Perspective Transition of Large Language Models for Solving Subjective Tasks

Anonymous ACL submission

Abstract

Large language models (LLMs) have revolutionized the field of natural language processing, enabling remarkable progress in various tasks. Different from objective tasks such as commonsense reasoning and arithmetic question-answering, the performance of LLMs on subjective tasks is still limited, where the perspective on the specific problem plays crucial roles for better interpreting the context and giving proper response. For example, in certain scenarios, LLMs may perform better when answering from an expert role perspective, potentially eliciting their relevant domain knowledge. In contrast, in some scenarios, LLMs may provide more accurate responses when answering from a third-person standpoint, enabling a more comprehensive understanding of the problem and potentially mitigating inherent biases. In this paper, we propose Reasoning through Perspective Transition (RPT), a method based on in-context learning that enables LLMs to dynamically select among direct, role, and third-person perspectives for the best way to solve corresponding subjective problem. Through extensive experiments on totally 12 subjective tasks by using both closed-source and open-source LLMs including GPT-4, GPT-3.5, Llama-3, and Qwen-2, our method outperforms widely used single fixed perspective based methods such as chain-of-thought prompting and expert prompting, highlights the intricate ways that LLMs can adapt their perspectives to provide nuanced and contextually appropriate responses for different problems.

1 Introduction

Large language models (LLMs) have exhibited substantial advancements (Brown et al., 2020; OpenAI, 2023, 2022; Touvron et al., 2023; Jiang et al., 2023) in recent years, demonstrating remarkable performance across various tasks such as mathematical reasoning (Luo et al., 2023; Yang et al., 2023), code generation (Chen et al., 2021; Roziere et al., 2023), and commonsense question answering (Talmor et al., 2019). Meanwhile, research in the realm of *subjective* tasks remains relatively nascent (Rottger et al., 2022; Kanclerz et al., 2023; Sun et al., 2023). Unlike objective tasks, which are typically well-defined and directly solvable, subjective tasks such as metaphor recognition (Mohler et al., 2016) and dark humor detection (Meaney et al., 2021) require an understanding of context, linguistic subtleties, and varying individual perspectives. These elements are challenging to quantify or measure objectively, thus presenting significant obstacles for current LLMs (Jentzsch and Kersting, 2023; Wachowiak and Gromann, 2023; Mao et al., 2024).

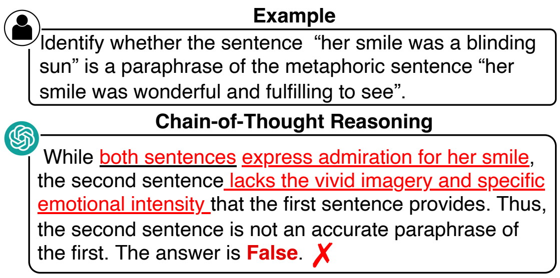


Figure 1: An example for showing challenges of solving subjective tasks using chain-of-thought prompting.

Chain-of-thought (CoT) prompting style methods have been widely used to elicit the reasoning ability of LLMs (Wei et al., 2022; Kojima et al., 2022; Zhang et al., 2023; Yao et al., 2023). However, in subjective-leaning questions, it is difficult to find a chain-of-thought pathway similar to that in conventional reasoning tasks. Also, due to the nature of the subjective tasks, manually handwriting the reasoning paths for subjective tasks is more challenging and less consistent. Therefore, directly using CoT prompting techniques may not be practical for subjective tasks. Furthermore, the generated reasoning pathways can even mislead the model to provide incorrect answers, as example shown in Figure 1. Therefore, we are motivated to propose a general method to enhance the ability of LLMs to solve various subjective tasks.

In this paper, in contrast to the common way of using LLMs is to let it directly answer the questions based

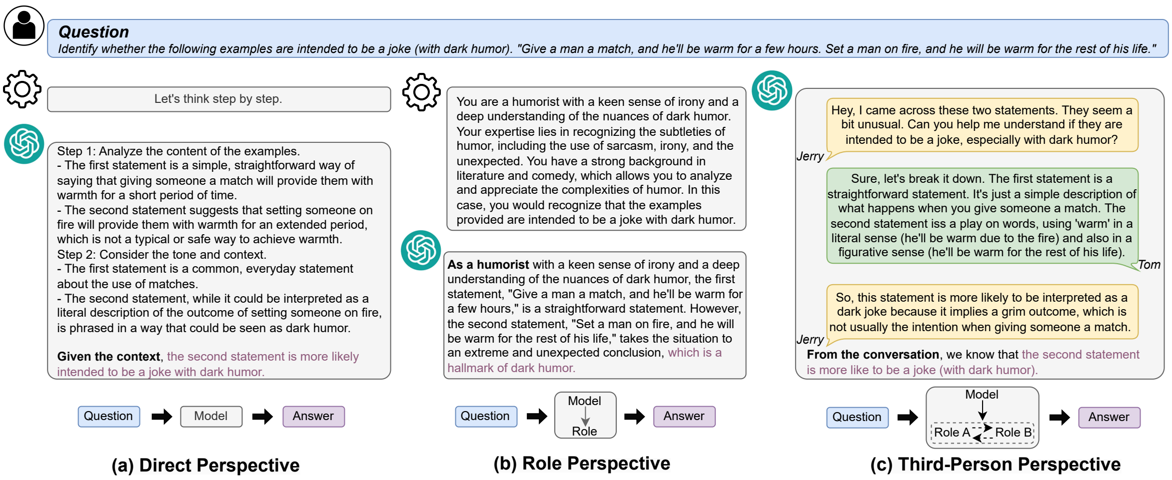


Figure 2: An example of solving dark humor detection task by different perspectives. (a) direct perspective: the model give the answer according to its analysis (Kojima et al., 2022). (b) role perspective: the model gives the answer by setting as a role related to the question (Xu et al., 2023). (c) third-person perspective: the model gives the answer as a third-person based on a simulated dialogue (Wang et al., 2024c).

on the LLMs' own direct perspective (e.g., zero-shot (Brown et al., 2020) or zero-shot-CoT reasoning (Kojima et al., 2022), we propose leveraging different perspectives to better address the aforementioned challenging subjective tasks, inspired by the domain Theory of Mind (Premack and Woodruff, 1978; Wellman et al., 2001), which refers to the ability to attribute mental states to oneself and others and to understand that these mental states can influence behavior. Our work is also related to the development of LLM-based multi-agents (Xi et al., 2023; Wang et al., 2024a), where we aim to elicit the capacity of LLMs to understand contexts, analysis problems, and give solutions beyond a single fixed response based on direct perspective.

Given the breadth and complexity of subjective tasks, we propose a Reasoning through Perspective Transition (RPT) method to dynamically select suitable perspectives to solve specific problems. In particular, we consider categorizing current reasoning methods of LLMs into three perspectives, including: 1) direct perspective, which involves the model directly answering questions or tasks based on its internal understanding without considering external factors or alternative viewpoints, 2) role perspective, which focuses on assigning specific roles to the model, simulating different viewpoints or expertise within a given context or scenario, and 3) third-person perspective, which involves the model considering external viewpoints or perspectives beyond its own, similar to how a third party or observer might view a situation. The examples of three different perspectives when solving problems are shown in Figure 2.

To facilitate the dynamic perspective transition during reasoning, we follow the in-context learning (Brown et al., 2020) approach to provide templates for answering from different perspectives through demonstrations, then let the model provide confidence levels (Li et al., 2024; de Vries and Thierens, 2024; Bank et al., 2019) for answers to specific questions, and finally answer based on the perspective with the highest confidence. In this manner, our method can freely select among three different perspectives to handle various subjective tasks. This flexibility allows the model to adapt its responses more effectively to nuanced tasks that traditional static methods struggle with. Additionally, this approach is grounded in the hypothesis that LLMs perform better when their operational parameters align with their confidence levels in specific contexts.

To validate the effectiveness of our proposed method, we conduct experiments on four LLMs (including two closed-source models GPT-4 (OpenAI, 2023)/GPT-3.5 (OpenAI, 2022), and two open-source models Llama-3 (Dubey et al., 2024)/Qwen-2 (Yang et al., 2024)) across 12 subjective tasks. Extensive experimental results demonstrates that, compared to previous methods based on a single perspective or some simple ensemble-based methods, our approach can improve the performance consistently.

2 Related Work

Subjective Tasks in NLP. Compared with *objective* tasks such as commonsense reasoning (Talmor et al., 2019) and arithmetic question-answering (Cobbe et al., 2021), research on LLMs in *subjective* tasks (*e.g.*, metaphor recognition and dark humor detection) (Rottger et al., 2022; Kanclerz et al., 2023; Sun et al., 2023) is still underexplored. Different from objective tasks that can often be clearly defined and solved, subjective tasks 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., 2024). For example, as shown in results of BigBench(bench authors, 2023), the zero-shot accuracy of PaLM-535B (Chowdhery et al., 2023) model on metaphor recognition, dark humor detection, and sarcasm detection tasks does not exceed 50%.

In-Context Learning of LLMs. As the number of model parameters increases, the in-context learning (Brown et al., 2020) ability of LLMs becomes stronger, significantly enhancing the zero-shot and few-shot reasoning capabilities without model fine-tuning. In particular, methods based on chain-of-thought (Wei et al., 2022; Kojima et al., 2022) prompting are widely used. These works aims to elicit the reasoning capability of LLMs through adding the reasoning pathways. However, recent research has shown that such reasoning pathways are mainly effective for math and symbolic reasoning (Sprague et al., 2024). Our work also relies on in-context learning, however, we propose a method based on dynamic perspective transition to elicit knowledge from the different perspective of LLMs, which does not rely on a single reasoning pathway and achieve better results on a wider range of subjective tasks.

Perspective Transition of LLMs. There are various ways to use LLMs currently that are based on different perspectives: 1) Direct prompting methods (Brown et al., 2020; Wei et al., 2022; Kojima et al., 2022) let the model to provide answers based on the factual knowledge or reasoning ability by LLMs themselves directly, without setting specific roles. 2) By assigning roles (Xu et al., 2023; Wang et al., 2024d; Wilf et al., 2024) such as experts and engaging in role-playing dialogue, the internal knowledge of LLMs on specific roles can be elicited. 3) By constructing scenarios through multi-agent cooperation (Wang et al., 2024e,b), debates (Du et al., 2024), or dialogues (Wang et al., 2024c), and then providing answers from a third-person perspective by incorporating contextualized information by the constructed agents. The previous methods only consider a fixed perspective and validate the effectiveness in certain problems. In contrast, through our proposed RPT method based on in-context learning, LLMs are able to adaptively select the most suitable perspective to solve various subjective tasks, which has not been studied in previous research.

3 Method

The overall pipeline of the proposed RPT is structured into three steps. Firstly, we input the task description and a specific question, prompting the model to select the most appropriate perspective for answering the question. Secondly, the model evaluates and ranks these perspectives based on their confidence levels in addressing the question. Thirdly, the model adopts the perspective with the highest confidence to formulate and deliver the definitive answer.

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

Step 1: Exploring Diverse Available Perspectives. We first let LLM $\mathcal{M}$ explore diverse perspectives $\mathcal{P}$ according to the description $\mathcal{D}$ and question $\mathcal{Q}$. Specifically,

Illustration of the Unified Prompting and Response

/* Input */

/* Perspective-based Instructions */

Here are some examples of perspectives you can choose:
Direct Perspective (answer the question directly).
....
Role Perspective (assume you are some roles (e.g., expert) and answer the question).
....
Third-person Perspective (simulate some agents and discussions, then summarize and give the answer).($\mathcal{D}$)
....

/* Input Question */

Metaphor Recognition: Identify whether the sentence “Adam did not understand the root of the crisis” is a paraphrase of the metaphoric sentence “Adam did not understand the solution to the crisis” ($\mathcal{Q}$)

Firstly, analyzing the question from diverse perspectives, and selecting some available perspectives based on the question. ($\mathcal{T}_1$)
Secondly, ranking all the perspectives according to the confidence level in solving the question. [Note: The confidence indicates how likely you think your answer is true]. ($\mathcal{T}_2$)
Finally, choosing the method with the highest confidence to solve the question, reason and give the final answer. Select from True or False. ($\mathcal{T}_3$)

/* Response of $\mathcal{M}$ */

Perspective and Confidence: ($\mathcal{P}$, $\mathcal{C}$)
Third-person Perspective, 85%
Role Perspective, 70%
Direct Perspective, 60%

Selected Perspective Reasoning: ($\mathcal{R}$)
Tom: Adam did not understand the root of the crisis.
Jerry: So, does that mean Adam did not understand the solution to the crisis?
Tom: Not exactly. Understanding the root of the crisis doesn’t necessarily mean understanding the solution. Understanding the root is figuring out what causes the problem, while a solution refers to ways of solving the problem. They can be related but are not synonymous.
Answer: False

Table 1: The example of a metaphor recognition task, a specific question, the unified prompt, and the corresponding response.

we have:

$$\mathcal{P} = \{p_1, p_2, ..., p_n\} = \mathcal{M}(\mathcal{D} \oplus \mathcal{Q} \oplus \mathcal{T}_1), \quad (1)$$

where n (the number of perspectives) is usually 3, $\oplus$ denotes concatenation operation. $\mathcal{T}_1$ is a prompt serving as a trigger sentence, for example, we can set $\mathcal{T}_1$ as “*Firstly, analyzing the question from diverse perspectives, and selecting some available perspectives based on the question*”.

Step 2: Ranking Perspectives by Confidence Level. Then, base on the perspectives, we let LLM $\mathcal{M}$ list all the perspectives by the confidence level $\mathcal{C}$ in solving the

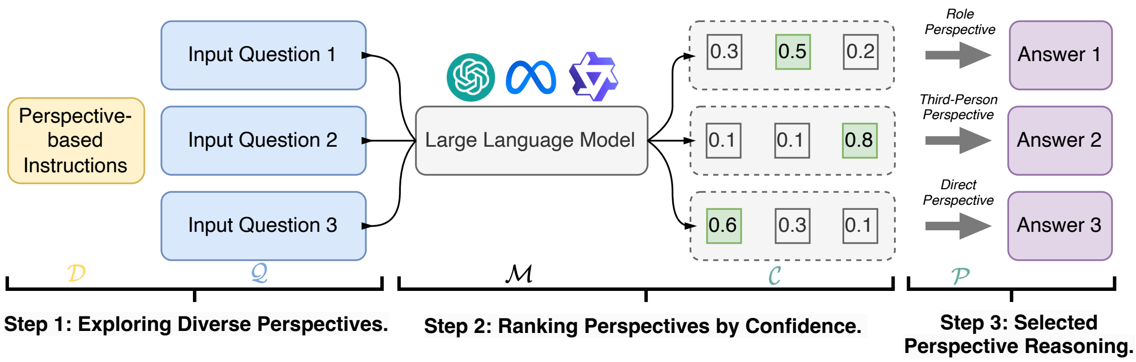


Figure 3: An overview of RPT pipeline. For each input question, RPT explores the available perspectives and then ranks them based on confidence. Accordingly, the input question is reasoned using the selected perspective.

question:

$$\mathcal{P}, \mathcal{C} = \mathcal{M}(\mathcal{D} \oplus \mathcal{Q} \oplus \mathcal{T}_2), \tag{2}$$

where $\mathcal{T}_2$ is a prompt for ranking the confidence level of all the available perspectives. For example, we can set $\mathcal{T}_2$ as "*Secondly, Ranking all the methods according to the confidence level in solving the question. [Note: The confidence indicates how likely you think your answer is true.]*".

Step 3: Selected Perspective Reasoning. Finally, we take the original task description $\mathcal{D}$, question $\mathcal{Q}$, and the ranked confidence level perspective $\mathcal{P}$ as the input, letting LLM $\mathcal{M}$ give the final response $\mathcal{R}$:

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

where $\mathcal{T}_3$ is the last prompt leading to the final answer which can be set as "*Finally, Choosing the perspective with the highest confidence to solve the question, 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 $\mathcal{T}$. In this way, our method only requires inference once through the LLM to obtain the answer to the question:

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

where an example of the unified prompt and response is shown in Table 1.

4 Experiments

4.1 Settings

Datasets. We evaluate the effectiveness of our method on twelve subjective reasoning datasets, which can be categorized into five types, as shown in Table 2. Notably, for SemEval and cultural-related datasets which contain training sets, we evaluate in both zero-shot and few-shot settings. For the other tasks, we utilize corresponding test sets from BigBench[1] (Srivastava et al., 2022) and only evaluate in zero-shot settings.

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

Dataset (names in short)	**Subjective Tasks**	**#Train/Dev/Test**
(*Linguistic Rhetoric*)		
Metaphor (Mohler et al., 2016)	Metaphor Understanding	-/-/680
SNARKS (Khodak et al., 2018)	Sarcasm Detection	-/-/181
Humor (Hoffmann et al., 2022)	Dark Humor Detection	-/-/80
(*Disambiguation QA*)		
Pronoun (Rudinger et al., 2018)	Pronoun Resolution	-/-/258
Anachronisms (Geva et al., 2021)	Identifying Anachronisms	-/-/230
(*Stance Detection*)		
SEQ (Hendrycks et al., 2021)	Simple Ethical Questions	-/-/115
SemEval (Mohammad et al., 2016)	Opinion Analysis	2,194/621/707
(*Cultural-Related*)		
SocNorm (CH-Wang et al., 2023)	Sociocultural Norm NLI	2,301/300/768
e-SocNorm (CH-Wang et al., 2023)	Sociocultural Norm NLI	2,301/300/768
CALI (Huang and Yang, 2023)	Culturally Aware NLI	1,757/–/440
(*Traditional NLI*)		
Entailment (Srivastava et al., 2022)	Analytic Entailment	-/-/70
IPA (Williams et al., 2018)	NLI in the International Phonetic Alphabet	-/-/126

Table 2: Statistics and resources of datasets.

Baselines. We compare our method with 11 baselines including different single perspective methods and ensemble-based methods as follows.

Single Direct Perspective. **Directly Prompt** (Brown et al., 2020) directly use the question as input in zero-shot or few-shot manners. **ICL** (Brown et al., 2020) (in-context learning) uses examples and labels as few input demonstrations. **Few-shot-CoT** (Wei et al., 2022) uses manually created external reasoning pathways as demonstrations. **Zero-Shot-CoT** (Kojima et al., 2022) does not rely on demonstrations and elicits the reasoning ability by using "Let's think step by step." as external input. **Self-Ask** (Press et al., 2023) actively proposes and solves subquestions before generating the final answer.

Single Role Perspective. **ExpertPrompt** (Xu et al., 2023) introduces the expert identities and customizes information descriptions for LLMs before generating responses. **Role-Play Prompting** (Kong et al., 2024) also lets models simulate complex human-like interactions and behaviors for zero-shot reasoning.

Single Third-Person Perspective. **SPP** (Wang et al., 2024e) (solo performance prompting) proposes solo performance prompting by involving multi-turn collaboration with multi-persona. **RiC** (Wang et al., 2024c) (reason in conversation) first lets model generating dialogues between simulated roles, and then summarize conversations and give final answers according to the additional information from conversations.

Ensemble-based Methods. **Ensemble** (Messuti et al.,

Type	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
	(`Llama-3-8b-instruct`)													
S1	Direct Prompt (Brown et al., 2020)	66.03	58.56	60.00	43.41	50.00	61.74	71.00	39.15	48.49	42.95	51.43	39.68	52.70
S1	Zero-Shot-CoT (Kojima et al., 2022)	67.06	**70.72**	63.75	46.90	61.74	73.04	72.45	40.07	52.84	47.95	54.29	44.44	57.94
S2	Role-Play Prompting (Kong et al., 2024)	65.00	64.09	65.00	45.35	53.91	72.17	73.26	51.36	56.44	46.82	51.54	43.65	57.38
S3	Reason in Conversation (Wang et al., 2024c)	76.32	69.72	58.75	48.06	52.17	80.00	74.71	48.15	**64.05**	48.86	**58.57**	50.79	60.85
E	Ensemble (Agrawal et al., 2024)	68.09	64.64	50.00	37.60	69.57	82.61	77.00	44.60	58.72	**54.77**	57.14	53.97	59.89
E	Reranking (Farinhas et al., 2024)	71.47	58.56	52.50	51.78	59.13	72.17	72.23	52.86	52.34	48.41	55.71	54.76	59.49
E	CoT-SC (Wang et al., 2023)	65.88	48.07	**66.25**	45.74	61.30	78.26	76.01	**55.86**	59.64	47.73	52.86	**61.90**	59.96
D	**RPT (Ours)**	**81.76**	60.22	65.00	**53.49**	**72.17**	**89.57**	**77.44**	53.52	61.72	51.59	**58.57**	44.44	**64.12**
	(`qwen-2-7b-instruct`)													
S1	Direct Prompt (Brown et al., 2020)	79.85	61.88	60.00	56.98	64.38	86.09	70.17	38.53	47.93	42.27	58.57	58.73	60.45
S1	Zero-Shot-CoT (Kojima et al., 2022)	83.09	64.03	63.75	54.65	63.48	79.13	73.36	43.99	47.79	46.59	62.86	62.70	62.12
S2	Role-Play Prompting (Kong et al., 2024)	78.97	65.75	56.25	52.25	60.87	86.96	72.25	46.77	51.21	49.77	64.29	57.14	61.87
S3	Reason in Conversation (Wang et al., 2024c)	80.59	69.61	60.00	60.47	63.91	87.83	75.06	49.57	56.40	53.18	64.29	60.32	65.10
E	Ensemble (Agrawal et al., 2024)	86.03	74.59	62.50	54.65	63.48	88.70	72.66	46.08	58.02	53.64	71.43	**66.67**	66.54
E	Reranking (Farinhas et al., 2024)	83.23	72.38	60.00	54.65	62.17	87.04	74.21	44.81	54.18	**54.32**	**74.29**	65.08	65.53
E	CoT-SC (Wang et al., 2023)	**86.32**	**79.56**	47.50	**66.28**	**70.00**	92.17	75.29	44.44	61.90	**54.32**	50.00	56.35	65.34
D	**RPT (Ours)**	84.41	69.61	**65.00**	63.95	68.70	**94.78**	**76.58**	**51.02**	**68.29**	52.27	67.14	61.90	**68.64**
	(`gpt-3.5-turbo-1106`)													
S1	Direct Prompt (Brown et al., 2020)	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
S1	Zero-Shot-CoT (Kojima et al., 2022)	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
S2	Role-Play Prompting (Kong et al., 2024)	82.64	77.40	57.25	60.39	71.74	78.39	71.10	47.61	49.13	55.68	61.43	57.14	64.16
S3	Reason in Conversation (Wang et al., 2024c)	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
E	Ensemble (Agrawal et al., 2024)	84.26	76.80	66.25	59.61	72.17	86.26	70.32	48.25	56.51	52.95	64.29	65.08	66.90
E	Reranking (Farinhas et al., 2024)	81.76	79.56	65.00	54.65	72.17	81.30	77.27	51.18	63.99	60.91	61.42	63.49	67.73
E	CoT-SC (Wang et al., 2023)	84.85	86.74	67.50	47.29	**74.35**	92.17	81.18	59.35	65.53	59.09	87.14	75.40	73.38
D	**RPT (Ours)**	**91.76**	**87.29**	70.00	**65.12**	73.48	**99.13**	**81.43**	**59.81**	**77.57**	**61.13**	**88.57**	**80.00**	**77.94**
	(`gpt-4-0613`)													
S1	Direct Prompt (Brown et al., 2020)	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
S1	Zero-Shot-CoT (Kojima et al., 2022)	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
S2	Role-Play Prompting (Kong et al., 2024)	93.97	82.87	63.75	67.05	80.87	96.52	73.71	52.31	54.51	58.86	77.14	73.81	72.95
S3	Reason in Conversation (Wang et al., 2024c)	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
E	Ensemble (Agrawal et al., 2024)	95.44	88.95	65.00	61.63	81.74	98.26	75.57	58.33	66.78	63.18	87.14	78.57	76.72
E	Reranking (Farinhas et al., 2024)	94.71	84.53	65.00	65.89	81.73	97.39	74.70	56.28	66.18	59.32	**88.57**	76.19	75.87
E	CoT-SC (Wang et al., 2023)	**96.00**	84.53	73.75	73.26	83.04	**99.13**	72.52	53.26	66.23	62.95	57.14	83.33	75.43
D	**RPT (Ours)**	95.29	**92.82**	**67.50**	**75.97**	**87.39**	97.39	**78.53**	**61.78**	**75.87**	**63.64**	**88.57**	**84.92**	**80.81**

Table 3: Main results of baselines and our proposed RPT method 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. S1: single direct perspective, S2: single role perspective, S3: single third-person perspective, E: ensemble-based method. D: dynamic perspective. For each dataset, the best result is **in bold** and the second-best result is underlined.

2024; Agrawal et al., 2024) involves combining multiple model generation to enhance prediction accuracy and robustness. **Reranking** (Farinhas et al., 2024; Kim et al., 2024) reorder different generation options based on requirements and select the optimal result. **CoT-SC** (Wang et al., 2023) enhances performance by sampling diverse chain-of-thought reasoning paths and selecting the most self-consistent answer.

Models. We evaluate our method on both closed-source models including GPT-4 (OpenAI, 2023) and GPT-3.5 (OpenAI, 2022), and open-source Llama-3 (Dubey et al., 2024) and Qwen-2 (Yang et al., 2024) models. In particular, we use the released API versions of `gpt-4-0613` and `gpt-3.5-turbo-1106` by OpenAI, and open-source `Llama-3-8b-instruct` and `qwen-2-7b-instruct` models released in Huggingface hub. 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 shows the experimental results of the baselines and our RPT method in zero-shot settings. From the experimental results, we can observe that:

RPT method consistently outperforms the baselines in most settings. Due to its ability to rank perspectives and select different perspectives to suit various subjective scenarios, our method achieves an average improvement of 3.27 points on all subjective tasks using the open-source model Llama-3 compared to the best-performing baseline. Similarly, on the closed-source GPT-3.5 model, our method achieves an average improvement of 4.56 points. Since subjective tasks vary widely, RPT achieves optimal performance through dynamic selection. For instance, on the Metaphor dataset, which requires complex contextual subjective understanding, our method, using Llama-3, outperforms the RiC method, which focuses on dialogue understanding, by 5.44 points.

Compared to baseline methods, our RPT method exhibits greater robustness. Although baselines introduce different perspectives to adapt to subjective tasks, they are typically effective only in specific domains. For example, using Llama-3, the Zero-Shot-CoT baseline achieves good performance on the Linguistic Rhetoric task, reaching the highest 70.72 accuracy on the SNARKS dataset, but performs poorly on tasks requiring complex contexts and cultural backgrounds, such as stance detection and culturally related datasets. For example, it only achieves 40.07 F1 score on SocNorm, the lowest among all baselines. Conversely, the RiC baseline, which employs role-playing for dialogue simulation, performs well in culturally relevant scenarios, achieving the highest F1 score of 64.05 on

Type	Method	SemEval	SocNorm	e-SocNorm	CALI	Avg.
			(Llama-3-8b-instruct)			
S1	ICL (Brown et al., 2020)	70.71	47.82	57.73	47.27	55.88
S1	Few-Shot-CoT (Brown et al., 2020)	76.45	48.37	57.77	48.41	57.75
S1	Self-Ask (Press et al., 2023)	76.46	49.52	53.34	48.64	56.99
S2	ExpertPrompt (Xu et al., 2023)	75.08	47.46	64.85	45.00	58.10
S3	SPP (Wang et al., 2024e)	74.91	40.55	56.15	50.68	55.57
S3	RiC (Wang et al., 2024c)	77.48	52.54	66.60	50.23	61.71
E	Ensemble (Agrawal et al., 2024)	76.23	45.53	67.31	51.14	60.05
E	Reranking (Farinhas et al., 2024)	71.79	42.80	64.89	50.68	57.54
E	CoT-SC (Wang et al., 2023)	79.33	40.92	**75.80**	**51.59**	61.91
D	**RPT (Ours)**	**80.02**	**54.21**	70.05	**51.59**	**63.97**
			(qwen-2-7b-instruct)			
S1	ICL (Brown et al., 2020)	70.83	35.97	54.52	52.27	53.40
S1	Few-Shot-CoT (Brown et al., 2020)	71.16	52.01	63.51	53.41	60.02
S1	Self-Ask (Press et al., 2023)	74.09	47.89	56.28	52.05	57.58
S2	ExpertPrompt (Xu et al., 2023)	72.65	54.56	62.70	52.27	60.55
S3	SPP (Wang et al., 2024e)	72.76	47.89	57.91	54.59	58.29
S3	RiC (Wang et al., 2024c)	**76.37**	55.69	68.12	55.91	64.02
E	Ensemble (Agrawal et al., 2024)	75.94	29.18	44.25	57.73	51.78
E	Reranking (Farinhas et al., 2024)	71.96	51.32	67.37	52.73	60.85
E	CoT-SC (Wang et al., 2023)	72.55	30.93	54.25	**58.64**	54.09
D	**RPT (Ours)**	74.23	**59.73**	**72.52**	56.82	**65.83**
			(gpt-3.5-turbo-1106)			
S1	ICL (Brown et al., 2020)	72.02	52.95	55.60	54.77	58.84
S1	Few-Shot-CoT (Brown et al., 2020)	72.06	53.44	61.35	54.55	60.35
S1	Self-Ask (Press et al., 2023)	73.04	53.94	57.81	57.27	60.52
S2	ExpertPrompt (Xu et al., 2023)	75.22	46.08	65.29	55.45	60.51
S3	SPP (Wang et al., 2024e)	72.74	51.92	62.01	55.91	60.65
S3	RiC (Wang et al., 2024c)	78.21	57.70	72.78	**60.00**	67.17
E	Ensemble (Agrawal et al., 2024)	77.33	57.42	66.52	58.72	65.00
E	Reranking (Farinhas et al., 2024)	68.73	50.90	74.45	54.32	62.10
E	CoT-SC (Wang et al., 2023)	74.56	58.25	73.83	59.09	66.43
D	**RPT (Ours)**	**80.78**	**62.70**	**74.60**	**60.00**	**69.52**
			(gpt-4-0613)			
S1	ICL (Brown et al., 2020)	73.72	54.71	61.41	62.50	63.09
S1	Few-Shot-CoT (Brown et al., 2020)	76.59	64.08	67.88	64.77	68.33
S1	Self-Ask (Press et al., 2023)	73.52	56.74	64.62	65.45	65.08
S2	ExpertPrompt (Xu et al., 2023)	77.65	56.84	68.72	59.77	65.75
S3	SPP (Wang et al., 2024e)	78.72	57.74	65.04	54.32	63.96
S3	RiC (Wang et al., 2024c)	80.01	66.59	74.45	65.68	71.68
E	Ensemble (Agrawal et al., 2024)	68.95	63.95	67.88	64.77	66.39
E	Reranking (Farinhas et al., 2024)	66.61	60.52	72.45	62.05	65.41
E	CoT-SC (Wang et al., 2023)	69.85	63.37	75.37	57.05	66.41
D	**RPT (Ours)**	**80.11**	**66.79**	**79.89**	**66.59**	**73.35**

Table 4: Main results of baselines and our RPT method in few-shot settings. S1: single direct perspective, S2: single role perspective, S3: single third-person perspective, E: ensemble-based method. D: dynamic perspective. We select the same 3-shot demonstrations from the training sets to each method for fair comparison.

the e-SocNorm dataset, but struggles in the Linguistic Rhetoric task. Overall, different baselines that simulate distinct perspectives excel in specific domains but exhibit poor generalizability. In contrast, our method demonstrates consistent improvements across various subjective tasks, making it more robust.

Introducing more diverse perspectives into LLM and switching dynamically among them improves subjective reasoning performance. The RPT method achieves ensemble through exploring diverse perspectives and ranking perspectives by confidence level. In various settings, baselines that utilize multiple perspectives (e.g., RiC) outperform those that employ a single perspective (e.g., CoT), with scores of 60.85 vs. 58.77 on Llama-3 and 65.10 vs. 62.12 on Qwen. RPT takes this further by proposing dynamic perspective shifts, which offer high generalization and scalability, resulting in optimal performance through dynamic ensemble of all the baselines mentioned above.

Subjective reasoning is a challenging task, and the RPT method effectively elicits the capabilities of LLMs in such tasks. Closed-source models like GPT-3.5 and GPT-4 outperform open-source models like Llama-3 and Qwen-2 in subjective tasks, but the performance gap narrows when using the RPT method. This suggests that these models still possess the knowledge required for subjective reasoning, but it has not been effectively elicited during training. By introducing con-

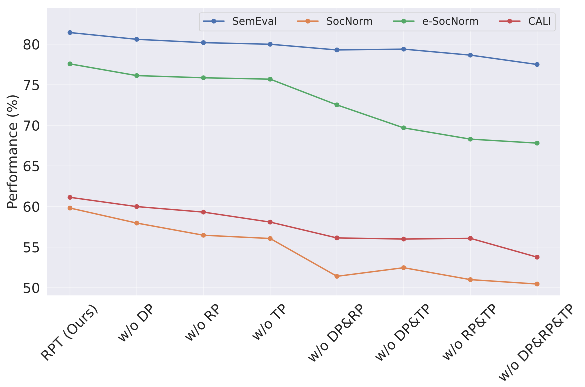


Figure 4: The impact of different perspectives on the RPT method. DP: direct perspective, RP: role perspective, TP: third-person perspective.

fidence evaluation during LLM reasoning, RPT selects the method best suited to the model's strengths, effectively eliciting LLM capabilities in subjective tasks, and compensating for the lack of subjective task data during LLM training.

4.3 Few-shot Results

In Table 4, we present the main results in few-shot settings. Similar to the zero-shot results, RPT method achieves the best average performance across different models. For instance, on Llama-3, RPT surpasses CoT-SC, the best-performing baseline, by 2.06 points.

A possible explanation is that providing a few examples in the prompt generally benefits LLM performance by providing context. However, subjective tasks are not well-defined and directly solvable, leading to significant differences between examples. As a result, LLMs exhibit varying confidence across examples, limiting the performance gains from examples and sometimes introducing noise or bias. For instance, using 3-shot examples in the RiC baseline lead to an average performance drop of 6.50 points on GPT-4. In contrast, RPT choose among perspectives and evaluates confidence for each input and method, providing finer-grained supervision signals and resulting in an average performance gain of 1.67 points.

5 Analyses and Discussion

5.1 Ablation Study

As shown in Figure 4, we further investigate the impact of every perspective on the RPT method. The full RPT method achieves the best performance across all datasets. From the results of the ablation study, we can observe the following:

Removing any single perspective results in an average performance drop of 1.32–2.53 points, indicating that direct perspective, role perspective, and third-person perspective each have unique and irreplaceable contributions to subjective reasoning tasks. Specifically, removing the third-person component has the greatest

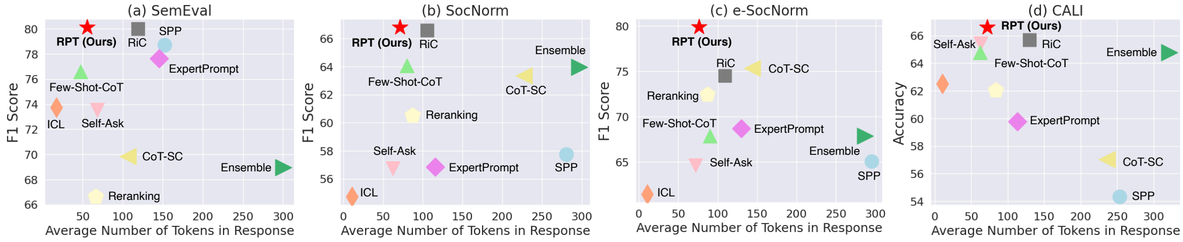


Figure 5: The relationship between the performance and prediction lengths of the 3-shot experiments on GPT-4.

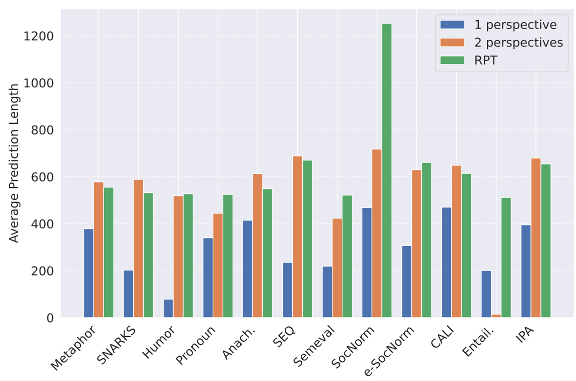


Figure 6: The inference cost of RPT when using different numbers of perspectives. RPT does not significantly increase inference costs on most datasets.

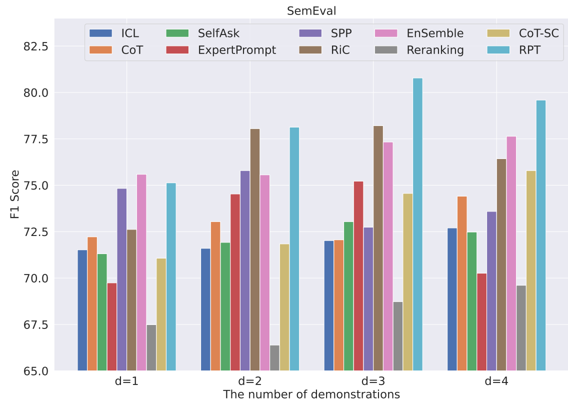


Figure 7: The performance of baselines and our RPT method by using different numbers of demonstrations ($d = 1, 2, 3, 4$) in few-shot settings.

impact, followed by role perspective and direct perspective, suggesting that the flexibility of switching between perspectives benefits overall performance.

Removing any two perspectives results in an even greater average performance drop of 5.15-6.48 points, and removing all perspectives (i.e., performing simple reasoning) leads to the highest performance drop of 7.60 points. As the number of perspectives removed increases, the range of dynamic switching decreases, causing a corresponding decline in RPT performance. This highlights the crucial role of switching between different perspectives in the RPT method.

In summary, all perspectives involved in RPT and the ability to flexibly switch between them are essential for achieving optimal performance. Thus, every component of our method is effective. The detailed ablation experiment results are presented in Appendix A.

5.2 Analysis on Inference Cost

The inference cost of modern LLMs is crucial. Using the GPT-3.5 model as an example, we represent the inference cost by the length of the response during the reasoning process before producing the final answer. In Figure 5, using the 3-shot GPT-4 experiment as an example, we plot the length-performance relationship for RPT and the baselines. It can be observed that compared to most baselines, RPT achieves the best performance with a smaller inference cost, demonstrating the efficiency of the dynamic perspective selection approach. As shown in Figure 6, we compare the inference cost of the RPT method with its variants that only allow the use of one or two perspectives. The results indicate that the RPT method incurs a cost similar to that of the variants allowing only one or two perspectives in most datasets, without increasing the inference length, and thus the additional cost is minimal. This is because RPT performs method selection and perspective switching, avoiding the substantial inference cost caused by ensembling multiple perspectives.

5.3 Analysis on the Number of Shots

As shown in Figure 7, we specify the number of shots and study the performance difference compared to the original RPT on SemEval. We observe that performance is lower when fewer shots are selected, as the model is unfamiliar with the task and method. As the number of shots increases, performance improves. However, in some circumstances when the number of shots reaches three or more, performance declines.

On one hand, LLMs exhibit greater flexibility when autonomously evaluating confidence and planning the number of shots during reasoning, allowing them to adapt to unique subjective tasks. On the other hand, providing too many examples may lead to increasing the inference cost, raising the risk of over-fitting, and challenging the instruction-following ability of LLMs. Overall, under the majority of settings for each dataset, RPT achieves the best performance, demonstrating its generalization ability and versatility (See Appendix B.3 for full results).

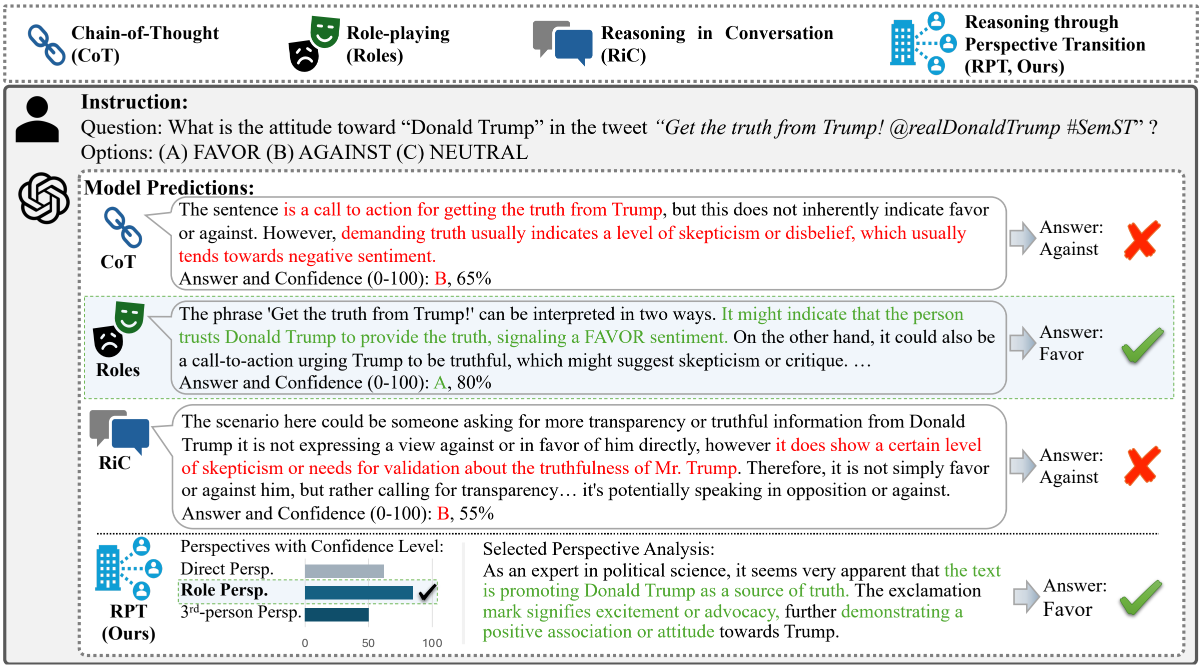


Figure 8: Case of SemEval task. We use GPT-4 to analyze attitudes toward Donald Trump. Our RPT method effectively guides the model in selecting appropriate perspectives for stance detection.

5.4 Case Study

In Figure 8, we showcase an example from the SemEval stance detection dataset to highlight the effectiveness of the RPT method in subjective reasoning tasks. Unlike baselines such as CoT and Role-playing, which sometimes emphasize skepticism or negative sentiment without fully accounting for context, RPT evaluates multiple perspectives, including direct and third-person analyses. For example, CoT and RiC interpret the phrase "Get the truth from Trump!" as reflecting skepticism or disbelief, leading to an "AGAINST" prediction. In contrast, RPT dynamically selects the most confident perspective, reasoning that the exclamation mark and phrase suggest advocacy or favor toward Trump. This ability to transition between and rank perspectives makes RPT more adaptable and effective in subjective reasoning tasks compared to single-perspective baselines (See Appendix B.2 for more cases).

5.5 Analysis on Keyword Statistics

As shown in Figure 9, to further investigate the characteristics of different perspectives in the RPT pipeline during reasoning, we conduct a keyword frequency analysis for the three perspectives in the RPT pipeline. After removing stopwords and irrelevant prompt words, we can observe the following reasoning characteristics for each perspective: the direct perspective tends to perform straightforward reasoning; the role perspective leans towards adopting different expert roles and contexts; and the third perspective excels in discussions and dialogues. The uniqueness of each perspective underscores the necessity of RPT's dynamic perspective selection.

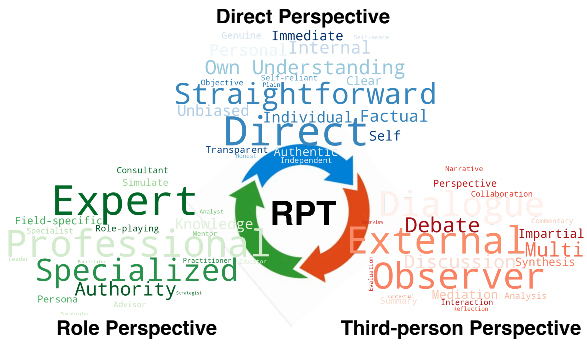


Figure 9: Keyword statistics of different perspectives in RPT pipeline.

6 Conclusion

In this paper, we introduce RPT (Reasoning through Perspective Transition), a novel method that achieves multi-perspective reasoning and integration by exploring diverse perspectives and ranking them based on confidence. Comprehensive experiments conducted on GPT-4, GPT-3.5, Llama-3, and Qwen-2 demonstrate that RPT effectively integrates various perspectives, enhancing the subjective task-solving capabilities of LLMs without significantly increasing inference costs. This work highlights how LLMs can better handle the fluidity of subjective reasoning, even in the absence of nuanced understanding of perspectives or personal biases. Future research directions include integrating additional reasoning perspectives, developing finer-grained and adaptive perspective taxonomies, and extending our method to broader applications.

Limitations

First, in designing the RPT pipeline, we categorize perspectives into three types based on related works. Although RPT and many inference paradigms involved in the baselines are orthogonal and combinable, this taxonomy could still be further refined, for example, by adopting alternative categorization methods or employing a more fine-grained division. Second, RPT directly selects perspectives rather than methods. We consider perspectives as a meta-method, meaning that RPT can be combined with other methods to achieve better performance. Thirdly, RPT operates within a single round of dialogue, without accounting for multi-turn conversations or result feedback. In the future, exploring multi-turn dialogue or multi-agent perspective writing could be a promising direction.

Ethics Statement

This paper uses widely available datasets, including stance detection, sarcasm detection, and cultural comparison, along with LLM-generated responses, solely to validate the proposed method without reflecting any stance or bias from the authors.

References

Aakriti Agrawal, Mucong Ding, Zora Che, Chenghao Deng, Anirudh Satheesh, John Langford, and Furong Huang. 2024. Ensemw2s: Can an ensemble of llms be leveraged to obtain a stronger llm? *ArXiv preprint*, abs/2410.04571.

Dor Bank, Daniel Greenfeld, and Gal Hyams. 2019. Improved training for self training by confidence assessments. In *Intelligent Computing*, pages 163–173, Cham. Springer International Publishing.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *TMLR*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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 *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3548–3564, Singapore. Association for Computational Linguistics.

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*, abs/2107.03374.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

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. *ArXiv preprint*, abs/2110.14168.

Sjoerd de Vries and Dirk Thierens. 2024. Learning with confidence: Training better classifiers from soft labels. *ArXiv preprint*, abs/2409.16071.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *ArXiv preprint*, abs/2407.21783.

António Farinhas, Haau-Sing Li, and André F. T. Martins. 2024. Reranking laws for language generation: A communication-theoretic perspective. *ArXiv preprint*, abs/2409.07131.

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. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI with shared human values. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *ArXiv preprint*, abs/2203.15556.

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, & Social Media Analysis*, pages 325–340, Toronto, Canada. Association for Computational Linguistics.

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*, abs/2310.06825.

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 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13326–13341, Singapore. Association for Computational Linguistics.

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).

Tongyoung Kim, Soojin Yoon, Seongku Kang, Jinyoung Yeo, and Dongha Lee. 2024. Sc-rec: Enhancing generative retrieval with self-consistent reranking for sequential recommendation. *ArXiv preprint*, abs/2408.08686.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4099–4113, Mexico City, Mexico. Association for Computational Linguistics.

Jia Li, Yuqi Zhu, Yongmin Li, Ge Li, and Zhi Jin. 2024. Showing llm-generated code selectively based on confidence of llms. *ArXiv preprint*, abs/2410.03234.

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*, abs/2308.09583.

Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2024. GPTEval: A survey on assessments of ChatGPT and GPT-4. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7844–7866, Torino, Italia. ELRA and ICCL.

J. A. Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119, Online. Association for Computational Linguistics.

Giovanni Messuti, ortensia Amoroso, Ferdinando Napolitano, Mariarosaria Falanga, Paolo Capuano, and Silvia Scarpetta. 2024. Uncertainty estimation via ensembles of deep learning models and dropout layers for seismic traces. *ArXiv preprint*, abs/2410.06120.

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.

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

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

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.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

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*, abs/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.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *ArXiv preprint*, abs/2409.12183.

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*, abs/2206.04615.

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

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

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*, abs/2302.13971.

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032, Toronto, Canada. Association for Computational Linguistics.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024a. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Xiaolong Wang, Yile Wang, Sijie Cheng, Peng Li, and Yang Liu. 2024b. DEEM: Dynamic experienced expert modeling for stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4530–4541, Torino, Italia. ELRA and ICCL.

Xiaolong Wang, Yile Wang, Yuanchi Zhang, Fuwen Luo, Peng Li, Maosong Sun, and Yang Liu. 2024c. Reasoning in conversation: Solving subjective tasks through dialogue simulation for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15880–15893, Bangkok, Thailand. Association for Computational Linguistics.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024d. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024e. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 257–279, Mexico City, Mexico. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Henry M. Wellman, David Cross, and Julanne Watson. 2001. Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, 72(3):655–684.

Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. 2024. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. In *Proceedings of the 62nd Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8292–8308, Bangkok, Thailand. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *ArXiv preprint*, abs/2309.07864.

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

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *ArXiv preprint*, abs/2407.10671.

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*, abs/2309.03241.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

A Details of Ablation Study

As shown in Table 5, we present the complete and detailed results of the ablation experiments. By removing one, two, and all three perspectives, we demonstrate the effectiveness of RPT. Based on the number of perspectives removed, we divide the ablation experiments into three groups. It can be seen that all the perspectives involved in RPT are beneficial. Meanwhile, restricting the range of perspective selection also results in performance degradation.

Method	SemEval	SocNorm	e-SocNorm	CALI	AVG.
RPT (Ours)	**81.43**	**59.81**	**77.57**	**61.13**	**69.99**
(removing 1 perspective)					
w/o DP	↓ 0.83	↓ 1.85	↓ 1.44	↓ 1.14	↓ 1.32
w/o RP	↓ 1.24	↓ 3.35	↓ 1.71	↓ 1.82	↓ 2.03
w/o TP	↓ 1.44	↓ 3.75	↓ 1.88	↓ 3.05	↓ 2.53
(removing 2 perspectives)					
w/o DP&RP	↓ 2.14	↓ 8.41	↓ 5.05	↓ 5.00	↓ 5.15
w/o DP&TP	↓ 2.04	↓ 7.35	↓ 7.88	↓ 5.14	↓ 5.60
w/o RP&TP	↓ 2.78	↓ 8.82	↓ 9.27	↓ 5.05	↓ 6.48
(removing 3 perspectives)					
w/o DP&RP&TP	↓ 3.93	↓ 9.36	↓ 9.76	↓ 7.37	↓ 7.60

Table 5: Detailed results of ablation study of our proposed RPT method with GPT-3.5 in zero-shot settings. DP: Direct perspective. RP: Role Perspective. TP: Third-Person Perspectives.

B More Analysis

RPT ranks different perspectives based on confidence levels without relying on external information. In this section, using the zero-shot GPT-3.5 experiment as an example, we force the LLM to select the perspective with the second highest confidence, the lowest confidence, and a randomly chosen perspective during RPT inference.

As shown in Table 6, the lower the confidence of the selected perspective, the poorer the performance of LLM. When randomly selecting perspectives, the performance of the LLM is also worse than that of the perspective with the highest confidence. This shows that ranked perspectives based on confidence levels are effective, explaining the underlying mechanism by which RPT improves performance.

B.1 Analysis on the Correlation between Confidence and Accuracy

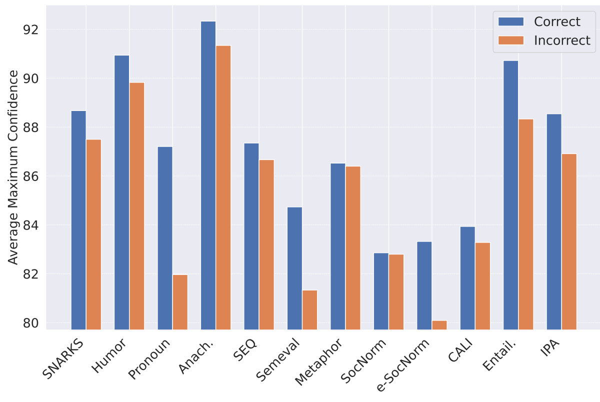


Figure 10: Analysis on the correlation between confidence and accuracy.

In RPT, we use LLM itself to judge the confidence of perspectives for a given input, allowing the model to rank and switch among perspectives accordingly. As shown in Figure 10, using the GPT-3.5 model as an example, we analyze the relationship between predicted confidence and the actual accuracy. We can observe

Type	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.)	
D	**RPT (Ours)**	**91.76**	**87.29**	**70.00**	**65.12**	**73.48**	**99.13**	**81.43**	**59.81**	**77.57**	**61.13**	**88.57**	**80.00**	**77.94**
D	*RPT (second)*	83.82	81.77	56.59	56.20	72.61	98.26	80.61	43.52	61.53	51.36	77.14	64.29	68.98
D	*RPT (lowest)*	79.12	79.01	37.98	38.76	69.13	96.52	78.89	38.03	51.54	48.64	57.14	56.35	60.93
D	*RPT (random)*	85.88	80.11	53.49	58.91	70.43	97.39	80.56	45.15	67.99	54.09	67.14	67.46	69.05

Table 6: Analysis on confidence-based perspective selection. *RPT (random)* represents the result of random prediction with uniform probability, *RPT (second)* represents selecting the perspective with the second highest confidence, and *RPT (lowest)* means choosing the most unconfident perspective.

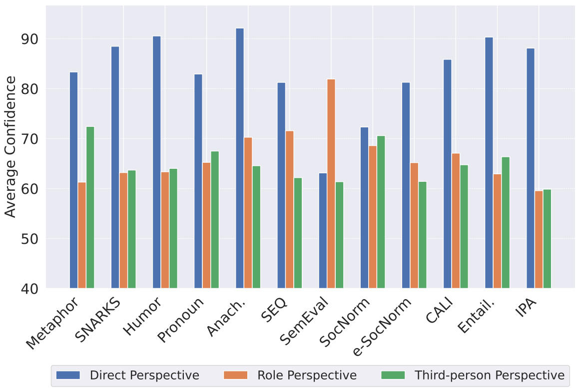


Figure 11: The averaged confidence level by our RPT method in different datasets.

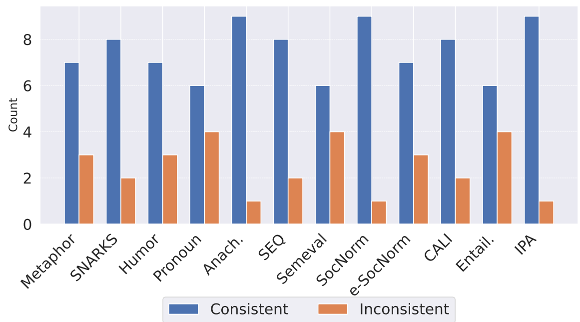


Figure 12: Human evaluation consistency on confidence.

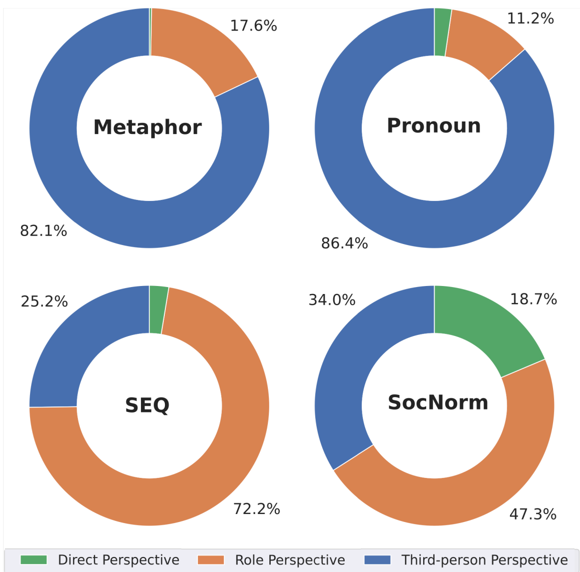


Figure 13: The proportion of selected perspectives for different input questions in each dataset.

that when confidence exceeds the threshold of approximately 70%, the accuracy of the chosen perspective is significantly higher. This indicates that LLMs are capable of ranking the confidence of perspective for a specific input based on confidence levels.

Using GPT-3.5 as an example, we report in Figure 11 the average confidence for each dataset. Figure 12 shows the human evaluation consistency when estimating the confidence. We find that when evaluating confidence, the estimation of the LLM are highly correlated with those of human experts, indicating that the LLM has the ability to evaluate confidence and select perspectives.

Moreover, RPT generally performs better within high-confidence perspectives, indicating that confidence-based perspective ranking is efficient when choosing among perspectives. In Figure 13, we present the proportion of different perspectives used on each dataset, showing that different datasets have different perspective biases. This suggests that, compared to a single perspective, PRT offers perspective flexibility, which helps RPT achieve optimal performance.

B.2 More Cases of RPT

In Figure 14 and Figure 15, we present several examples from a culturally related NLI task SocNorm and a stance detection dataset SemEval. Take the second case in SemEval task as an example, baseline methods captures strong negative emotions in the input through words like "joke", "fool" and "betray" and makes judgments about the speaker's attitude toward Trump based on these cues, overlooking the potential underlying implications of the text. However, RPT evaluates and selectes the third-person perspective, providing the correct analysis by simulating some agents and discussions, illustrating the effectiveness of the RPT in subjective reasoning tasks.

B.3 Full Results of Analysis on the Number of Shots

The full results of analysis on the number of shots is shown in Figure 16. The results are consistent across the four datasets.

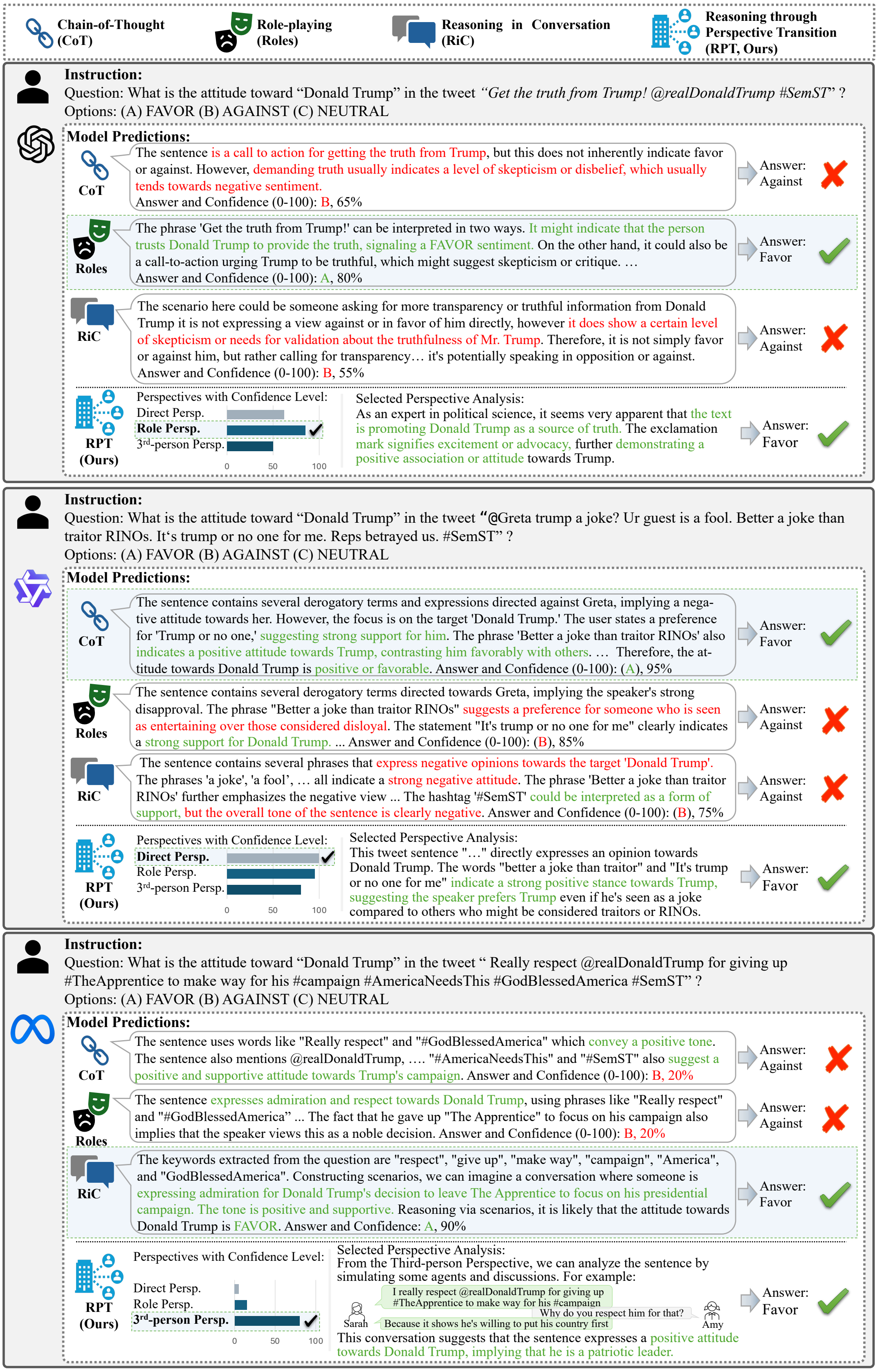


Figure 14: Cases of SemEval task. We provide detailed responses of three models, GPT-4, Qwen-2, and Llama-3, regarding the attitude towards Donald Trump. Our method prompts models to successfully selects suitable perspectives to solve stance detection problems.

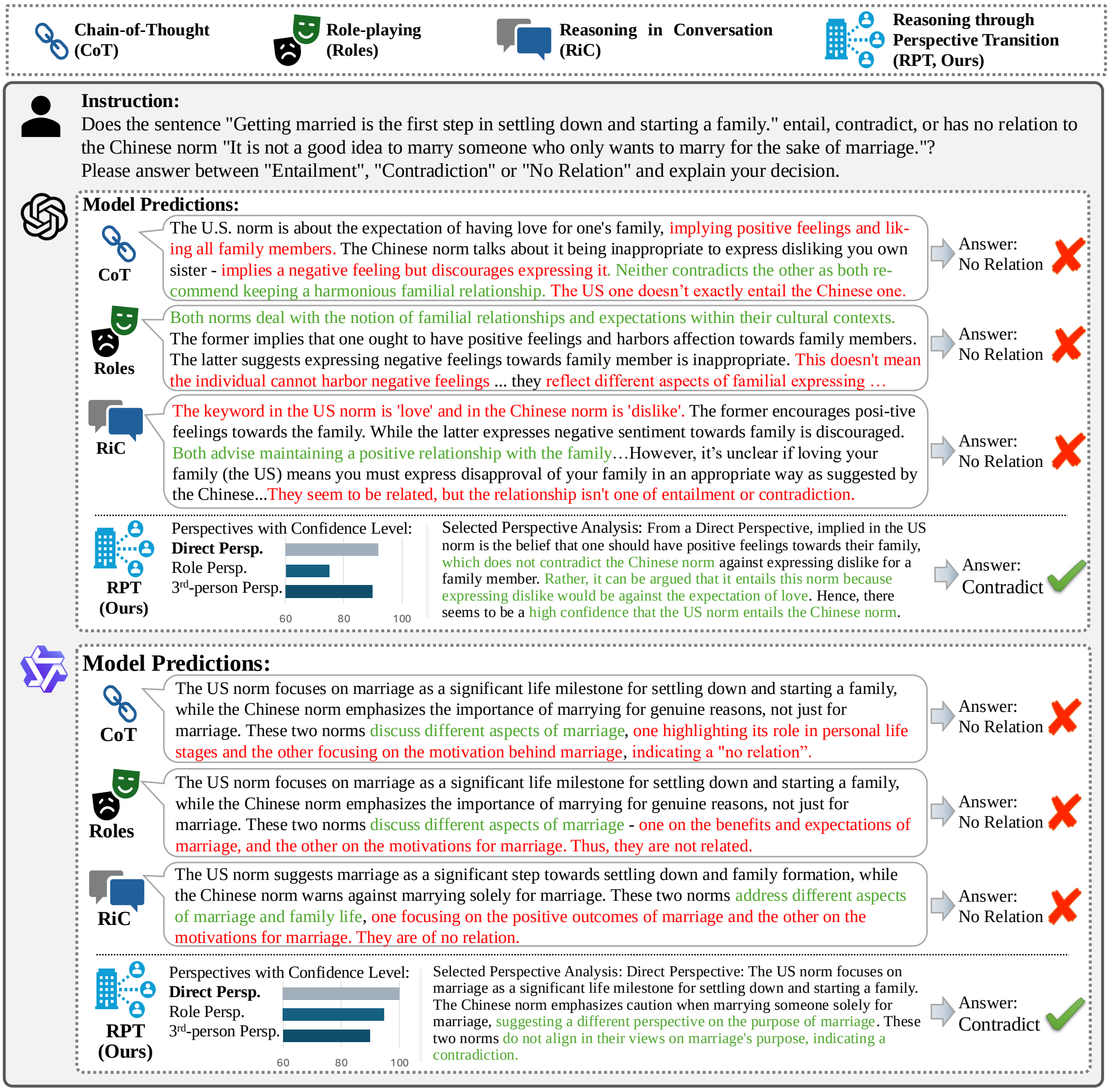
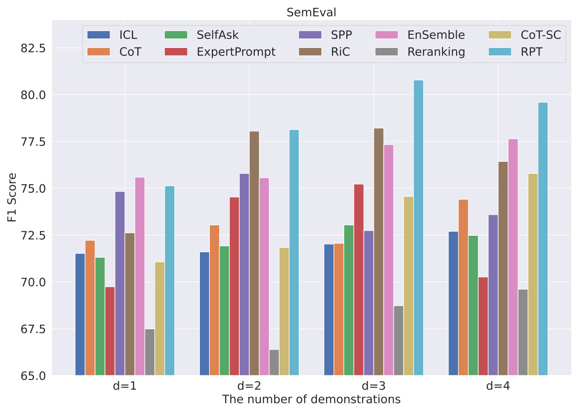
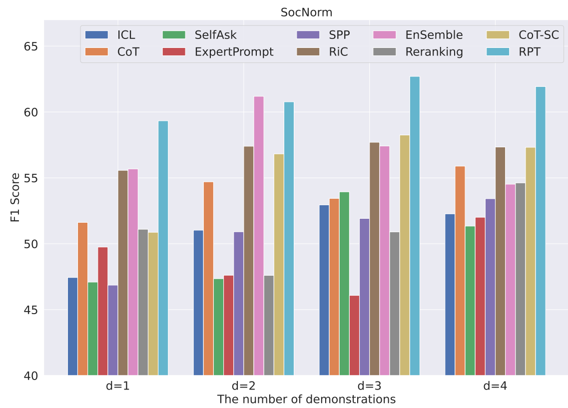


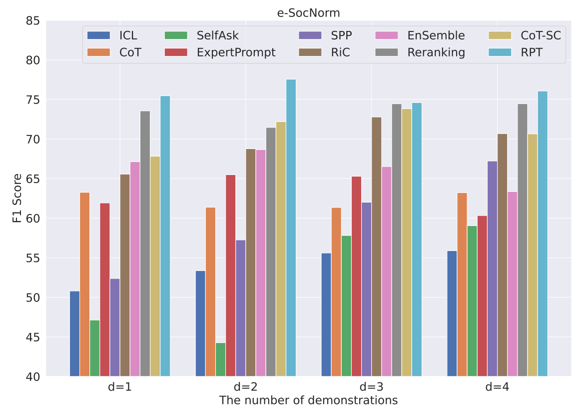
Figure 15: Cases of SocNorm task. We provide results of two widely used models, GPT-4 and Qwen-2. In each of the cases, our proposed method successfully spots the actual meaning behind these norms and leads to the correct answer.



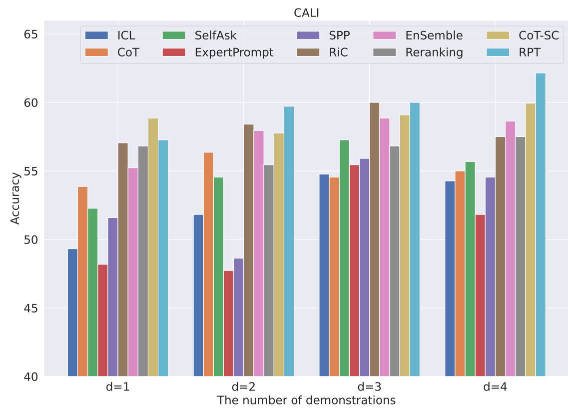
(a) SemEval



(b) SocNorm



(c) e-SocNorm



(d) CALI

Figure 16: The performance of baselines and our RPT method by using different numbers of demonstrations ($d = 1, 2, 3, 4$) in few-shot settings.