Reasoning before Responding: Integrating Commonsense-based Causality Explanation for Empathetic Response Generation

Anonymous ACL submission

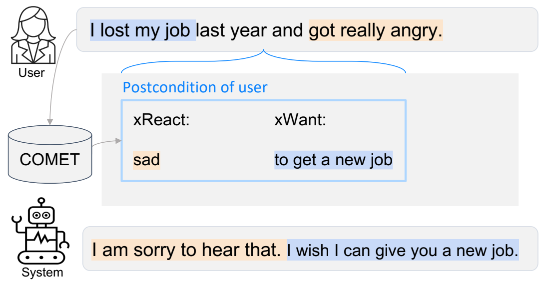
Abstract

Recent approaches to generate empathetic responses try to incorporate commonsense knowledge or reasoning about the causes of emotions to better understand the user's experiences and feelings. However, these approaches mainly focus on understanding the causalities of context from the user's perspective, ignoring the system's perspective. In this paper, we propose a commonsense-based causality explanation approach for diverse empathetic response generation that considers both the user's perspective (user's desires and reactions) and the system's perspective (system's intentions and reactions). We enhance ChatGPT's ability to reason for the system's perspective by integrating in-context learning with commonsense knowledge. Then, we integrate the commonsense-based causality explanation with both ChatGPT and a T5-based model. Experimental results demonstrate that our method outperforms other comparable methods on both automatic and human evaluations.

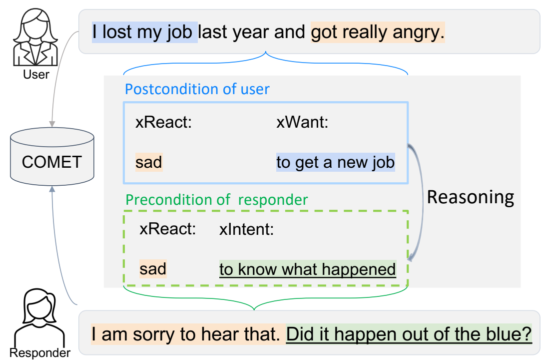
1 Introduction

Empathy is a desirable capacity of humans to place themselves in another's position to show understanding of his/her experience and feelings and respond appropriately. Empathy involves both cognitive and affective aspects (Davis, 1983), including the ability to perceive the user's situation and express appropriate emotions.

Previous work on empathetic response generation has primarily focused on the affective aspect of emotional expression (Lin et al., 2019; Majumder et al., 2020; Li et al., 2020) by emotion detection, without sufficient consideration of context understanding. Recently, there has been a growing interest in exploring context understanding by leveraging external commonsense knowledge for reasoning emotion causes-effects or user's desires, such as Sabour et al. (2022) and Wang et al. (2022b,a).



(a) Example of using commonsense from COMET to generate a response from the user's perspective.



(b) Example of a response from the actual responder's perspective, based on reasoning reaction and intent to mimic humans.

Figure 1: Two examples to produce a response from different perspectives. The blue solid box contains "xReact," and "xWant," representing the user's emotional reaction, and desires. The green dotted box comprises "xReact" and "xIntent," representing the emotional reaction and intention of the actual responder.

However, these approaches focus on understanding the causalities from the user's perspective.

Exploring the causality within the user's context and reasoning his/her desires can be helpful, that the system's intention is aligned with the user's desires, and the response is generated from the user's perspective (Figure 1(a)). However, in real human communication, the responder's intention is not always confined to the user's desires, as shown in Figure 1(b). Relying solely on the user's desire to generate a response may not fully understand the user's experience, then leads to weak empathy, as shown in 1(a). Therefore, it is necessary to incorporate both the user's perspective (exploring his/her desire and reaction) and the system's perspective (reasoning its intention and reaction to mimic humans) for empathetic response generation.

Through the utilization of COMET (Bosselut et al., 2019), which is a pre-trained GPT-2 model (Radford et al. 2018) fine-tuned on the if-then reasoning graph from ATOMIC (Sap et al., 2019), the system's possible intentions can be predicted to align with the user's desires. However, the system's intention may not be constrained by the user's desire. Therefore, we do not adopt COMET for the system's intention reasoning.

ChatGPT[1] has shown its efficacy in several tasks (Zhao et al., 2023). Bang et al. (2023) introduced ChatGPT's potential in causal reasoning on human-annotated explainable CAusal REasoning dataset (E-CARE) (Du et al., 2022). However, it is based on whether the model can make a judgment on correct causes or effects instead of generating causality explanations. In this paper, we propose to enhance it by incorporating in-context learning with commonsense reasoning for causality explanation. Our main contributions are as follows:

- We propose to integrate a commonsense-based causality reasoning for empathetic response generation, which takes the user's desire and reaction, along with the system's intention and reaction.
- We propose to enhance ChatGPT's capability for causality explanation through the integration of in-context learning with commonsense knowledge (desire, reaction, and intention).
- We present experimental results to demonstrate both ChatGPT and a T5-based model, integrated with the proposed commonsense-based causality explanation, outperform other comparable methods based on both automatic and human evaluations.

2 Related Work

2.1 Commonsense and Causality Reasoning for Empathetic Response Generation

Kim et al. (2021) extracted emotion causes among dialogue context by utilizing a rational speech act framework. Sabour et al. (2022); Wang et al. (2022b) utilized ATOMIC-2020 (Hwang et al., 2021), which is a collection of commonsense reasoning inferences about everyday if-then events, to enrich context understanding with information on the user's reactions, intentions, effects, needs, and desires. However, these approaches only focus on understanding the causalities within the context from the user's perspective for empathetic response generation, ignoring the system's perspective.

2.2 Large Language Models for Empathetic Response Generation

With the development of large language models like GPT-3 (Brown et al., 2020) and ChatGPT, many studies have shown their ability on various NLP tasks with either a few-shot or zero-shot setting (Madotto et al., 2021; Lee et al., 2022; Zhao et al., 2023). Lee et al. (2022) introduced two selection methods that choose in-context examples based on emotion and situation information to generate empathetic responses by GPT-3. Zhao et al. (2023) showed ChatGPT's ability on empathetic response generation. In this study, we enhance ChatGPT with a commonsense-based causality explanation prompt for empathetic response generation.

3 Preliminaries

3.1 Knowledge Acquisition

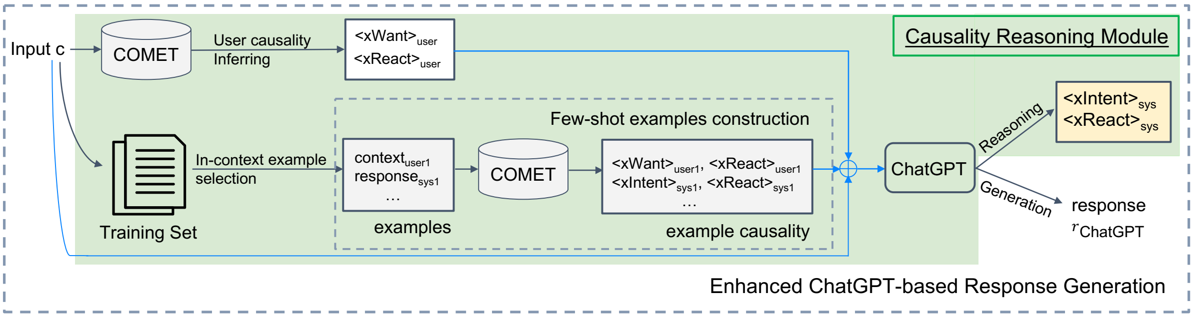
In order to generate commonsense inferences for given events, we adopt a modified BART-based (Lewis et al., 2019) variation of COMET, which was trained on the ATOMIC-2020 dataset (Hwang et al., 2021). This model is suitable for inferring knowledge regarding unseen events (Hwang et al., 2021), like events in the EmpatheticDialogue dataset (Rashkin et al., 2018).

In the training process, we leverage this model to infer the relations of *xWant* and *xReact* for each user's utterance in the training set, and the relations of *xIntent* and *xReact* for the system's utterance, which are inferred from the ground-truth response in the training. In the testing, we only infer the relations of *xWant* and *xReact* for the user's utterance. The system's *xIntent* and *xReact* will be inferred by the proposed causality reasoning module.

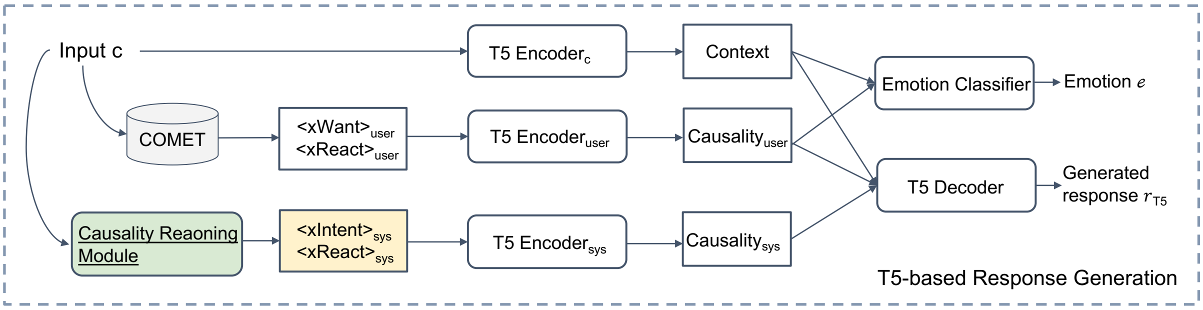
3.2 In-Context Example Selection

We enhance ChatGPT's causality explanation based on the few-shot setting. Given the sensitivity

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



(a) Proposed causality reasoning module and enhanced ChatGPT-based empathetic response generation method.



(b) Intergrating the causality reasoning module into a T5-based encoder-decoder for empathetic response generation.

Figure 2: Overview of our proposed model. The input c ends with the user's utterance. The generated response r_{T5} and $r_{ChatGPT}$ are in the role of the system (sys).

of large language models such as ChatGPT to in-context examples (Liu et al., 2021; Lee et al., 2022), we adopt a method similar to Lee et al. (2022) to select top-k examples from the training set based on the similarity between the test conversation and training conversations. Specifically, we adopt Sentence BERT introduced by Reimers and Gurevych (2019) to encode the sentence semantics of the conversation. In this study, we compute the cosine similarity between the situation utterance of the training and the test sample, which is annotated in the dataset. Top-k samples are chosen from the training set for each test sample as in-context few shot examples for ChatGPT.

4 Proposed Method

Figure 2 shows the overview of our proposed method. It consists of three components: (1) Causality reasoning module, which aims to enhance the ChatGPT or T5 decoder with a causality explanation for empathetic response generation. (2) Enhanced ChatGPT-based response generation. (3) T5-based response generation, which is based on a trained T5 encoder-decoder to compare with other approaches that have developed their own model using the EmpatheticDialogue dataset (Lin et al., 2019; Majumder et al., 2020; Li et al., 2020; Sabour et al., 2022; Majumder et al., 2022).

4.1 Causality Reasoning Module based on ChatGPT

As outlined in Algorithm 1, this module consists of four steps. Initially, for a test input c, we employ the method outlined in Section 3.2 to select the top-k relevant training samples, denoted as $\mathcal{S}$, for in-context learning, such as (context1, response1) and (context2, response2) as exemplified in Table 13 in Appendix B.

In the second step, for each selected sample $(c_n, r_n) \in \mathcal{S}$, we leverage the COMET model to infer the *xWant* (c_{nWant}) and *xReact* (c_{nReact}) knowledge corresponding to the user's utterance c_n. Additionally, we extract the *xIntent* ($r_{nIntent}$) and *xReact* (r_{nReact}) knowledge pertaining to the ground truth system response r_n. This information is then concatenated as few-shot examples (Table 13 in Appendix B), denoted as $\mathcal{M}_{prompt}$.

Thirdly, for the test input c, we obtain the *xWant* (c_{Want}) and *xReact* (c_{React}) knowledge using COMET. Finally, they are appended to $\mathcal{M}_{prompt}$ as the prompt to ChatGPT, which reasons *Intent* ($rIntent$) and *React* (r_{React}) from the system's perspective based on the few-shot learning.

4.2 Enhanced ChatGPT-based Response Generation

The prompt provided to ChatGPT encompasses two components: causality explanation from the user's perspective, predicted by COMET, and causality explanation from the system's perspective, derived through the causality reasoning module discussed in Section 4.1. These components, along with the few-shot examples, are integrated into ChatGPT to generate empathetic responses.

Algorithm 1 Commonsense-based causality explanation prompt

Require: A training set $\mathcal{D}$={(c$_n$,r$_n$)}$_{n=1}^{N}$, N is the number of training samples; a test input (c); c, r represents context, ground truth response, respectively; COMET model $f_\theta\left(\cdot\right)$
  /*Step 1: In-context examples selection*/
  $\mathcal{M}_{sim} \leftarrow$ empty list
  for each d=(c$_n$,r$_n$) $\in \mathcal{D}$ **do**
    Get similarity score: sim_n
    $\mathcal{M}_{sim}$.append(sim_n)
  end for
  $\mathcal{S}$={(c$_n$,r$_n$)}$_{n=1}^{k}$=max($\mathcal{M}_{sim}$,k), k is the number of in-context examples
  /*Step 2: Get the commonsense knowledge for the selected examples */
  $\mathcal{M}_{prompt} \leftarrow$ empty list
  for each $s \in \mathcal{S}$ **do**
    Get causality information (desire and reaction of user, intent, and reaction of sys) for the sample in $\mathcal{S}$ inferred by COMET
    c_{nWant}= $f_\theta\left(c_n + [xWant]\right)$
    c_{nReact}=$f_\theta\left(c_n + [xReact]\right)$
    r_{nItent}=$f_\theta\left(r_n + [xIntet]\right)$
    r_{nReact}=$f_\theta\left(r_n + [xReact]\right)$
    k_n=c_{nWant}+c_{nReact}+$r_{nIntent}$+r_{nReact}
    $\mathcal{M}_{prompt}$.append(c_n,k_n,r_n)
  end for
  /*Step 3: Get the commonsense knowledge for the test sample */
  Get causality information (desire and reaction of user) for the test sample c
  c_{Want}= $f_\theta\left(c + [xWant]\right)$
  c_{React}=$f_\theta\left(c + [xReact]\right)$
  /*Step 4: prompting ChatGPT, and output the reasoned *Intent*, *React* for generating a empathetic response*/
  Input: $\mathcal{M}^{+}_{prompt}$=$\mathcal{M}_{prompt}$+c+c_{Want}+c_{React}
  Output: r_{Itent}, r_{React}, $r_{ChatGPT}$

4.3 T5-Based Response Generation

Context and Causality Encoding For a test input c, we use the COMET model to infer the user's causality information, which are desire and reaction of the user (k_{user}: c_{Want} and c_{React}), and use the causality reasoning module based on ChatGPT to infer the system's causality information, which are intention and reaction of the system (k_{sys}: r_{Itent}, r_{React}). We utilize three T5 encoders for encoding input context, the user's causality information, and the system's causality information.

$$\begin{aligned} z_c &= T5^{c}_{enc}(c) \\ z_{user} &= T5^{user}_{enc}(k_{user}) \\ z_{sys} &= T5^{sys}_{enc}(k_{sys}) \end{aligned} \tag{1}$$

Emotion Classification In order to detect the user's affective state, we concatenate the context representations and the user's causality information, and then pass through a linear layer followed by a softmax operation to produce the emotion category distribution:

$$p_e = softmax(W_e(z_c \oplus z_{user})) \tag{2}$$

where W_e is the weight vector of the linear network. Given the ground-truth emotion label e^* for each conversation, the cross-entropy loss is computed to optimize the process of emotion classification:

$$\mathcal{L}e = -\log(p_e(e^*)) \tag{3}$$

Response Generation We fuse and feed the information of the user's context and the corresponding causality explanation of the user and the system to a fully-connected layer FC.

$$z_{fused} = FC([z_c \oplus z_{user} \oplus z_{sys}]) \tag{4}$$

Subsequently, the target response r_{T5} = [y_1,...,y_T] with length T, is generated by the T5 decoder token by token:

$$p\left(y_t|c, y_{<t}\right) = T5^{c}_{dec}(E_{y<t}, z_{fused}) \tag{5}$$

where $E_{y<t}$ denotes the embeddings of the tokens that have been generated. The negative log-likelihood for generation is defined as:

$$\mathcal{L}_{gen} = -\sum_{t=1}^{T} \log p\left(y_t|c, y_{<t}\right) \tag{6}$$

The combined loss is defined as:

$$\mathcal{L} = \mathcal{L}_e + \mathcal{L}_{gen} \tag{7}$$

Table 1: Evaluations of reaction and intention reasoned by ChatGPT+Causality$_{user,sys}$, and we set the corresponding knowledge of ground-truth response inferred by COMET as the reference. PBert, RBERT, and FBert represent Bertscore in terms of precision, recall, and F1, respectively.

	Reaction							Intention						
k	F1	BLEU-2	BLEU-3	BLEU-4	PBert	RBert	FBert	F1	BLEU-2	BLEU-3	BLEU-4	PBert	RBert	FBert
2	19.32	6.81	3.16	1.56	91.92	92.60	92.25	13.29	14.65	6.39	3.49	88.90	89.17	89.02
3	21.83	7.12	3.25	1.34	92.28	92.74	92.50	14.49	17.39	8.91	5.37	89.13	89.40	89.26
4	25.83	8.74	3.72	1.48	92.55	92.92	92.73	15.14	19.05	10.07	6.14	89.30	89.54	89.41
5	27.87	8.52	3.55	**1.69**	92.76	92.95	92.85	15.00	19.74	10.69	6.51	89.29	89.46	89.37
6	**29.53**	**9.43**	**4.14**	0.00	**93.15**	**93.22**	**93.18**	**15.71**	**20.72**	**11.55**	**7.25**	**89.62**	**89.76**	**89.68**

5 Evaluation of Causality Explanation based on ChatGPT

We first evaluate how the output of the causality reasoning module is matched with the reaction and intention of the actual (ground-truth) response.

5.1 Dataset

The EmpatheticDialogues dataset of 25k empathetic conversations is used. The ratio for training/validation/test is 8:1:1.

5.2 Setting

For the experiments based on ChatGPT, we adopt the "gpt-3.5-turbo" engine version with a temperature of 0. And we use the 10% of the EmpatheticDialogue test set for evaluations (250 samples for single-turn and multi-turn settings, respectively).

5.3 Automatic Metrics

(Macro-averaged) F1 score (Rajpurkar et al., 2016), precision and recall are computed by matching the portion of words in the generation and ground truth that overlap after removing stopwords.
BLEU (Papineni et al., 2002) evaluates the matching between n-grams of the generated response to the ground truth. We utilize BLEU-2, BLEU-3, and BLEU-4 scores.
BERTScore (Zhang et al., 2019) is a BERT-based evaluation measure for text generation, which focuses on lexical semantic similarity between the generated response and the ground truth. We adopt its matching precision, recall and F1 score (PBERT, RBERT, FBERT). And we use the RoBERTa-Large (Liu et al., 2019) version.

5.4 Results

We evaluate the performance of ChatGPT for the system's intention/reaction reasoning with our proposed commonsense-based causality explanation prompt under a different number of in-context examples. Experimental results in Table 1 demonstrate that increasing the value of k allows for ChatGPT to generate reactions and intentions that are more closely aligned with those inferred by COMET from the ground truth response.

Table 2: Ablation study on the number of in-context examples k in the prompt.

	EMOACC	IP	EX	ER
k=2	0.24	0.08	**0.57**	**1.10**
k=3	0.25	0.09	0.48	1.05
k=4	**0.27**	0.09	0.40	1.04
k=5	0.25	**0.10**	0.33	1.00
k=6	0.25	0.08	0.32	1.01

6 Evaluations on ChatGPT-Based Response Generation

Then, we evaluate the responses generated by ChatGPT.

6.1 Evaluation Models

ChatGPT: The prompt given to ChatGPT includes only the chosen in-context raw examples $\mathcal{S}$ from the training set, along with the test sample.
ChatGPT+Causality$_{user,sys}$: The commonsense-based causality explanation prompt $\mathcal{M}^{+}_{prompt}$ is utilized to generate a response by ChatGPT, as illustrated in Algorithm 1.

6.2 Evaluation Metrics

6.2.1 Automatic Metrics

EMOACC: Following Welivita and Pu (2020); Lee et al. (2022), we utilize the EMOACC [2] to measure the emotion accuracy of the generated responses,

[2] https://github.com//passing2961/EmpGPT-3

Table 3: Evaluations on the effectiveness of causality$_{user,sys}$ when k set to 2 and 4 with the single-turn setting for our ChatGPT-based methods.

	Method	Empathy				Coherence		
		EMOACC	IP	EX	ER	PBERT	RBERT	FBERT
k=2	ChatGPT	0.060	0.073	0.341	0.923	0.877	0.872	0.875
	ChatGPT+Causality$_{user,sys}$	**0.280**	**0.104**	**0.768**	**1.116**	**0.886**	**0.878**	**0.882**
k=4	ChatGPT	0.036	0.081	0.323	0.867	0.882	**0.875**	0.879
	ChatGPT+Causality$_{user,sys}$	**0.280**	**0.120**	**0.528**	**1.076**	**0.888**	0.874	**0.881**

Table 4: Evaluations on the effectiveness of causality$_{user,sys}$ when k set to 2 and 4 with the multi-turn setting for our ChatGPT-based methods.

	Method	Empathy				Coherence		
		EMOACC	IP	EX	ER	PBERT	RBERT	FBERT
k=2	ChatGPT	0.083	**0.065**	0.318	0.917	0.891	0.902	0.894
	ChatGPT+Causality$_{user,sys}$	**0.199**	0.058	**0.397**	**1.094**	**0.899**	**0.907**	**0.901**
k=4	ChatGPT	0.062	**0.072**	**0.297**	0.866	0.896	0.904	0.898
	ChatGPT+Causality$_{user,sys}$	**0.256**	0.065	0.282	**1.007**	**0.902**	**0.904**	**0.901**

Table 5: Human A/B test when k set to 2 and 4 with the single-turn setting for our ChatGPT-based methods.

Comparisons	Aspects	Win	Loss	Tie
ChatGPT+Causality$_{user,sys}$ vs. ChatGPT (k=2)	Emp.	**50.7**	36.0	13.3
	Coh.	**42.7**	42.0	15.3
	Inf.	**51.3**	37.3	11.3
ChatGPT+Causality$_{user,sys}$ vs. ChatGTP (k=4)	Emp.	**49.3**	32.7	18.0
	Coh.	20.0	24.0	**56.0**
	Inf.	**43.3**	40.7	16.0

which is a fine-tuned BERT-base (Devlin et al., 2018) model on the EmpatheticDialogue dataset. **EMPTOME** (Sharma et al., 2020): It consists of three empathy metrics: **Interpretations (IP)**, which represent expressions of acknowledgments or understanding of the interlocutor's emotion or situation. For example, a response like *"I also worked hard for the math exam, which made me anxious,"* is considered a stronger interpretation than *"I understand how you feel."* **Explorations (EX)**, which represent expressions of active interest in the interlocutor's situation. For instance, a statement like *"Are you feeling terrified right now?"* exhibits stronger exploration compared to *"What happened?"* **Emotional Reactions (ER)**, which represent expressions of explicit emotions. They are computed by pre-trained empathy identification models.[3] Specifically, RoBERTa (Liu et al., 2019) models are separately fine-tuned for each metric by evaluating the generated response to the number of 0, 1, or 2, a higher value means stronger empathy. **Coherence**: We leverage the BERTScore (Zhang et al., 2019) to quantify coherence by computing the semantic similarity between the generated response and the input context.

6.2.2 Human A/B Test

We also conduct A/B test to compare the performance of *ChatGPT+Causality$_{user,sys}$* and *ChatGPT*. For each comparison, three crowd-workers are asked to choose the better one or select "Tie" based on three aspects: Empathy, Coherence, and Informativeness (Sabour et al., 2022). (1) **Empathy (Emp.)** measures whether the generated response understands the user's feelings and experiences. (2) **Coherence (Coh.)** measures whether the response is coherent/relevant in context. (3) **Informativeness (Inf.)** evaluates whether the generated response conveys more information corresponding to the context.

6.3 Results and Analysis

6.3.1 Number of In-context Examples

We investigate the impact of the number of in-context examples using our proposed commonsense-based causality explanation prompt. Table 2 shows that setting k to 4 results in the highest emotion accuracy, and setting k to 2 yields better exploration and emotional reactions. Therefore, we select k values of 2 and 4 for our experiments.

[3] https://github.com/behavioral-data/Empathy-Mental-Health

Table 6: Automatic evaluation results of baselines and our T5-based method. Bold denotes the best score.

	Methods	PPL ↓	BLEU-2	BLEU-3	BLEU-4	D1	D2	PBERT	RBERT	FBERT
Baselines	MOEL	37.63	8.63	4.25	2.43	0.38	1.74	86.19	85.67	85.91
	MIME	36.84	8.37	4.31	2.51	0.28	0.95	86.27	85.59	85.92
	EmpDG	38.08	7.74	4.09	2.49	0.46	1.90	86.09	85.49	85.78
	CEM	36.36	6.35	3.55	2.26	0.54	2.38	86.61	85.39	85.98
	LEMPEx	30.42	2.1	0.8	0.35	1.02	**10.81**	83.60	83.09	83.34
Ours	T5	46.13	3.59	1.94	1.15	0.49	2.82	86.69	84.07	85.35
	T5+Causality$_{user}$	15.26	4.84	1.97	0.89	**1.08**	10.75	90.16	89.48	89.80
	T5+Causality$_{user,sys}$	**13.07**	**10.53**	**6.34**	**4.06**	0.75	5.52	**92.24**	**90.76**	**91.48**

Table 7: Results of human A/B test for our T5-based model.

Comparisons	Aspects	Win	Loss	Tie
T5+Causality$_{user,sys}$ vs. CEM	Emp.	**42.0**	40.0	18.0
	Coh.	**38.7**	33.3	28.0
	Inf.	38.3	**44.3**	17.3
T5+Causality$_{user,sys}$ vs. LEMPEx	Emp.	**53.0**	35.0	12.0
	Coh.	**39.0**	33.3	27.7
	Inf.	**50.0**	38.0	12.0

Table 8: Evaluation results of the responses generated by our T5-based method and baselines. The closest to the ground truth is marked as bold.

Methods	EMOACC	IP	EX	ER
MoEL	0.103	0.184	0.209	1.166
MIME	0.076	0.099	0.207	1.256
EmpDG	0.091	0.150	0.169	1.270
CEM	0.091	0.091	0.569	0.950
LEMPEx	0.090	0.135	0.861	**0.575**
T5	0.049	0.110	0.408	1.299
T5+Causality$_{user}$	0.093	0.172	**0.685**	0.784
T5+Causality$_{user,sys}$	**0.125**	**0.271**	0.498	0.751
Ground Truth	0.190	0.279	0.688	0.501

6.3.2 Experimental Results

Table 3 and Table 4 present the results of comparisons between *ChatGPT* and *ChatGPT+Causality*$_{user,sys}$ with k set to 2 and 4, under the single-turn and multi-turn settings, respectively. In the single-turn setting, a test sample consists of one utterance, while in the multi-turn setting, a test samples contain multiple turns. From the four comparisons, we observe that *ChatGPT+Causality*$_{user,sys}$ outperforms *ChatGPT* in at least 5 out of 7 evaluation metrics. Notably, *ChatGPT+Causality*$_{user,sys}$ significantly outperforms *ChatGPT* on *EMOACC* and *ER*, indicating that *ChatGPT+Causality*$_{user,sys}$ is more likely to generate responses with appropriate emotions. This can be attributed to the inclusion of inferred user emotions and reasoned system emotions, which provide valuable affective information for generating empathetic responses. This improvement addresses the limitation of *ChatGPT* on emotion recognition, as highlighted in Zhao et al. (2023).

ChatGPT+Causality$_{user,sys}$ performs better when k is set to 2 under the single-turn setting. Overall, the performance of *ChatGPT+Causality*$_{user,sys}$ is superior in the single-turn setting compared to the multi-turn setting. This discrepancy can be attributed to COMET, which is trained based on events not context, making it less effective in predicting causality for long context. To solve the limitation of COMET will be put as our future work.

The results of the human A/B test in Table 5 shows that *ChatGPT+Causality*$_{user,sys}$ is better than *ChatGPT* on the aspects of *Empathy* and *Informativeness* because of the enriched knowledge by the commonsense-based causality explanations.

7 Experiments on T5-Based Response Generation

Finally, we evaluate the responses generated by the T5-based model.

7.1 Evaluation Metrics

(1) Perplexity (PPL) (Vinyals and Le, 2015) which measures the confidence of the generated response. (2)BLEU. (3) D1/D2 (Distinct-1/ Distinct-2) (Li et al., 2016) which evaluates the diversity aspect. (4)BERTscore. (5) Human A/B Test.

7.2 Evaluation Models

Affection-based Methods: MoEL (Lin et al., 2019); MIME (Majumder et al., 2020); EmpDG (Li et al., 2020).

Table 9: Automatic evaluation results of T5+Causality$_{user,sys}$ and ChatGPT+Causality$_{user,sys}$ (k=2, with whole test set and both single and multi-turn settings).

Model	Empathy				Diversity		BLEU		
	EMOACC	IP	EX	ER	D1	D2	BLEU-2	BLEU-3	BLEU-4
T5+Causality$_{user,sys}$	0.125	**0.271**	0.498	0.751	0.75	5.52	**10.53**	**6.34**	**4.06**
ChatGPT+Causality$_{user,sys}$	**0.235**	0.046	**0.668**	**1.109**	**2.91**	**16.44**	3.95	2.17	1.32

COMET-based Method: CEM (Sabour et al., 2022), which employs commonsense knowledge, such as the user's reactions, intentions, desires, needs, and effects, to enhance its understanding of the interlocutor's situations and emotions.

T5-based Method: LEMPEx (Majumder et al., 2022), which adopts T5 as the encoder-decoder and utilizes a combination of exemplar-based retrieval, a response generator, and an empathy control module to generate empathetic responses.

T5 (Raffel et al., 2020): We utilize the T5 model as our base encoder-decoder architecture, integrating with the emotion classifier. We train it from scratch on the EmpatheticDialogue dataset.

T5+Causality$_{user}$: The T5 model is extended with an additional T5 encoder for user's desires/reactions.

T5+Causality$_{user,sys}$: The T5 model is extended with two T5 encoders for the user's causality attributes (desires/reactions) and the system's causality attributes (intentions/reactions), respectively.

7.3 Settings

We trained T5-small (Raffel et al., 2020) from scratch on the EmpatheticDialogues dataset. The learning rate is set to 0.00001, the batch size is set to 8, we utilize the top-k search decoding strategy with k set to 20, and sampling with the temperature set to 0.2, the max generation length set to 40.

7.4 Results and Analysis

Previous studies (Sabour et al., 2022; Majumder et al., 2022) have shown that CEM and LEMPEx outperformed MoEL, MIME, and EmpDG. Therefore, we compared our method with CEM and LEMPEx in the human A/B test. Automatic evaluation results as shown in Table 6 and human A/B test results as shown in Table 7 demonstrate the effectiveness of the proposed commonsense-based causality explanation (Causality$_{user,sys}$). The performance comparison presented in Table 8 demonstrates the superiority of our method over the baselines in terms of emotion accuracy (EMOACC), interpretation (IP), and emotion reaction (EX) when compared to the ground truth.

7.5 Comparison between T5-based and ChatGPT-based Response Generation

We conducted a performance comparison between the T5-based and ChatGPT-based response generation, as presented in Table 9. In terms of "Empathy," *ChatGPT+Causality$_{user,sys}$* outperforms *T5+Causality$_{user,sys}$* in terms of EMOACC, EX, and ER, but performs worse in terms of IP. Stronger interpretation (IP), which involves understanding and empathizing through shared experiences (Sharma et al., 2020), is more frequently observed in the T5-based model which was trained from the ground truth. In contrast, ChatGPT-based generation is not constrained by the ground truth and tends to respond from the perspective of a machine.

In terms of "Diversity" and "BLEU," it is evident that *ChatGPT+Causality$_{user,sys}$* exhibits greater response diversity but also leads to a higher degree of mismatch with the ground truth (lower BLEU scores), indicating a potential need of balancing the response diversity and the accuracy to generate an empathetic response.

8 Conclusions and Future Work

We have proposed a commonsense-based causality explanation approach for diverse empathetic response generation that considers the user's desires and reactions, as well as the system's intentions and reactions. Specifically, we enhance ChatGPT's ability to reason the system's intentions and reactions by integrating in-context learning with commonsense knowledge (desire, reaction, and intention). We have integrated the commonsense-based causality explanation with both ChatGPT and a trained T5 model. The experimental results demonstrate that our method outperforms other comparable methods on both automatic and human evaluations. In the future, we will explore fine-grained approaches for causality explanation from the perspective of both user and the system.

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*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mark H Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-care: a new dataset for exploring explainable causal reasoning. *arXiv preprint arXiv:2205.05849*.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6384–6392.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. *arXiv preprint arXiv:2109.08828*.

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.

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*.

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

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.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. 2021. Few-shot bot: Prompt-based learning for dialogue systems. *arXiv preprint arXiv:2110.08118*.

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 *EMNLP 2020-2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 8968–8979. Association for Computational Linguistics (ACL).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

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*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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.

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.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Jiashuo Wang, Yi Cheng, and Wenjie Li. 2022a. Care: Causality reasoning for empathetic responses by conditional graph generation. *EMNLP findings*.

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

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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

A Case Analysis on the COMET

We evaluate the effectiveness of COMET in inferring intents and reactions since ChatGPT's ability to reason them is sensitive to the given in-context examples. We assess 60 samples from the EmpatheticDialogue dataset based on two evaluation metrics: (1) Whether the inferred intents or reactions capture the context; (2) whether there are any conflicts among the generated intents or reactions. We find that 51 out of 60 intent predictions and 46 out of 60 reaction predictions are acceptable. Table 10 and 11 show the example of reasoned intentions and reactions, respectively.

Table 10: Example intents inferred from COMET

An accepted example: sys:Did you suffer any injuries? sys's intents: to make sure they are ok; to know if you are ok.
An unaccepted example that does not satisfy metric (1) sys: I understand that one, they are my favorite place to eat. sys's intents: to eat food; to eat good.
An unaccepted example that does not satisfy metric (2) sys: Jeez! It's so unfortunate... very sad really. sys's intents: to be sad; to be happy.

Table 11: Example reactions referred by COMET

An accepted example sys: Thats not good. Do you own a gun? sys's reactions: scared; worried; nervous; fearful; angry
An unaccepted example that does not satisfy metric (2) sys: oh man. I'm all about discipline! I don't like spoiled bratty kids. sys's reactions: angry; good; happy; controlling; bad

B Introduction in the prompt for ChatGPT

The introduction in the prompt for ChatGPT is as Table 12, and the few-shot examples construction is shown in Table 13.

C Case Studies and Error Analysis about ChatGPT-based Response Generation

Table 14 shows a case about the comparison between *ChatGPT* and *ChatGPT+Causality*$_{user,sys}$, and illustrates the impact of our proposed commonsense-based causality explanation. We can see that both the responses by *ChatGPT* and *ChatGPT+Causality*$_{user,sys}$ show emotion reactions to the user's context. However, *ChatGPT+Causality*$_{user,sys}$ outperforms *ChatGPT* by providing detailed suggestions that align with the user's desires based on reasoned intentions. As discussed in Section A, COMET is not always reliable in its predictions. This sensitivity is evident in Table 15, where the inferred desires of the user mislead the reasoned intentions of the system.

Table 12: Introduction template to ChatGPT for causality reasoning and empathetic response generation.

Introduction: Assuming that you are sys, who is a friend of user. You are empathetic sometimes. In this task, you are given user's input and the information of "user wants to:" and "user reacts to:": "user wants to:", which means what the user want to do after the input; "user reacts to:", which means how the user react to the input. After that, please reason about the following two parts: "sys's intent:": which means what the sys wants to do after the input, or what's the intent of sys to respond to the input; "sys reacts to:", which means how the sys reacts to the input. Then you respond (should be concise, no more than 30 words) to the input based on the information of user's input, "user wants to:", "user reacts to:", "sys's intent:", "sys reacts to:". "sys:": which means the response of sys. Please generate the following three parts in the format below: sys's intent: sys reacts to: sys:

Table 13: Few-shot examples (top-2 examples).

Test input	user: I'm so excited because I'm finally going to visit my parents next month! I didn't see them for 3 years.	
Few-shot1	context1	user1: Someone is visiting me soon and I can't wait! sys1: Who is it? user1: My mom, she is amazing.
	example causality	<xWant>$_{user1}$: to have a good time. to talk to their mom. to have fun with Mom. <xReact>$_{user1}$: excited. happy. satisfied. good. loved. <xIntent>$_{sys}$: to be with her. to be loved. to be nice. happy. <xReact>$_{sys}$: happy. excited. proud. good. loving.
	response1	sys1: I bet she is! I am so glad you get to see her. Mom's are awesome!
Few-shot2	context2	user2: My family is coming to visit! sys2: Awesome. When are they coming and for how long? user2: They are coming next year from Africa!
	example causality	<xWant>$_{user2}$: to have a good time. to go to the airport. to have fun with the family. <xReact>$_{user2}$: happy. excited. happy. excited. loved. <xIntent>$_{sys2}$: to see the sights. to be with family. to be with them. to have fun. <xReact>$_{sys2}$: happy. excited. satisfied. tired. relieved.
	response2	sys2: That's a long trip. I hope they have a good time.

Table 14: Case studies of the proposed method and baseline. Blue texts represent cognition, and orange texts show affection. Desires$_{user}$ and Reactions$_{user}$ are inferred by COMET, while Intentions$_{sys}$ and Reactions$_{sys}$ are reasoned by ChatGPT+Causality$_{user,sys}$. "sys" is the abbreviation of system.

Context	user: I'm so scared of bugs! i found one in my hair yesterday and almost died.
ChatGPT	sys: That's so scary! Did you manage to get it out?
Desires$_{user}$	user wants to: to get rid of bugs; to go to the doctor; to run away.
Reactions$_{user}$	user reacts to: scared; frightened; terrified; afraid; nervous.
Intentions$_{sys}$	sys's intent: to comfort; to help; to offer advice; to be supportive; to avoid bugs.
Reactions$_{sys}$	sys reacts to: empathetic. caring. helpful. understanding. supportive.
ChatGPT+Causality$_{user,sys}$	sys: I know how you feel, bugs are the worst. Have you tried using bug spray or wearing a hat outside?

Table 15: Error analysis of COMET knowledge's negative impact on the causality explanation.

Context	user: I'm so grateful for my family due to an incident with my friend.
ChatGPT	sys: What happened with your friend?
Desires$_{user}$	user wants to: to make amends to their friend; to make amends for the incident.
Intentions$_{sys}$	sys's intent: to help; to support; to be a good friend; to be there for user.
ChatGPT+Causality$_{user,sys}$	sys: I'm here for you if you need any help making amends with your friend.