancy between its prediction standard and the dataset's annotation standard. Additionally, ChatGPT's performance decreases when given a specific example, highlighting the importance of a thorough understanding of the complex annotation standards for tasks such as identifying the causes of emotions. Unlocking the full potential of ChatGPT's performance requires a deep understanding of the dataset's specifications, allowing it to overcome its own language model prior and achieve results that better align with downstream testing data.

4.3 Dialog Act Classification

Dialogue Act Classification (DAC) is a classification task that involves labeling the dialogue acts for each utterance in a dialogue. The input for this task is a continuous segment of dialogue consisting of several utterances, and the output is the labeling of dialogue acts for all utterances. It is assumed that each round in a dialogue involves a particular dialogue act, and therefore, the label set does not include a neutral label. An example of DAC is shown in Table 4. Similar to ERC, DAC is not a simple sentence classification problem; instead, it requires considering the context of the conversation.

Categorizing dialogue acts enables a better understanding of a conversation's intentions, emotions, and context. This task has extensive applications, including dialogue systems, customer service chatbots, smart speakers, and sentiment analysis. Dialogue act classification in dialogue systems and customer service chatbots can improve robots' comprehension of users' needs and intentions, resulting in better service and responses. In sentiment analysis, dialogue action classification can facilitate an improved understanding of dialogue behaviors under various emotional states, allowing for a more thorough analysis and comprehension of users' emotional states.

4.3.1 Task Definition

Given a conversation with several utterances, the goal of DAC is to detect the dialog acts of all utterances. Formally, given a conversation $C = \{u_1, u_2, ..., u_N\}$ consisting of a sequence of N utterances, the task is to map the utterance sequence to corresponding dialog act label sequence $Z = \{z_1, z_2, ..., z_N\}$.

4.3.2 Dataset and Evaluation Metrics

The experiments were carried out on the DailyDialog (Li et al., 2017) dataset, as previously introduced in the ERC section. This study solely employed the act tags available in the dataset. To evaluate the classification task, both weighted-F1 and macro-F1 were used as performance metrics. However, ChatGPT assigned meaningless labels beyond the four categories, which greatly affected the macro-average value. To address this issue, the evaluation metric used in this task was the weighted average F1 score.

Speaker	Dialogue Content	Annotation	Prediction
A	When can we expect you for dinner? Can you come tonight?	Directive	Question
B	Not tonight. I promised to go to a concert with my sister.	Commissive	Inform
A	Well ... How about Friday then?	Directive	Question
B	That sounds fine.	Commissive	Commissive

Table 4: An example of ChatGPT prediction on DAC task.

Model	Acc	Weighted-F1
CASA	-	0.78
DCR-Net + Co-Attention	-	0.79
Co-GAT	-	0.79
WEAKDAP	0.84	-
ChatGPT 1-shot	0.67	0.65
ChatGPT 1-shot + P.E	0.71	0.70
ChatGPT 3-shot	0.73	0.71
ChatGPT 3-shot + P.E	0.73	0.72

Table 5: Comparison of ChatGPT and other baselines on DAC task. Experimental results are partly cited from Co-GAT(Qin et al., 2021). P.E stands for prompt engineering.

4.3.3 Main Results

We select the current state-of-the-art (SOTA) model **WEAKDAP** (Chen et al., 2022) and some other strong baselines **CASA** (Raheja and Tetreault, 2019), **DCR-Net** (Qin et al., 2020) and **Co-GAT** (Qin et al., 2021). Then we teste the performance of the baseline models and ChatGPT on Dialog Act Classification(DAC) task.

The experimental results are shown in Table 5, from which it can be seen that ChatGPT generally has a performance gap of 11-17 percentage points compared to the baseline fine-tuned models.

4.3.4 Case Study

We show a dialog from the DailyDialogue dataset, which simulates a conversation scenario between a doctor and a patient, as shown in Table 4.

4.3.5 Analysis and Discussion

ChatGPT has demonstrated limited understanding of labels such as directives and promises. Specifically, it tends to conflate question with directive and inform with commissive, as illustrated in the case study. Due to the lack of clear definitions, these two labels have a semantic overlap, making it difficult to distinguish between them. For example, "Can you come today?" is a directive question, while "I promised to go to the concert with my sister" is a commissive inform. It should be noted that this issue does not necessarily indicate ChatGPT's poor understanding of conversational actions, but rather highlights a disparity between the model's labeling system and that of the dataset. To address this challenge, incorporating comprehensive label explanations into prompts, referred to as prompt engineering, can significantly enhance evaluation metrics.

The experimental results demonstrate that the few-shot setting is the most effective prompt enhancement method for ChatGPT in this task. The approach does not require complex prompt engineering for the commissive and directive labels, yet it can significantly improve the evaluation metrics. The experiment employed three samples for few-shot and a simple prompt engineering design. We have reasons to believe that carefully selecting more examples and refining the prompt engineering can further decrease the difference between ChatGPT's label system and the original label system of the dataset, leading to an improvement in ChatGPT's performance in this task. Nevertheless, as mentioned in the previous two tasks, the evaluation system's alignment with the dataset label system warrants further consideration.

The results of the experiment indicate that prompt engineering can enhance ChatGPT's performance in a specific task. However, since the prompts used in this study were not intricate, the upper limit of ChatGPT's performance was not tested. We believe that a more sophisticated prompt engineering approach could potentially improve ChatGPT's performance on this task. Developing more customized and high-quality prompts that are well-suited to the task and maximizing the potential of large models may represent a new paradigm for task-solving.

Model	D-1	D-2	B-1	B-2	B-3	B-4	R-L
Multi-TRS	0.44	1.98	21.42	7.60	3.67	2.13	21.66
MoEL	0.58	2.91	21.70	7.75	3.58	1.96	**22.08**
MIME	0.42	1.71	**22.20**	**8.07**	3.89	2.22	21.54
EmpDG	0.44	1.91	21.99	8.02	3.86	2.19	22.03
CEM	0.65	3.03	18.69	6.84	3.37	1.92	21.65
EmpSOA	0.69	3.87	21.41	8.06	**4.14**	**2.41**	21.64
SEEK	0.66	2.74	15.21	4.40	1.81	0.85	19.44
ChatGPT 2-shot	**4.17**	**23.23**	12.99	4.07	1.90	1.05	12.70

Table 6: Comparison of ChatGPT against state-of-the-art baselines in terms of the automatic evaluation. The best results among all models are highlighted in bold.

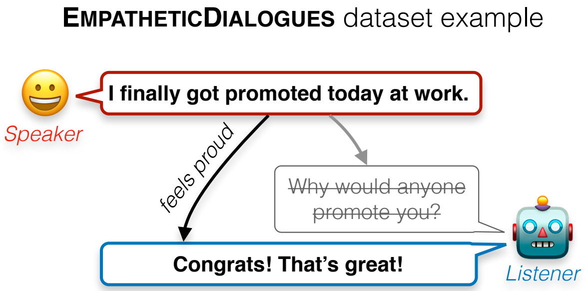

Figure 4: An example of empathetic conversation from EMPATHETICDIALOGUE dataset (Rashkin et al., 2018).

ChatGPT vs.	EmpSOA		
	Win	Lose	Tie
Coherence	**49.00**	6.33	44.67
Empathy	**54.33**	8.00	37.67
Informativeness	**78.67**	6.00	15.33

Table 7: The results of the human evaluation.

5 Generation Capability

5.1 Empathetic Response Generation

Empathy is a highly valued characteristic in engaging human conversation, and it is also a crucial element in the development of human-like chatbots. We concentrate on the task of generating empathetic responses (Rashkin et al., 2018) that accurately comprehend the user's emotions and circumstances and provide appropriate feedback. An example is shown in Figure 4.

5.1.1 Task Definition

We formally define the task of empathetic response generation. Specifically, we consider a dialogue history denoted by $D = [X_1, X_2, \cdots, X_N]$, where N represents the number of utterances exchanged between the user and the system. Each utterance $X_i = [w_1^i, w_2^i, \cdots, w_m^i]$ consists of a sequence of m words. Additionally, the conversation is annotated with an emotion label e from a set of 32 available emotions, indicating the emotional tone of the user. The objective of this task is to generate a coherent response Y from the perspective of the system that not only takes into account the dialogue history D but also demonstrates empathy towards the user's situation and emotions.

5.1.2 Dataset and Evaluation Metrics

We conduct our experiments on EMPATHETICDIALOGUES[7] dataset (Rashkin et al., 2018), which is a large-scale, multi-turn corpus consisting of 25k empathetic conversations, collected via Amazon Mechanical Turk. It features conversations between a speaker and a listener. Notably, the dataset includes 32 emotion labels, evenly distributed to indicate the personal emotional states of the user.

Automatic Evaluation. We apply two kinds of automatic metrics for evaluation: (1) BLEU-n (**B-1**, **B-2**, **B-3**, **B-4**) (Papineni et al., 2002) and ROUGE-L (**R-L**) (Lin, 2004) evaluate the lexical and semantic aspects of the generated responses; (2) Distinct-n (**Dist-n**) (Li et al., 2015) evaluates the diversity of the generated responses by measuring the ratio of unique n-grams.

Human Evaluation. We conduct an aspect-based pairwise preference test to evaluate the quality of responses generated by ChatGPT and EmpSOA. Specifically, we randomly sample 100 pairs

[7]https://github.com/facebookresearch/EmpatheticDialogues

Emotion	Devastated
Context	My dog I had for five years just passed away. He was hit by a car last night.
MIME	Oh no! I am so sorry to hear that.
EmpDG	Oh no! I am so sorry to hear that.
CEM	Oh no! I am so sorry to hear that.
EmpSOA	Oh no! I am so sorry to hear that.
ChatGPT	I'm so sorry for your loss. Losing a beloved pet can be incredibly hard. Would you like to talk about it more?
Ground-Truth	Oh I am so sorry. That must be fresh on your heart, may he rest in peace.
Emotion	Content
Context	My wife made me pancakes for breakfast. I have a full belly and feel rather happy now. Yum! I love pancakes. Is it your favorite food too? One of my favorite breakfast foods. She also made some thick sliced bacon. She treats me so well.
MIME	That is a great idea. I love my girlfriend too.
EmpDG	That is a great attitude to have!
CEM	That is good, I love it!
EmpSOA	That sounds so sweet. I am glad you have a great time!
ChatGPT	It's so wonderful to have someone who cares for and treats us well. Enjoy your delicious breakfast!
Ground-Truth	That is great you have a sweet wife!

Table 8: Case study of the generated empathetic responses by ChatGPT and the baselines.

of responses and ask three professional annotators to assess which response is better based on three criteria: **Coherence**, **Empathy**, and **Informativeness**. The Coherence criterion assesses the coherence and relevance of the response to the dialogue history, while the Empathy criterion evaluates the degree of empathy displayed towards the user's feelings and situations. Finally, the Informativeness criterion gauges the amount of information related to the dialogue history contained in the response.

5.1.3 Main Results

We compare ChatGPT with the following competitive baselines: **Multi-TRS** (Rashkin et al., 2018), **MoEL** (Lin et al., 2019), **MIME** (Majumder et al., 2020), **EmpDG** (Li et al., 2019), **CEM** (Sabour et al., 2022), **EmpSOA** (Zhao et al., 2022c) and **SEEK** (Wang et al., 2022).

The experimental results presented in Table 6 demonstrate that the responses generated by ChatGPT exhibit greater diversity, with an average length of 21.77 tokens, compared to the baseline methods, which have an average length of 14.89 tokens. It is important to note, however, that this increased diversity may lead to a higher degree of mismatch with the golden, indicating a potential trade-off between response diversity and accuracy.

5.1.4 Analysis and Discussion

During empathetic response generation, ChatGPT tends to produce longer and more diverse responses, particularly when the user is in a negative emotional state. ChatGPT tends to suggest solutions to address the problems users are facing, leading to a deviation from real responses. This is also the reason why ChatGPT performs significantly worse than the baseline method on word overlap-based automatic evaluation metrics. Furthermore, based on human evaluation, the baseline method's coherence and empathy ability are comparable to ChatGPT, but the amount of information provided in the responses differs significantly. ChatGPT's generated responses are able to fully understand the user's situation, expand on the user's topic, and provide more effective information to the user. As shown in the cases in Table 8, regardless of whether the user's emotion is positive or negative, ChatGPT's responses are more specific to the context of the conversation, rather than generating generic responses. However, in terms of empathy expression, ChatGPT often repeats the pattern of restating the user's emotion before expanding on the information, which can make the user feel bored.

In light of future directions for this task, several considerations emerge. First and foremost, investing in expanding the parameter and data volume of dialogue models has demonstrated its effectiveness in improving performance, as exemplified by ChatGPT. Second, enhancing the model's ability to empathize on a personalized level remains crucial, as it is apparent that relying on template-like expressions for empathy does not accurately align with

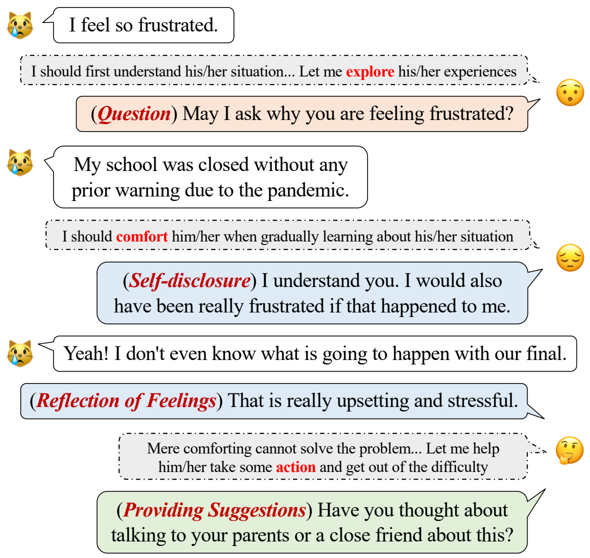

Figure 5: An example of emotional support conversation from ESConv dataset (Liu et al., 2021).

authentic human empathetic conversation. Lastly, differences in model performance revealed through automatic and human evaluation underscore the current lack of a suitable evaluation metric for assessing the quality of empathetic dialogue systems.

5.2 Emotional Support Conversation

Emotional support conversation (ESC) is a dialogue response generation task that aims to provide assistance to seekers who approach for help while experiencing negative emotions. The task requires input of the dialogue history between the seeker and supporter, and outputs an emotional support response from the supporter. An example is shown in Figure 5. It consists of three stages: first, the supporter needs to explore the situations and identify the issue the seeker is facing; second, the supporter should offer comfort to the seeker, and finally, take action to provide advice or information to help the seekers address their problem. The supporter may utilize eight distinct strategies, including questioning, restatement or paraphrasing, reflection of feelings, self-disclosure, affirmation and reassurance, providing suggestions, information, and others. Please refer to Appendix A for the detailed definitions of these strategies.

5.2.1 Task Definition

Formally, let $D = [X_1, X_2, \cdots, X_N]$ denote a dialogue history consisting of N utterances between a seeker and a supporter. Each utterance $X_i = [w_1^i, w_2^i \cdots, w_m^i]$ is a sequence of m words, and utterances from the supporter's turns are associated with the support strategy S_i. Our objective is to generate the next coherent and supportive utterance Y from the supporter's perspective, with the aim of alleviating the seeker's distress.

5.2.2 Dataset and Evaluation Metrics

Our experiments are conducted on the ESConv[8] dataset (Liu et al., 2021). It is an English dataset. To construct the dataset, they recruited crowdworkers, who had learned the common procedures and strategies for providing emotional support, to converse with volunteers that needed emotion support through an online platform. It contains 1,300 long dialogues with 38,350 utterances. There is an average of 29.5 utterances per dialogue and an average 16.7 tokens per utterance.

Automatic Evaluation. We employ the same automatic metrics as those used in the empathetic response generation task. In addition, Accuracy (**Acc**) of the strategy prediction is utilised to evaluate the capability to choose the supportive strategy.

Human Evaluation. Following Liu et al. (2021), we recruit one professional annotators to interact with the models for human evaluation. Specifically, we recruit a professional annotator to evaluate 89 dialogues randomly sampled from the test set of ESConv. The annotator assumes the role of a seeker and engages in conversations with both models, and evaluates them based on five criteria: (1) **Fluency**: the coherence and smoothness of the generated responses, (2) **Identification**: how effectively the models identify and address the seeker's problems, (3) **Empathy**: the level of empathetic understanding displayed by the models towards the seeker's feelings and situation, (4) **Suggestion**: the quality of the suggestions offered by the models, and (5) **Overall**: the overall effectiveness of the models in providing emotional support.

5.2.3 Main Results

We compare ChatGPT with the following competitive baselines on ESC task: **BlenderBot-Joint** (Liu et al., 2021), **MISC** (Tu et al., 2022), **GLHG** (Peng et al., 2022) and **MultiESC** (Cheng et al., 2022). They are based on generative pretrained model BlenderBot (Roller et al., 2020) and BART (Lewis et al., 2019).

[8]https://github.com/thu-coai/Emotional-Support-Conversation.

Model	Acc	D-1	D-2	B-1	B-2	B-3	B-4	R-L
BlenderBot-Joint	17.69	2.96	17.87	18.78	7.02	3.20	1.63	14.92
MISC	31.67	4.62	20.17	16.31	6.57	3.26	1.83	17.24
GLHG	-	3.50	21.61	19.66	7.57	3.74	2.13	16.37
MultiESC	**42.01**	-	-	**21.65**	**9.18**	**4.99**	**3.09**	**20.41**
ChatGPT 1-shot	17.0	**5.92**	**31.38**	13.91	4.53	1.96	1.02	13.19

Table 9: Comparison of ChatGPT against state-of-the-art baselines in terms of the automatic evaluation on ESC task. The best results among all models are highlighted in bold.

ChatGPT vs.	MISC		
	Win	Lose	Tie
Fluency	**61**	6	22
Identification	**68**	6	15
Empathy	16	**40**	33
Suggestion	**73**	3	13
Overall	**65**	12	12

Table 10: The results of the human interaction evaluation between ChatGPT and MISC.

5.2.4 Analysis and Discussion

As shown in Table 9, the responses generated by ChatGPT exhibits both long and diverse characteristics, leading to superior performance over SOTA methods in terms of the automatic evaluation metric Distinct-n. However, such diversity can also introduce deviations from golden responses.

And for human evaluation displayed in Table 10, since one of the characteristics of ESC is to provide users with suggestions and effective information to help them get out of their dilemma, this happens to be consistent with the generation preference of ChatGPT of generating informative responses. Therefore it demonstrates excellent performance in this task. However, in terms of empathy, the reason why SOTA methods outperform ChatGPT is that ChatGPT is too eager to give corresponding advice and coping strategies once it confirms the dilemma the user faces, ignoring the comfort and care of the user's emotions. But this does not imply that ChatGPT lacks the ability of eliciting empathy. Its excellent performance in empathetic response generation tasks can prove that it can comfort users empathetically (Section 5.1). Through appropriate prompt engineering, we believe that ChatGPT can "slow down" and carry out sufficient emotional guidance before giving users advice. Compared with MISC, ChatGPT can display more diverse and effective supportive responses. However, MISC cannot learn this point from existing datasets because real advice in the corpus itself is limited.

Future research on ESC should focus on how to make the model adaptively control the rhythm of emotion support to avoid giving advice too hastily or repeating ineffective comfort. Furthermore, exploring more reasonable automatic evaluation metrics that align with human evaluation is a research direction that deserves further exploration.

6 Conclusion

In this report, we conduct a preliminary exploration of the emotional conversation capability of ChatGPT. It should be noted that our experimental results may not fully reflect the optimal performance of ChatGPT in the corresponding task. With more refined prompt engineering and context example selection, we believe that the performance of ChatGPT can be further improved. One of the future directions for emotional dialogue understanding is to explore the alignment of ChatGPT with labeling standards. As for emotional dialogue generation, it is important to investigate reasonable automatic evaluation metrics to measure model performance, as the results obtained from widely used automatic and human evaluations may differ.

Limitations

Limitations of Model Selection. Due to resource constraints, we have limited our evaluation to a single representative language model (LLM), namely the gpt-3.5-turbo variant of ChatGPT. However, the field of LLMs is rapidly advancing, with numerous other notable models such as the GPT-3.5 series (including text-davinci-002, code-davinci-002, textdavinci-003), as well as the recently announced GPT-4. It is our belief that a comprehensive analysis of the abilities of various LLMs in the emotion dialogue capability will be necessary in the future.

Limitations of Automic Evaluation. Owing to limited resources, we only employ simple prompt engineering and conduct few-shot prompting under a low-resource setting (no more than 3-shot). However, this approach may not accurately reflect the optimal performance of ChatGPT on the corresponding downstream tasks.

Limitations of Human Evaluation. When it comes to evaluating the performance of dialogue emotion generation tasks, relying solely on a small group of volunteers may not provide a comprehensive understanding of the model's capabilities. The preferences of this limited group towards the SOTA model and ChatGPT's generation might not be reflective of a more general audience. Therefore, there is a need for a more objective and universal evaluation method to provide a more accurate assessment of the model's performance.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Cen Chen, Xiaolu Zhang, Sheng Ju, Chilin Fu, Caizhi Tang, Jun Zhou, and Xiaolong Li. 2019. Antprophet: an intention mining system behind alipay's intelligent customer service bot. In *IJCAI*, volume 8, pages 6497–6499.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Andy Rosenbaum, Seokhwan Kim, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2022. Weakly supervised data augmentation through prompting for dialogue understanding. *arXiv preprint arXiv:2210.14169*.

Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. *arXiv preprint arXiv:2210.04242*.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv preprint arXiv:2010.02795*.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao K. Huang, and Lun-Wei Ku. 2018. Emotionlines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. *arXiv preprint arXiv:2106.01978*.

Veton Kepuska and Gamal Bohouta. 2018. Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home). In *2018 IEEE 8th annual computing and communication workshop and conference (CCWC)*, pages 99–103. IEEE.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. Chatgpt: Jack of all trades, master of none. *arXiv preprint arXiv:2302.10724*.

Joosung Lee and Wooin Lee. 2022. CoMPM: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5669–5679, Seattle, United States. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jiangnan Li, Fandong Meng, Zheng Lin, Rui Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. 2022. Neutral utterances are also causes: Enhancing conversational causal emotion entailment with social commonsense knowledge. *arXiv preprint arXiv:2205.00759*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

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

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

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

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

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Xin Lu, Yanyan Zhao, Yang Wu, Yijian Tian, Huipeng Chen, and Bing Qin. 2020. An iterative emotion interaction network for emotion recognition in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4078–4088.

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

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. *arXiv preprint arXiv:2204.12749*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13:1317–1332.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Libo Qin, Wanxiang Che, Yangming Li, Mingheng Ni, and Ting Liu. 2020. Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8665–8672.

Libo Qin, Zhouyang Li, Wanxiang Che, Minheng Ni, and Ting Liu. 2021. Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13709–13717.

Vipul Raheja and Joel Tetreault. 2019. Dialogue act classification with context-aware self-attention. *arXiv preprint arXiv:1904.02594*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021a. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13789–13797.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021b. Directed acyclic graph network for conversational emotion recognition. *arXiv preprint arXiv:2105.12907*.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5197–5206, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2203.13560*.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Hao-jun Huang, Wei Ye, Xiubo Geng, et al. 2023. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*.

Lanrui Wang, Jiangnan Li, Zheng Lin, Fandong Meng, Chenxu Yang, Weiping Wang, and Jie Zhou. 2022. Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection. *arXiv preprint arXiv:2210.11715*.

Sayyed M Zahiri and Jinho D Choi. 2017. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. *arXiv preprint arXiv:1708.04299*.

Duzhen Zhang, Zhen Yang, Fandong Meng, Xiuyi Chen, and Jie Zhou. 2022. Tsam: A two-stream attention model for causal emotion entailment. *arXiv preprint arXiv:2203.00819*.

Weixiang Zhao, Yanyan Zhao, Zhuojun Li, and Bing Qin. 2022a. Knowledge-bridged causal interaction network for causal emotion entailment. *arXiv preprint arXiv:2212.02995*.

Weixiang Zhao, Yanyan Zhao, and Xin Lu. 2022b. Cauain: Causal aware interaction network for emotion recognition in conversations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4524–4530.

Weixiang Zhao, Yanyan Zhao, Xin Lu, and Bing Qin. 2022c. Don't lose yourself! empathetic response generation via explicit self-other awareness. *arXiv preprint arXiv:2210.03884*.

Weixiang Zhao, Yanyan Zhao, and Bing Qin. 2022d. Mucdn: Mutual conversational detachment network for emotion recognition in multi-party conversations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7020–7030.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

A Definitions of Strategies in ESC

There are overall 8 types of support strategies that are originally annotated in the ESConv dataset:

- **Question**: ask for information related to the problem to help the help-seeker articulate the issues that they face.
- **Restatement or Paraphrasing**: a simple, more concise rephrasing of the support-seeker's statements that could help them see their situation more clearly.
- **Reflection of Feelings**: describe the help-seeker's feelings to show the understanding of the situation and empathy.
- **Self-disclosure**: share similar experiences or emotions that the supporter has also experienced to express your empathy.
- **Affirmation and Reassurance**: affirm the help-seeker's ideas, motivations, and strengths to give reassurance and encouragement.
- **Providing Suggestions**: provide suggestions about how to get over the tough and change the current situation.
- **Information**: provide useful information to the help-seeker, for example with data, facts, opinions, resources, or by answering questions.
- **Others**: other support strategies that do not fall into the above categories.