Sibyl: Sensible Empathetic Dialogue Generation with Visionary Commonsense Knowledge

Anonymous ACL submission

Abstract

Recently, there has been a heightened interest in building chatbots based on Large Language Models (LLMs) to emulate human-like qualities in dialogues, including expressing empathy and offering emotional support. Despite having access to commonsense knowledge to better understand the psychological aspects and causality of dialogue context, even these powerful LLMs struggle to achieve the goals of empathy and emotional support. As current approaches do not adequately anticipate dialogue future, they may mislead language models to ignore complex dialogue goals of empathy and emotional support, resulting in unsupportive responses lacking empathy. To address this issue, we present an innovative framework named Sensible Empathetic Dialogue Generation with Visionary Commonsense Knowledge (*Sibyl*). Designed to concentrate on the imminent dialogue future, this paradigm directs LLMs toward the implicit requirements of the conversation, aiming to provide more sensible responses. Experimental results demonstrate that incorporating our paradigm for acquiring commonsense knowledge into LLMs comprehensively enhances the quality of their responses.[1]

1 Introduction

Empathy, in its most comprehensive definition, is the reaction of one individual to the observed experiences of another (Davis, 1983). Given the inherent complexity of conversation, recent works focus on integrating commonsense knowledge to aid in unraveling the implicit psychological motivations and causality within utterances (Wang et al., 2022; Tu et al., 2022; Peng et al., 2022; Zhou et al., 2023; Zhao et al., 2023a). Meanwhile, sophisticated abilities of Large Language Models (LLMs) (Chowdhery et al., 2023; Touvron et al., 2023) in dialogue understanding and response generation

[1] The code will be released at GitHub upon publication.

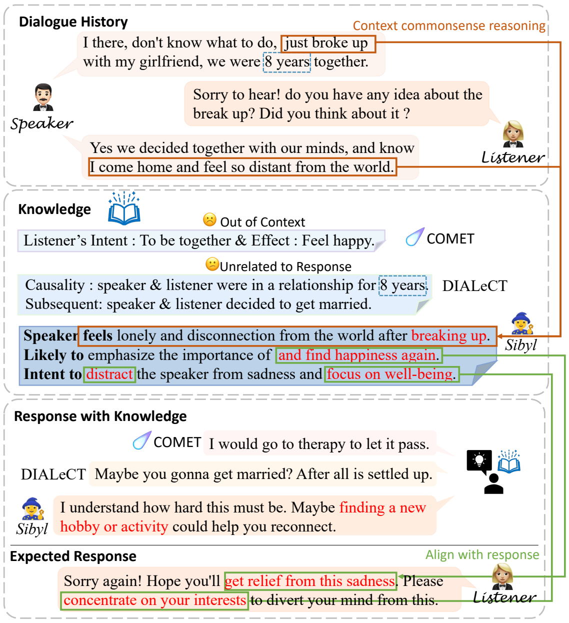

Figure 1: An example from the EMPATHETICDIALOGUES dataset reveals that the commonsense reasoning deduced by COMET and DIALeCT demonstrates notable limitations.

have ignited a new zeitgeist for building a powerful dialogue system (OpenAI, 2022, 2023). These sophisticated models demonstrate strong performance when directly prompted in a dialogue role (Brown et al., 2020), and their responses can be further improved by incorporating explicit intermediate reasoning steps (Wei et al., 2023; Wang et al., 2023).

Despite their achievements, these advanced LLMs are still struggling with generating empathetic responses and providing emotional support (Zhao et al., 2023b). Figure 1 shows that the commonsense inference derived from COMET (Bosselut et al., 2019) primarily concentrates on the last utterance of the Speaker. This narrow focus fails to correspond with the full context of the multi-turn conversation and inaccurately captures the Speaker's emotional state, leading to cascade errors in generating responses. Meanwhile, Shen et al. (2022) employs commonsense reasoning for a complete and static dialogue. This limitation increases the risk of inaccuracies, stemming from its sole focus on dialogue history. As illustrated in Figure 1, DIALeCT (Shen et al., 2022) deduces disadvantaged commonsense inference unrelated to the response and even misunderstands the background information participants.

Investigating the above phenomenon, we suggest that the issue arises since **current approaches do not adequately anticipate dialogue future**. Due to the one-to-many nature of dialogue generation, the existence of multiple distinct responses that can appropriately answer the same dialogue history suggests that within a given context, there are diverse dialogue commonsense inferences associated with each possible response (Liu et al., 2022; Zhou et al., 2022). Exclusively deduced from dialogue history, contemporary methods integrating commonsense inferences into dialogues overlook the future intent of interlocutors and the potential development of the conversation. These methods are prone to introducing noisy information and confusing language models to ignore the demand for empathy and emotional support.

In response to these challenges, this paper presents a new paradigm that dynamically deduces commonsense knowledge relevant to the prospective future of dialogue, called Sensible Empathetic Dialogue Generation with Visionary Commonsense Knowledge (*Sibyl*). We argue that the dialogue history does not encompass enough information to generate the intended response. By deriving plausible future-aware commonsense knowledge from prophetic powerful LLMs, we empower open-source language models to generate these visionary inferences solely based on dialogue history. Essentially, these visionary inferences act as a form of chain-of-thought (CoT) prompts, aiding LLMs in effectively dealing with complex dialogue contexts, bridging the gap between dialogue history and potential response, and ultimately promoting empathy and emotional support. They furnish crucial implicit information regarding emotional states, intentions, subsequent events, and the scope of dialogue context that can elicit the desired response in the conversation. In-depth experiments on the EmpatheticDialogues and Emotional Support Conversation datasets (Rashkin et al., 2019; Liu et al., 2021) demonstrate the superiority of *Sibyl* over competitive categories of commonsense knowledge when applied to LLMs under multiple settings.

In summary, our contributions are as follows:

- We concentrate on addressing the inadequacy of current commonsense inference in anticipating dialogue future. Due to the one-to-many problem, the existence of multiple commonsense knowledge related to a single context potentially confuses LLMs, leading them to inadvertently ignore the goals of achieving empathy and providing emotional support.
- We propose *Sibyl*, an innovative paradigm that encompasses psychological, emotional, and causality factors in commonsense inference, which is pertinent to dialogue future.
- Extensive experiments demonstrate the effectiveness of our paradigm and detailed analysis validates the effectiveness of our method under multiple scenarios, showing significant improvements in automated metrics and evaluations by human and powerful LLM assessors.

2 Related Work

Empathy refers to the capacity to anticipate and understand the reactions of others (Keskin, 2014). Early studies concentrated on producing empathetic dialogues by leveraging the Speaker's emotional signals (Lin et al., 2019; Majumder et al., 2020) within the EMPATHETICDIALOGUES dataset (Rashkin et al., 2019). To enhance the ability to understand, perceive, and respond appropriately to the situation and feelings of others, commonsense knowledge is widely incorporated into empathetic chatbots (Sabour et al., 2021; Li et al., 2020; Wang et al., 2022; Zhou et al., 2023). Recently, several research efforts have explored the application of LLMs in generating empathetic responses within a prompt-based framework revealing the limitations of LLMs in accomplishing this task (Zhao et al., 2023b; Qian et al., 2023).

Empathy has also been related to several other variables such as helping, introversion, and affiliative tendency (Chlopan et al., 1985). **Emotional Support Conversation** is a benchmark focusing on exploring the problem of help seekers and generating more supportive responses. COMET (Bosselut et al., 2019), a pre-trained generative commonsense reasoning model is employed to obtain commonsense knowledge of the dialogue (Tu et al., 2022;

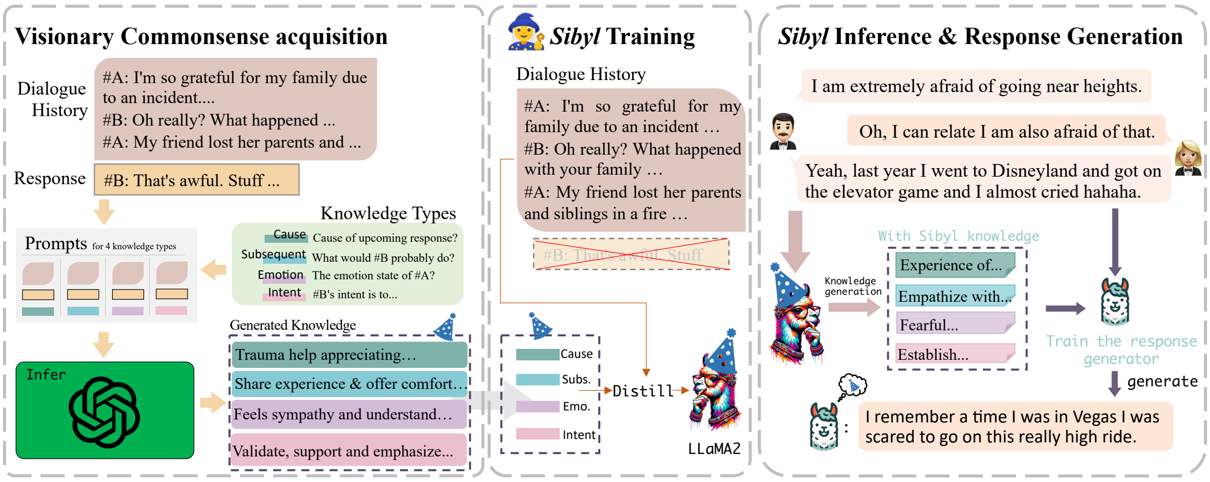

Figure 2: The overview of our proposed paradigm of Commonsense Inference, *Sibyl*. Incorporating both dialogue history and ground truth responses, the powerful LLM first deduces four categories of visionary commonsense. These inferences serve as a guiding oracle, aiding LLaMA2 models in inferring from dialogue history alone during the training stage. Subsequently, these trained models function as experts in inferring four categories of commonsense knowledge.

Peng et al., 2022; Deng et al., 2023). However, in the absence of harmonious knowledge selection, external information might trigger logical conflicts in dialogue (Yang et al., 2022; Wang et al., 2022).

Commonsense knowledge plays a vital role in dialogue systems, with many studies focusing on improving the techniques for acquiring commonsense knowledge. Ghosal et al. (2022); Shen et al. (2022) train language models to produce context-aware commonsense knowledge through natural language generation (NLG) and multi-choice answer selection (MCQ) tasks. This advances the application of commonsense knowledge in dialogue for further research. Recently, numerous research indicates that commonsense reasoning, obtained via **multi-step** methodologies, markedly surpasses the strategy of prompting LLMs to concurrently deduce implicit information and generate responses (Wang et al., 2023; Santra et al., 2023). By appending commonsense knowledge to the dialogue context (Wang et al., 2023; Chae et al., 2023), these inferences of dialogue context serve as intermediate reasoning to trigger LLMs analysis and compose high-quality responses.

3 Preliminaries

3.1 Problem Formulation

In the task of dialogue response generation, we employ θ to signify a dialogue model, while $C = [u_1, u_2, ..., u_{n-1}]$ indicates the context utterances, and K corresponds to commonsense knowledge. The objective here is to predict the forthcoming response Y based on the given context C from the $n-1$ turn, supplemented with the external commonsense knowledge K.

$$Y \sim P_\theta(\cdot \mid K, C) \tag{1}$$

3.2 Categories of Commonsense Inference

This study incorporates four categories of commonsense inferences within dialogues, which include: 1) **Cause**: Identifying the possible cause in the dialogue history for the forthcoming response. 2) **Subsequent Event**: Events that might take place in the dialogue future. 3) **Emotion state**: The user's emotional state as indicated in their latest utterance. 4) **Intention**: The probable dialogue intent behind the assistants next response. The overarching goal is to enrich our understanding of the dialogue history and meticulously project potential traits of the possible upcoming responses. These inferences operate as crucial intermediate reasoning steps that assist language models in enhancing dialogue comprehension and producing empathetic and supportive responses, with further details in Appendix A.

4 Method

In this section, we propose a novel paradigm for obtaining visionary commonsense knowledge, named *Sibyl*, as demonstrated in Figure 2.

4.1 Visionary Commonsense acquisition

The advanced LLMs which are aligned with human intention, exhibit robust logical deduction abilities. Initially, we utilize ChatGPT (*gpt-3.5-turbo*) (OpenAI, 2022) to generate four categories of plausible commonsense inferences $\mathcal{K}$, using inputs that include dialogue history $\mathcal{C}$ and the response $\mathcal{Y}$. We randomly selected a sample as demonstration to guide the powerful LLM in generating a visionary commonsense inference, considering dialogue history and response.

$$K = \arg\max_{K} P_{LLM}(\mathcal{C}; \mathcal{Y}) \quad (2)$$

The details of prompt templates are illustrated in Appendix B.1. To confirm the reasonableness of the four knowledge categories, we employ five highly educated postgraduates to perform a binary evaluation on 400 randomly chosen samples of commonsense knowledge. The average scores for the knowledge categories are all exceeding **0.87**[2]. To prevent information leakage, all dialogue samples mentioned in this section are sourced exclusively from the training sets.

4.2 *Sibyl* Training

To independently generate visionary commonsense inferences based on dialogue history, we further undertake Supervised Finetuning (SFT) of open-source LLMs to learn how to cultivate their prophetic abilities. Given the constraints of computational resources, we opt for LLaMA2-7B as visionary models.

Prompts of LLMs are carefully designed as hints to guide these models to understand the purpose of performing commonsense inference. Similar to prompting LLMs to generate oracle commonsense inference, we describe the aim of deducing a certain aspect of commonsense knowledge first and give one example of dialogue for tunable Language Models to grasp the demand of reasoning implicitly. Inspired by instruction tuning, the final template of our input consists of 1) Task Definition and instruction; 2) Examples and Answers; and 3) Dialogue context to be inferred.

The training loss is the standard negative log-likelihood (NLL) loss on the commonsense knowledge inferred by LLMs:

$$\mathcal{L}_{Infer} = -\sum_{m=1}^{M} log(P(k_m|C, k_{<m})) \quad (3)$$

where M is the length of commonsense inference generated by powerful LLMs, $K = [k_1, ..., k_M]$.

4.3 *Sibyl* Inference and Response Generation

After the training phase of visionary language models, we apply these models to deduce the mentioned four categories of commonsense knowledge focusing on dialogue future. Notably, differing from the process outlined in Sec. 4.1, these aspect-specialized models are presented with input that encompasses **solely the dialogue history**. In other words, they are trained to anticipate the imminent dialogue future, under the instruction of powerful LLMs that possess prior knowledge about the possible response.

Denoted as Ψ, these well-trained models are capable of analyzing causality, psychology, subsequency, and intent aspects of unseen conversations. In practice, we take the prompt C_p as the input of models Ψ, and we obtain four types of visionary commonsense inference $\mathcal{K}_p$.

$$C_p = Prompt_{template}(C) \quad (4)$$

$$\mathcal{K}_p = \Psi(C_p) \quad (5)$$

Where C indicates dialogue context, the prompt template is detailed in Appendix B.2, which is consistent with the template used in the training stage, as mentioned in Sec. 4.2.

Response Generation. For response generation, we append all four categories of visionary commonsense inferences $\mathcal{K}_p$ to the corresponding context to compose the input of LLMs. These inferences act as a bridge between dialogue history and the next response, aiding the foundation models to envision the future based on these cues for the probable response.

We conduct experiments using two strategies for creating the response generator: a finetuned approach and a prompt-based approach using LLMs. The finetuned approach involves two prominent open-source models: LLaMA2-7B (Touvron et al., 2023), and *Flan-t5-xl* (Chung et al., 2022). Standard NLL loss is adopted for the ground truth response Y during the finetuning process:

$$\mathcal{L}_{gen} = -\sum_{g=1}^{G} log(P(y_g|C; \mathcal{K}_p, y_{<g})) \quad (6)$$

[2] The Fleiss's **Kappa** measure among annotators stands at 0.52, signifying a moderate level of agreement.

where G stands for the length of the ground truth response of the dialogue, y_g specifies the g-th token in target response Y.

In the prompt-based approach, we directly engage an LLM to generate the subsequent response. The prompt provided to the LLM includes the dialogue history C, along with the four types of commonsense inferences $\mathcal{K}_p$.

5 Experimentals

5.1 Datasets

Our experiments are conducted on the EMPATHETICDIALOGUES (Rashkin et al., 2019) (ED) and the Emotional Support Conversation (Liu et al., 2021) (ESConv). ED is a vast multi-turn dialogue dataset encompassing 25,000 empathetic conversations between a speaker and a listener. ESConv comprises approximately 1,053 multi-turn dialogues between a help seeker experiencing emotional distress and a professional supporter.

5.2 Implementation Details

For the implementation of finetuning LLaMA2-7B and *Flan-t5-xl* models, we utilize the open-source Hugging Face transformers (Wolf et al., 2020). Due to the constraints on GPU resources, we employ LoRA-Tuning for training the LLaMA2-7B models. In terms of LoRA-Tuning, the LoRA's rank is set as 8, the $alpha$ is 16, the dropout rate of LoRA is assigned to 0.05, and the target modules are Q and V. We set the learning rate to 3e-5 and training batch size to 16, train up to 5 epochs, and select the best checkpoints based on performance on the validation sets. The whole model is optimized with the Adam (Kingma and Ba, 2015) algorithm. All of the experiments are performed on a single NVIDIA A800 GPU.

5.3 Baseline Methods

We compare *Sibyl* with several state-of-the-art methods and commonsense knowledge deduced by other baseline frameworks:

CASE (Zhou et al., 2023): A model trained from scratch with vanilla transformers (Vaswani et al., 2017) on ED dataset. This work utilizes a conditional graph to represent all plausible causalities between the user's emotions and experience.

M-Cue CoT (Wang et al., 2023): A multi-step prompting mechanism to trace the status of users during the conversation, performing complex reasoning and planning before generating the final response.

LLaMA2 (Touvron et al., 2023): To test the performance of vanilla open-source foundation models, we apply LLaMA2-7B[3] which only responds based on dialogue context.

+ COMET (Bosselut et al., 2019): A foundation model enhanced by external knowledge comes from ATOMIC (Hwang et al., 2021) which makes inferences based on the last utterance of context.

+ DOCTOR (Chae et al., 2023): A dialogue Chain-of-Thought commonsense reasoner which integrates implicit information in dialogue into rationale for generating responses.

+ DIALeCT (Shen et al., 2022): Trained on a variety of dialogue-related tasks, DIALeCT is a pretrained transformer for commonsense inference in dialogues which expert in leveraging the structural information from the dialogues.

5.4 Automatic Evaluation

The generated responses are evaluated using several automatic metrics, namely BLEU (Papineni et al., 2002), ROUGE-L (**ROU-L.**) (Lin, 2004), METEOR (**MET**) (Lavie and Agarwal, 2007), Distinct-n (**Dist-n**) (Li et al., 2016), and **CIDEr** (Vedantam et al., 2015). Additionally, we employ Average (**Ave.**) and Extrema (**Ext.**) Cosine Scores to assess embedding-based semantic similarity.

Supervised Finetuning (SFT) plays a crucial role in applying LLMs to specific tasks. Our approach significantly outperforms the mentioned baseline methods in generating **empathetic** responses on both Decoder-Only and Encoder-Decoder models (LLaMA2 and *Flan-t5*). As shown in the upper portion of Table 1, the similarity scores (**BLEU-n**, **ROU-L**. and **MET.**) of responses generated by LLaMA2-7B enhanced with *Sibyl* exceed those of all baseline methods by a significant margin, suggesting that the more sensible responses stem from the paradigms ability to deduce commonsense knowledge. However, for extrema score (**Ext.**), *Sibyl* performs slightly worse than the baselines. Equipped with *Sibyl*, LLaMA2 excels in achieving the highest scores in both average embedding similarity (**Avg.**) and **CIDEr**, further proving its effectiveness in empathetic response generation. The performance of the Finetuned model on *Flan-t5-xl*, as depicted in Table 3, additionally shows significant improvement when enhanced by *Sibyl*, espe-

[3]The version of LLaMA2 used in this paper: https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

Generation Paradigm	Model	BLEU-1/2/3/4	Dist-1/2/3	ROU_L.	MET.	Ave.	Ext.	CIDEr
Finetuned	CASE	15.99/7.41/3.90/2.29	0.64/3.02/5.98	18	7.77	87	**51.02**	18.12
	LLaMA2	16.8/5.94/2.67/1.38	**5.63/36.57/72.06**	15.09	7.59	87.3	48.05	13.72
	+ COMET	17.34/6.3/2.86/1.53	5.59/35.83/70.74	15.21	7.69	87.26	48.35	14.38
	+ DOCTOR	17.37/6.26/2.85/1.50	5.57/35.80/70.91	15.09	7.5	86.95	48.2	13.51
	+ DIALeCT	19.56/7.98/4.07/2.37	5.52/35.98/70.80	17.33	8.55	87.66	49.77	22.19
	+ *Sibyl*	**21.34/9.25/4.89/2.84***	5.61/36.07/71.17	**19***	**9.54***	**88.29**	50.85	**26.89***
Prompt-based	GPT-3.5	14.08/4.91/2.20/1.11	9.14/39.29/62.85	14.67	8.75	88.79	45.27	8.01
	+ M-Cue CoT	13.01/4.32/1.89/0.95	9.30/39.78/62.47	13.99	8.9	88.86	44.75	4.8
	+ COMET	14.07/5.06/2.43/1.34	9.36/40.13/64.12	14.89	9.13	88.94	45.69	7.54
	+ DOCTOR	14.43/5.34/2.63/1.48	9.68/**41.92/64.40**	15.65	9.3	89.29	46.24	8.38
	+ DIALeCT	15.36/5.67/2.64/1.39	8.98/38.07/60.13	16.23	9.46	89.29	47.47	10.48
	+ *Sibyl*	**16.20/6.43/3.21/1.81***	**9.70**/39.86/62.69	**17.62***	**10.05***	**89.8**	**47.99***	**14***

Table 1: Automatic Evaluation results on EMPATHETICDIALOGUES dataset. The best results are highlighted with **bold**. "*" denotes that the improvement to the best baseline is statistically significant (t-test with p-value < 0.01).

Generation Paradigm	Model	BLEU-2/3/4	Dist-1/2/3	ROU_L.	MET.	Ave.	Ext.	CIDEr
Finetuned	LLaMA2	6.73/2.9/1.4	6.24/40.34/75.6	15.62	9.02	88.44	44.6	8.32
	+ COMET	6.48/2.78/1.35	6.22/39.81/75.18	15.58	**9.04**	89.19	45	9.34
	+ DOCTOR	6.58/2.83/1.39	6.68/41.32/75.82	15.78	8.23	89.24	45.04	9.64
	+ DIALeCT	6.78/2.79/1.29	6.35/40.46/76.29	16.02	8.22	88.25	44.86	10.44
	+ *Sibyl*	**6.97/3.04/1.52***	**6.84/41.59/76.41***	**16.23**	8.53	**89.55***	**45.86**	**10.92***
Prompt-based	GPT-3.5	5.06/2.01/0.93	6.43/31.39/56.38	14.86	8.5	90.14	41.9	4.01
	+ M-Cue CoT	5.03/1.89/0.92	6.32/30.97/55.78	14.99	9.27	89.76	42.43	**4.92**
	+ COMET	5.06/1.99/0.91	5.98/29.56/52.89	14.87	9.44	90.66	**42.98**	4.14
	+ DOCTOR	4.46/1.72/0.79	6.36/31.76/56.48	13.98	8.73	90.24	40.93	3.39
	+ DIALeCT	4.95/1.82/0.81	6.42/31.14/54.24	14.97	9.1	90.6	42.56	4.15
	+ *Sibyl*	**5.19/2.21/1.10***	**6.52/32.09/56.72**	**15.2***	**9.65**	**90.7***	41.9	4.85

Table 2: Automatic Evaluation results on ESConv dataset. The best results are highlighted with **bold**. "*" denotes that the improvement to the best baseline is statistically significant (t-test with p-value < 0.01).

cially in the areas of overlap and embedding similarity scores. Impressively, the **CIDEr** score improvement of our method over the standard model by about 13 points highlights the critical role of anticipating dialogue futures and the distinct effectiveness of our proposed paradigm.

In the context of ESConv, we compared *Sibyl* paradigm to the baseline methods for commonsense knowledge. As shown in Table 2, *Sibyl* enhances foundation models' performance in emotional support scenarios. With *Sibyl* integration, LLMs outshine all other categories of commonsense knowledge under diversity metrics (**Dist-n**), underscoring the critical role of prophetic abilities in response generation.

Given that In-context Learning (ICL) is widely regarded as a key strength of Large Language Models (LLMs), our study assesses the effect of various commonsense inferences on LLMs' response generation without finetuning (Prompted-based). We mainly selected *gpt-3.5-turbo* from OpenAI's API as our LLM base. As outlined in the lower part of Table 1 and Table 2, the diversity scores of the content of our methodology generated are competitive with baselines and markedly superior in other metrics for empathetic dialogues. In the realm of emotional support, *Sibyl* catalyzes LLMs potential to provide empathetic and supportive responses. Through our proposed visionary commonsense inference, LLMs attain scores in Extrema (**Ext.**) and **CIDEr** that are on par with the best, while exceeding baseline models in all other diversity-driven and overlapping metrics. Superior performance under the setting of ICL underscores the effectiveness of our response-focused paradigm and demonstrates the viability of employing this commonsense knowledge as Chain-of-Thoughts in dialogue generation.

Model	BLEU-3/4	ROU_L.	MET.	Ave.	CIDEr
Flan-t5-xl	5.82/3.78	20.73	8.92	88.35	30.44
+ COMET	2.49/1.29	14.96	7.05	86.82	12.92
+ DOCTOR	2.58/1.33	14.78	6.97	86.92	23.41
+ DIALeCT	3.90/2.26	17.17	8.03	87.61	13.16
+ *Sibyl*	**7.71/5.24**	**23.09**	**10.39**	**88.53**	**43.36**

Table 3: Automatic Evaluation results on EMPATHETICDIALOGUES dataset. The foundation model is Flan-t5-xl. The best results are highlighted with **bold**.

Comparisons	Aspects	Win	Lose	Tie
+ *Sibyl* vs. CASE	Coh.	**53.2**	5.4	41.4
	Emp.	**41.7**	12.6	45.7
	Inf.	**46.4**	5.4	48.2
+ *Sibyl* vs. LLaMA2	Coh.	**19.3**	15.6	65.1
	Emp.	**30**	16.6	53.4
	Inf.	**21.8**	21.4	56.8
+ *Sibyl* vs. + COMET	Coh.	**19.8**	15	65.2
	Emp.	**25.2**	21.2	53.6
	Inf.	**24.9**	23.8	51.3
+ *Sibyl* vs. + DOCTOR	Coh.	**30.2**	6.4	63.4
	Emp.	**31.7**	8.5	59.8
	Inf.	**46.7**	32.6	20.7
+ *Sibyl* vs. + DIALeCT	Coh.	**17.1**	7.6	75.3
	Emp.	**49.4**	29.3	21.3
	Inf.	**40.7**	26.8	32.5

Table 4: Human A/B test (%) of EMPATHETICDIALOGUES. The inter-annotator agreement is evaluated by Fleiss's **Kappa** (denoted as κ), where $0.4 < \kappa < 0.6$ indicates moderate agreement.

Comparisons	Aspects	Win	Lose	Tie
+ *Sibyl* vs. LLaMA2	Flu.	**27.2**	18.4	54.4
	Com.	**28.5**	20.3	51.2
	Sup.	**32.5**	29.5	38
	All.	**36.7**	30.2	33.1
+ *Sibyl* vs. + COMET	Flu.	**23.5**	17.2	59.3
	Com.	**31.9**	24.3	43.8
	Sup.	**31.3**	28.6	40.1
	All.	**38.7**	29.9	31.4
+ *Sibyl* vs. + DOCTOR	Flu.	**51.3**	29.8	18.9
	Com.	**54.2**	31.8	14
	Sup.	**45.6**	37.7	16.7
	All.	**56.4**	37.2	6.4
+ *Sibyl* vs. + DIALeCT	Flu.	**13.5**	10	76.5
	Com.	**51.5**	40.1	8.4
	Sup.	**53.3**	33.8	12.9
	All.	**47.6**	28.2	24.2

Table 5: The human A/B test results for ESConv (%). **Kappa** (κ) fall between 0.4 and 0.6, suggesting moderate agreement.

5.5 Human Interactive Evaluation

The human evaluation on the ED dataset adheres to methodologies established in prior studies (Sabour et al., 2021; Wang et al., 2022), conducting a human evaluation based on three aspects 1) *Coherence* (**Coh.**): which models response is more coherent and relevant to the dialogue context? 2) *Empathy* (**Emp.**): which model has more appropriate emotional reactions, such as warmth, compassion, and concern? *Informativeness* (**Inf.**): which models response incorporates more information related to the context? In the realm of ESConv, we consider four aspects: 1) *Fluency* (**Flu.**): Evaluating the models based on the fluency of their responses. 2) *Comforting* (**Com.**): Assessing the models' skill in providing comfort. 3) *Supportive* (**Sup.**): Determining which model offers more supportive or helpful responses. 4) *Overall* (**All.**): Analyzing which model provides more effective overall emotional support.

We randomly select 200 dialogue samples and engage five professional annotators to evaluate the responses generated by finetuned LLaMA2-7B models for both the ED and ESConv datasets. Considering the variation between individuals, we conduct human A/B tests to compare our paradigm with other baselines directly. Annotators score the questionnaire of the response pairs to choose one of the responses in random order or select "Tie" when the quality of those provided sentences is difficult to distinguish. Fleiss's **kappa** is employed to analyze the evaluations. Table 4 demonstrates *Sibyl*'s significant advantage over CASE across all metrics. Compared to commonsense inference obtained from COMET, DOCTOR, and DIALeCT, our paradigm exhibits considerable progress, highlighting our approach's effectiveness in incorporating commonsense knowledge. These comparisons emphasize our paradigm's superior performance compared to the three baseline commonsense knowledge. Similarly, results from Table 5 strongly highlight the effectiveness of *Sibyl* within emotional support scenarios. The considerable lead in the overall score over the baselines indicates a more substantial influence, demonstrating the greater supportiveness of the knowledge, acting as cues that guide LLMs to be more helpful.

5.6 Ablation Study

To assess the influence of different categories of commonsense knowledge on response generation, we systematically remove each of these four categories of commonsense knowledge to facilitate a performance comparison on the ED dataset with *Sibyl*, as illustrated in Table 6. Excluding any of the four commonsense knowledge categories leads to a reduction in the quality of the generated response. Although some variants perform better than the complete method in particular metrics, the overall performance shows a notable decrease. Clearly, the causality of the conversation holds less significance in the generation of empathetic responses, whereas emotional cues provide greater insight into future information for understanding the user's situation.

Model	BLEU-1/2/3/4	Dist-1/2/3	ROU_L.	MET.	Ave.	Ext.	CIDEr
+ *Sibyl*	**21.34/9.25/4.89/2.84**	**5.61/36.07/71.17**	**19**	**9.54**	**88.29**	50.85	**26.89**
w/o Cause	20.89/9.06/4.78/2.78	5.35/34.52/68.48	18.69	9.38	88.01	**50.9**	25.87
w/o Intent	18.72/7.05/3.35/1.82	5.29/33.67/67.44	16.18	8.17	87.34	49.12	16.46
w/o Subs	20.69/8.89/4.66/2.71	5.37/34.16/67.91	18.23	9.2	87.83	50.45	24.39
w/o Emo	21.18/9.12/4.79/2.74	5.41/34.47/68.4	18.63	9.25	87.92	50.82	25.35

Table 6: Ablation study on the ED dataset.

	ED			ESConv		
	Nat.	**Emp.**	**Coh.**	**Nat.**	**Sup.**	**Coh.**
CASE	2.053	1.539	1.995	-	-	-
MultiESC	-	-	-	2.092	1.23	1.812
LLaMA2	2.512	1.849	2.635	2.332	1.376	2.214
+ COMET	2.464	1.747	2.646	2.368	1.944	2.465
+ DOCTOR	2.503	2.088	2.653	2.349	1.408	2.496
+ DIALeCT	2.441	1.115	2.644	2.381	1.867	2.526
+ *Sibyl*	**2.568**	**2.396**	**2.774**	**2.387**	**1.958**	**2.599**

Table 7: LLMs based Evaluation results on EPATHETICDIALOGUES (ED) and ESConv dataset under Supervised Finetyuning.

Furthermore, the conspicuous disparity between the variant (*w/o intent*) and our proposed complete method highlights the importance of predicting the potential intent of future responses, aligning with earlier studies (Chen et al., 2022; Wang et al., 2022).

5.7 LLMs-based Evaluation

We apply G-Eval (Liu et al., 2023; Chiang and yi Lee, 2023) to assess the Naturalness (**Nat.**) and Coherence (**Coh.**) of responses from baseline approaches that utilize commonsense knowledge in diverse ways. For task-specific requirements, we compare Empathy (**Emp.**) in the context of EMPATHETICDIALOGUES and Supportiveness (**Sup.**) for ESConv. Strictly following the rating strategy (Liu et al., 2023; Chiang and yi Lee, 2023), we prompt *gpt-4-0314* to discretely rate 1 to 3 points to these generated responses. Specifically, we require the LLMs to rate 1 when the generated response fails to meet a certain aspect. Rating a '2-point' means the response is totally ok, and meets the certain requirement to some extent. For responses that actually meet the desired demands, LLM is asked to give a '3-point' rating.

Notably, we prompt LLM to first explain/analyze before rating the target response for better correlation for human ratings (Chiang and yi Lee, 2023). From each of the ED and ESConv datasets, we randomly selected 200 data samples to conduct the G-Eval evaluation. Calculating the average weighted score of sampled data, the comparison result is shown in Table 7 and Table 8, *Sibyl* outperforms all strong baseline of commonsense inference in all aspects. Notably, in terms of Empathy (**Emp.**) and supportiveness (**Sup.**) scores, *Sibyl* significantly outpaces other commonsense knowledge frameworks and models under finetuned generators.

5.8 Case Study

To better evaluate the performance of response generation, we selected an example generated by our proposed paradigm and baselines for comparison. The example in Table 9 demonstrates that baseline models employing COMET and DIALeCT to derive commonsense knowledge struggled to identify the future direction of the dialogue. Although DOCTOR was able to partially recognize the potential information about the future to some extent, these three kinds of inferences still led to responses that were deficient in coherence and empathy. In contrast, *Sibyl* concentrates on crucial information, such as the possibility of the speaker having regular interactions with children. The visionary red-highlighted words accurately identify this detailed information, leading to a more sensible and suggestive response.

6 Conclusion

Even when enhanced with commonsense knowledge, LLMs still struggle with providing sensible and empathetic responses when providing support. This paper posits that the underlying issue stems from the one-to-many nature of dialogue generation and commonsense inference. We introduce a novel paradigm named *Sibyl*, highlighting the critical role of anticipating future information and distilling the visionary abilities of powerful LLMs into small tunable models. Through rigorous evaluation, *Sibyl* has proven its superiority, marked by notable improvements in automated metrics and assessments conducted by human evaluators and advanced LLMs.

Limitations

In this paper, we explore a new paradigm of acquiring visionary commonsense knowledge named *Sibyl*. However, we acknowledge the limitations of this work from the following perspectives:

Shortage of data. One of the limitations of our work stems from the shortage of datasets in the task of empathetic and emotional support dialogue generation. Although these two benchmarks have been proposed for a long time, most of the research also focuses on these two datasets.

Evaluations. The scores of automatic evaluation metrics are not fully consistent with human evaluations for the tasks of dialogue generation, as depicted by Liu et al. (2016). Employing LLMs as professional assessors alleviates the problem of the lack of labour-free and task-specific evaluation metrics. However, these approaches (Liu et al., 2023; Chiang and yi Lee, 2023; Fu et al., 2023) can only regarded as a reference, the usage of human evaluation metric still takes the most cathedratic place. Therefore, there still exists trouble evaluating the empathy and supportiveness of the generated content automatically and convincingly. To address this, we employ all three aforementioned methods to thoroughly assess the response, aiming to validate the efficacy of our proposed approach.

Ethics Statement

The datasets (Rashkin et al., 2019; Liu et al., 2021) utilized in our study are widely recognized and sourced exclusively from open-source repositories. The conversations of the ED dataset are around given emotions and carried out by employed crowdsourced workers, with no personal privacy issues involved. For our human evaluation, all participants were volunteers provided with comprehensive information about the researchs purpose, ensuring informed consent. Moreover, participants were provided with fair and appropriate compensation for their involvement. The call of the OpenAI API for this paper was conducted during a period when the authors were on vacation in Singapore.

References

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4762–4779. Association for Computational Linguistics.

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.

Hyungjoo Chae, Yongho Song, Kai Tzu iunn Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. 2023. Dialogue chain-of-thought distillation for commonsense-aware conversational agents.

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, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1063–1074. Association for Computational Linguistics.

Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3014–3026. Association for Computational Linguistics.

Cheng-Han Chiang and Hung yi Lee. 2023. A closer look into automatic evaluation using large language models.

Bruce E Chlopan, Marianne L McCain, Joyce L Carbonell, and Richard L Hagen. 1985. Empathy: Review of available measures. *Journal of personality and social psychology*, 48(3):635.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

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

Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. 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 4079–4095. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *CoRR*, abs/2302.04166.

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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5010–5028. Association for Computational Linguistics.

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 *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6384–6392. AAAI Press.

Sevgi Coşkun Keskin. 2014. From what isnt empathy to empathic learning process. *Procedia-Social and Behavioral Sciences*, 116:4932–4938.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation, WMT@ACL 2007, Prague, Czech Republic, June 23, 2007*, pages 228–231. Association for Computational Linguistics.

Jiangnan Li, Fandong Meng, Zheng Lin, Rui Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. 2022. Neutral utterances are also causes: Enhancing conversational causal emotion entailment with social commonsense knowledge. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4209–4215. ijcai.org.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119. The Association for Computational Linguistics.

Qintong Li, Piji Li, Zhumin Chen, and Zhaochun Ren. 2020. Empathetic dialogue generation via knowledge enhancing and emotion dependency modeling. *CoRR*, abs/2009.09708.

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, Hong Kong, China. Association for Computational Linguistics.

Chang Liu, Xu Tan, Chongyang Tao, Zhenxin Fu, Dongyan Zhao, Tie-Yan Liu, and Rui Yan. 2022. Prophetchat: Enhancing dialogue generation with simulation of future conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 962–973. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian 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. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2122–2132. The Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie

Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3469–3483. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using GPT-4 with better human alignment. *CoRR*, abs/2303.16634.

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, Online. Association for Computational Linguistics.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. Accessed on January 10, 2023.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

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

Yushan Qian, Weinan Zhang, and Ting Liu. 2023. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6516–6528, Singapore. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2021. CEM: commonsense-aware empathetic response generation. *CoRR*, abs/2109.05739.

Bishal Santra, Sakya Basak, Abhinandan De, Manish Gupta, and Pawan Goyal. 2023. Frugal prompting for dialog models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 4383–4407. Association for Computational Linguistics.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. MISC: A mixed strategy-aware model integrating COMET for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 308–319. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Hongru Wang, Rui Wang, Fei Mi, Zezhong Wang, Ruifeng Xu, and Kam-Fai Wong. 2023. Chain-of-thought prompting for responding to in-depth dialogue questions with LLM. *CoRR*, abs/2305.11792.

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. *CoRR*, abs/2210.11715.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

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, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4886–4899. International Committee on Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Chenxu Yang, Zheng Lin, Jiangnan Li, Fandong Meng, Weiping Wang, Lanrui Wang, and Jie Zhou. 2022. TAKE: topic-shift aware knowledge selection for dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 253–265. International Committee on Computational Linguistics.

Weixiang Zhao, Yanyan Zhao, Xin Lu, and Bing Qin. 2023a. Don't lose yourself! empathetic response generation via explicit self-other awareness. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13331–13344. Association for Computational Linguistics.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023b. Is chatgpt equipped with emotional dialogue capabilities? *CoRR*, abs/2304.09582.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 730–739. AAAI Press.

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. Association for Computational Linguistics.

Pei Zhou, Hyundong Cho, Pegah Jandaghi, Dong-Ho Lee, Bill Yuchen Lin, Jay Pujara, and Xiang Ren. 2022. Reflect, not reflex: Inference-based common ground improves dialogue response quality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 10450–10468. Association for Computational Linguistics.

A Four Categories of Commonsense Knowledge

We mainly employ four categories of commonsense knowledge of our proposed paradigm, which is as follows.

Cause *What is the cause of the assistant to post the last utterance?* We emphasize the crucial role of causality within the dialogue context. Similar to the approach outlined by Shen et al. (2022) and previous investigations (Li et al., 2022; Cheng et al., 2022), we delve into potential words or phrases that could lead to the desired response.

Subsequent Event *What will be the potential subsequent events involving the assistant that may occur after the user's last utterance?* Conversations demonstrate a causal connection between past utterances to the ensuing responses. Dialogues contain a cause-and-effect connection between the context and the target response. Following (Ghosal et al., 2022), we employ a language model to project potential scenarios that follow the dialogue history, which is a key factor in determining the assistant's response.

Emotion reaction *What is the emotional reaction of the user in their last utterance?* Emotion is a fundamental element in human conversation (Zhou et al., 2018), acting as a natural means for individuals to express their feelings during dialogues. With explicit emotion traits, it is easier for chatbots to grasp a more profound understanding of the dialogue and anticipate the potential emotional content within the target response.

Intention *What is the assistant's intent to post the last utterance according to the emotional reaction of the user?* Dialogue intention is a focal point in the realm of dialogue generation (Welivita and Pu, 2020). It comprises the underlying logic and objectives guiding the forthcoming conversation, thus forming a vital aspect in contextual understanding and response generation.

The above four categories of commonsense inference are all used in our paradigm, acting as intermediate reasoning steps for steering language models for better dialogue comprehension and more empathetic responses.

B Detailed Prompts

B.1 Prompts for Visionary Commonsense acquisition

The template input for prompting Large Language Models generating prophetic commonsense inference is as follows:

Given a dyadic dialogue clip between a listener and a speaker, the objective is to comprehend the dialogue and make inferences to identify the underlying cause of the latest utterance stated by the listener (the reason contributing to the utterance stated by the listener).

I will provide an example of a conversation clip and the explanation of causes, which is as follows:

(1)Speaker: Job interviews always make me sweat bullets, makes me uncomfortable in general to be looked at under a microscope like that.
(2)Listener: Don't be nervous. Just be prepared.
(3)Speaker: I feel like getting prepared and then having a curve ball thrown at you throws you off.
(4)Listener: Yes but if you stay calm it will be ok.

What is the cause of the listener to post the next response? Please make inferences based on the utterances before the last utterance of the conversation. Please generate the answer like this: Answer: The cause of the listener's last utterance is to reassure and encourage the speaker, emphasizing the importance of staying calm despite unexpected challenges during a job interview.

Now, generate one concise and relevant inference (no more than 40 words) of the cause of the last utterance. The conversation clip is:

{context}

What is the cause of the listener to post the next response?

Answer:

B.2 Prompts for *Sibyl* Training

The prompt we designed as hints to guide tunable models to understand the purpose of performing commonsense inference is as follows:

	ED			ESConv		
	Nat.	**Emp.**	**Coh.**	**Nat.**	**Sup.**	**Coh.**
GPT-4	2.19	2.171	**2.192**	1.838	1.983	1.713
+ COMET	2.188	2.176	2.188	1.842	1.979	1.712
+ DIALeCT	2.126	1.793	2.186	1.841	1.793	1.71
+ M-Cue CoT	2.189	1.792	2.124	1.841	1.982	1.716
+ *Sibyl*	**2.191**	**2.176**	2.191	**1.846**	**1.984**	**1.717**

Table 8: LLMs based Evaluation results on EPATHETICDIALOGUES (ED) and ESConv dataset under In-Context Learning.

1) Task Definition and instruction:
You are an expert in the theory of empathy and conversational contextual reasoning.
Given a dyadic dialogue clip between a listener and a speaker, the objective is to comprehend the dialogue and make inferences to identify the underlying cause of the latest utterance stated by the listener (the reason contributing to the utterance stated by the listener).
2) Example and Answers:
I will provide an example of a conversation clip and the explanation of causes, which is as follows:

{example}

What is the cause of the speaker to post the last utterance?
Please make inferences based on the utterances before the last utterance of the conversation. Please generate the answer like this: Answer: {example answer}.
3) Dialogue context to be inferred:
Now, generate one concise and relevant inference (no more than 40 words) of the cause of the last utterance.
The conversation clip is:

{context}

Answer:

At the training stage, we append the oracle commonsense inference generated by powerful LLMs to the prompt above.

C Details of LLMs-based evaluation

The absence of labor-free and practical evaluation metrics has been a persistent challenge within the field of NLP research. Thanks to the rise of LLMs, several studies have explored the utilization of LLMs in assessing content generated by neural models. (Fu et al., 2023) propose a direct approach, using LLMs as reference-free evaluators for Natural Language Generation (NLG), viewing the evaluation process as a probability calculation. Moreover, (Liu et al., 2023) and (Chiang and yi Lee, 2023) introduce a prompt-based framework for LLMs, ensuring adherence to the generated instructions and offering a more detailed continuous score by adjusting the discrete scores based on their token probabilities.

We apply G-Eval (Liu et al., 2023; Chiang and yi Lee, 2023) to assess the Naturalness (**Nat.**) and Coherence (**Coh.**) of responses from baseline models that utilize commonsense knowledge in diverse ways. For task-specific requirements, we compare Empathy (**Emp.**) in the context of EMPATHETICDIALOGUES and Supportiveness (**Sup.**) for ESConv. As the token probabilities of ChatGPT (OpenAI, 2022) are unavailable, we set '$n = 20, temperature = 1, top_p = 1$' to sample 20 times to estimate the token probabilities.

Strictly following the rating strategy (Liu et al., 2023), we prompt *gpt-4-0314* to discretely rate 1 to 3 points to these generated responses. Specifically, we require the LLMs to rate 1 when the generated response fails to meet a certain aspect. Rating a '2-point' means the response is totally ok, and meets the certain requirement to some extent. For responses that actually meet the desired demands, LLM is asked to give a '3-point' rating.

The results of the average weighted score are demonstrated in Table 7 and Table 8, *Sibyl* outperforms all strong baselines of commonsense inference in all aspects.

We randomly selected 200 data from both ED and ESConv datasets to perform G-Eval evaluation. Calculating the average weighted score of sampled data, the comparison result is shown in Table 7 and Table 8, *Sibyl* outperforms all strong baseline of commonsense inference in all aspects.

Dialogue Context	*Speaker:* Hi, I feel so lonely sometimes because all my friends live in a different country.
	Listener: Oh, I'm sure you are lonely. Maybe you can join some kind of club that lets you meet new friends?
	Speaker: I was thinking about it! I wanted to join a group for local moms.
	Response: That's a good idea! This way you can also meet friends for yourself, but also maybe meet new friends for your children to hang out with while you do with their moms!
+ COMET	That would be fab! Do you live in a big city or a small town?
+ DOCTOR	The speaker wants to join a club that allows them to meet new friends. The listener suggests that the speaker join a local club that would be suitable for their area.
+ DIALeCT	That would be great, I'm sure you will have great luck!
Visionary Commonsense	**Subsequent events**: The listener is likely to suggest specific activities or events that the speaker can participate in to meet new friends, showing a proactive and helpful approach to the conversation.
	Emotion state: The speaker feels hopeful and appreciates the listener's suggestion to join a group for local moms, as it aligns with their desire to meet new friends.
	Cause: The listener is motivated by empathy and the desire to offer practical solutions, encouraging the speaker to pursue social connections .
	Intent: To encourage the speaker, acknowledging the potential benefits of joining a group for local moms and expressing hope that it will lead to positive outcomes for both the speaker and their children.
+ *Sibyl* (Ours)	That would be a great idea. You can make friends for yourself and for your children.

Table 9: An example involving responses from different versions of LLaMA2 models which are enhanced with different commonsense knowledge. The words relating to commonsense knowledge are highlighted in red, while phrases in red signify the connection with knowledge and dialogue history.