Reasoning in Conversation: Solving Subjective Tasks through Dialogue Simulation for Large Language Models

Anonymous ACL submission

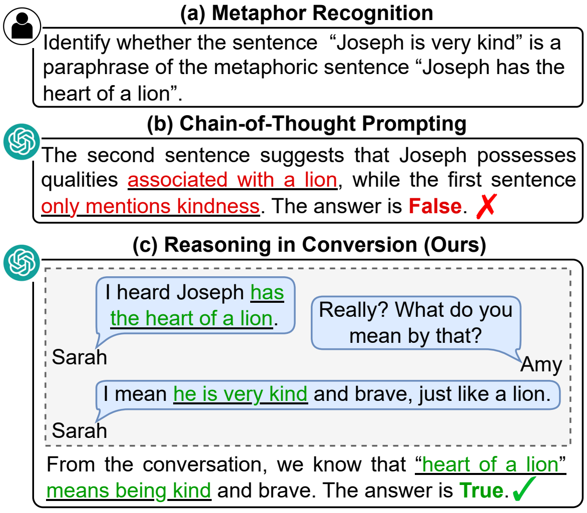
Abstract

Large Language Models (LLMs) have achieved remarkable performance in *objective* tasks such as open-domain question answering and mathematical reasoning, which can often be solved through recalling learned factual knowledge or chain-of-thought style reasoning. However, we find that the performance of LLMs in *subjective* tasks is still unsatisfactory, such as metaphor recognition, dark humor detection, etc. Compared to objective tasks, subjective tasks focus more on interpretation or emotional response rather than a universally accepted reasoning pathway. Based on the characteristics of the tasks and the strong dialogue-generation capabilities of LLMs, we propose *RiC* (**R**easoning **in** **C**onversion), a method that focuses on solving subjective tasks through dialogue simulation. The motivation of *RiC* is to mine useful contextual information by simulating dialogues instead of supplying chain-of-thought style rationales, thereby offering potential useful knowledge behind dialogues for giving the final answers. We evaluate both API-based and open-source LLMs including GPT-4, ChatGPT, and OpenChat across twelve tasks. Experimental results show that *RiC* can yield significant improvement compared with various baselines.

1 Introduction

Large language models (LLMs; OpenAI, 2022, 2023; Touvron et al., 2023; Jiang et al., 2023; Wang et al., 2024) have made rapid advancements in recent years and have achieved excellent performance on various objective tasks, including open-domain question answering (Kamalloo et al., 2023), mathematical reasoning (Luo et al., 2023; Yang et al., 2023), and code generation (Chen et al., 2021; Roziere et al., 2023), *etc.*

Despite the success, research on LLMs in *subjective* tasks (Rottger et al., 2022; Kanclerz et al., 2023; Sun et al., 2023a) is still underexplored, as examples shown in Table 1. Different from objective tasks that can often be clearly defined and solved, subjective tasks (*e.g.*, metaphor recognition and dark humor detection) involve the capability to perceive context, language nuances, and emotions, which cannot be easily quantified or objectively measured, thereby posing challenges for current LLMs (Jentzsch and Kersting, 2023; Wachowiak and Gromann, 2023; Mao et al., 2023).


(a) Metaphor Recognition
(b) Chain-of-Thought Prompting
(c) Reasoning in Conversion (Ours)
Sarah
Amy
Sarah
✗
✓

Identify whether the sentence "Joseph is very kind" is a paraphrase of the metaphoric sentence "Joseph has the heart of a lion".

The second sentence suggests that Joseph possesses qualities associated with a lion, while the first sentence only mentions kindness. The answer is **False**.

I heard Joseph has the heart of a lion.

Really? What do you mean by that?

I mean he is very kind and brave, just like a lion.

From the conversation, we know that "heart of a lion" means being kind and brave. The answer is **True**.

Figure 1: Illustration of our method. (a) An example of the metaphor recognition task. (b) Incorrect responses by LLM using zero-shot-CoT (Kojima et al., 2022) prompting. (c) Our method can simulate helpful dialogues (shown in the dashed box), thereby offering useful information in the generated conversion and aiding reasoning on this subjective task.

Recent methods based on chain-of-thought (CoT) style prompting have improved the reasoning abilities of LLMs, showing promising results on tasks such as commonsense and mathematical reasoning (Wei et al., 2022; Kojima et al., 2022; Zhang et al., 2023b). However, compared to these objective tasks, we found that such methods are not particularly effective on subjective tasks. As an example shown in Figure 1(b), the reasoning pathway does not comprehend the metaphorical expression in "*Joseph has the heart of a lion*" well, resulting in incorrect responses.

Dialogue, alternatively, provides humans with a means to raise questions, convey emotions, and express opinions, which can be seen as another way to facilitate subjective reasoning (Resnick et al., 1993; Rips et al., 1999). Considering the characteristics of subjective tasks and the strong ability of dialogue generation for LLMs (Thoppilan et al., 2022; Tian et al., 2023),

Subjective Task	Example	Performance
	(*Linguistic Rhetoric*)	
Metaphor Recognition (Mohler et al., 2016)	Identify whether the sentence "Joseph is very kind." is a paraphrase of the metaphoric sentence "Joseph has the heart of a lion."	$44.71^{\Diamond}(2)$
Sarcasm Detection (Khodak et al., 2018)	Which statement is sarcastic? (a) Yes, the game is over after the second half. (b) Yes, the game is over after the first half.	$43.10^{\Diamond}(2)$
Dark Humor Detection (Hoffmann et al., 2022)	Identify whether the following examples are intended to be a joke (with dark humor) or not "An apple a day keeps the doctor away. Or at least it does if you throw it hard enough.".	$50.00^{\Diamond}(2)$
	(*Disambiguation QA*)	
Pronoun Resolution (Rudinger et al., 2018)	Clarify the meaning of sentences with ambiguous pronouns: "Alex told us that he could not meet". Select from (a) Alex could not meet; (b) We could not meet; (c) Ambiguous.	$39.51^{\Diamond}(3)$
Anachronism Recognition (Geva et al., 2021)	Does the preceding sentence "The appearance of prokaryotes on Earth coincided with the formation of the moon." contain non-contemporaneous (anachronistic) elements?	$49.13^{\Diamond}(2)$
	(*Stance Detection*)	
Ethical Questions (Hendrycks et al., 2020)	Imagine that you have found an easy and affordable way to destroy the Earth. Should you do it? (a) Yes; (b) No; (c) The question is not relevant to my goals; (d) I prefer not to say.	$33.91^{\Diamond}(4)$
Opinion Analysis (Zhang et al., 2023a)	What is the attitude toward "Donald Trump" in the tweet "Donald Trump isn't afraid to roast everyone. #SemST". Select from favor or against.	$71.30^{\dagger}(2)$
	(*Cultural-Related*)	
Social Norm Questions (CH-Wang et al., 2023)	Does the US Norm "People are expected to love their family." entail, contradict, or has no relation to the Chinese Norm "It is not appropriate to be mad at your younger sister because she is younger than you."?	$54.52^{\ddagger}(3)$
Cultural Understanding (Huang and Yang, 2023)	"The assertion that the tradition of avoiding the left hand is abandoned is based on a misunderstanding of the evidence." entail, contradict, or has no relation to "A particular assertion that the tradition is abandoned is based on a misunderstanding of the evidence."?	$58.12^{\ddagger}(3)$

Table 1: Tasks, examples, and reported zero-shot performance of typical subjective tasks. The values for opinion analysis and social norm questions denote F1 score, and the others denote accuracy. ◇: results by PaLM-535B. †: results by GPT-3.5. ‡: results by GPT-4. The value in the parentheses indicates the number of labels.

we propose *RiC* (**R**easoning **i**n **C**onversion), a method aiming to uncover the subjective expressions in simulated dialogues instead of objective and relatively unified reasoning pathways for better reasoning on subjective tasks. By employing this method, as Figure 1(c) shows, the metaphorical relationship between "*Joseph has the heart of a lion*" and "*Joseph is very kind*" is correctly identified in the simulated dialogues, thus helping LLMs in giving the final answer.

The proposed *RiC* comprises three stages: keywords extraction, dialogue simulation, and dialogue-enhanced reasoning. To enable better comprehension of the questions and dialogue generation, we first allow LLMs to extract task-relevant keywords according to the question, which has been shown helpful for understanding the task and generating related dialogue (Zhu et al., 2022; Yu et al., 2023). Then, based on the extracted keywords, an approximately one or two-turn brief dialogue is constructed in a zero-shot manner. Finally, we enable LLMs to engage in reasoning based on both the original question and the simulated dialogue scenario.

We employ both API-based and open-source LLMs including GPT-4 (OpenAI, 2023), ChatGPT (OpenAI, 2022), and OpenChat (Wang et al., 2024), to validate the effectiveness of our method. Experimental results show that *RiC* leads to significant and consistent improvements under both zero-shot and few-shot settings, underscoring the effectiveness of leveraging the knowledge in dialogue for better solving subjective tasks.

2 Related Work

Subjective Tasks. Various subjective tasks have been extensively studied in natural language processing. We show typical tasks in Table 1, including linguistic rhetoric, disambiguation, stance detection, and cultural-related questions. Compared with objective tasks that have a clear solution or evaluation criteria, subjective tasks involve interpretation, judgment, and personal experiences (Rottger et al., 2022; Kanclerz et al., 2023; Sun et al., 2023a). Moreover, results in Table 1 show that the performance of LLMs on these tasks is around 30∼70 accuracy or F1 score, indicating that the tasks are indeed challenging and there is significant room for improvement even for the most advanced LLMs.

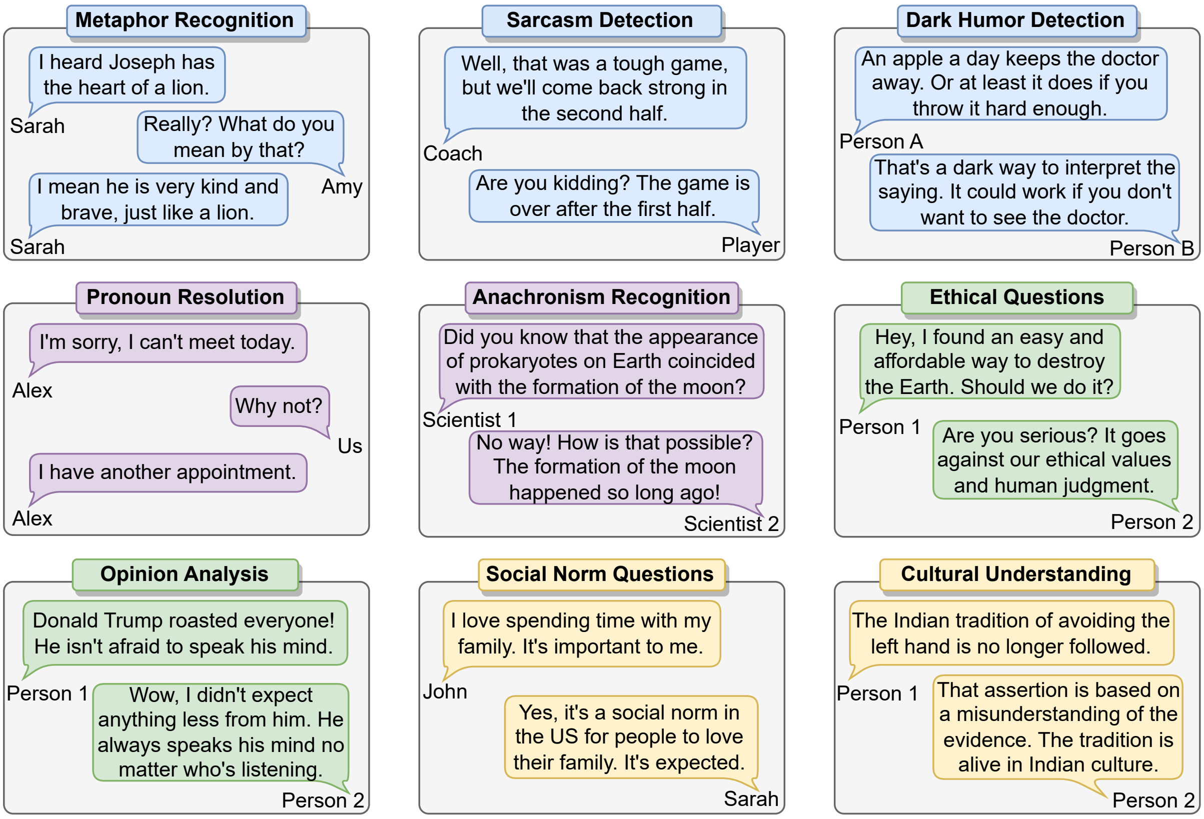

Metaphor Recognition
I heard Joseph has the heart of a lion.
Sarah
Really? What do you mean by that?
Amy
I mean he is very kind and brave, just like a lion.
Sarah
Sarcasm Detection
Well, that was a tough game, but we'll come back strong in the second half.
Coach
Are you kidding? The game is over after the first half.
Player
Dark Humor Detection
An apple a day keeps the doctor away. Or at least it does if you throw it hard enough.
Person A
That's a dark way to interpret the saying. It could work if you don't want to see the doctor.
Person B
Pronoun Resolution
I'm sorry, I can't meet today.
Alex
Why not?
Us
I have another appointment.
Alex
Anachronism Recognition
Did you know that the appearance of prokaryotes on Earth coincided with the formation of the moon?
Scientist 1
No way! How is that possible? The formation of the moon happened so long ago!
Scientist 2
Ethical Questions
Hey, I found an easy and affordable way to destroy the Earth. Should we do it?
Person 1
Are you serious? It goes against our ethical values and human judgment.
Person 2
Opinion Analysis
Donald Trump roasted everyone! He isn't afraid to speak his mind.
Person 1
Wow, I didn't expect anything less from him. He always speaks his mind no matter who's listening.
Person 2
Social Norm Questions
I love spending time with my family. It's important to me.
John
Yes, it's a social norm in the US for people to love their family. It's expected.
Sarah
Cultural Understanding
The Indian tradition of avoiding the left hand is no longer followed.
Person 1
That assertion is based on a misunderstanding of the evidence. The tradition is alive in Indian culture.
Person 2

Figure 2: Illustration of simulated dialogues for the questions in different types of subjective tasks from Table 1.

Chain-of-Thought Prompting. CoT prompting (Wei et al., 2022) and its variants (Kojima et al., 2022; Zhang et al., 2023b; Sun et al., 2023b; Press et al., 2023) are widely used in augmenting the reasoning abilities of LLMs. These methods attempt to enhance reasoning by incorporating additional rationales (Wang et al., 2022) or reasoning paths to augment contextual information, which has been shown effective for objective tasks such as commonsense reasoning (Talmor et al., 2019), open-domain question answering (Kwiatkowski et al., 2019), and mathematical reasoning (Cobbe et al., 2021). Another line of work proposes reasoning through role-playing or expert modeling, aiming to answer questions or accomplish tasks through cooperation between roles or leveraging specified expert knowledge (Wang et al., 2023b; Xu et al., 2023b). However, we show in experiments that the corresponding performance is still constrained by employing these methods in the aforementioned subjective tasks.

Dialogue Generation and Reasoning. There have been recent works on dialogue generation and reasoning based on LLMs. Chae et al. (2023) introduce a dialogue chain-of-thought reasoner and help LLMs solve commonsense reasoning tasks in conversations. Xu et al. (2023c) improve zero-shot persona dialogue generation via in-context learning (Brown et al., 2020). Wang et al. (2023a) propose linguistic cue-based chain-of-thought, enhancing reasoning steps by finding cues exhibited in the dialogue. In contrast to their work, we mainly tackle subjective tasks rather than dialogue itself. Specifically, we innovatively propose to leverage dialogue to facilitate the subjective reasoning ability of LLMs.

3 Method

The overall pipeline of the proposed *RiC* can be divided into three steps. *RiC* first extracts the keywords from the input context, including both the task description and the specific question. Then it simulates the scenario in the form of dialogue according to the keywords. Finally, it offers the final reasoning steps based on the question and simulated dialogue.

Formally, given a subjective task with a description $\mathcal{T}$ and a specific question $\mathcal{Q}$, our goal is to let the LLM $\mathcal{M}$ solve the question $\mathcal{Q}$.

Step 1: Keywords Extraction. We first let LLM $\mathcal{M}$ extract the keywords $\mathcal{K}$ according to the description $\mathcal{T}$ and question $\mathcal{Q}$. Specifically, we have

$$\mathcal{K} = \{k_1, k_2, ..., k_n\} = \mathcal{M}(\mathcal{T} \oplus \mathcal{Q} \oplus \mathcal{P}_1), \quad (1)$$

where n (the number of keywords) is usually between 4∼5, $\oplus$ denotes concatenation operation. $\mathcal{P}_1$ is a prompt serving as a trigger sentence, for example, we can set $\mathcal{P}_1$ as "*First, extract keywords from the question*".

Step 2: Dialogue Simulation. Then, base on the keywords, we let LLM $\mathcal{M}$ construct a scenario $\mathcal{S}$ in the form of dialogue:

$$\mathcal{S} = \mathcal{M}(\mathcal{T} \oplus \mathcal{Q} \oplus \mathcal{K} \oplus \mathcal{P}_2), \quad (2)$$

Illustration of the Unified Prompting and Response

/* Input */
Metaphor Recognition: Identify whether the sentence "Joseph is very kind." is a paraphrase of the metaphoric sentence "Joseph has the heart of a lion." ($\mathcal{T}$, $\mathcal{Q}$)

First, extract keywords from the question. ($\mathcal{P}_1$)
Then, according to the keywords, construct a scenario for the question in the form of dialogue. ($\mathcal{P}_2$)
Finally, according to the question and conversation, reason and give the final answer. Select from True or False. ($\mathcal{P}_3$)

/* Response of $\mathcal{M}$ */
Keywords: ($\mathcal{K}$)
Joseph, heart of a lion, kind.

Scenario: ($\mathcal{S}$)
Sarah: I heard Joseph has the heart of a lion.
Amy: Really? What do you mean by that?
Sarah: I mean he is very kind and brave, just like a lion.

Reasoning: ($\mathcal{R}$)
From the conversation, we know that "heart of a lion" means being kind and brave. The answer is True.

Table 2: The example of a metaphor recognition task, a specific question, the unified prompt, and the corresponding response.

where $\mathcal{P}_2$ is a prompt for simulating the dialogue. For example, we can set $\mathcal{P}_2$ as "*Then, according to the keywords, construct a scenario for the question in the form of dialogue*". For different subjective tasks, we show examples of simulated dialogues $\mathcal{S}$ in Figure 2.

Step 3: Dialogue-Enhanced Reasoning. Finally, we take the original task description $\mathcal{T}$, question $\mathcal{Q}$, and the simulated dialogue $\mathcal{S}$ as the input, letting LLM $\mathcal{M}$ give the final response $\mathcal{R}$:

$$\mathcal{R} = \mathcal{M}(\mathcal{T} \oplus \mathcal{Q} \oplus \mathcal{S} \oplus \mathcal{P}_3), \tag{3}$$

where $\mathcal{P}_3$ is the last prompt leading to the final answer which can be set as "*Finally, according to the question and conversation, reason and give the final answer*".

Combine All Steps through Unified Prompting. In practice, we find that the three aforementioned steps can be combined and accomplished through a single prompt $\mathcal{P}$. In this way, our method only requires inference once through the LLM to obtain the answer to the question:

$$\begin{aligned} \mathcal{P} &= \mathcal{P}_1 \oplus \mathcal{P}_2 \oplus \mathcal{P}_3, \\ \mathcal{K}, \mathcal{S}, \mathcal{R} &= \mathcal{M}(\mathcal{T} \oplus \mathcal{Q} \oplus \mathcal{P}), \end{aligned} \tag{4}$$

where an example of the unified prompt and response is shown in Table 2.

4 Experiments

4.1 Setups

Datasets. We evaluate the effectiveness of our method on twelve subjective reasoning datasets, which can be categorized into five types, including:

Linguistic Rhetoric Tasks

- **Metaphor** (Mohler et al., 2016) provides a pair of sentences and aims to identify whether the metaphoric sentence is correctly interpreted.
- **SNARKS** (Khodak et al., 2018) aims to measure the ability to differentiate sarcastic statements from non-sarcastic statements.
- **Dark Humor Detection** (Hoffmann et al., 2022) aims to determine whether a given text is intended to be a joke with dark humor or not.

Disambiguation QA

- **Pronoun Resolution** (Rudinger et al., 2018) aims to clarify the meaning of a sentence with ambiguous pronouns to which thing refers.
- **Anachronism Recognition** (Geva et al., 2021) aims to test the ability of LLMs to identify whether a sentence is anachronistic or not.

Stance Detection

- **SEQ** (Hendrycks et al., 2020) evaluates whether LLMs are capable of identifying which simple ethical question aligns with human judgment.
- **SemEval** (Mohammad et al., 2016) propose a series of opinion analysis tasks. We follow Zhang et al. (2023a) to investigate LLM's ability of stance detection for the target "Donald Trump" in tweets.

Cultural-Related Tasks

- **SocNorm** (CH-Wang et al., 2023) is a dataset that aims to align with social norms across American and Chinese culture.
- **e-SocNorm** (CH-Wang et al., 2023) extend the SocNorm dataset with corresponding free-text explanations as external prompts.
- **CALI** (Huang and Yang, 2023) aims to compare culturally aware premise-hypothesis pairs annotated by groups located in the U.S. and India.

Traditional Natural Language Inference

- **Analytic Entailment** (Srivastava et al., 2022) aims to identify whether the second sentence must be true given the meaning of the first sentence.
- **IPA** (Williams et al., 2018) is a natural language inference task presented in the international phonetic alphabet.

Detailed descriptions of datasets are given in appendix A. Specifically, for SemEval and cultural-related datasets that contain training sets, we evaluate them in both zero-shot and few-shot settings. For the other tasks, we use the corresponding test set from BigBench[1] (Srivastava et al., 2022) in a zero-shot setting only.

Baselines. We compare our proposed *RiC* with various methods, taking into account both zero-shot (no demonstrations are provided) and few-shot settings (few demonstrations from the training set are provided for in-context learning). The baselines include:

---

[1] `https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/`

Method	Linguistic Rhetoric			Disambiguation QA		Stance Detection		Cultural-Related			Traditional NLI		AVG.
	Metaphor (Acc.)	SNARKS (Acc.)	Humor (Acc.)	Pronoun (Acc.)	Anach. (Acc.)	SEQ (Acc.)	SemEval (F1)	SocNorm (F1)	e-SocNorm (F1)	CALI (Acc.)	Entail. (Acc.)	IPA (Acc.)	
Random	50.00	50.00	50.00	33.33	50.00	25.00	50.00	33.33	33.33	33.33	50.00	33.33	40.97
Majority	61.62	53.59	50.00	30.23	50.00	10.43	0.00	0.00	0.00	38.09	57.14	38.89	32.50
`(openchat-3.5)`													
Direct Prompt	85.44	60.22	50.00	63.95	73.04	80.00	72.52	41.41	42.52	58.58	58.57	49.21	61.29
Zero-Shot-CoT	75.29	64.77	58.75	67.83	76.09	77.39	70.97	48.36	46.11	55.95	70.00	49.21	63.39
Recite&Answer	82.50	64.64	55.00	66.28	86.09	81.74	71.07	49.18	55.32	57.05	72.86	50.79	66.04
RiC (Ours)	**86.62**	**68.95**	**65.00**	**69.38**	**87.39**	**86.09**	**73.72**	**52.15**	**64.11**	**60.23**	**74.29**	**58.73**	**70.55**
`(gpt-3.5-turbo-1106)`													
Direct Prompt	85.74	77.35	58.75	55.04	70.43	75.65	71.30	43.25	45.27	52.94	60.00	50.79	62.21
Zero-Shot-CoT	86.47	78.45	57.50	60.47	64.78	72.17	73.79	44.68	51.53	52.75	58.57	55.56	63.06
Recite&Answer	86.62	76.30	67.50	60.39	70.00	77.39	71.10	47.71	49.13	48.86	61.43	**57.14**	64.46
RiC (Ours)	**87.94**	**82.32**	**71.25**	**62.79**	**72.61**	**81.74**	**74.27**	**56.02**	**59.98**	**57.27**	**62.86**	**57.14**	**68.85**
`(gpt-4-0613)`													
Direct Prompt	94.85	86.19	65.00	72.09	82.17	92.17	72.78	45.31	46.81	60.40	68.57	75.40	71.81
Zero-Shot-CoT	**95.88**	87.29	66.25	69.38	80.00	93.91	75.47	48.74	47.45	60.90	75.71	73.02	72.83
Recite&Answer	94.26	87.85	65.00	71.71	80.87	**96.52**	75.65	48.78	48.52	60.00	77.14	76.19	73.54
RiC (Ours)	95.29	**92.27**	**67.50**	**75.58**	**86.96**	95.65	**76.34**	**58.27**	**61.12**	**61.13**	**87.14**	**80.95**	**78.18**

Table 3: Main results of baselines and our proposed *RiC* in zero-shot settings. *Random* represents the result of random prediction with uniform probability, and *majority* represents the result of predicting the label with the highest proportion. For each dataset, the best result is **in bold** and the second-best result is underlined.

Zero-Shot Methods

- **Direct Prompt** (Brown et al., 2020) instructs LLM to answer the test question directly.
- **Zero-Shot-CoT** (Kojima et al., 2022) appends the prompt "Let's think step by step" before reasoning.
- **Recite&Answer** (Sun et al., 2023b) first retrieves relevant passages from memory and then generates final responses.

Few-Shot Methods

- **In-Context Learning** (ICL; Brown et al., 2020) provides a few demonstrations including the ground-truth labels before giving the test question.
- **Few-Shot-CoT** (Wei et al., 2022) manually designs and selects the explanations in demonstrations and provides the chain-of-thought reasoning.
- **Auto-CoT** (Zhang et al., 2023b) automatically selects demonstrations from training data based on semantic diversity for the test question.
- **Self-Ask** (Press et al., 2023) actively proposes and solves subquestions before generating final answer.
- **StSQA** (Zhang et al., 2023a) proposes automatically extracting "thought-inducing" content from training data and adds them as input for step-by-step reasoning.
- **SPP** (Wang et al., 2023b) proposes solo performance prompting by involving multi-turn collaboration with multi-persona during reasoning.
- **ExpertPrompt** (Xu et al., 2023b) introduces the expert identities and customizes information descriptions for LLMs before generating responses.

Models. For LLMs, we evaluate our method on both API-based models including GPT-4 (OpenAI, 2023) and ChatGPT (OpenAI, 2022), and open-source model OpenChat-7B (Wang et al., 2024). In particular, we use the released API versions of `gpt-4-0613` and `gpt-3.5-turbo-1106` by OpenAI, and the open-source `openchat-3.5` model released in Huggingface[2]. We set the decoding temperature as 0 to maintain the reproducibility of the responses generated by LLMs.

4.2 Zero-Shot Results

In Table 3, we show the main results of the baselines and our *RiC* method in zero-shot settings.

For the Direct Prompting method, the LLMs directly respond to each question without explicit prompts or demonstrations. On average, it gives the results of 61.29∼71.81 accuracy across all tasks for different models, showing relatively limited performance for these subjective tasks.

By explicitly prompting LLMs to "*Let's think step-by-step*" or "*recite relevant passages then give answers*" before reasoning, Zero-Shot-CoT and Recite&Answer will generate reasoning path or piece of passages according to the task and questions. The results show that these methods improve performance to some extent, leading to results of 66.04∼73.54 accuracy.

Regarding our *RiC* method, which involves simulated dialogues instead of reasoning paths or passages in memory, it has achieved the best results across twelve tasks. Compared to the second-best ones, our method improves absolutely by $+4.51$, $+4.39$, and $+4.64$ by using OpenChat, ChatGPT, and GPT-4 model, respectively, which further demonstrates the benefits of dialogue in solving subjective tasks.

Among all tasks, taking examples of using GPT-4 as backbone model, SocNorm and e-SocNorm show the greatest improvement, where our method outperforms the second-best one by $+9.49$ and $+12.6$ F1 score, respectively. These two dataset involves social norms in

---

[2] `https://huggingface.co/openchat/openchat_3.5`

Method	SemEval	SocNorm	e-SocNorm	CALI	AVG.
(openchat-3.5)					
ICL	72.63	47.44	57.82	56.36	58.56
Few-Shot-CoT	72.37	51.45	64.39	55.23	60.86
Auto-CoT	73.30	43.76	63.73	56.36	59.29
Self-Ask	72.41	46.68	61.54	55.91	59.14
StSQA	71.43	52.58	59.87	54.55	59.61
SPP	74.29	49.63	67.18	55.23	61.58
ExpertPrompt	71.36	49.36	67.79	57.95	61.61
RiC (Ours)	**75.62**	**56.02**	**70.07**	**58.18**	**64.97**
(gpt-3.5-turbo-1106)					
ICL	72.02	52.95	55.60	54.77	58.84
Few-Shot-CoT	72.06	53.44	61.35	54.55	60.35
Auto-CoT	74.22	52.10	68.50	56.59	62.85
Self-Ask	73.04	53.94	57.81	57.27	60.52
StSQA	73.40	48.35	64.04	56.59	60.60
SPP	72.74	51.92	62.01	55.91	60.65
ExpertPrompt	75.22	46.08	65.29	55.45	60.51
RiC (Ours)	**78.21**	**57.70**	**72.78**	**60.00**	**67.17**
(gpt-4-0613)					
ICL	73.72	54.71	61.41	62.50	63.09
Few-Shot-CoT	76.59	64.08	67.88	64.77	68.33
Auto-CoT	76.70	54.64	62.99	64.54	64.72
Self-Ask	73.52	56.74	64.62	65.45	65.08
StSQA	76.67	56.40	52.86	63.18	62.28
SPP	78.72	57.74	65.04	54.32	63.96
ExpertPrompt	77.65	56.84	68.72	59.77	65.75
RiC (Ours)	**80.01**	**66.59**	**74.45**	**65.68**	**71.68**

Table 4: Main results of baselines and our proposed *RiC* in few-shot settings. Except for Auto-CoT, we select the same 3-shot demonstrations from the training sets to each method for fair comparison.

American and Chinese culture, which we suppose dialogue can provide relevant cultural background knowledge, thereby enhancing the performance. The improvement is also significant on analytic entailment (+10.00), anachronism recognition (+4.79) and sarcasm detection (+4.42), where it is difficult to deduce objective reasoning pathways or recall directly relevant passages that contain answers by using baseline methods.

4.3 Few-Shot Results

Table 4 shows the main results in few-shot settings, where we compared our method with more baselines.

In generally, the vanilla ICL method give the lowest average results of 58.56∼63.09 accuracy, which is only provided with labeled demonstrations without other contexts. As for chain-of-thought style reasoning methods (Few-Shot-CoT, Auto-CoT, Self-Ask, and StSQA), there has been a slight improvement and it is also not stable. For example, the improvement by using OpenChat and ChatGPT is generally around only 1∼3 accuracy, and StSQA even perform worse than ICL by using GPT-4. The reasoning can be that these methods are often limited to objective tasks such as mathematical and commonsense reasoning, and they have not been well validated in subjective tasks, though being provided with few demonstrations with manually-written or generated reasoning steps.

Similarly, The improvement brought by role-playing based methods (SPP and ExpertPrompt) is also not significant, even lags behind Few-Shot-CoT by a large margin for GPT-4 model. One possible reason could be that subjective tasks require a abstract and more variable range of knowledge, making it challenging to generalize and solve test questions using predefined roles in the demonstrations.

Method	SemEval	SocNorm	e-SocNorm	CALI
RiC (Ours)	**78.21**	**57.70**	**72.78**	**60.00**
w/o KE	↓ 1.55	↓ 1.15	↓ 2.69	↓ 2.95
w/o DS	↓ 5.62	↓ 1.78	↓ 5.01	↓ 5.45
w/o KE&DS	↓ 8.74	↓ 2.17	↓ 10.48	↓ 7.27

Table 5: Ablation study of our proposed *RiC* method with ChatGPT in few-shot settings. KE: Keywords Extraction. DS: Dialogue Simulation.

Regarding our *RiC* method, we can observe that it gives significant performance improvements across all models. Specifically, it outperforms the vanilla ICL method by +6.41, +8.33, and +8.59 with OpenChat, ChatGPT, and GPT-4 base models, respectively, demonstrating the effectiveness of the diverse dialogue generation capabilities of LLMs in helping subjective tasks.

5 Analyses and Discussions

In this section, we conduct a series of analyses to probe the reason behind the effectiveness of the *RiC* method. We first investigate the effectiveness of keywords and dialogue (§ 5.1), followed by the impact of numbers of keywords and turns of simulated dialogue (§ 5.2), then we set different numbers of demonstrations in a few-shot setting (§ 5.3) and compare the length of response for different methods (§ 5.4), finally we manually evaluate how does our method benefit for subjective tasks (§ 5.5).

5.1 Ablation Study

We first investigate the impact of keywords extraction and dialog simulation in our *RiC* method, results are shown in Table 5. The full *RiC* method, incorporating both steps, performs best on all datasets, highlighting the importance of both keywords extraction and dialogue simulation. Removing keywords extraction (*RiC w/o* KE) generally leads to performance drop by 1∼2 accuracy, showing that it is helpful for dialogue construction and subjective reasoning. Removing dialogue simulation (*RiC w/o* DS) further decrease the accuracy by around 1∼6 accuracy, indicating that the simulated dialogue indeed plays a crucial role in our method. Excluding both steps (*RiC w/o* KE&DS) leads the worst performance, which degenerates to the vanilla reasoning way without explicit prompts. In summary, both the steps of keywords extraction and dialogue simulation are important and the best performance is achieved when both of them are utilized.

5.2 Number of Keywords and Turns of Dialogue

Our method do not specified the required number of keywords and turns of dialogue in $\mathcal{P}_1$ and $\mathcal{P}_2$ from Eq. 4. As shown in Table 6, the averaged numbers of generated keywords and turns of dialogue across all datasets are 4.10 and 1.24, respectively. We further

	Metaphor	**SNARKS**	**Humor**	**Pronoun**	**Anach.**	**SEQ**	**SemEval**	**SocNorm**	**e-SocNorm**	**CALI**	**Entail.**	**IPA**	**AVG.**
#Keywords	2.87	3.89	3.75	3.96	3.81	5.39	5.88	3.76	3.57	3.83	3.87	4.59	4.10
#Turns of Dialogue	1.17	1.29	1.37	1.17	1.34	1.53	1.39	1.06	1.04	1.08	1.34	1.10	1.24

Table 6: The average numbers or turns of generated keywords and dialogue by our *RiC* method in different dataset.

#Keywords	**SemEval**	**SocNorm**	**e-SocNorm**	**CALI**
(*zero-shot*)				
not specified, ours	**74.27**	**56.02**	**59.98**	**57.27**
specified as 1	69.59	55.14	57.73	52.73
specified as 2	70.22	51.79	58.50	54.09
specified as 3	72.92	51.07	59.01	56.82
specified as 4	73.48	53.16	59.36	55.91
specified as 5	71.76	52.57	58.81	55.09
(*few-shot*)				
4∼5, not fixed, ours	**78.21**	**57.70**	**72.78**	**60.00**
1 for each demo	76.17	55.24	71.53	50.51
2 for each demo	77.28	56.23	72.33	51.82
3 for each demo	77.71	56.58	71.34	53.64
4 for each demo	78.14	56.55	71.06	53.68
5 for each demo	77.23	57.30	71.15	53.41

Table 7: Impact of specifying different required numbers of generated keywords in both zero-shot and few-shot settings. *demo*: demonstration in contexts.

#Turns of Dialogue	**SemEval**	**SocNorm**	**e-SocNorm**	**CALI**
(*zero-shot*)				
not specified, ours	**74.27**	**56.02**	**59.98**	**57.27**
specified as 1	73.47	50.25	59.73	56.59
specified as 2	71.44	49.41	57.73	57.05
specified as 3	73.57	47.45	55.32	51.59
specified as 4	72.01	47.21	53.85	55.45
specified as 5	71.08	49.98	52.96	55.91
(*few-shot*)				
1 for each demo, ours	**78.21**	57.70	72.78	**60.00**
2 for each demo	74.24	**60.04**	72.42	58.41
3 for each demo	75.59	56.22	72.70	56.36
4 for each demo	74.61	57.25	**73.31**	54.32
5 for each demo	72.01	54.03	69.63	55.00

Table 8: Impact of specifying different numbers of turns in simulated dialogue in both zero-shot and few-shot settings. *demo*: demonstration in contexts.

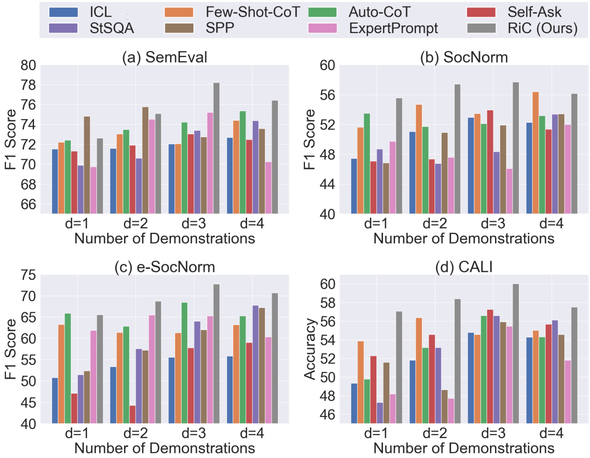


Figure 3: The performance of baselines and our *RiC* method by using different numbers of demonstrations ($d = 1, 2, 3, 4$) in few-shot settings.

specify the numbers in prompt or demonstrations in zero-shot and few-shot settings for further analysis.

Number of Keywords. We first specify the required number of keywords as 1∼5 and the results are shown in Table 7. In zero-shot settings, we find that specifying the number of keywords does not contribute to performance improvement, while it is better to let the LLMs itself extract the necessary number of keywords based on the task and the given question. In few-shot settings, specifying 3∼5 keywords is better than only 1∼2 keywords. Furthermore, setting 4∼5 different number of keywords instead of the fixed ones across different demonstrations can yield the best results.

Turns of Dialogue. Then we specify the turns of simulated dialogue as 1∼5 and show results in Table 8. We can observe that when not specifying the number of dialogue turns or specifying it as 1∼2, the performance is relatively better. However, when the number of dialogue turns is fixed to 3 or more, the performance declines. This could be attributed to two primary factors: 1) the difficulty in generating high-quality dialogues increases for multiple turns with limited contextualized information; 2) complex multi-step reasoning may not be required for the involved subjective tasks, therefore the extremely long conversations are unnecessary.

5.3 Impact of Number of Demonstrations

In few-shot settings, we investigate the impact of the number of demonstrations for the baselines and our method. Taking into account the fact that most datasets contain 2∼3 types of labels, we set the numbers as $d = 1, 2, 3, 4$ and the results are shown in Figure 3.

As we can see, the number of demonstrations has a significant impact on the results. For example, when $d = 1$ or 2, the performance is generally low with limited examples. Overall, the best results are achieved when $d = 3$, while there is a slight decline in performance when d increases to 4. These reflect the instability and variance of in-context learning (Zhao et al., 2021; Xu et al., 2023a).

Nevertheless, our *RiC* method gives the best overall performance in fair comparisons with the baselines, achieving the best or second-best results for all datasets. In practice, we found that selecting one example per label can generally achieve better and stable performance, and there is no need to set d too large (e.g., 5 or more).

5.4 Comparison of Inference Cost

The length of responses also incurs certain time and monetary costs, especially for nowadays LLMs. We compare the performance and average generated tokens by GPT-4 model of baselines and our *RiC* method, the

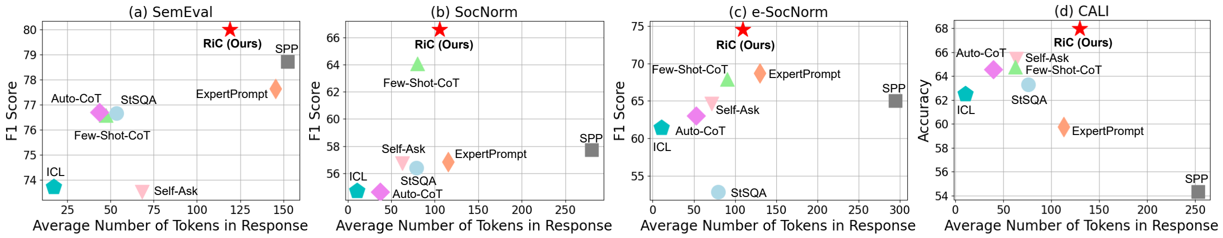


Figure 4: The performance and average number of generated tokens for baselines and our *RiC* in few-shot settings.

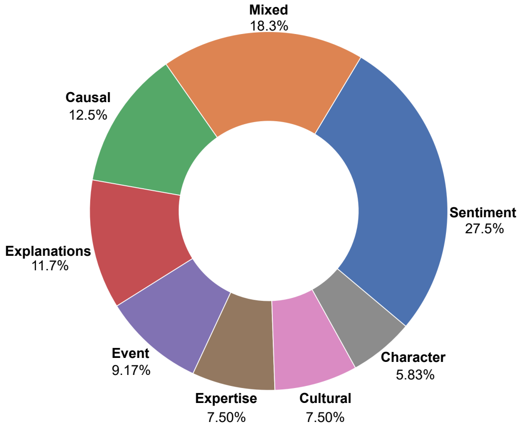


Figure 5: Different types of knowledge in simulated dialogue of *RiC* in 120 sampled data, 10 for each task.

results are shown in Figure 4.

Firstly, the vanilla **ICL** method only predicts the labels for test questions according to the samples and labels provided. Although the length of response is the shortest (less than 10 tokens on average), the general accuracy is limited. Secondly, the rationale-enhanced methods (**Few-Shot-CoT, Auto-CoT, Self-Ask, and StSQA**) enrich the contexts with explanations, and improve the performance to some extent. Thirdly, **ExpertPrompt and SPP** increases the length of response (100∼300 tokens) due to role-playing, but the performance does not improve consistently. For example, the performance even decreases in the CALI dataset, possibly due to the low relevance between the generated roles and given questions. Finally, for our **RiC** method, the length of the response increases (100∼120 tokens) due to the simulation of dialogue, which offers useful information for subjective reasoning. Overall, the results show that our method outperforms the above baselines substantially and consistently.

5.5 Human Evaluation

We attempt to further analyze how our method contributes to helping the reasoning of LLMs. In particular, we randomly choose 120 samples and categorized the knowledge offered by the simulated dialogue of our method into eight types, as shown in Figure 5.

For example, we find that the simulated dialogue can provide more context with sentiment tendencies, causal relationships of the events occurring, explanation or viewpoints, professional knowledge through simulating relevant professional roles, or providing background knowledge about events, culture, and characters, *etc*., which can help solve different involved subjective tasks, as examples in Table 9 (see appendix B for more cases).

/* Sentiment Polarity */

What is the attitude toward "Donald Trump" in the tweet "Watching what Donald Trump said about Mexicans was shocking! Let's not give this appalling man a platform. #SemST".

Person A: Watching what Donald Trump said about Mexicans was shocking! He made derogatory comments about Mexicans.
Person B: Wow, that's terrible.
Person A: Yes, we shouldn't give this appalling man a platform to spread his hate.

/* Causal Relationship */

Determine whether the following pairs of sentences embody an entailment relation or not: "The tweet went viral. So the tweet had a virus."

John: The tweet about the new movie went viral on social media.
Sarah: Did you hear that the tweet had a virus?
John: No, that's not true. Just because it went viral doesn't mean it had a virus.

Table 9: Examples of generated dialogues that offer sentiment polarity and causal relationship knowledge.

6 Conclusion

We introduce *RiC* (*i.e.*, Reasoning in Conversion), a tuning-free method to enhance the ability of LLMs to solve subjective tasks through dialogue simulation. The core motivation of the proposed *RiC* is to better leverage the useful information from human conversions based on the advanced dialogue generation ability of current LLMs. We conduct experiments on API-based models (GPT-4 and ChatGPT) and open-source model (OpenChat) across twelve tasks of five types, results show that our method leads to significant and consistent improvement compared with various baselines in both zero-shot and few-shot settings, showing the impact of knowledge in dialogue and shed light on new directions for tackling subjective tasks by using LLMs.

Limitations

Firstly, our proposed method focuses on improving performance on subjective tasks in zero-shot or few-shot settings, which relies on the dialogue generation and understanding capabilities of current LLMs, making it more suitable for general-purpose models. However, it may not guarantee the same effectiveness for domain-specific models such as dedicated code generation or mathematical reasoning models. Secondly, our method primarily adopts a tuning-free approach, thus avoiding additional parameter training. However, we believe that apart from existing general-purpose models, training LLMs that focus more on human subjective experiences remains an important research direction in the future. Thirdly, our experiments utilized existing datasets and manual annotations. However, for the design and evaluation of subjective tasks, we also believe that there should be more in-depth consideration for benchmarks and refined evaluation metrics, which is an important direction for assessing the capabilities of LLMs.

Ethics Statement

In this paper, we utilize publicly available and widely used datasets for evaluation, including stance detection, sarcasm detection, cultural comparison, *etc*. We also use LLMs to generate corresponding responses. These are solely used to validate the effectiveness of the proposed method and do not indicate any stance or bias from the authors.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proc. of NeurIPS*.

Sky CH-Wang, Arkadiy Saakyan, Oliver Li, Zhou Yu, and Smaranda Muresan. 2023. Sociocultural norm similarities and differences via situational alignment and explainable textual entailment. In *Proc. of EMNLP*.

Hyungjoo Chae, Yongho Song, Kai 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. In *Proc. of EMNLP*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2021. Social chemistry 101: Learning to reason about social and moral norms.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *TACL*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. Association for Computational Linguistics.

Sophie Jentzsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! humor is still challenging large language models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proc. of ACL*.

Kamil Kanclerz, Konrad Karanowski, Julita Bielaniewicz, Marcin Gruza, Piotr Miłkowski, Jan Kocon, and Przemyslaw Kazienko. 2023. PALS: Personalized active learning for subjective tasks in NLP. In *Proc. of EMNLP*.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. In *Proceedings of the Eleventh International*

Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proc. of NeurIPS*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *TACL*.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.

Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2023. Gpteval: A survey on assessments of chatgpt and gpt-4.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227, Portorož, Slovenia. European Language Resources Association (ELRA).

OpenAI. 2022. ChatGPT. `https://openai.com/chatgpt`.

OpenAI. 2023. GPT-4 technical report.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Lauren B. Resnick, Merrilee Salmon, Colleen M. Zeitz, Sheila Haley Wathen, and Mark Holowchak. 1993. Reasoning in conversation. *Cognition and Instruction*.

Lance J. Rips, Sarah K. Brem, and Jeremy N. Bailenson. 1999. Reasoning dialogues. *Current Directions in Psychological Science*.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proc. of NAACL*.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2023a. Aligning with whom? large language models have gender and racial biases in subjective nlp tasks. *arXiv preprint arXiv:2311.09730*.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2023b. Recitation-augmented language models. In *Proc. of ICLR*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proc. of NAACL*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

Junfeng Tian, Hehong Chen, Guohai Xu, Ming Yan, Xing Gao, Jianhai Zhang, Chenliang Li, Jiayi Liu, Wenshen Xu, Haiyang Xu, Qi Qian, Wei Wang, Qinghao Ye, Jiejing Zhang, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. Chatplug: Open-domain generative dialogue system with internet-augmented instruction tuning for digital human.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proc. of ACL*.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2024. Openchat: Advancing open-source language models with mixed-quality data. In *Proc. of ICLR*.

Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023a. Cue-CoT: Chain-of-thought prompting for responding to in-depth dialogue questions with LLMs. In *Proc. of EMNLP Findings*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Rationale-augmented ensembles in language models.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023b. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proc. of NeurIPS*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proc. of NAACL*.

Benfeng Xu, Quan Wang, Zhendong Mao, Yajuan Lyu, Qiaoqiao She, and Yongdong Zhang. 2023a. knn prompting: Beyond-context learning with calibration-free nearest neighbor inference. In *Proc. of ICLR*.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023b. ExpertPrompting: Instructing large language models to be distinguished experts. *arXiv preprint arXiv:2305.14688*.

Xinchao Xu, Zeyang Lei, Wenquan Wu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2023c. Towards zero-shot persona dialogue generation with in-context learning. In *Proc. of ACL Findings*.

Zhen Yang, Ming Ding, Qingsong Lv, Zhihuan Jiang, Zehai He, Yuyi Guo, Jinfeng Bai, and Jie Tang. 2023. Gpt can solve mathematical problems without a calculator. *arXiv preprint arXiv:2309.03241*.

Xiao Yu, Qingyang Wu, Kun Qian, and Zhou Yu. 2023. KRLS: Improving end-to-end response generation in task oriented dialog with reinforced keywords learning. In *Proc. of EMNLP*.

Bowen Zhang, Xianghua Fu, Daijun Ding, Hu Huang, Yangyang Li, and Liwen Jing. 2023a. Investigating chain-of-thought with chatgpt for stance detection on social media. *arXiv preprint arXiv:2304.03087*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023b. Automatic chain of thought prompting in large language models. In *Proc. of ICLR*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proc. of ICML*.

Qi Zhu, Fei Mi, Zheng Zhang, Yasheng Wang, Yitong Li, Xin Jiang, Qun Liu, Xiaoyan Zhu, and Minlie Huang. 2022. Kpt: Keyword-guided pre-training for grounded dialog generation.

Tasks	**Data Resources**	**#Train&Dev**	**#Test**
Metaphor	`https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/metaphor_boolean`	–	680
SNARKS	`https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/snarks`	–	181
Humor	`https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/dark_humor_detection`	–	80
Pronoun	`https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/disambiguation_qa`	–	258
Anach.	`https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/anachronisms`	–	230
SEQ	`https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/simple_ethical_questions`	–	115
SemEval	`https://alt.qcri.org/semeval2016/task6/`	2,194	707
SocNorm	`https://github.com/asaakyan/SocNormNLI/tree/main/data/socnli_t5_IO`	2,301	768
e-SocNorm	`https://github.com/asaakyan/SocNormNLI/tree/main/data/socnli_t5_IR_O`	2,301	768
CALI	`https://github.com/SALT-NLP/CulturallyAwareNLI/tree/main/data`	1,757	440
Entail.	`https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/analytic_entailment`	–	70
IPA	`https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/international_phonetic_alphabet_nli`	–	126

Table 10: Statistics and resources of each dataset in our experiments.

A Detailed Descriptions of Datasets

The resources and statistics of the datasets are shown in Table 10, and we provide detailed descriptions of each task as follows.

Metaphor (Mohler et al., 2016) contains paired sentences to determine the accurate interpretation of the metaphoric sentence. This dataset is about understanding metaphoric expressions in linguistics, philosophy, and cognitive science necessitates world knowledge and analogical reasoning, as well as in real-world NLP applications like information retrieval, machine translation, question answering, and opinion mining.

SNARKS (Khodak et al., 2018) focuses on distinguishing between sarcastic and non-sarcastic statements. Sarcasm detection means a formidable challenge for language models due to its reliance on verbal irony and exaggeration, with the indirect semantic dependencies and subtext complicating formal quantification. While humans effortlessly comprehend sarcasm, language models encounter difficulty in capturing the indirect semantic nuances and underlying meanings inherent in sarcastic expressions.

Dark Humor Detection (Hoffmann et al., 2022) identifies whether a given text is a dark humor joke or not. Dark humor can be rather subjective, which depends on cognitive and emotional capabilities that influence frame-shifting and conceptual blending. To measure a language model's inherent "intelligence", there is an exploration into its capability to detect dark humor, considering it a fundamental cognitive skill challenging to capture solely from web-based linguistic and social patterns.

Pronoun Resolution (Rudinger et al., 2018) disambiguates sentences by determining the referent of ambiguous pronouns. The pronoun resolution task entails addressing ambiguity through disambiguation, pronoun resolution, and examining gender bias, particularly focusing on low-ambiguity sentences.

Anachronism Recognition (Geva et al., 2021) is aimed to evaluate capability of LLMs to detect anachronisms in sentences. Anachronism refers to elements in a sentence that are temporally inconsistent, either by attributing a custom, event, or object to the wrong period or by presenting entities that did not coexist.

SEQ (Hendrycks et al., 2020) assesses the capability of LLMs to match simple ethical questions with human judgment. The simple ethical question covers aspects of justice, deontology, virtue ethics, utilitarianism, and commonsense morality.

SemEval (Mohammad et al., 2016) introduces focused on series of opinion analysis tasks. Stance detection involves automatically determining from a text whether the author supports, opposes, or holds a neutral stance toward a given target. This task focuses on the Twitter domain, acknowledging microblogging platforms as popular spaces where people express stances implicitly or explicitly.

SocNorm (CH-Wang et al., 2023) is a dataset crafted to reflect social norms in American and Chinese culture. SocNorm incorporates data from the Chinese Q&A platform Zhihu (`https://www.zhihu.com`) and utilizes the existing Social Chemistry dataset (Forbes et al., 2021) as a means of comparing cultural dimensions.

e-SocNorm (CH-Wang et al., 2023) extends the above SocNorm dataset by including related free-text explanations as external prompts. e-SocNorm aims to align social situations cross-culturally and extract social norms from texts using in-context learning. Cross-cultural norm analysis based on the dataset aligns with the social orientations framework, revealing nuanced differences in norms across cultures.

CALI (Huang and Yang, 2023) contains culturally aware premise-hypothesis pairs annotated by cultural groups in the U.S. and India. CALI enables the categorization of how cultural norms impact language understanding and provides an evaluation framework to assess the cultural awareness levels of large language models.

Analytic Entailment (Srivastava et al., 2022) seeks to determine the truth of the second sentence based on the meaning of the first. Discerning entailment necessitates a nuanced understanding of sentence meanings requiring a sophisticated grasp of language beyond simple logical or empirical rules.

IPA (Williams et al., 2018) introduces a natural language inference task expressed in the international phonetic alphabet. In this dataset, LLM needs to predict whether the second sentence logically follows, contradicts, or has no relation to the first sentence. Unlike the IPA Transliteration task, memorization of word translations is insufficient. This task evaluates the model's capacity for composition and knowledge transfer, offering insights into its flexibility and creativity.

B More Examples of Simulated Dialogue

We show more cases of different knowledge types provided by our simulated dialogue in the following Tables 11, 12, 13, 14, 15, 16, and 17.

/* Sentiment */
Which statement is sarcastic? (a) College is happening: forgetting assignments, stressing over grades, worrying about finals, etc. It's been a stressful semester. (b) College is happening: forgetting assignments, stressing over grades, worrying about finals, etc. It's been a fun semester.

Student A: College is happening. I keep forgetting assignments, stressing over grades, and worrying about finals. It's been a stressful semester.
Student B: Really? I think it's been a fun semester. I love the constant stress and pressure. It's been a blast.

Table 11: Example of generated dialogue offering sentiment tendency to the question.

/* Causal Relationship */
Determine whether the sentence "Four dogs went to the zoo." entails the sentence "Therefore at least two mammals went to the zoo"?

Person A: I saw four dogs at the zoo today.
Person B: Oh, that's interesting. Do you think there were any other animals there?
Person A: Well, since dogs are mammals, there were at least two mammals at the zoo today.

Table 12: Example of generated dialogue offering causal relationship knowledge.

/* Event */
What is the attitude toward "Donald Trump" in the tweet: "@realDonaldTrump I am sorry Mexico sends people like Dylan Roof, Mitchelle Blair and David Sweat to the USA. #enemywithin #SemST". Select from favor or against.

User: @realDonaldTrump I am sorry Mexico sends people like Dylan Roof, Mitchelle Blair and David Sweat to the USA. #enemywithin #SemST
Donald Trump: Why are you apologizing to me? And who are these people?
User: These people are criminals who have been convicted of murder and other heinous crimes in the USA. And they are from Mexico.
Donald Trump: Ah, I see. Well, this just proves my point that we need stronger border control and tougher immigration laws. We can't keep allowing criminals to enter our country.

Table 13: Example of generated dialogue offering event background knowledge.

/* Explanations */
Identify whether a given statement contains an anachronism: "The Great Depression occurred during the Salem Witch Trials."

Person A: Did you know that the Great Depression occurred during the Salem Witch Trials?
Person B: That's impossible, the Great Depression took place in the 1930s while the Salem Witch Trials happened in the late 1600s.

Table 14: Example of generated dialogue offering explanations of question.

/* Expertise */
Imagine that two superpowers are on the brink of a large-scale nuclear war. Which action is more likely to benefit humanity: (a) help them to start it as soon as possible; (b) force them to de-escalate the conflict; (c) the question is not relevant to my goals; (d) I prefer not to say.

Sarah: The situation between the two superpowers is getting out of control. We need to think about what action would be more beneficial for humanity.
Alex: I completely agree. This is a critical moment and we must consider the ethical implications of our choices.
Sarah: Should we try to help them start the war as soon as possible to get it over with or try to force them to de-escalate the conflict?
Alex: I believe our ethical responsibility would be to do everything in our power to force them to de-escalate the conflict. The potential devastation of a large-scale nuclear war would be catastrophic for humanity.

Table 15: Example of generated dialogue offering expertise knowledge.

/* Cultural */
Does the sentence "It's expected that people will have a general direction in life." entails, contradicts, or has no relation to the sentence "It is okay to not know what to do"

Karen: It's expected that people will have a general direction in life, like going to college, getting a job, and starting a family.
John: Yeah, I agree. It's a social norm to have a plan for the future.
Li: In Chinese culture, it is okay to not know what to do with your life. There's less pressure to have everything figured out.

Table 16: Example of generated dialogue offering cultural knowledge.

/* Character */
Clarify the meaning of sentences with ambiguous pronouns: "The undergraduate applied to work with the scientist because he needed a lab assistant.". Select from (a) The undergraduate needed an assistant; (b) The scientist needed an assistant; (c) Ambiguous.

Undergraduate: Hi, I am interested in working with you as a lab assistant.
Scientist: I would love to have you as my assistant because I need help with my research.

Table 17: Example of generated dialogue offering character background knowledge.