

# Enhancing Empathetic Response Generation by Augmenting LLMs with Small-scale Empathetic Models

Anonymous ACL submission

## Abstract

Empathetic response generation is increasingly significant in AI, necessitating nuanced emotional and cognitive understanding coupled with articulate response expression. Current large language models (LLMs) excel in response expression; however, they lack the ability to deeply understand emotional and cognitive nuances, particularly in pinpointing fine-grained emotions and their triggers. Conversely, small-scale empathetic models (SEMs) offer strength in fine-grained emotion detection and detailed emotion cause identification. To harness the complementary strengths of both LLMs and SEMs, we introduce a Hybrid Empathetic Framework (HEF). HEF regards SEMs as flexible plugins to improve LLM’s nuanced emotional and cognitive understanding. Regarding emotional understanding, HEF implements a two-stage emotion prediction strategy, encouraging LLMs to prioritize primary emotions emphasized by SEMs, followed by other categories, substantially alleviates the difficulties for LLMs in fine-grained emotion detection. Regarding cognitive understanding, HEF employs an emotion cause perception strategy, prompting LLMs to focus on crucial emotion-eliciting words identified by SEMs, thus boosting LLMs’ capabilities in identifying emotion causes. This collaborative approach enables LLMs to discern emotions more precisely and formulate empathetic responses. We validate HEF on the Empathetic-Dialogue dataset, and the findings indicate that our framework enhances the refined understanding of LLMs and their ability to convey empathetic responses.

## 1 Introduction

As an important hot topic in dialogue tasks, empathetic response generation aims to finely understand the dialogue context from both emotional and cognitive perspectives, and express appropriate responses (Rashkin et al., 2019; Sabour et al., 2022; Yang et al., 2023b; Zhao et al., 2022; Zhou et al.,

Complementary Capabilities for Empathy	SEMs	LLMs
Fine-grained emotion detection (Affection)	Stronger	Weaker
Detailed emotion cause identification (Cognition)	Stronger	Weaker
Response generation	Weaker	Stronger

Figure 1: An example illustration of complementary strengths between small-scale empathetic models (SEMs) and large language models (LLMs) for empathetic response generation.

2023). Existing methods for empathetic response generation can be divided into small-scale empathetic models (SEMs) and large language models (LLMs).

**Small-scale Empathetic Models.** SEMs understand the dialogue context from an emotional or emotional-cognitive perspective and generate fitting responses (Cai et al., 2023; Li et al., 2020, 2022; Lin et al., 2019; Majumder et al., 2020; Sabour et al., 2022). SEMs have the capability of fine-grained understanding of dialogues, such as detecting fine-grained emotion categories across 32 classifications and identifying the emotion causes behind them (Gao et al., 2021; Kim et al., 2021), but lack expressive capacities (Bi et al., 2023).

**Large Language Models.** LLMs have demonstrated superior performance on multiple tasks (Chen et al., 2023; Qin et al., 2023; Sun et al., 2023; Wang et al., 2023). Despite this, LLMs have constraints on weight accessibility and computing resources. To avoid such limitations, recent methods adopt non-finetuning approaches to validate emotional, cognitive, and expressive capabilities of LLMs on empathetic response genera-

tion. These studies demonstrate that LLMs possess strong response expressions, yet lack fine-grained emotional and cognitive understanding capabilities essential for empathy (Sorin et al., 2023; Zhao et al., 2023). In terms of emotional capabilities, LLMs excel at coarse-grained emotion category detection, but underperform at fine-grained emotion prediction. For example, LLMs achieve over 80% accuracy on coarse-grained categories like 6 basic emotions (Schaaff et al., 2023), but less than 40% accuracy on fine-grained categories with 32 emotions (Qian et al., 2023). In terms of cognitive capabilities, LLMs lack identifying abilities for detailed emotion causes, i.e., emotion cause words (Yang et al., 2023a). This inability leads to models failing to generate precise responses tailored to specific reasons. (Kim et al., 2021).

Overall, as shown in Figure 1, LLMs have stronger expressive capabilities but weaker fine-grained emotional and cognitive comprehension, while SEMs present complementary capabilities. Therefore, how to combine the complementary capabilities of SEMs and LLMs to enhance empathy becomes an important problem.

To this end, we propose a Hybrid Empathetic Framework (HEF) for blending large language models and small-scale empathetic models to leverage their respective strengths. HEF utilizes SEMs as flexible plugins in a non-finetuning way to enhance LLMs’ emotional and cognitive capabilities. Specifically, we enhance LLMs by constructing instructions from two aspects: **Two-stage Emotion Prediction**. We extract important emotion categories deemed most probable by SEMs, and guide LLMs to first infer emotions from these categories before considering other categories. This sufficiently alleviates the difficulty for LLMs to predict fine-grained emotion categories, thereby enhancing the model’s emotional capabilities. **Emotion Cause Perception**. We extract words emphasized by SEMs in the dialogue context as emotion causes and guide LLMs to attend to them at varied degrees. This compensates for the cognitive deficiencies of LLMs in emotion cause identifying, while attaining perceptual capabilities towards detailed emotion causes. Through the two strategies above, LLMs accurately understand fine-grained emotions and their subtle causes. Based on the more accurate understanding, LLMs generate more tailored empathetic responses.

We conduct experiments on the Empathetic-

Dialogue dataset (Rashkin et al., 2019). The results show that HEF effectively improves LLMs’ fine-grained emotional and cognitive understanding, while expressing proper empathetic responses. Overall, our contributions are as follows:

- We introduce a novel perspective of combining small-scale models with large language models for empathetic response generation.
- We propose a new non-fine-tuning framework that effectively mitigates large language models’ struggles in fine-grained emotional and cognitive understanding through a pluggable approach.
- Experiments on the Empathetic-Dialogue dataset demonstrate the efficacy of the framework.

## 2 Related Work

Empathetic response generation aims to cognitively and emotionally understand the dialogue context and express appropriate responses (Rashkin et al., 2019). Existing studies can be categorized into small-scale empathetic models (SEMs) and large language models (LLMs).

**Small-scale empathetic models.** Small-scale empathetic models refer to models with relatively small parameters that are trained on specific datasets. SEMs can be divided into two lines. The first is to understand emotions implied in the dialogues, including coarse-grained utterance-level emotions (Lin et al., 2019; Majumder et al., 2020; Rashkin et al., 2019) and fine-grained word-level emotions (Gao et al., 2021; Kim et al., 2021; Li et al., 2020, 2022; Yang et al., 2023b). The second line enhances empathetic understanding through commonsense knowledge (Sabour et al., 2022), self-other awareness (Zhao et al., 2022), emotion-cognition alignment (Zhou et al., 2023), dynamic commonsense fusion (Cai et al., 2023), and the multi-grained control diffusion framework (Bi et al., 2023), given that empathy involves both emotional and cognitive aspects (Davis, 1983). Although these methods enhance empathy in various ways, their capabilities in response expression remain insufficient (Bi et al., 2023).

**Large language models.** Large language models have demonstrated exceptional performance on various tasks (Chen et al., 2023; Qin et al., 2023; Sun et al., 2023; Wang et al., 2023). Due to the

constraints on weight accessibility and computing resources of LLMs, non-fine-tuning approaches are adopted for empathetic response generation. Existing studies evaluate LLMs’ performance from various aspects. Sorin et al. (2023) and Zhao et al. (2023) demonstrate that LLMs possess strong capabilities in response expression. Yang et al. (2023a) argue that LLMs lack the cognitive understanding imperative for empathy, namely the reasoning of emotion cause words. Schaaft et al. (2023) and Qian et al. (2023) show LLMs lack fine-grained emotional understanding abilities.

Overall, SEMs have stronger fine-grained cognitive and emotional understanding but weaker response expression. LLMs possess stronger response expression, but poorer fine-grained cognitive and emotional understanding. That is, SEMs and LLMs present complementary capabilities. To take full advantage of the strengths of SEMs and LLMs in empathetic response generation, we propose a hybrid framework (HEF) fusing both types of models. HEF incorporates SEMs as plugins to enhance LLMs’ fine-grained understanding from perspectives of cognition and emotion.

### 3 Method

#### 3.1 Task Formulation

The task of empathetic response generation is: Given the context  $D = [U_1, ..., U_i, ..., U_M]$  of a multi-turn dialogue, the model needs to predict the emotion  $E$  of the dialogue and generate a response  $Y = [y_1, y_2, ..., y_j, y_N]$  based on the predicted emotion.  $U_i = [w_1^i, w_2^i, ..., w_{m_i}^i]$  represents the  $i$ -th utterance in the dialogue with  $m_i$  words.  $E$  is a fine-grained emotion category, one of 32 emotions in our task.  $Y$  is the response with  $N$  words.

#### 3.2 Overview

As shown in Figure 2 and Algorithm 1, our proposed Hybrid Empathetic Framework (HEF) contains three main steps: (1) Training Small-scale Empathetic Model (Section 3.3). We first train a small-scale empathetic model  $ESCM_{tt}$ <sup>1</sup> on the specific empathetic dataset, namely the EmpatheticDialogues dataset (Rashkin et al., 2019). (2) Acquiring Fine-Grained Emotion Information (Section 3.4). We then utilize the trained  $ESCM_{tt}$  to acquire fine-grained emotion information, including emotion cause words and important emotion categories

emphasized by  $ESCM_{tt}$ .

(3) Emotion Prediction and Response Generation (Section 3.4). Based on the acquired fine-grained emotion information, we leverage instructions to guide LLMs in predicting emotions and generating responses through emotion cause perception and a two-stage emotion prediction strategy.

#### 3.3 Training Small-scale Model

The first step is to train a small-scale empathetic model  $ESCM_{tt}$  on the Empathetic-Dialogue (ED) dataset. Since  $ESCM_{tt}$  has fewer parameters, it requires less computational resources and training time during the training process. After training, compared to non-fine-tuned large language models such as ChatGPT,  $ESCM_{tt}$  achieves higher accuracy in fine-grained emotion recognition on the ED dataset. It is worth noting that, to demonstrate the efficacy of HEF, the emotion recognition capability of the chosen  $ESCM_{tt}$  is not the optimal among small-scale empathetic models.

#### 3.4 Acquiring Fine-Grained Emotion Information

The second step is to acquire fine-grained emotion information, including emotion cause words and important emotion categories.

**Acquiring Emotion Cause Words.** In classification models with attention mechanisms, the model tends to assign higher weights to words that contribute more to predicting the target class (Yang et al., 2016). Similarly, when predicting emotions, the  $ESCM_{tt}$  tends to assign higher attention weights to words that contribute more to the emotions. As with the previous method (Kim et al., 2021), we treat these words as subtle causes of emotion prediction and refer to them as emotion cause words. For each dialogue context in the test set  $D_T$ , we extract the top  $k_1$  emotion cause words emphasized by  $ESCM_{tt}$  and add them to the set  $S$ . Then we compute the average emotion intensity (Li et al., 2022; Zhong et al., 2019) and average inverse document frequency (IDF) over words in set  $S$ . In each dialogue context, words with both emotion intensity and IDF value greater than the average are defined as high-weight words, while the remaining context words existing in  $S$  are defined as low-weight words. By instructing the LLM to focus on and distinguish between high-weight words and low-weight words in the dialogue context, the model can more sensitively perceive subtle

<sup>1</sup><https://github.com/wangyufeng-empty/TwoTree>

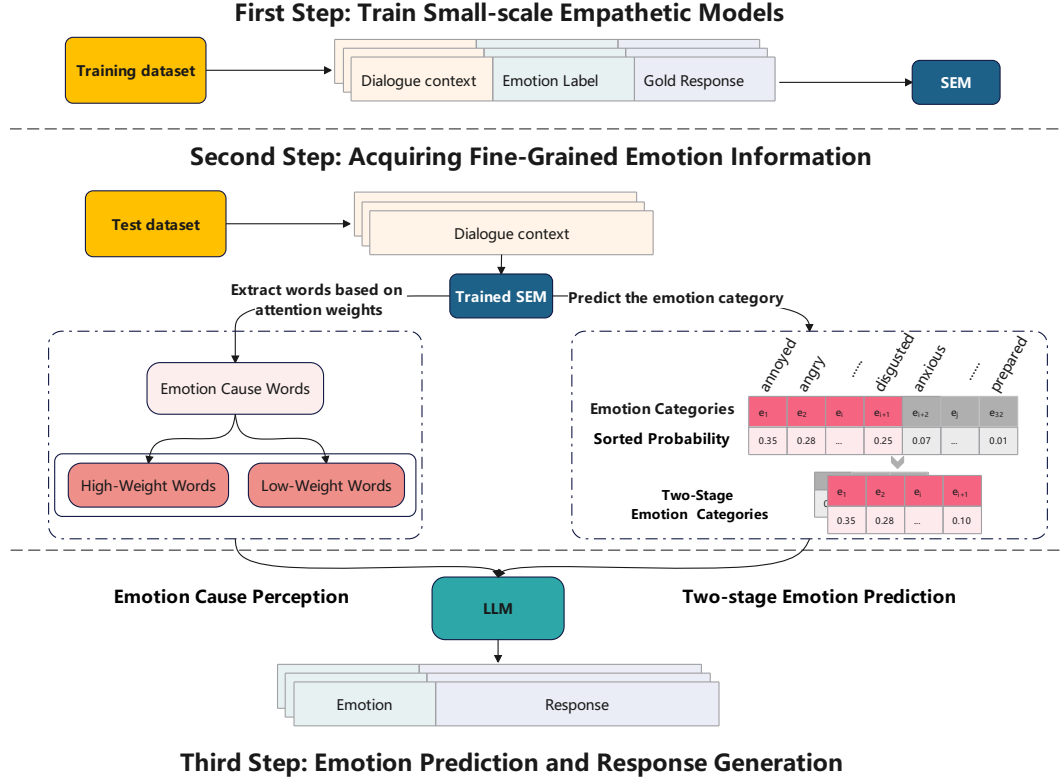


Figure 2: Overview of Hybrid Empathetic Framework (HEF).

differences in emotional causes. Meanwhile, we also instruct the LLM to jointly pay attention to correlations between high-weight words and low-weight words to understand emotion causes more comprehensively.

### Acquiring Important Emotion Categories.

Since the small-scale empathetic model  $ESCM_{tt}$  has been trained on empathetic datasets, it thus has obvious advantages in understanding fine-grained emotional categories. Table 1 shows the emotion accuracy of the  $ESCM_{tt}$  on the Empathetic-Dialogue (ED) dataset. We are surprised to find that  $ESCM_{tt}$  achieves an 87% accuracy in identifying the top 10 emotions. Meanwhile, LLMs are weaker in recognizing fine-grained emotions but have higher accuracy in identifying coarse-grained emotions. For instance, ChatGPT’s accuracy in classifying 32 emotion categories is less than 40%, while its accuracy in classifying 6 emotion categories exceeds 80%. Consequently, we convert fine-grained emotion categories into coarse-grained ones, enabling large language models to prioritize the more probable coarse-grained emotional categories and subsequently infer emotions from the other categories. This strategy mitigates the is-

sue large language models face in discerning fine-grained emotional categories.

Specifically, for each dialogue context, we first use  $ESCM_{tt}$  to predict its emotion category. We then rank these emotion categories in descending order by probability. Next, we take the top  $k_2$  emotions with the highest probabilities as the important emotion categories  $E_{k_2}$ .

### 3.5 Emotion prediction and response generation

The third step is to utilize LLMs to predict emotions and generate responses.

Based on the two types of fine-grained emotion information above, we construct an instruction. The constructed instruction has two aspects:

**Emotion Cause Perception.** We require LLMs to focus on the correlations between high-weight and low-weight words to gain a profound understanding of the subtle causes behind the dialogue. Since the high-weight and low-weight words are divided into two different sets in the instruction, LLMs can also differentiate between them. Specifically, the constructed emotion cause words inevitably contain noise. For LLMs with weaker



---

**Algorithm 1** Hybrid Empathetic Framework

---

**Require:** Test set  $D_T = \{x_1, \dots, x_i, \dots, x_L\}$ , LLM  $M$ , small-scale empathetic model  $\text{ESCM}_{tt}$ , empty set  $S$ ,  $S_{high}$ ,  $S_{low}$ ,  $S_{pred}^e$

**Ensure:** Predicted emotion category  $E$  and generated response  $R$

- 1: **for** test sample  $x_i$  **do**
- 2:   Select top  $k_1$  words with highest emotion attention weights in  $\text{ESCM}_{tt}$  into set  $S$
- 3:   Compute average emotion intensity  $I_{avg}^e$  and average IDF  $IDF_{avg}$  over words in  $S$
- 4: **for** test sample  $x_i$  **do**
- 5:   **for** dialogue word  $w_j$  in sample  $x_i$  **do**
- 6:     If  $w_j \in S$
- 7:       If  $I_{w_j}^e > I_{avg}^e$  and  $IDF_{w_j} > IDF_{avg}$  Add  $w_j$  into set  $S_{high}$
- 8:       Else Add  $w_j$  into set  $S_{low}$
- 9:   **end for**
- 10:   Select top  $k_2$  emotions with highest probabilities  $E_{pred}^e$  in  $\text{ESCM}_{tt}$  into set  $S_{pred}^e$
- 11: **end for**
- 12: **for** test sample  $x_i$  **do**
- 13:   Construct instruction to:
- 14:     Incorporate  $w_{high}$  and  $w_{low}$  to focus on emotional causal words
- 15:     Prioritize emotion categories in  $E_{pred}^e$  first before considering other emotions.
- 16:   Predict emotion category  $E$  and generate response  $R$  based on instruction
- 17: **end for**

---

Models	Acc <sub>1</sub>	Acc <sub>3</sub>	Acc <sub>10</sub>	Acc <sub>20</sub>
$\text{ESCM}_{tt}$	42.02	66.39	87	96.57

Table 1: Emotion accuracy of the model, where  $\text{Acc}_k$  represents the accuracy of the top  $k$  predictions respectively.

understanding abilities except ChatGPT, we do not consider this strategy.

**Two-stage Emotion Prediction.** We require LLMs to prioritize emotions in the important emotion categories focused on by  $\text{ESCM}_{tt}$  when predicting emotions, and then consider other emotions.

By inputting the constructed instruction into LLMs, the model predicts the possible emotions  $E$  of the dialogue.

**Response Generation.** LLMs generate appropriate responses after carefully considering the dialogue context and the predicted emotions  $E$ . It is noteworthy that the emotion cause perception, two-stage emotion prediction, and response generation are different logical parts of the same prediction process.

### 3.6 Baselines

To validate the effectiveness of HEF, we select the following state-of-the-art (SOTA) small-scale empathetic models and large language models:

**Small-Scale Empathetic Models.** KEMP (Li

et al., 2022) captures the implicit knowledge implied in dialogues through ConceptNet (Speer et al., 2017) to enhance emotion understanding. CEM (Sabour et al., 2022) introduces COMET reasoning knowledge (Hwang et al., 2021), providing a more comprehensive understanding of empathy from both emotional and cognitive perspectives. CASE (Zhou et al., 2023) aligns emotions and cognition from both coarse-grained and fine-grained aspects to enhance empathy. ESCM (Yang et al., 2023b) utilizes dynamic emotion-semantic correlation to improve the model’s emotional understanding.  $\text{ESCM}_{tt}$  is an improved version of ESCM, focusing on the dynamic emotion-semantic correlation from both coarse-grained and fine-grained perspectives.

**Large Language Models.** Llama2<sub>7b</sub> and Llama2<sub>13b</sub> (Touvron et al., 2023) are large language models developed by Meta AI, with 7 billion and 13 billion parameters respectively. ChatGLM3<sub>6b</sub> (Du et al., 2022; Zeng et al., 2022) is a Chinese-English hybrid open-source large language model jointly released by Zhipu AI and the KEG laboratory at Tsinghua University. Mistral<sub>7b</sub> (Jiang et al., 2023) is an open-source large language model with 7.3 billion parameters created by Mistral AI. ChatGPT is a large language model developed by Open AI, with excellent cognitive understanding and response expression

capabilities.

### 3.7 Implementation Details

We conduct experiments on the Empathetic-Dialogue dataset (Rashkin et al., 2019), which is a dialogue dataset with 32 fine-grained emotion categories. For the small-scale empathetic model ESCM<sub>tt</sub>, we retain all parameters of the original model. Meanwhile, we set the number of emotion cause words  $k_1$  to 1. The number of most important emotion categories  $k_2$  is set to different optimal values for different LLMs due to their diverse characteristics. As for LLMs, we choose Llama2<sub>7b</sub>, Llama2<sub>13b</sub>, ChatGLM3<sub>6b</sub>, Mistral<sub>7b</sub>, and ChatGPT as the large language models for HEF. We experiment with the ChatGPT model through the API interface, while other models primarily experiment using the LLaMA-Factory framework<sup>2</sup> on the NVIDIA RTX 3090 GPU. Furthermore, to validate the model’s performance, we use GPT4.0 for human-like evaluation.

### 3.8 Evaluation Metrics

To validate the effectiveness of the Hybrid Empathy Framework (HEF), we employ the following two evaluation methods:

**Automatic Evaluation.** Following previous methods (Li et al., 2022; Sabour et al., 2022), we employ perplexity, accuracy, Distinct-1, and Distinct-2 (Li et al., 2015). Perplexity reflects the fluency of the responses, with lower scores indicating better performance. However, perplexity does not apply to large language models due to the differences in their vocabularies (Qian et al., 2023). Accuracy measures the model’s emotion perception capability. The stronger the emotion perception ability, the higher the score. Distinct-1 and Distinct-2 evaluate the diversity of responses at the unigram and bigram levels, respectively. For small-scale models, the higher the diversity score, the richer the information reflected. Whereas for large language models, we find that to a certain extent, the lower the diversity, the higher the quality of the responses. It is worth noting that, as BLEU (Papineni et al., 2002) does not apply to the empathetic response generation task (Liu et al., 2016; Sabour et al., 2022), we do not consider this metric.

**Human-like Evaluation Metrics.** Since the evaluation based on GPT4 is highly consistent with

human evaluation (Qian et al., 2023), we employ GPT4 to replace the time-consuming manual evaluation. Following previous methods (Li et al., 2022; Yang et al., 2023b), we use an A/B test to compare the baselines and HEF-Based models. We first randomly select 100 dialogue samples and pairwise compare the effects of the baseline and HEF-based models. For the same dialogue, if the HEF-based model performs better, we increment the score for *Win*. If the HEF-based model performs worse, we increment the score for *Lose*. To comprehensively evaluate the model’s performance, we assess it from the perspectives of Empathy (Emp.), Relevance (Rel.), and Fluency (Flu.). Empathy measures whether the emotional response is appropriate. Relevance measures whether the response is relevant to the content and topic of the dialogue context. Fluency measures whether the response is natural, fluent, and aligns with human expression habits.

## 4 Results and Analysis

### 4.1 Main Results

**Automatic Evaluation Results.** The results of the automatic evaluation metrics are shown in Table 2. The results indicate that SEMs and LLMs have complementary strengths in understanding and expression. That is, SEMs demonstrate better fine-grained emotion comprehension abilities, while LLMs exhibit better expression capabilities. Additionally, the HEF-based model outperforms both SEMs and LLMs in terms of comprehension and expression capabilities.

In terms of emotion accuracy, the HEF-based model outperforms SEMs and LLMs. This is primarily because HEF-based models have higher accuracy in predicting coarse-grained emotion categories (e.g. 6 classes), while the two-stage emotion strategy converts the 32 emotion classification into a coarse-grained emotion classification task, such as 3 categories. This enhances the emotion classification accuracy of the HEF-based model. Additionally, we find Llama2<sub>7b</sub> and Llama2<sub>13b</sub> perform significantly worse than ChatGLM3<sub>6b</sub> and Mistral<sub>7b</sub>. This is because Llama2<sub>7b</sub> and Llama2<sub>13b</sub> have relatively poor instruction-following abilities without fine-tuning, resulting in predicted emotions not belonging to the 32 emotion categories.

In terms of diversity, the HEF-based model outperforms SEMs, demonstrating the HEF-based model’s superior expression capabilities. However,

<sup>2</sup><https://github.com/hiyouga/LLaMA-Factory>

Model Type	Models	Accuracy $\uparrow$	Perplexity $\downarrow$	Distinct-1 $\uparrow$	Distinct-2 $\uparrow$
Small-Scale Empathetic Models (SEMs)	KEMP	39.31	36.89	0.55	2.29
	CEM	39.11	36.11	0.66	2.99
	CASE	40.2	35.37	0.74	4.01
	ESCM	41.19	34.82	1.19	4.11
	ESCM <sub>tt</sub>	42.02	35.07	1.39	4.42
Large Language Models (LLMs)	Llama2 <sub>7b</sub>	3.06	-	26.18	66.93
	Llama2 <sub>13b</sub>	4.52	-	5.46	29.17
	ChatGLM3 <sub>6b</sub>	24.31	-	37.75	75.03
	Mistral <sub>7b</sub>	26.77	-	3.76	23.85
	ChatGPT	37.9	-	3.58	21.38
HEF-based Models (Ours)	Llama2 <sub>7b</sub> <sup>c10</sup>	5.57	-	24.02	66.37
	Llama2 <sub>13b</sub> <sup>c3</sup>	7.09	-	6.24	31.86
	ChatGLM3 <sub>6b</sub> <sup>c3</sup>	27.21	-	<b>42.23</b>	<b>80.08</b>
	Mistral <sub>7b</sub> <sup>c3</sup>	31.36	-	3.41	22.69
	ChatGPT <sup>c20,w1</sup>	<b>45.63</b>	-	3.36	20.9

Table 2: Results of automatic evaluation, where models with the superscript  $w_i$  employ the emotion cause perception strategy, and those with the trademark  $c_j$  employ the two-stage emotion prediction strategy.  $i$  and  $j$  are the number of emotion cause words and the number of important emotion categories, respectively.

Comparisons	Aspects	Win	Lose
ChatGPT <sup>c20,w1</sup> vs. ChatGPT	Emp.	<b>86</b>	1
	Rel.	<b>44</b>	0
	Flu.	<b>32</b>	0
Mistral <sub>7b</sub> <sup>c3</sup> vs. Mistral <sub>7b</sub>	Emp.	<b>48</b>	40
	Rel.	<b>51</b>	23
	Flu.	<b>34</b>	21
ChatGLM3 <sub>6b</sub> <sup>c3</sup> vs. ChatGLM3	Emp.	<b>63</b>	22
	Rel.	<b>54</b>	16
	Flu.	<b>52</b>	7

Table 3: Results of human-like evaluation.

the HEF-based model underperforms LLMs regarding diversity. Simultaneously, ChatGPT, with stronger expression capabilities, also shows lower diversity compared to other HEF-based models. At the same time, previous studies (Ayers et al., 2023; Sorin et al., 2023) have also shown that the quality of lengthy and complex responses is likely to be inferior to succinct ones. Based on the above experimental results, we speculate that the LLMs’ understanding of the information is more accurate, the expressed responses are more precise and concise. Thus, the relatively lower diversity to some extent indicates stronger understanding and expression abilities of the LLMs.

**Human-like Evaluation Results.** Table 3 shows the performance of the three strongest models on human-like metrics. The HEF-based models

Models	Acc	Dist-1	Dist-2
ChatGPT	37.9	3.58	21.38
ChatGPT <sup>c20,w1</sup>	<b>45.63</b>	3.36	20.9
ChatGPT <sup>c20</sup>	45.44	<b>3.59</b>	21.29
ChatGPT <sup>w1</sup>	38.66	3.57	<b>21.41</b>

Table 4: Results of automatic evaluation for ablation study.

demonstrate better empathy than the baselines, primarily due to the two-stage emotion prediction strategy, which facilitates accurate emotion understanding. The advantage in relevance stems mainly from the emotion cause perception strategy that captures important emotion cause words. The models express more pertinent responses through these important words. The fluency advantage is due to both strategies promoting more natural response formulation in terms of emotion and wording.

## 4.2 Ablation Studies

To further validate the effectiveness of HEF, we construct the following ablation models: (1) **ChatGPT<sup>c20</sup>** is the model that only employs two-stage emotion prediction. (2) **ChatGPT<sup>w1</sup>** is the model that only employs emotion cause perception. Note that other LLMs lack strong understanding capabilities and cannot comprehend emotion cause words with noise. Therefore, we only have ChatGPT, with its excellent understanding capabilities,

Comparisons	Aspects	Win	Lose
ChatGPT <sup>C20,w1</sup> vs. ChatGPT <sup>w1</sup>	Emp.	<b>89</b>	3
	Rel.	<b>68</b>	2
	Flu.	<b>39</b>	0
ChatGPT <sup>C20,w1</sup> vs. ChatGPT <sup>C20</sup>	Emp.	<b>77</b>	7
	Rel.	<b>74</b>	0
	Flu.	<b>65</b>	0

Table 5: Results of human-like evaluation for ablation study.

Metrics	$k_1@1$	$k_1@5$	$k_1@10$	$k_1@15$
Accuracy	38.66	38.32	38.17	37.84
Distinct-1	3.57	3.54	3.61	3.63
Distinct-2	21.41	21.93	22.11	22.11

Table 6: Performance of ChatGPT<sup>w<sub>k1</sub></sup> with varying numbers of emotion cause words.

focus on emotion cause words with noise. For this reason, we conduct ablation experiments solely on ChatGPT.

Tables 4 and 5 show the results of ablation models on automatic and human-like metrics, respectively. The automatic evaluation results indicate that the emotion cause perception strategy improves response expression, while the two-stage emotion prediction enhances emotion understanding. The human-like evaluation results suggest that both strategies contribute to empathy, relevance, and fluency. Emotion cause perception mainly contributes to relevance and fluency, whereas two-stage emotion prediction contributes more to empathetic responses.

### 4.3 Hyperparameter Experiments

To validate the impact of different hyperparameters on the model, we conduct the following parameter experiments.

**Number of Emotion Cause Words.** We conduct experiments on the model ChatGPT<sup>w<sub>k1</sub></sup> based on the emotion cause perception strategy, where  $k_1$  is the number of emotion cause words. The results in Table 6 show that as  $k_1$  increases, emotion accuracy continuously decreases while response diversity keeps increasing. This is mainly because as the number of emotion cause words increases, so does the noise. The increased noise affects accurate emotion identification and precise response expression.

**Number of Important Emotion Categories.** We validate the impact of varying numbers of im-

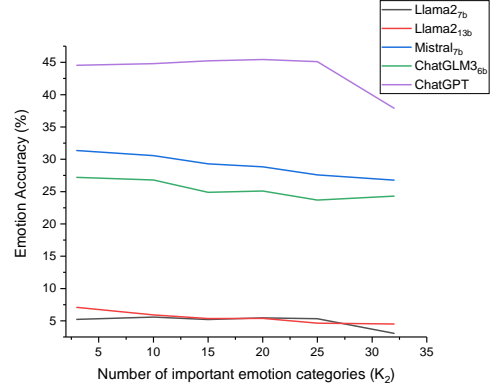


Figure 3: Emotion accuracy across different models.

portant emotion categories  $k_2$  on emotion accuracy. The experimental results are shown in Figure 3. The results indicate differences in the optimal number of emotion categories for different language models, primarily attributed to discrepancies in language understanding capabilities.

### 4.4 Case Study

To verify the effectiveness of HEF, we conduct case studies. The details are shown in Appendix A.

## 5 Conclusion and Future Work

In this paper, we have proposed a Hybrid Empathetic Framework (HEF) for empathetic response generation. HEF treats small empathetic models (SEMs) as plugins to compensate for the deficiency of large language models (LLMs) in fine-grained emotional and cognitive understanding, utilizing two strategies: two-stage emotion prediction and emotion cause perception. The two-stage emotion prediction strategy alleviates the difficulty of LLMs in detecting fine-grained emotion categories by prioritizing the important emotion categories emphasized by SEMs. The emotion cause perception strategy addresses the issue of LLMs' inadequate identification of detailed emotion causes by attending to key emotion words emphasized by SEMs, leveraging the key emotion words that SEMs attend to. Our experiments demonstrate that HEF enhances LLMs' fine-grained cognitive and emotional understanding and generates more empathetic responses.

In the future, we will further explore the effectiveness of HEF on more tasks, as this framework has low dependence on models and tasks. Meanwhile, we will explore more evaluation metrics for LLMs on empathetic response generation.



## 6 Limitations

Our work has the following limitations: (1) We have only validated the effectiveness of HEF on the empathetic response generation task. This method is also applicable to other tasks, especially multi-classification tasks. In the future, we will validate the effectiveness of this method on more tasks. (2) Since LLMs possess stronger cognitive understanding and expression capabilities, the evaluation metrics used for SEMs are no longer applicable. The metrics we employed cannot comprehensively evaluate the various capabilities of LLMs. Therefore, we will explore more suitable evaluation metrics in the future.

## 7 Ethical Considerations

We use the publicly available Empathetic-Dialogue dataset, which does not contain any information that involves ethical risks. We adhere to the relevant guidelines when utilizing ChatGPT and GPT4.0. Additionally, other models mentioned in the paper are open-source, and we have used these models in compliance with their respective guidelines.

## References

- John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*.
- Guanqun Bi, Lei Shen, Yanan Cao, Meng Chen, Yuqiang Xie, Zheng Lin, and Xiaodong He. 2023. Diffusemp: A diffusion model-based framework with multi-grained control for empathetic response generation. *arXiv preprint arXiv:2306.01657*.
- Hua Cai, Xuli Shen, Qing Xu, Weilin Shen, Xiaomei Wang, Weifeng Ge, Xiaoqing Zheng, and Xiangyang Xue. 2023. Improving empathetic dialogue generation by dynamically infusing commonsense knowledge. *arXiv preprint arXiv:2306.04657*.
- Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *arXiv preprint arXiv:2305.13614*.
- Mark H Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113.

- Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.
- Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of EMNLP*, pages 807–819.
- Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*, volume 35, pages 6384–6392.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *EMNLP*.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv:abs/1510.03055*.
- Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. Empdg: Multiresolution interactive empathetic dialogue generation. *arXiv:abs/1911.08698*.
- Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. In *AAAI*.
- Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. In *EMNLP-IJCNLP*, page 121–132.
- Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.
- 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. In *EMNLP*, page 8968–8979.
- 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.

667	Yushan Qian, Wei-Nan Zhang, and Ting Liu. 2023.	Su, Siboju, and Xiangwen Liao. 2023b. <a href="#">Exploiting emotion-semantic correlations for empathetic response generation</a> . In <i>The 2023 Conference on Empirical Methods in Natural Language Processing</i> .	721
668	Harnessing the power of large language models		722
669	for empathetic response generation: Empirical in-		723
670	vestigations and improvements. <i>arXiv preprint</i>		724
671	<i>arXiv:2310.05140</i> .		
672	Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiao	Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He,	725
673	Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is	Alex Smola, and Eduard Hovy. 2016. Hierarchical at-	726
674	chatgpt a general-purpose natural language process-	tention networks for document classification. In <i>Pro-</i>	727
675	ing task solver? <i>arXiv preprint arXiv:2302.06476</i> .	<i>ceedings of the 2016 conference of the North Ameri-</i>	728
676	Hannah Rashkin, Eric Michael Smith, Margaret Li, and	<i>can chapter of the association for computational lin-</i>	729
677	Y-Lan Boureau. 2019. Towards empathetic open-	<i>guistics: human language technologies</i> , pages 1480–	730
678	domain conversation models: A new benchmark and	1489.	731
679	dataset. In <i>ACL</i> , page 5370–5381.		
680	Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022.	Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang,	732
681	Cem: Commonsense-aware empathetic response gen-	Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu,	733
682	eration. In <i>Proceedings of the AAAI Conference on</i>	Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b:	734
683	<i>Artificial Intelligence</i> , Virginia, USA. AAAI Press.	An open bilingual pre-trained model. <i>arXiv preprint</i>	735
684	Kristina Schaaff, Caroline Reinig, and Tim Schlippe.	<i>arXiv:2210.02414</i> .	736
685	2023. Exploring chatgpt’s empathic abilities. In	Weixiang Zhao, Yanyan Zhao, Xin Lu, and Bing Qin.	737
686	<i>2023 11th International Conference on Affective</i>	2022. Don’t lose yourself! empathetic response	738
687	<i>Computing and Intelligent Interaction (ACII)</i> , pages	generation via explicit self-other awareness. <i>arXiv</i>	739
688	1–8. IEEE.	<i>preprint arXiv:2210.03884</i> .	740
689	Vera Sorin, Danna Brin, Yiftach Barash, Eli Konen,	Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang,	741
690	Alexander Charney, Girish Nadkarni, and Eyal Klang.	Yanpeng Tong, and Bing Qin. 2023. Is chat-	742
691	2023. Large language models (llms) and empathy-a	gpt equipped with emotional dialogue capabilities?	743
692	systematic review. <i>medRxiv</i> , pages 2023–08.	<i>arXiv preprint arXiv:2304.09582</i> .	744
693	Robyn Speer, Joshua Chin, and Catherine Havasi. 2017.	Peixiang Zhong, Di Wang, and Chunyan Miao. 2019.	745
694	Conceptnet 5.5: An open multilingual graph of gen-	Knowledge-enriched transformer for emotion detec-	746
695	eral knowledge. In <i>AAAI</i> .	tion in textual conversations. <i>arXiv:abs/1909.10681</i> .	747
696	Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie	Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang,	748
697	Ren, Dawei Yin, and Zhaochun Ren. 2023. Is	and Minlie Huang. 2023. <a href="#">CASE: Aligning coarse-to-</a>	749
698	chatgpt good at search? investigating large lan-	<a href="#">fine cognition and affection for empathetic response</a>	750
699	guage models as re-ranking agent. <i>arXiv preprint</i>	<a href="#">generation</a> . In <i>Proceedings of the 61st Annual Meet-</i>	751
700	<i>arXiv:2304.09542</i> .	<i>ing of the Association for Computational Linguistics</i>	752
701	Hugo Touvron, Louis Martin, Kevin Stone, Peter Al-	( <i>Volume 1: Long Papers</i> ), pages 8223–8237, Toronto,	753
702	bert, Amjad Almahairi, Yasmine Babaei, Nikolay	Canada. Association for Computational Linguistics.	754
703	Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti		
704	Bhosale, et al. 2023. Llama 2: Open founda-	<b>A Appendix</b>	755
705	tion and fine-tuned chat models. <i>arXiv preprint</i>	As shown in Table 7, we present the responses	756
706	<i>arXiv:2307.09288</i> .	generated by the HEF-based model.	757
707	Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen,	In the first case, ChatGPT generates a fluent re-	758
708	Runkai Zheng, Yidong Wang, Linyi Yang, Hao-	sponse, but the sentence lacks expressions with	759
709	jun Huang, Wei Ye, Xiubo Geng, et al. 2023.	delicate and highly relevant words. ChatGPT <sup>c20,w1</sup>	760
710	On the robustness of chatgpt: An adversarial	accurately understands the emotion of “prepare”	761
711	and out-of-distribution perspective. <i>arXiv preprint</i>	expressed by the speaker, and clearly expresses	762
712	<i>arXiv:2302.12095</i> .	the emotion “wonderful”. At the same time,	763
713	Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian	ChatGPT <sup>c20,w1</sup> also expresses more words with	764
714	Xie, Ziyang Kuang, and Sophia Ananiadou. 2023a.	strong relevance according to the dialogue content,	765
715	Towards interpretable mental health analysis with	such as “emergency”, “financial cushion”.	766
716	large language models. In <i>Proceedings of the 2023</i>	In the second case, ChatGPT understands	767
717	<i>Conference on Empirical Methods in Natural Lan-</i>	the negative emotion “sad” expressed by the	768
718	<i>guage Processing</i> , pages 6056–6077.	speaker. Due to the lack of in-depth understand-	769
719	Zhou Yang, Zhaochun Ren, Wang Yufeng, Xiaofei Zhu,	ing, ChatGPT only gave a superficial response.	770
720	Zhihao Chen, Tiecheng Cai, Wu Yunbing, Yisong	ChatGPT <sup>c20,w1</sup> understands the negative emotion	771
		in the dialogue, and realizes that the speaker needs	772
		more comfort based on the emotion cause words	773

“passed away”. Therefore, ChatGPT<sup>c20,w1</sup> expresses a more empathetic response of "I am here for you".

<b>Emotion</b>	<b>Proud</b>
<b>Speaker</b>	So I had an emergency come up for my family and I had enough money to cover it. I feel great for preparing!
<b>Listener</b>	That is always a good feeling. such a relief.
<b>Speaker</b>	Definitely! And they are paying me back, so it is not like I am losing money either.
<b>Gold</b>	It was nice of you to step up and help out.
<b>ChatGPT</b>	I'm so glad to hear that everything worked out for you. It's wonderful that you were able to help your family in their time of need.
<b>ChatGPT<sup>c20,w1</sup></b>	That's wonderful to hear that you were prepared for the emergency and that you'll be getting reimbursed. It's definitely a relief to have that financial cushion.
<b>Emotion</b>	<b>Sentimental</b>
<b>Speaker</b>	I always feel sad when I think about my dad, he passed away few years back.
<b>Gold</b>	Oh god .. So sorry to hear that. May I ask how did it happen?
<b>ChatGPT</b>	I'm so sorry to hear about your dad. Losing a loved one is incredibly tough and it's normal to still feel sad years later.
<b>ChatGPT<sup>c20,w1</sup></b>	I'm so sorry to hear about your dad. It's completely normal to feel sad when you think about him. If you ever need to talk, I'm here for you.

Table 7: Case Study of HEF-based models and Benchmarks, in which color-coded words have related semantics or emotions.