EmpCRL: Controllable Empathetic Response Generation via In-Context Commonsense Reasoning and Reinforcement Learning

Mingxiu Cai, Daling Wang*, Shi Feng, Yifei Zhang
Northeastern University, Shenyang, China
2201760@stu.neu.edu.cn, {wangdaling,fengshi,zhangyifei}@cse.neu.edu.cn

Abstract

Empathetic response generation aims to understand the user's feelings emotionally and generate responses with appropriate emotion. According to psychological theories, empathy consists of two main aspects: affection and cognition. However, existing works lack the perception of fine-grained dialogue emotion propagation, as well as have limitations in reasoning about the intentions of users on cognition, which affect the quality of empathetic response. To this end, we propose to generate **Emp**athetic response based on in-context **C**ommonsense reasoning and **R**einforcement **L**earning (**EmpCRL**). First, we use a current popular large language model combined with multi-view contextual reasoning to broaden the cognitive boundaries through in-context learning. Furthermore, we infer the response emotion by jointly modeling the dialogue history and emotion flow, and achieve the control of response emotion and diversity through reinforcement learning. Extensive experiments on Empathetic-Dialogues dataset show that our model outperforms state-of-the-art models in both automatic and human evaluation.

Keywords: Empathetic Response Generation, Commonsense Reasoning, Reinforcement Learning

1. Introduction

With the rapid development of human-machine dialogue systems, empathetic dialogue plays an important role in social interaction and user experience (Hoffman, 2001). The empathetic response generation task aims to enable machines to adaptively generate responses with consistent emotional expressions based on the user's emotional state and dialogue history, thus enhancing empathetic connections with users. Early empathetic dialogue generation research focused on sensing users' emotions and incorporating them into responses, such as mixture of empathetic listeners (Lin et al., 2019), mimicking emotions for empathetic response generation (Majumder et al., 2020), generating empathetic responses with human-like intents (Chen et al., 2022). Recent works have included commonsense reasoning as an important factor, such as leveraging commonsense to draw more information (Sabour et al., 2022), sensitive emotion recognizing and sensible knowledge selecting (Wang et al., 2022), aligning coarse-to-fine cognition and affection (Zhou et al., 2023).

In psychological research, two important elements of empathy are affection and cognition (Westbrook et al., 2011). Affection is primarily concerned with the understanding and expression of emotions, while cognition is primarily concerned with understanding and reasoning about the thoughts and intentions of others (Keskin, 2014). Some recent work has modeled empathy in terms of cognition and affection. However, empathetic dialogue involves complex processes of human cognition and affection in action (Davis, 1983), and existing approaches fail to adequately incorporate these factors into their models. On one hand, previous models usually only consider the current utterance in commonsense reasoning and ignore contextual information (Ghosal et al., 2022), while the weak reasoning ability of the model leads to limited information richness and truthfulness of the responses. On the other hand,

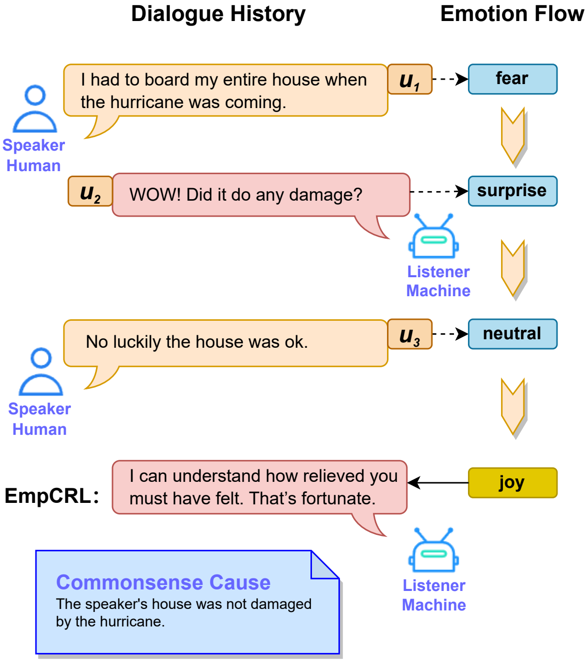
Dialogue History
Emotion Flow
I had to board my entire house when the hurricane was coming.
u_1
fear
Speaker Human
u_2
WOW! Did it do any damage?
surprise
Listener Machine
No luckily the house was ok.
u_3
neutral
Speaker Human
EmpCRL:
I can understand how relieved you must have felt. That's fortunate.
joy
Commonsense Cause
The speaker's house was not damaged by the hurricane.
Listener Machine

Figure 1: An example from the EmpatheticDialogues dataset.

* Corresponding author.

LREC-COLING 2024, pages 5734–5746
20-25 May, 2024.

emotion inference and emotion expression play key roles in empathetic response generation (Li et al., 2021), but existing studies tend to ignore the fine-grained modeling of dialogue emotions as well as the contagious nature of emotion propagation in dialogues, leading to the limited emotion expression ability of the model in empathetic response generation. Figure 1 shows an example of a real empathetic dialogue. Here we define the emotional process as **emotion flow**. By generating commonsense cause through in-context commonsense reasoning, while modeling fine-grained (utterance-level emotions) dialogue emotion flow, we can perceive the speaker's epiphenomenal and implicit emotional states, followed by controlled emotion response generation from the emotion signals obtained from emotion foresight.

To achieve the above objective, we propose to generate **Emp**athetic response based on in-context **C**ommonsense reasoning and **R**einforcement **L**earning (**EmpCRL**). First we construct in-context examples for multi-view contextual reasoning and use ChatGPT[1] to reason about commonsense cause sentences. Then we predict the emotions required for responses by jointly modeling dialogue history and fine-grained emotion flow. Meanwhile, we train the activation model by reinforcement learning approach to achieve accurate control of the emotions. Finally, using the predicted emotions as control signals, commonsense cause sentences are injected into the decoder along with the dialogue history to generate empathetic responses from both cognitive and affective perspectives. Extensive experimental results on a widely used benchmark EmpatheticDialogues dataset show that EmpCRL outperforms strong baselines on both automatic and human evaluation metrics. Our contributions are summarized as follows:

- We leverage the reasoning ability of ChatGPT to generate commonsense cause sentences through in-context learning and multi-view contextual reasoning to enhance the cognitive capabilities of the dialogue model.
- We model the dialogue emotion flow to infer the future emotion and use reinforcement learning to enable controlled emotion response generation.
- Automatic and human evaluation on the EmpatheticDialogues dataset shows that our proposed model EmpCRL outperforms strong baselines and is capable of generating more reasonable and diverse empathetic responses.

2. Related Work

2.1. Empathetic Dialogue Generation

The empathetic dialogue generation task aims to enable systems to generate responses that resonate with the emotions of human users, establishing an emotional connection with them. Since the release of the EmpatheticDialogues dataset by Rashkin et al. (2019), the research in recent years has proposed a variety of approaches to this task. Lin et al. (2019) use mixture of empathetic listeners to generate empathetic responses. Majumder et al. (2022) use exemplars to cue the generative model on fine stylistic properties that signal empathy to the interlocutor. More recently, Zhao et al. (2023) propose EmpSOA to generate empathetic responses via explicit self-other awareness.

In order to improve the performance of empathetic dialogue generation, several studies have better modeled the emotional state of generated responses through fine-grained emotion perception. These models make generated responses more consistent with the predefined emotions by taking emotion labels as additional inputs or post-processing the generated responses through emotion classifiers. Li et al. (2020) exploit both coarse-grained dialogue-level emotions and fine-grained token-level emotions. Wang et al. (2022) propose a sequence coding and emotion knowledge interaction approach for empathetic dialogue generation.

However, these approaches still lack the consideration of contagious nature of dialogue emotions as well as future emotional states.

2.2. Controllable Text Generation

Controlled text generation is an important technique aimed at enabling control over the attributes and features of generated text. Some of the previous studies control the generated attributes of text by post-processing. Pascual et al. (2021) combine pre-trained language models (PLMs) with simple attribute classifiers to guide text generation without further training. Krause et al. (2021) use a generative discriminator to instantly classify candidate next tokens in the inference process.

In addition, some studies have used reinforcement learning to achieve fine-grained control of generated text by training an intelligent agent to control decisions during the generation process. Ziegler et al. (2019) use reinforcement learning techniques to fine-tune PLMs with reward models trained on human preferences. Wu et al. (2023) propose a framework that allows LMs to learn from multiple fine-grained reward models trained on human feedback. Reinforcement learning methods can achieve better control of generated text

[1] https://chat.openai.com/

Utterance-level Emotion Label	Dialogue-level Emotion Label
Anger	Angry, Annoyed, Furious
Fear	Afraid, Terrified, Anxious, Apprehensive
Joy	Excited, Proud, Grateful, Hopeful, Confident, Joyful, Content, Prepared, Anticipating
Sadness	Sad, Lonely, Guilty, Disappointed, Devastated, Embarrassed, Ashamed
Surprise	Surprised, Impressed
Disgust	Disgusted, Jealous
Neutral	Caring, Sentimental, Trusting, Faithful, Nostalgic

Table 1: The mapping of utterance-level emotions and dialogue-level emotions.

attributes than post-processing methods (Zhang et al., 2022).

In order to achieve controllable response emotions, we guide the generation process by setting empathy reward signals so that the generated responses are more closely aligned with the specified emotions.

2.3. Dialogue Commonsense Reasoning

Dialogue commonsense reasoning is an important aspect of empathetic dialogue generation, aiming to introduce commonsense knowledge to enhance the rationality of generated responses. In past studies, many approaches have used external knowledge bases such as ConceptNet (Speer et al., 2017) and ATOMIC (Sap et al., 2019) to enrich the commonsense knowledge of dialogue models. Li et al. (2022) proposed to leverage external knowledge to explicitly understand and express emotions in empathetic dialogue generation. Sabour et al. (2022) utilized common knowledge to obtain more information about the user's situation and used additional information to further enhance empathetic expression in the generated responses. Zhou et al. (2023) designed a two-level strategy to align coarse-grained and fine-grained cognition and affection for responding empathetically.

However, existing approaches tend to perform commonsense reasoning at the sentence level, failing to incorporate dialogue contextual information well. Meanwhile, due to the limitation of model size, the truthfulness and reasonableness of the reasoning results will be affected to some extent. Therefore, we use the large language model ChatGPT combined with multi-view contextual reasoning to broaden the cognitive boundaries through in-context learning.

3. Methodology

The overview of our proposed model EmpCRL is shown in Figure 2. It consists of four modules: commonsense reasoner, emotion perceiver, empathy driver and response generator. The commonsense reasoner uses ChatGPT to generate commonsense cause sentences through in-context learning and multi-view contextual reasoning. The emotion perceiver infers future emotions by jointly modeling dialogue history and emotion flow. The empathy driver enables control of response emotions and diversity through reinforcement learning. The response generator integrates the information gained from the above three modules and generates appropriate empathetic responses.

3.1. Task Formulation

We define the task of empathetic response generation as follows. The dialogue history is a set of alternating utterance sequences $D = [u_1, u_2, ..., u_{n-1}]$, where $u_i = [w_1, w_2, ..., w_m]$ represents that the i-th utterance consists of m words. We map the 32 classes of emotions at the dialogue-level of the original dataset into 7 classes of high-frequency emotions, including anger, disgust, fear, joy, neutral, sadness and surprise. The specific mapping relationships are in Table 1. Then we label the dataset with fine-grained emotions at the utterance-level. Our goal is to generate empathetic response Y based on dialogue history with commonsense reasoning and emotion inference.

3.2. Commonsense Reasoner

The commonsense reasoner provides the basis for subsequent response generation by modeling commonsense cause information in the context

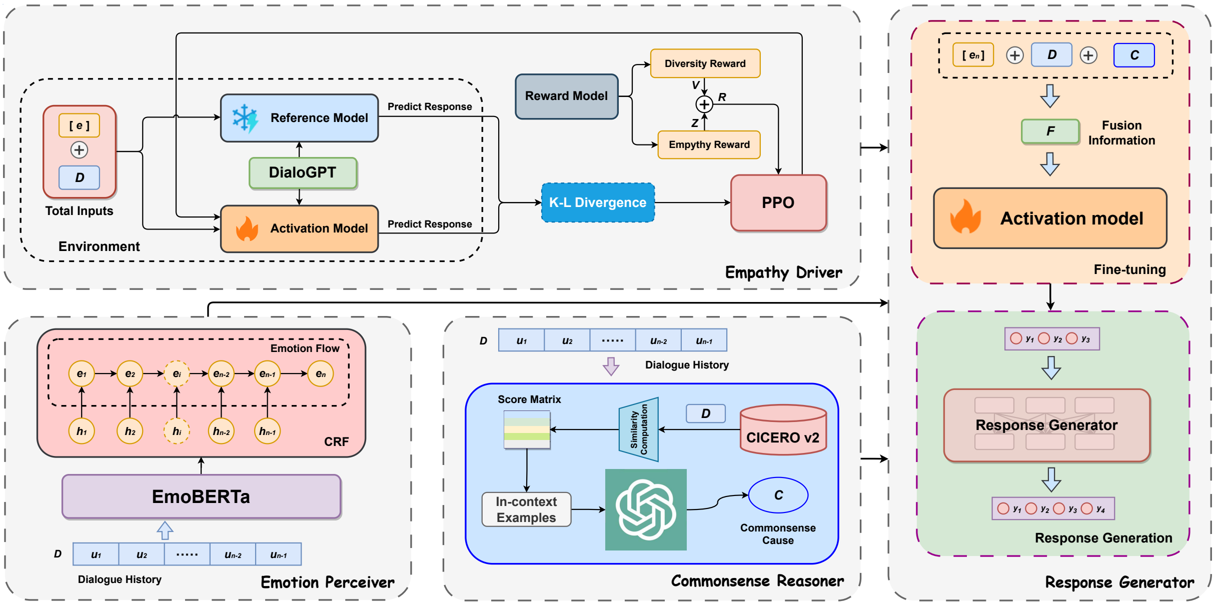

Figure 2: The overall architecture of our proposed EmpCRL model.

of the dialogue. First we select the CICERO v2 (Shen et al., 2022) dataset as the in-context example repository. CICERO v2 is a contextual reasoning dataset that performs reasoning from multi-view in order to determine commonsense explanations around events in binary dialogues. Based on the current dialogue context D, we select K example dialogues in the example repository that are most similar to it as in-context examples (Yang et al., 2022). Assuming that the example repository contains r example dialogues, and each example dialogue consists of l sentences, we use $P = \{p_1, p_2, ..., p_r\}$ to denote the example repository, where $p_i = \{p_{i_1}, p_{i_2}, ..., p_{i_l}\}$ denotes the i-th dialogue example and $1 \leq i \leq r$.

To compute the semantic similarity score between the current dialogue and the in-context example dialogue, we use the sentence encoder[2]. First, we vectorize the representation of the current dialogue and example dialogue by using the sentence encoder to get the representation vector of the dialogue context $D_{emb} = \{u_{emb_1}, u_{emb_2}, ..., u_{emb_{n-1}}\}$ and the representation vector of the example dialogue $P_{i,emb} = \{p_{i,emb_1}, p_{i,emb_2}, ..., p_{i,emb_l}\}$. Then, we compute the semantic similarity scores between each sentence of the current dialogue and the example dialogue to obtain the score matrix S, where $S \in R^{(n-1)\times l}$ and S_{ij} denotes the score between the i-th sentence of the dialogue context and the j-th sentence of the example dialogue:

$$S_{ij} = Sim(u_{emb_i}, p_{i,emb_j}) \quad (1)$$

where Sim is the semantic similarity score function. Based on the score matrix S, we select K example dialogues that are most similar to the dialogue context as in-context examples I. We sort the score matrix and select the top-K example dialogues:

$$I_i = TopK(S, K), \forall i \in [1, N] \quad (2)$$

where i denotes the index of the K example dialogues that are most similar to the dialogue context. Based on previous work (Han et al., 2023), we take N to be 4. We then input in-context examples I into the ChatGPT model with specific templates to perform commonsense reasoning and generate commonsense cause sentences. The specific prompts are in Appendix A. ChatGPT is a powerful large language model (LLM) that can generate coherent texts based on the input contents. Then the ChatGPT model generates commonsense cause sentence C:

$$C = ChatGPT(I) \quad (3)$$

The commonsense cause sentences generated by ChatGPT are used as the cognitive bases for empathetic response generation.

3.3. Emotion Perceiver

The emotion perceiver captures the sequential nature of emotions based on fine-grained utterance-level emotion classification and dialogue history modeling to enable inference of the future emotion. First, we use the pre-trained model EmoBERTa (Kim and Vossen, 2021) to encode each utterance u_i into its corresponding sentence feature vector representation h_i. EmoBERTa is a utterance-level

[2] https://huggingface.co/cross-encoder/stsb-roberta-large

emotion classifier built on multiple dialogue emotion classification datasets based on the RoBERTa model, which is able to capture the emotional details of the dialogues. We define the vector representation of the dialogue history D as $H = \{h_1, h_2, ..., h_{n-1}\}$. The purpose is to synthesize the emotion information in the dialogue history so that emotion inference can utilize the context better.

To perform inference on emotion sequences, we design a conditional random field (CRF) model based on previous work (Song et al., 2022). We denote the emotion sequence as $E = \{e_1, e_2, ..., e_{n-1}\}$, where e_i represents the emotion label corresponding to the i-th utterance. In the CRF model, we define the emotion transfer feature function $f_{tran}(e_{i-1}, e_i)$, which is used to characterize the transfer from the previous emotion e_{i-1} to the current emotion e_i. In addition, we define the dialogue utterance feature function $f_{join}(e_i, h_i)$, which jointly models the dialogue utterance representation vector h_i with the current emotion label e_i to capture the relationship between the dialogue utterance and the current emotion. The score $O(E|H)$ for each emotion label sequence E is defined as:

$$O(E|H) = \sum_{i=1}^{n-1} \left(\sum_{f_{join}} w_{join} \cdot f_{join}(e_i, h_i) \right) + \sum_{i=2}^{n-1} w_{tran} \cdot f_{tran}(e_{i-1}, e_i) \quad (4)$$

where w_{join} and w_{tran} are weights of the corresponding feature functions. By calculating the score $O(E|H)$ for all possible emotion sequences E, we infer the most likely future emotion e_n:

$$e_n = arg \underset{E}{max}\, O(E|H) \quad (5)$$

Finally, based on the future emotion e_n obtained by inference, we use it as a signal for future emotion supervision to generate responses.

3.4. Empathy Driver

The empathy driver aims to use reinforcement learning to train dialogue generation model for accurate control of emotional text generation. Specifically, we use a combination of reference and activation models. The KL divergence is used to prevent policy bias. The reward model combining empathy scoring and diversity scoring is used to update the model by using the PPO algorithm (Schulman et al., 2017).

3.4.1. Environment

We consider the empathetic dialogue system as the environment. In the environment, our model receives a given input spliced with emotion signals and dialogue history, and then outputs the corresponding empathetic responses. The state s of the environment can be represented as a tuple (x, a), where x denotes the total input and a denotes the current response.

3.4.2. PPO Algorithm

We use PPO-based policy optimization algorithm to update the dialogue generation model. Specifically, we use a combination of the reference and activation models. The reference model does not perform parameter updates during policy updates and is used as a reference for policies. The activation model performs parameter updates using the PPO algorithm for policy generation. The DialoGPT model (Zhang et al., 2020) is chosen for both the initial reference and activation models.

At first, the two models receive the input x and output the responses a separately. Then we use the KL divergence to measure the similarity between the responses of the two models to prevent the activation model from deviating excessively. The KL divergence is defined as:

$$D_{KL}(\pi_{ref}(\cdot|s)||\pi(\cdot|s)) = \mathbb{E}_{a\sim\pi_{ref}(\cdot|s)}\left[log\frac{\pi_{ref}(a|s)}{\pi(a|s)}\right] \quad (6)$$

where $\pi_{ref}(\cdot|s)$ denotes the policy distribution of the reference model and $\pi(\cdot|s)$ denotes the policy distribution of the activation model. In the policy update phase, we use the PPO algorithm to update the policy distribution of the activation model. PPO is an algorithm based on the gradient of the policy, which updates the policy by evaluating the computed ratio of superiority and inferiority.

3.4.3. Reward Model

In the reward model, we introduce empathy reward and diversity reward. The empathy reward function is scored using the pre-trained model DistilRoBERTa[3], and is used to measure how well the emotions of the responses generated by the model match the given emotions. The diversity reward is scored using the pre-trained model DialogRPT[4] and is used to measure how diverse the model generated responses are. We set a combined reward R:

$$R(s, a) = \alpha V(a) + \beta Z(a) \quad (7)$$

where $V(a)$ denotes diversity rewards and $Z(a)$ denotes empathy rewards. α and β are hyperparameters to balance the contribution of the two rewards.

[3] https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

[4] https://huggingface.co/microsoft/DialogRPT-updown

3.5. Response Generator

The response generator fuses commonsense cause information C, future emotion information e_n, and dialogue history D as inputs. Fine-tuning the activation model obtained from the empathy driver ultimately generates empathetic responses that match the target emotion.

First, we connect the three information using the concatenation operation to obtain the fusion information $F = [e_n] \oplus D \oplus C$. Next, we fine-tuning the activation model. Assuming that the target response $Y = [y_1, y_2, ..., y_k]$ is generated from the response generator by using the generated token together with the embedding of the fusion message, where k represents the length. At each decoding time step t, it reads the word embedding $y_{j<t}$ and fusion information for decoding. We use the standard negative log-likelihood as the loss function for response generation:

$$L = -\sum_{t=1}^{k} log(p(y_t | y_{j<t}, F)) \tag{8}$$

where k is the response length, L is the predicted response loss, and $y_{j<t}$ denotes the embedding of the generated tokens. By minimizing the loss function using gradient descent and incrementally improving the model's performance through parameter updates, we achieved the model's ability to generate empathetic responses that are more resonant with the target's emotions.

4. Experiments

4.1. Datasets

Our experiments are conducted on the EmpatheticDialogues dataset (Rashkin et al., 2019). It is a large-scale multi-turn empathetic dialogue dataset that contains 25k dialogue sessions with 3-5 turns of dialogue in each session. The EmpatheticDialogues dataset defines 32 emotion classes, and every dialogue is created based on an emotion class and a situation. We label the dataset with fine-grained emotions at the utterance-level. Based on the original dataset definition, we used 80%, 10% and 10% of the dataset for training, validation and testing.

4.2. Baselines

We compared our model to three groups of representative baseline models and adapted the baseline models' dialogue-level emotion classification to utterance-level future emotion prediction:

Transformer-based Models

- **Multi-TRS** (Rashkin et al., 2019): A variant of the Transformer with an additional unit to predict emotions.
- **EmpDG** (Li et al., 2020): A model that generates empathetic responses using both coarse-grained dialogue-level and fine-grained token-level emotions.
- **KEMP** (Li et al., 2022): A model that generates empathetic responses using both commonsense knowledge and knowledge of the emotion vocabulary.
- **CEM** (Sabour et al., 2022): A model that utilizes commonsense to augment the expression of empathy in the generated response.
- **CASE** (Zhou et al., 2023): A model that aligns coarse-grained and fine-grained cognitions and affects to enhance empathy.
- **EmpSOA** (Zhao et al., 2023): A model that generates empathetic responses through explicit self-other awareness.

PLM-based Models

- **BlenderBot** (Roller et al., 2021): An dialogue agent for pre-trained communication skills. We chose the 400M version.
- **DialoGPT** (Zhang et al., 2020): A dialogue generation model trained specifically for dialogue application situations. We chose the medium version.
- **LEMPEx** (Majumder et al., 2022): A model that uses human communication elements to generate empathetic responses.

LLM-based Model

- **EmpGPT-3** (Lee et al., 2022): A model that generates empathetic responses from GPT-3 by using prompt-based in-context learning.

4.3. Implementation Details

We implemented all the models using the PyTorch framework. The response generator of EmpCRL is based on the medium version of DialoGPT, and the ChatGPT version of the commonsense reasoner is GPT-3.5-turbo. The AdamW optimizer with β_1 = 0.9 and β_2 = 0.9 is used for training. The training sets the mini-batch size to 32 and during inference we use a batch size of 1 and up to 40 decoding steps. We use Top-p sampling with p = 0.9, temperature τ = 0.7. We set α = 0.5 and β = 0.5 for the combined reward R. All the models are trained on NVIDIA RTX 3090 GPUs.

Models	PPL↓	Dist-1↑	Dist-2↑	EAD-1↑	EAD-2↑	I-ACC↑	E-ACC↑
Transformer-based Models							
Multi-TRS	39.15	0.32	1.24	0.96	2.87	20.06	14.83
EmpDG	36.45	0.47	1.89	1.41	3.97	24.51	17.11
KEMP	37.96	0.51	2.12	1.09	3.48	29.15	21.68
CEM	37.47	0.65	2.76	1.13	3.69	26.81	18.79
CASE	35.79	0.71	3.85	1.47	4.96	32.41	21.44
EmpSOA	35.98	0.65	3.51	1.44	4.21	30.99	19.21
PLM-based Models							
LEMPEx	26.37	1.41	14.66	3.51	13.85	-	19.85
BrenderBot	16.71	2.58	11.55	2.24	16.80	-	24.45
DialoGPT	18.74	2.71	12.01	2.87	16.51	-	21.51
LLM-based Model							
EmpGPT-3	-	3.15	**18.63**	4.25	17.50	-	26.93
EmpCRL (Ours)	**15.70**	**4.27**	16.11	**5.39**	**22.63**	**41.57**	**32.76**

Table 2: Results of automatic evaluation. The best results among all models are highlighted in bold.

4.4. Evaluation Metrics

4.4.1. Automatic Evaluation

Most existing automated assessments directly compare the generated responses to the gold responses. However, there may be many reasonable empathetic responses for the same dialogue history. Previous research (Liu et al., 2016) has shown that metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) do not correlate with human judgment. To this end, we did not use these metrics in our experiments, instead we considered the following metrics:

- **PPL** (Serban et al., 2015): The perplexity (PPL) measures how well the probabilistic model predicts a given sample.

- **Dist-n** (Li et al., 2016): It measures the proportion of the distinct n-grams in all the generated results to indicate diversity.

- **EAD-n** (Liu et al., 2022): It refers to Expectation Adjusted Difference score. It is used to evaluate the diversity of long sentences.

- **INF ACC (I-ACC)** (Li et al., 2021): It measures the accuracy of fine-grained emotion inference. Higher INF ACC indicates that the model has better response emotion inference.

- **EXP ACC (E-ACC)** (Pascual et al., 2021): It measures the accuracy of fine-grained emotion expression. Higher EXP ACC indicates that the model has a better ability to respond to emotion adherence.

4.4.2. Human Evaluation

We evaluate the quality of responses in the following aspects:

- **Coherence (Coh.):** which response is more coherent in content and more relevant to the context.

- **Empathy (Emp.):** which response is more understanding of the user's situation and shows more appropriate emotions.

- **Informativeness (Inf.):** which response conveys more information.

We randomly selected 100 dialogues and assigned the responses generated by the model to 3 crowdsourcing workers for evaluation. Every aspect has a scale of 1 to 5.

5. Results and Analysis

5.1. Overall Results

5.1.1. Automatic Evaluation Results

Table 2 reports the results of the evaluation on the automatic metric values. We observe that EmpCRL significantly exceeds the baselines for most of the automatically evaluated metrics. The results for Dist-n and EAD-n show that EmpCRL generates more diverse responses than the baselines. Our model EmpCRL achieves the lowest perplexity, suggesting that the overall quality of our generated responses are higher than the baselines. In addition, INF ACC and EXP ACC scores indicate that EmpCRL has favourable emotion inference and controlled response generation. Since

Models	PPL↓	Dist-1↑	Dist-2↑	EAD-1↑	EAD-2↑	E-ACC↑
EmpCRL	15.70	**4.27**	**16.11**	**5.39**	**22.63**	**32.76**
w/o Com	**15.36**	4.25	15.23	5.15	21.05	29.44
w/o Emo	15.48	4.16	15.78	5.16	20.82	26.81
w/o Emp	15.37	4.03	15.32	5.12	20.61	23.97
w/o Div	15.39	3.89	14.18	4.84	19.96	30.14

Table 3: Results of ablation studies.

EmpGPT-3 uses GPT-3 to generate responses by prompting, the PPL is calculated differently from other baselines. Therefore its PPL is not shown. Furthermore, EmpGPT-3 performs worse than EmpCRL on most metrics because it uses prompt-based in-context learning instead of fine-tuning.

Comparisons	Aspects	Win	Lose	κ
vs. CASE	Coh.	**51.9**‡	34.1	0.56
	Emp.	**54.8**‡	36.5	0.49
	Inf.	**55.1**‡	34.9	0.57
vs. EmpSOA	Coh.	**53.4**‡	37.4	0.51
	Emp.	**51.2**‡	32.5	0.54
	Inf.	**51.5**‡	36.8	0.56
vs. EmpGPT-3	Coh.	**49.5**‡	43.4	0.51
	Emp.	**47.6**‡	42.3	0.44
	Inf.	**49.1**‡	41.7	0.49

Table 4: Human evaluation results (%). ‡ represents significant improvement with p-value $<$ 0.05.

5.1.2. Human Evaluation Results

Table 4 presents the results of the human evaluation. EmpCRL outperforms the baselines in all three aspects, which suggests that EmpCRL achieved the best performance in coherence, empathy, and informativeness scores. This verifies that EmpCRL can produce more empathetic, informative and coherent responses guided by in-context commonsense reasoning and emotion controlling. Meanwhile, we used Fleiss' kappa (κ) (Fleiss, 1971) to measure overall inter-annotator agreement. Consistency ratios in the range of [0.41,0.6] indicate moderate agreement.

5.2. Ablation Studies

To better study EmpCRL, we conduct the ablation study. The results of the ablation study are shown in Table 3. Specifically, we designed four variants of EmpCRL:

- **w/o Com**: Removing the in-context commonsense reasoner and not using commonsense cause information.
- **w/o Emo**: Removing the emotion perceiver and not using emotion inference information.
- **w/o Emp**: Removing the empathy reward of reinforcement learning and using only diversity reward.
- **w/o Div**: Removing diversity reward for reinforcement learning and using only empathy reward.

The results in Table 3 show that every component contributes to the final performance. Since the same emotion perceiver is used for several variants, I-ACC scores are not shown. Both commonsense reasoning and emotion inference have an effect on the results of the study, suggesting that both cognitive and affective perceptions are necessary to understand the speaker's emotions. Commonsense reasoning improves the diversity of responses, but makes the PPL slightly worse. The removal of empathy reward leads to limited emotional expressiveness. In addition, the removal of diversity reward leads to significant decreases in Dist-n and EAD-n scores, suggesting the effectiveness of the empathy driver's diversity reward in producing more diverse responses.

5.3. Case Study

In Table 5, we show two comparisons of responses generated by EmpCRL and three strong baselines. It can be seen that CASE and EmpSOA provide only poor empathy, with generalized responses that do not contain much information. EmpGPT-3 generates responses that are diverse and reasonable but not emotionally rich enough. Intuitively, our model achieves the best performance. Our model EmpCRL shows strong empathy and is more specific about the message. In contrast, the other baselines fail to perceive users' emotional changes, resulting in ineffective responses.

6. Conclusions

In this paper, we propose the EmpCRL model to generate controllable empathetic responses through in-context commonsense reasoning and reinforcement learning. We construct in-context

Emotion Flow	neutral → surprise → joy
Context	Speaker: I have never cheated on my wife. Listener: And thats something you should never do, good on you. Speaker: Yeah it hasn't been easy but I am proud I haven't.
CASE	That is very nice of you .
EmpSOA	That is a good thing to do.
EmpGPT-3	Absolutely, maintaining fidelity in a relationship is essential for its longevity.
Future Emotion	***joy***
Cause	The speaker has never cheated on his wife.
EmpCRL	Your loyalty is commendable. Keep up the good work!
Gold	What do you mean it hasn't been easy? How close have you come to cheating?
Emotion Flow	anger → sadness → neutral
Context	Speaker: I lost my job last year and got really angry. Listener: I am sorry to hear that. Did it happen out of the blue? Speaker: Yes, it was a complete surprise.
CASE	I am sorry to hear that.
EmpSOA	I understand, I know how you feel.
EmpGPT-3	Unexpected unemployment can be disheartening and overwhelming.
Future Emotion	***sadness***
Cause	The speaker lost his job last year.
EmpCRL	I can imagine how angry and shocked you must have felt.
Gold	I am sorry to hear that. I hope it turned out to be a blessing in decise.

Table 5: Case study of the generated responses by EmpCRL and strong baselines.

samples for multi-view contextual reasoning by using ChatGPT. In addition, we jointly model dialogue history and emotion flow to predict response emotions. We also use reinforcement learning to achieve controlled emotion response generation. Extensive experiments based on benchmark metrics and case studies show that EmpCRL outperforms state-of-the-art baseline models, proving the superiority of EmpCRL in generating more empathetic responses. In the future work, we will research the deeper combination of large language models and empathetic dialogue generation.

7. Limitations

The main limitation of our work is the problem of automatic evaluation metrics. The scores of the automatic evaluation metrics are not aligned with the results of the human evaluations. The automatic evaluation metrics focus on the quality of generated responses and the accuracy of emotion prediction and expression. The lack of a generalized empathy evaluation method makes it difficult to evaluate the generation of empathetic dialogue. Meanwhile, it is a valuable direction on how to use multi-source knowledge for commonsense reasoning to generate better empathetic responses. In addition, the scale of the model is an important factor that limits the effectiveness of response generation. In the future, we will explore the use of parameter-efficient tuning techniques on large language models such as LLaMA (Touvron et al., 2023), to generate better empathetic responses.

8. Ethics Statement

Since our work deals with subjects related to human dialogues, we make sure that all the experiments do not cause any harm to human beings. We use EmpatheticDialogues (Rashkin et al., 2019) dataset and CICERO v2 (Shen et al., 2022) dataset. All the above datasets are well-established and publicly available. In addition, the dataset providers filter all personal and sensitive information during the dataset construction process. It is important to clarify that our work is only a study of open-domain dialogue with empathy.

9. Acknowledgments

The work was supported by National Natural Science Foundation of China (62172086, 62272092).

10. Bibliographical References

Mao Yan Chen, Siheng Li, and Yujiu Yang. 2022. Emphi: Generating empathetic responses with human-like intents. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1063–1074.

Mark H Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44:113.

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

Deepanway Ghosal, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2022. Cicero: A dataset for contextualized commonsense inference in dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5010–5028.

Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz, and Tianlu Wang. 2023. Understanding in-context learning via supportive pretraining data. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12660–12673.

Martin L Hoffman. 2001. *Empathy and moral development: Implications for caring and justice*. Cambridge University Press.

Sevgi Coşkun Keskin. 2014. From what isn't empathy to empathic learning process. *Procedia-Social and Behavioral Sciences*, 116:4932–4938.

Taewoon Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *CoRR*, abs/2108.12009.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq R. Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. Gedi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 4929–4952.

Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2022. Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 669–683.

Dayu Li, Xiaodan Zhu, Yang Li, Suge Wang, Deyu Li, Jian Liao, and Jianxing Zheng. 2021. Emotion inference in multi-turn conversations with addressee-aware module and ensemble strategy. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3935–3941.

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

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. Empdg: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4454–4466.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10993–11001.

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. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike 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. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Siyang Liu, Sahand Sabour, Yinhe Zheng, Pei Ke, Xiaoyan Zhu, and Minlie Huang. 2022. Rethinking and refining the distinct metric. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 762–770.

Navonil Majumder, Deepanway Ghosal, Devamanyu Hazarika, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2022. Exemplars-guided empathetic response generation controlled by the elements of human communication. *IEEE Access*, 10:77176–77190.

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 *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 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.

Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3973–3997.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. CEM: commonsense-aware empathetic response generation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11229–11237.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2015. Hierarchical neural network generative models for movie dialogues. *arXiv preprint arXiv:1507.04808*, 7:434–441.

Siqi Shen, Deepanway Ghosal, Navonil Majumder, Henry Lim, Rada Mihalcea, and Soujanya Poria. 2022. Multiview contextual commonsense inference: A new dataset and task. *CoRR*, abs/2210.02890.

Xiaohui Song, Liangjun Zang, Rong Zhang, Songlin Hu, and Longtao Huang. 2022. Emotionflow: Capture the dialogue level emotion transitions. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8542–8546. IEEE.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4634–4645.

David Westbrook, Helen Kennerley, and Joan Kirk. 2011. *An introduction to cognitive behaviour therapy: Skills and applications*. Sage.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. *CoRR*, abs/2306.01693.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. *CoRR*, abs/2201.05337.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Weixiang Zhao, Yanyan Zhao, Xin Lu, and Bing Qin. 2023. Don’t lose yourself! empathetic response generation via explicit self-other awareness. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13331–13344, Toronto, Canada. Association for Computational Linguistics.

Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang, and Minlie Huang. 2023. CASE: aligning coarse-to-fine cognition and affection for empathetic response generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8223–8237.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593.

A. Prompt For Commonsense Reasoning

I will provide some examples of a dialogue and the commonsense cause of the speaker as follow:

Speaker: Excuse me , could I ask a favour?
Listener: Sure , go ahead .
Speaker: Could you tell me where the canteen is ?
Listener: Sure , I can take you there actually.
Speaker: Oh , I don't want to trouble you.
Listener: It's fine . I was heading there anyway.

The commonsense cause of the speaker emotion is: The speaker is kind and wants to help the listener.

Now, make a inference about the commonsense cause for the dialogue. The context of the dialogue is:

Speaker: I was in a bind not too long ago and I trusted my parents to help me out.
Listener: Did they help you?
Speaker: Yes! I knew I could depend on them.
Listener: I'm glad things worked out for you.
Speaker: Thank you. So am I. I don't know what I would have done otherwise.

What is the commonsense cause of the speaker?

Table 6: The prompt and an in-context example used by ChatGPT for commonsense reasoning.