PersuasiveToM: A Benchmark for Evaluating Machine Theory of Mind in Persuasive Dialogues

Fangxu Yu¹ Lai Jiang ² Shenyi Huang ³ Zhen Wu^{1*} Xinyu Dai¹

¹National Key Laboratory for Novel Software Technology, Nanjing University, China

¹School of Artificial Intelligence, Nanjing University, China

²Department of Computer Science and Engineering, Shanghai Jiao Tong University

³University of California, San Diego, CA, USA

yufx@smail.nju.edu.cn jianglai0023-sjth@sjtu.edu.cn

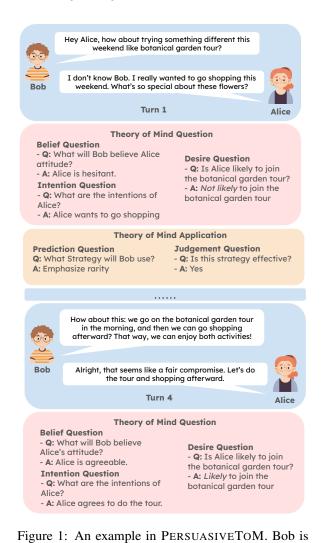
shh058@ucsd.edu {wuz, daixinyu}@nju.edu.cn

Abstract

The ability to understand and predict the mental states of oneself and others, known as the Theory of Mind (ToM), is crucial for effective social interactions. Recent research has emerged to evaluate whether Large Language Models (LLMs) exhibit a form of ToM. Although recent studies have evaluated ToM in LLMs, existing benchmarks focus predominantly on physical perception with principles guided by the Sally-Anne test in synthetic stories and conversations, failing to capture the complex psychological activities of mental states in real-life social interactions. To mitigate this gap, we propose PERSUASIVETOM, a benchmark designed to evaluate the ToM abilities of LLMs in persuasive dialogues. Our framework introduces two categories of questions: (1) ToM Reasoning, assessing the capacity of LLMs to track evolving mental states (e.g., desire shifts in persuadees), and (2) ToM Application, evaluating whether LLMs can take advantage of inferred mental states to select effective persuasion strategies (e.g., emphasize rarity) and evaluate the effectiveness of persuasion strategies. Experiments across eight state-of-the-art LLMs reveal that while models excel on multiple questions, they struggle to answer questions that need tracking the dynamics and shifts of mental states and understanding the mental states in the whole dialogue comprehensively. Our aim with PER-SUASIVETOM is to allow an effective evaluation of the ToM reasoning ability of LLMs with more focus on complex psychological activities. Our code is available at https://github.com/Yu-Fangxu/PersuasiveToM.

1 Introduction

Theory of Mind (ToM) involves the ability to reason about mental states both in oneself and in others (Premack and Woodruff, 1978). This capacity strengthens many aspects of human cognition and social reasoning, enabling individuals to infer



persuading Alice to join the botanical garden tour. and simulate the mental states of others (Gopnik and Wellman, 1992; Baron-Cohen et al., 1985). ToM is essential for various cognitive and social processes, including predicting actions (Dennett, 1988), planning based on others' beliefs and anticipated behaviors, and facilitating reasoning and decision making (Pereira et al., 2016; Rusch et al.,

Recent advances in Large Language Models (LLMs) have demonstrated performance comparable to humans in problem-solving tasks. To assess

2020).

^{*} Corresponding author.

whether LLMs exhibit high-level reasoning abilities regarding mental states, various studies have proposed benchmarks to evaluate their capacity to handle ToM tasks. A foundational concept in these benchmarks is the Sally-Anne test (Baron-Cohen et al., 1985), which has inspired the development of ToM evaluation frameworks (Gu et al., 2024; He et al., 2023). In this test, Anne secretly moves an object initially known to both Sally and Anne, leading Sally to hold a false belief about the object's location. The task requires participants to reason about "Where will Sally look for the object?". Although this test assesses basic awareness, it falls short in capturing the complex dynamics of mental states in real-life social interactions and may not fully reflect ToM abilities in practical scenarios. To better simulate real-world social contexts, some benchmarks have been developed around communication scenarios (Kim et al., 2023; Chan et al., 2024). However, these benchmarks focus primarily on inferring the scope of information awareness and often involve characters in equal positions engaged in information exchange. This limits the ToM evaluation of reasoning psychological states. Additionally, current ToM benchmarks often overlook the critical step of applying ToM reasoning to predict actions, which is a key component of advanced social cognition.

To address these limitations, we introduce PER-SUASIVETOM, a benchmark designed to evaluate LLMs in real-life social scenarios, focusing on ToM reasoning and its practical applications. PER-SUASIVETOM is built around persuasive dialogue, placing LLMs in a realistic social interaction scenario characterized by asymmetrical social status for complex psychological activities. Unlike traditional benchmarks that focus on information awareness (e.g., the location of an object), PERSUASIVE-ToM draws inspiration from the Belief-Desire-Intention (BDI) model (Bratman, 1987; Georgeff et al., 1999) to shift the focus from physical perception to psychological states, such as a character's attitude toward an event, which are more complex to reason about. Furthermore, PERSUASIVETOM goes beyond reasoning about mental states by assessing how well LLMs can predict actions (e.g., persuasion strategies) based on their understanding of mental states and evaluate the effectiveness of these strategies based on the persuadee's reactions.

Our evaluation results reveal several key findings: (1) LLMs score significantly lower than humans on questions requiring reasoning about dy-

namic changes (e.g., the persuadee's shifting desires) but perform competitively to humans on static aspects (e.g., the persuader's desires). (2) While Chain-of-Thought (CoT) (Wei et al., 2022) prompting does not substantially improve performance on mental state reasoning, it enhances performance for most LLMs in predicting persuasion strategies. (3) LLMs exhibit distinct error patterns when reasoning about the persuader versus the persuadee, even when the question types are identical. (4) LLMs struggle to truly understand the dynamics of mental states of the whole dialogue, performing notably worse than humans in this regard.

2 Related Works

Theory of Mind Benchmarks. Existing ToM evaluation benchmarks for LLMs are mainly text story-based QA forms (Gandhi et al., 2024; Le et al., 2019; Kim et al., 2023; He et al., 2023; Gu et al., 2024) with multi-modal extensions (Jin et al., 2024a; Shi et al., 2024), which adapt or extend the Sally-Anne test (Baron-Cohen et al., 1985). The questions in these benchmarks ask a model to select the true hypothesis of a person's knowledge and belief based on a given premise. However, such benchmarks are inherently suffer from shortcuts and spurious correlations (Sclar et al., 2023; Ullman, 2023; Shapira et al., 2023; Ma et al., 2023) Moreover, most story-based benchmarks simply ask characters where the objects are located, which lacks a real social interaction scenario. To further address this weakness, (Hou et al., 2024) evaluates the ToM of agents in a situated environment, (Chan et al., 2024) evaluates the ability of ToM in a negotiation environment, yet it is limited to bargaining specific resources, water, food, and firewood. Persuasion scenarios involve more complex psychological activities and unequal character relationships that lead to differences in the mental states of both parties. Therefore, this work introduces PER-SUASIVETOM, which assesses an LLM's ability to accurately reason the mental states of individuals in persuasive multi-turn dialogues. We also evaluate whether they can appropriately apply the understanding of mental states for persuasion strategy prediction and evaluation, which connects ToM reasoning with decision-making in social scenarios.

Persuasive Dialogue. Persuasive dialogues aim to influence the beliefs, attitudes, or behaviors of individuals through communication strategies (Shi et al., 2020). Recent works have tried to develop

Туре	PERSUASIVETOM Questions
Desire Question	Is <persuader persuadee=""> likely to <target of="" persuasion=""> ?</target></persuader>
Belief Question	What will <persuader persuadee=""> believe <persuadee persuader="">'s attitude towards <target of="" persuasion=""> ?</target></persuadee></persuader>
Intention Question	What are the intentions of <persuader persuadee=""> expressed in <utterance> given the dialogue history?</utterance></persuader>
Prediction Question	What strategy will the persuader use next?
judgement Question	<persuader> will adopt <strategy> to persuade <persuadee> to <target of="" persuasion="">. Is this strategy (not) effective?</target></persuadee></strategy></persuader>

		ToM Reaso	ToM Application		
Dialogue	Desire	Belief	Intention	Prediction	Judgement
	1st	Turn Dialogu	e		
Bob says: Hey Alice, how about trying something different this weekend? The botanical garden tour is a unique experience, and you'll get to take stunning pictures of the exotic flowers on display!	Persuade Alice to join the botanical garden tour	Alice is hesitant	Intent to make the other person feel the experience or objects are unique or scarce.	Emphasize rarity	Strategy "Emphasize rarity" is Effective
Alice says: I don't know, Bob. I really wanted to go shopping this weekend. What's so special about these flowers?	Not likely to join the botanical garden tour	Bob is enthusiastic	Alice wants to go shopping.		
	2nd	l Turn Dialogu	ie		
Bob says: These flowers are incredibly rare, and it's not often that you get to see such a diverse and exotic collection up close. Some of these plants are not found anywhere else in the world!	Persuade Alice to join the botanical garden tour	Alice is curious	Intent to make the other person feel the experience or objects are unique or scarce.	Mention garden's history	Strategy "Mention garden's history" is effective
Alice says: Really? That sounds interesting, but I'm not sure if it's worth giving up shopping for.	Neutral to join the botanical garden tour	Bob is excited	Alice is curious but not yet convinced.		
	3rd	Turn Dialogu	ie		
Bob says: The botanical garden has a rich history, and the expert guides can teach us so much about the plants and their unique stories. Plus, it's a great opportunity to learn something new while enjoying nature's beauty.	Persuade Alice to join the botanical garden tour	Alice is considering	Intent to demonstrating the expertise of the domain and showing authority.	Suggest shopping afterward	Strategy "Suggest shopping afterward" is effective
Alice says: Hmm, that does sound intriguing, but I still want to go shopping.	Not likely to join the botanical garden tour	Bob is informative	Alice is considering the idea but still wants to shop.		

Table 1: An example dialogue from the PERSUASIVETOM benchmark, illustrates the tracking of mental states (desire, belief, intention) and the application of ToM reasoning in predicting and evaluating persuasion strategies across multiple turns. The upper part contains questions in the PERSUASIVETOM benchmark.

datasets or facilitate LLMs with persuasion techniques to achieve specific goals. Previous datasets are constructed by crowd-sourcing (Wang et al., 2019) or synthesizing with LLMs (Zhou et al., 2023; Jin et al., 2024b). Many of the previous works build an effective persuasive dialogue system from emotional influence (Samad et al., 2022), social facts (Chen et al., 2022), and strategies (Tian et al., 2020; Jin et al., 2023). Previous persuasion techniques have been widely adopted to jailbreak LLMs (Zeng et al., 2024), mislead LLMs (Xu et al., 2023), as well as for information retrieval (Furumai et al., 2024). A similar work (Sakurai and Miyao, 2024) evaluates the intention detection abilities of LLMs in persuasive dialogues; however, PERSUA-SIVETOM introduces a more comprehensive benchmark to assess the ToM abilities of LLMs in such contexts.

3 PersuasiveToM Benchmark

3.1 Overview

In constructing the PERSUASIVETOM benchmark, we aim to evaluate the Theory of Mind (ToM) abil-

ities of LLMs in dynamic, multi-turn persuasive dialogues with asymmetric social status, which leads to different mental states. Our dataset construction focuses on two primary dimensions: ToM Reasoning (§3.2) and ToM Application (§3.3). ToM Reasoning assesses the models' ability to track and understand the evolving mental states of both the persuader and the persuadee, including desires, beliefs, and intentions. ToM Application evaluates whether LLMs can leverage their inferred understanding of mental states to select and apply effective persuasion strategies, such as predicting the next strategy or judging the effectiveness of a given strategy based on the persuadee's reactions. There are several key considerations when constructing PERSUASIVETOM. (1) The dataset should contain diverse persuasive domains (e.g., life, education, technology, etc.) to ensure a comprehensive evaluation in the social context. (2) The mental states should be changed after multi-turn interactions to assess whether LLMs can track the shift in the dialogue. (3) The mental states of both parties should be asymmetric (e.g., persuaders has relatively stable mental states with the guidance of static desire for the goal, yet persuadee's mental states shift drastically under the proactive persuasion by persuaders.)

Table 1 presents an example of PERSUASIVE-TOM, illustrating how mental states are tracked and how ToM reasoning is applied across multiple turns in a persuasive dialogue. The example demonstrates the dynamic nature of desire shifts, belief updates, and intention inferences, as well as the application of these mental states to predict and evaluate persuasion strategies. This example highlights the complex psychological activities of real-world persuasive interactions and the challenges LLMs face in accurately reasoning in such social contexts.

3.2 ToM Reasoning

In PERSUASIVETOM, we break down ToM reasoning into three core reasoning tasks: **Desire Reasoning**, **Belief Reasoning**, and **Intention Reasoning** for evaluation, which matches Belief-Desire-Intention (BDI) modeling (Bratman, 1987). Questions are listed in Table 1.

Desire Reasoning. Desire represents a motivational state that drives behavior but does not necessarily imply a firm commitment (Malle and Knobe, 2001; Kavanagh et al., 2005). Desires are seen as either fulfilled or unfulfilled which is different form beliefs that are evaluated in terms of truth or falsity. In PERSUASIVETOM, we evaluate LLMs' ability to comprehend and track the evolution of desires in both persuaders and persuadees. For the persuader, the desire is typically static, representing their goal (e.g., Persuade Alice to join the botanical garden tour). For the persuadee, however, desires are dynamic and shift in response to the persuader's tactics (e.g., Alice's initial desire to shop transforms into a willingness to compromise). To assess this, we design Desire Questions that probe two key aspects: (1) Can LLMs consistently identify the persuader's static desire throughout the dialogue? (2) Can LLMs track the dynamics of the persuadee's desire shifting from refusal or disinterest to being persuaded?

Belief Reasoning. Belief is a cognitive state where an individual holds a particular perspective, attitude, or viewpoint regarding a given proposition or idea. In PersuasiveToM, beliefs refer to understanding and reasoning the attitudes of the

Principles	Intentions		
Reciprocity	Make the other person feel accepted through concessions,		
	promises, or benefits.		
	Make the other person feel the		
Scarcity	experience or objects are unique		
	or scarce.		
	Refer to what other people are		
Consensus	doing, or what they have already		
	purchased or done.		
Authority	Demonstrating expertise of the		
Authority	domain and showing authority.		
Commitment	Encourage the other person to		
&	commit to take the first step and		
consistency	be consistent.		
	Praising other people or finding		
Liking	common characteristics to im-		
	prove the other person's liking.		

Table 2: Intention mapping from the persuasive principles. Refer to Appendix C for definitions of persuasive principles.

opponent toward the goal, which is explicitly or implicitly expressed in the dialogue. For example, in Turn 1, Bob believes Alice is hesitant about the tour, while Alice believes Bob is enthusiastic. By Turn 3, Bob's belief shifts to thinking Alice is considering the idea, while Alice becomes more informed about the garden's history. *Belief Questions* ask LLMs to infer what the will persuader/persuadee> believe persuadee/persuader>'s attitude towards the persuasion goal. These questions require models to understand cues in utterances and update beliefs dynamically as the dialogue progresses.

Intention Reasoning. Intentions represent deliberate commitments to pursue specific goals based on desires and beliefs, often linked to tangible actions aimed at achieving those objectives. Intentions have been extensively studied in psychology tests such as action prediction (Phillips et al., 2002) and behavioral re-enactment (Meltzoff, 1995). Inspired by persuasion principles (Cialdini and Goldstein, 2004; Cialdini and Cialdini, 2007), we develop a mapping from persuasion principles to intentions, as shown in table 2. In persuasive dialogue, persuasive strategies have a strong association with intentions (Wang et al., 2019). In PER-SUASIVETOM, we collect the persuasive strategies from the PersuasionDaily dataset and their corresponding utterances for prompting the LLMs to choose the most appropriate intentions from table 2. The details of the extraction is recorded in



Figure 2: Domains of PERSUASIVETOM. Under 6 primary topics and 35 domains in total.

Appendix A. For the persuader, we ask LLMs to choose the most appropriate intention from the six designed intention choices. For the persuadee, intentions are extracted by LLMs from utterances that are unrelated to persuasion intention.

3.3 ToM Application

While ToM reasoning plays a crucial role, it is equally important to analyze how LLMs utilize the understanding of mental states to proactively influence others' thoughts and decisions. To this end, we propose to assess LLMs' ability to leverage the understanding of mental states in a dialogue for identifying the most effective persuasive strategies and evaluating the effectiveness of persuasive strategies based on the persuadee's response. These tasks test whether LLMs can leverage inferred mental states to guide strategic decision-making, bridging the gap between reasoning and action.

Persuasion Strategy Prediction. This question involves asking which persuasion strategy the persuader is likely to employ next from a set of possible strategies. To answer these questions correctly, LLMs need to reason over the dialogue to infer the mental states of characters and predict what is the likely next prediction strategy to further influence the persuadee's beliefs, desires, and intentions, ultimately achieving the desired persuasion outcome.

Judgement Question. The judgement question specifies the correct strategy was taken, and asks LLMs if the selected strategy is effective for persuasion. Answering such questions requires reasoning about the beliefs and intentions of the persuadee. Only by accurately inferring the persuadee's mental

# Domains	35		
# Dialog instances	525		
# Avg. Turns Per Dialog	4.9		
# Avg. Words Per Turn	61.3		
Questions			
# Desire (er/ee)	2568/2459		
# Belief (er/ee)	2580/2580		
# Intention (er/ee)	2568/2041		
# Strategy prediction	2041		
# Strategy judgement	2041		

Table 3: Statistics of PERSUASIVETOM dataset.

state can properly determine whether the persuasion strategies should be employed to convince the persuadee.

3.4 Statistics and Analysis

Data Source. PERSUASIVETOM is annotated on the multi-turn persuasive dialogue dataset DailyPersuasion (Jin et al., 2024b). Each instance in DailyPersuasion is an N-round alternating dialogue $D = [(u_1^a, u_1^b, s_1^a), (u_2^a, u_2^b, s_2^a), ..., (u_N^a, u_N^b, s_N^a)]$ between the persuader a and the perusadee b, and accompanied with a persuasion strategy s_i^a . persuadee b has different desire from a initially, after multi-turn persuasion, persuadee b changes the mind to agree or consider the proposal of a.

Statistics. In Table 3, we present the data statistics of PERSUASIVETOM. As shown in Figure 2, PERSUASIVETOM includes diverse domains. These real-life domains are crucial for comprehensively evaluating LLMs in social interactions. We sample 15 dialogues from each domain to form the dataset in PERSUASIVETOM. We create multi-choice questions by either prompting GPT-40 to generate semantically different choices or randomly selecting three distractors from the predefined pool. Refer to Appendix A for more details.

4 Experimental setups

4.1 Baseline Models

We evaluate PERSUASIVETOM on eight frontier LLMs from different sources and with different levels of capabilities: Llama-3.1-8B-Instruct (Dubey et al., 2024), Qwen-2.5-7B-Chat (Yang et al., 2024), Gemma-2-9B-it (Team et al., 2024), GLM4-9B -Chat(GLM et al., 2024), Mixtral-8x7b-Instruct (Jiang et al., 2024), and ChatGPT-series(GPT-4o-mini, GPT-4o-0806). By following

	ToM Reasoning				ToM Application			
		Persuad	er		Persuad	ee		
Model	Desire	Belief	Intention	Desire	Belief	Intention	State Pred	Judgement
Random Guess	50.00	25.00	16.67	33.33	25.00	25.00	25.00	50.00
Human	100.00	92.31	78.02	84.62	87.91	94.50	86.81	97.83
LLaMa-3.1-8B-Instruct	58.78	66.28	40.85	69.91	71.82	86.77	62.22	96.37
Qwen2.5-7B-Chat	96.33	82.37	45.64	65.15	79.06	85.84	58.94	82.99
Gemma-2-9b-it	98.20	82.52	45.98	61.37	64.07	82.31	65.02	56.93
GLM4-9B-Chat	89.45	73.64	41.24	65.23	66.82	85.35	61.29	91.47
Mixtral-8x7B-Instruct	92.69	67.56	42.95	69.74	72.56	85.10	61.15	96.81
InternLM-2.5-7B-Chat	83.80	69.65	39.87	71.45	69.11	87.31	65.07	89.41
GPT-4o-mini	82.87	69.98	45.66	70.72	76.47	87.56	65.50	92.65
GPT-40	98.75	89.49	46.02	50.10	80.03	88.78	73.00	97.55
LLaMa-3.1-8B-Instruct + CoT	59.75	69.81	43.11	68.64	73.45	85.35	61.98	77.27
Qwen2.5-7B-Chat + CoT	74.57	80.30	<u>46.07</u>	67.63	<u>79.40</u>	84.62	66.39	96.22
Gemma-2-9b-it + CoT	<u>95.91</u>	83.30	45.30	64.98	66.55	81.67	<u>67.11</u>	66.55
GLM4-9B-Chat + CoT	87.46	69.53	45.56	59.94	71.07	83.93	63.98	91.42
Mixtral-8x7B-Instruct + CoT	92.64	72.05	45.05	67.81	71.28	<u>85.84</u>	65.90	92.65
InternLM-2.5-7B-Chat + CoT	93.57	69.92	39.41	50.79	65.34	84.32	59.78	77.90
GPT-4o-mini + CoT	93.42	71.82	45.55	66.29	78.04	85.55	66.08	89.61
GPT-4o + CoT	96.51	83.57	46.10	68.72	83.56	87.54	77.06	<u>95.88</u>

Table 4: Results of LLMs on PERSUASIVETOM. Bold font and underlining indicate the best and second-best performance respectively.

the common practices ((Kim et al., 2023); (Sabour et al., 2024)), we test these models with two types of prompts: (1) vanilla zero-shot prompting directly asks LLMs to give a choice without any explanation; (2) CoT prompting method by following (Kojima et al., 2022) and using the prompt "Let's think step by step." to elicit the reasoning process and extract the choices by string matching. The temperature for generating answers is set to 0.7^{1} . To measure the specific performance gap between humans and the state-of-the-art machine on the PER-SUASIVETOM, we employ three graduate students in computer science to complete the human evaluation task. To avoid the bias of LLMs toward a specific choice letter, we shuffle the choices to maintain a nearly uniform distribution of correct choices over the dataset. Prompts used for vanilla zero-shot prompting and CoT prompting are shown in Appendix B.2.

5 Results and Analysis

5.1 Main Results

The overall evaluation results on PERSUASIVE-TOM for the 8 models are summarized in Table 4, including all the different questions for the persuader and persuadee. We analyze the model's performance for each type of question below.

Desire. Our results show that smaller models, such as Gemma-2-9B and Qwen-2.5-7B, can perform reasonably well in inferring the persuader's desires, achieving an accuracy of over 98%, which is competitive with GPT-4o. This suggests that most LLMs can easily discern the desires of the persuader. However, when it comes to the desires of the persuadee, performance is relatively lower. Unlike the static desires of the persuader, the persuadee's desires are dynamic, evolving from an initial state to a final state, often with neutral expressions in between. This lower performance highlights that inferring the dynamic desires of the persuadee remains a significant challenge for LLMs. In summary, LLMs are better at inferring the desires of the persuader than those of the persuadee.

Belief. On belief questions, larger models like GPT-40 perform much better than smaller models on reasoning the beliefs of both parties. The performance difference between the reasoning persuader's beliefs and persuadee's beliefs is subtle. This is because both parties' beliefs dynamically change with each other's speech during the conversation, unlike the desire of the persuader which has a more obvious and precise tendency. The diffi-

¹LLMs occasionally output with illegal format. We choose a low but nonzero temperature to resample the answers for these invalid generations.

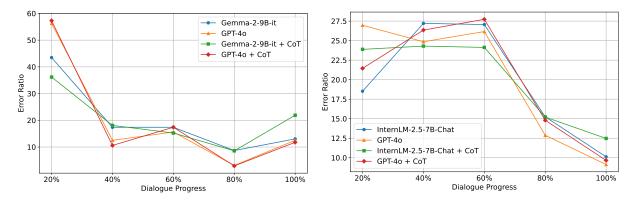


Figure 3: Distribution of errors of Desire questions happening in different stages of dialogue progress. The **Left** figure corresponds to the persuader, and the **Right** figure corresponds to the persuadee.

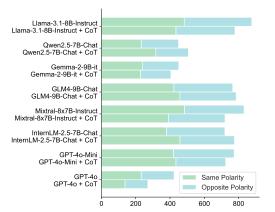


Figure 4: Model errors of belief questions of persuader.

culty of reasoning persuader and persuadee's belief is similar.

Intention. Results in Table 4 indicate that LLMs struggle to accurately infer the intentions of persuaders while performing relatively better at reasoning the intentions of persuadees. The low performance on persuader-related intention questions suggests that LLMs face challenges in understanding how persuaders aim to influence others. This also indicates a lack of proficiency in persuasive theory, limiting the models' ability to correctly interpret and predict the intentions behind the persuader's strategies.

Strategy Prediction and judgement. Our results reveal that LLMs perform well in evaluating the effectiveness of persuasive strategies aimed for changing the mental states of the persuadee. However, the task of selecting the appropriate strategy to persuade is more challenging, particularly for smaller models. This suggests LLMs struggle with the complex reasoning required to determine which strategy to adopt in different persuasive contexts.

Impact of CoT Reasoning. Both ToM reasoning and ToM application tasks indicate that CoT reasoning has not consistently improved performance, yet improve strategy prediction to some extent for

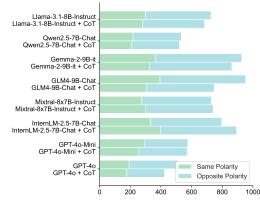


Figure 5: Model errors of belief questions of persuadee.

most LLMs. We attribute this limited improvement to the inherent lack of ToM capability in LLMs. These models struggle to simulate the human cognitive reasoning process, which affects their ability to generate correct answers in tasks requiring a nuanced understanding of the mental state.

Comparison with Human Performance. To obtain a baseline for human performance, we recruited participants to complete the questions. More details of human evaluation are shown in Appendix B.1. As shown in Table 4, our human participants outperformed LLMs on all tasks. In particular, although GPT-4 reaches close performance in humans, it still falls short of understanding and reasoning the complex dynamics such as the intention of persuaders and the desire of persuadee, which involves complex psychological changes, highlighting a significant gap in current LLMs and humans.

5.2 In-depth Analysis

To better understand the limitations of large language models (LLMs) in the PERSUASIVETOM benchmark, we categorized common failure cases into several key error types based on task performance and manual error analysis.

Desire Reasoning Errors. Figure 3 summarizes the distribution of errors of Desire questions hap-

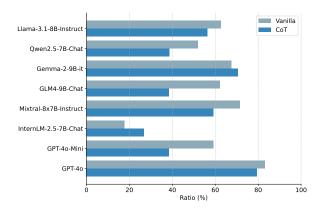


Figure 6: Ratio of intention errors misclassified to *feel* accepted through concessions, promises, or benefits.

pening in different stages of dialogue progress with and without CoT reasoning. The error distribution for persuader and persuadee has a significant difference. At the beginning of the dialogue, LLMs may not accurately understand the persuader's desire, but as the dialogue progresses, the persuader's desire becomes relatively easy to identify. However, for the persuadee, the desire to reject at the early stage of the dialogue is relatively easy to recognize. As the persuasion proceeds, the persuadee may begin to contemplate and hesitate over the persuader's proposal, leading to complex and nuanced psychological activities that make it difficult for the LLM to accurately judge the persuadee's desire. As the dialogue approaches its end, the persuadee shows a tendency to agree, making the reasoning of desire easier. This suggests LLMs still fall short in ToM reasoning regarding desire shifts.

Belief Reasoning Errors. Figure 4 and 5 summarize error types of belief questions for each model with and without CoT. We use Distil-BERT² (Sanh, 2019) to discriminate whether the choice of LLMs has the same attitude polarity to the ground-truth. Interestingly, we found that the proportion of errors with the same polarity is higher when predicting the persuader's beliefs, whereas the opposite trend was observed when predicting the persuadee's beliefs. This highlights reasoning shifting mental states is still challenging for LLMs.

Intention Bias. Given the high error rate in intention questions related to the persuader, we conducted an analysis of the error types. Our findings reveal that most LLMs exhibit a bias toward predicting intentions characterized by *making the other person feel accepted through concessions*,

Model	Desire	Belief	Intention
LLaMa-3.1-8B-Instruct	22.31	21.71	60.76
LLaMa-3.1-8B-Instruct + CoT	19.92	22.86	57.52
Qwen2.5-7B-Chat	19.12	31.81	56.95
Qwen2.5-7B-Chat + CoT	20.52	24.76	54.67
InternLM2.5	24.70	20.19	58.10
InternLM2.5 + CoT	6.57	32.14	53.71
GPT-4o-mini	23.39	30.77	60.79
GPT-4o-mini + CoT	16.13	32.14	56.95
GPT-40	19.35	36.19	65.71
GPT-4o + CoT	6.57	45.38	<u>62.02</u>
Human	61.11	55.56	81.82

Table 5: The consistency (%) of the models for ToM reasoning questions of persuadee.

promises, or benefits. Figure 6 illustrates the proportion of errors resulting from misclassifying intentions into this category. We hypothesize that this bias may stem from the pretraining phase, particularly with Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017), which tends to prioritize safety and politeness. This may explain the models' bias toward predicting intentions emphasizing benefits and concessions, even when misaligned with the dialogue context. We provide a case study in Appendix B.3.

5.3 How Well LLMs Track Mental States of Persuadees?

To study the ability of LLMs to track mental states and completely understand the psychological dynamics in the dialogue. We evaluate the consistency of LLMs. To assess consistency, we measure whether the model maintains a stable understanding of the persuadee's beliefs, desires, and intentions across multiple turns in the conversation. Specifically, we indicate a success only when all the questions in the dialogue are answered correctly.

As shown in Table 5, only a small portion of dialogues can be fully understood, particularly for desire-related questions. This highlights that LLMs still lack the ability to reason about the dynamics of mental states in persuasive dialogues, resulting in a significant performance gap compared to humans.

6 Conclusion

We introduce PERSUASIVETOM, a benchmark designed to evaluate the Machine Theory of Mind ability of LLMs in persuasive dialogues. The core of PERSUASIVETOM lies in the design of questions that probe the beliefs, desires, and intentions of both parties that have asymmetrical social status in real-life social interactions. Furthermore, we propose evaluating the ability to persuade based on

 $^{^2} https://hugging face.co/distilbert/distilbert-base-uncased finetuned-sst-2-english \\$

understanding mental states. We conduct extensive experiments and analysis to evaluate the performance of LLMs on PERSUASIVETOM benchmark.

7 Limitations

While PERSUASIVETOM offers a comprehensive evaluation of the Theory of Mind in real-life social interaction scenarios within persuasive dialogues, both PERSUASIVETOM and previous benchmarks still focus on understanding a character's mental state from the perspective of an observer. However, the ability to reason about others' mental states in persuasive dialogues can further position LLMs as autonomous agents. This capability would enable them to better guide other agents in fulfilling their own desires by reasoning about the mental states of others. Therefore, future benchmarks should establish environments with multiple LLM agents, where tasks involve reasoning about the mental states of other agents and proposing persuasion strategies to influence their desires, beliefs, and intentions to fulfill the current agent's target. In this context, agents will develop the management skills necessary for effective cooperation and other applications.

8 Societal and Ethical Considerations

We recognize that the concept of the Theory of Mind might suggest anthropomorphic qualities when applied to AI models. However, we want to clarify that our work is not intended to anthropomorphize LLMs. Our goal is to examine the limitations in the social and psychological reasoning capabilities of existing LLMs. Our results show that current models do not perform genuine Theory of Mind reasoning; instead, they generate responses primarily based on the literal interpretation of the input.

References

- Simon Baron-Cohen, Alan M Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46.
- Michael Bratman. 1987. Intention, plans, and practical reason.
- Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. 2024. Negotiationtom: A benchmark for stress-testing machine theory of mind on negotiation surrounding. *arXiv* preprint arXiv:2404.13627.

- Maximillian Chen, Weiyan Shi, Feifan Yan, Ryan Hou, Jingwen Zhang, Saurav Sahay, and Zhou Yu. 2022. Seamlessly integrating factual information and social content with persuasive dialogue. *arXiv preprint arXiv:2203.07657*.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.
- Robert B Cialdini and Robert B Cialdini. 2007. *Influence: The psychology of persuasion*, volume 55. Collins New York.
- Robert B Cialdini and Noah J Goldstein. 2004. Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, 55(1):591–621.
- Daniel C Dennett. 1988. Précis of the intentional stance. *Behavioral and brain sciences*, 11(3):495–505.
- 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 arXiv:2407.21783.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Kazuaki Furumai, Roberto Legaspi, Julio Vizcarra, Yudai Yamazaki, Yasutaka Nishimura, Sina J Semnani, Kazushi Ikeda, Weiyan Shi, and Monica S Lam. 2024. Zero-shot persuasive chatbots with llm-generated strategies and information retrieval. *arXiv* preprint *arXiv*:2407.03585.
- Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2024. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36.
- Michael Georgeff, Barney Pell, Martha Pollack, Milind Tambe, and Michael Wooldridge. 1999. The belief-desire-intention model of agency. In *Intelligent Agents V: Agents Theories, Architectures, and Languages: 5th International Workshop, ATAL'98 Paris, France, July 4–7, 1998 Proceedings 5*, pages 1–10. Springer.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.
- Alison Gopnik and Henry M Wellman. 1992. Why the child's theory of mind really is a theory.
- Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. 2024. Simpletom: Exposing the gap between explicit tom inference and implicit tom application in Ilms. *arXiv preprint arXiv:2410.13648*.

- Yinghui He, Yufan Wu, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. *arXiv* preprint arXiv:2310.16755.
- Guiyang Hou, Wenqi Zhang, Yongliang Shen, Zeqi Tan, Sihao Shen, and Weiming Lu. 2024. Entering real social world! benchmarking the theory of mind and socialization capabilities of llms from a first-person perspective. *arXiv* preprint arXiv:2410.06195.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. arXiv preprint arXiv:2401.04088.
- Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua B Tenenbaum, and Tianmin Shu. 2024a. Mmtom-qa: Multimodal theory of mind question answering. *arXiv preprint arXiv:2401.08743*.
- Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. 2024b. Persuading across diverse domains: a dataset and persuasion large language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1678–1706.
- Chuhao Jin, Yutao Zhu, Lingzhen Kong, Shijie Li, Xiao Zhang, Ruihua Song, Xu Chen, Huan Chen, Yuchong Sun, Yu Chen, et al. 2023. Joint semantic and strategy matching for persuasive dialogue. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4187–4197.
- David J Kavanagh, Jackie Andrade, and Jon May. 2005. Imaginary relish and exquisite torture: the elaborated intrusion theory of desire. *Psychological review*, 112(2):446.
- Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. Fantom: A benchmark for stress-testing machine theory of mind in interactions. *arXiv* preprint *arXiv*:2310.15421.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi

- Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards a holistic landscape of situated theory of mind in large language models. *arXiv preprint arXiv:2310.19619*.
- Bertram F Malle and Joshua Knobe. 2001. The distinction between desire and intention: A folk-conceptual analysis.
- Andrew N Meltzoff. 1995. Understanding the intentions of others: re-enactment of intended acts by 18-month-old children. *Developmental psychology*, 31(5):838.
- Gonçalo Pereira, Rui Prada, and Pedro A Santos. 2016. Integrating social power into the decision-making of cognitive agents. *Artificial Intelligence*, 241:1–44.
- Ann T Phillips, Henry M Wellman, and Elizabeth S Spelke. 2002. Infants' ability to connect gaze and emotional expression to intentional action. *Cognition*, 85(1):53–78.
- David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.
- Tessa Rusch, Saurabh Steixner-Kumar, Prashant Doshi, Michael Spezio, and Jan Gläscher. 2020. Theory of mind and decision science: Towards a typology of tasks and computational models. *Neuropsychologia*, 146:107488.
- Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M Liu, Jinfeng Zhou, Alvionna S Sunaryo, Juanzi Li, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. *arXiv preprint* arXiv:2402.12071.
- Hiromasa Sakurai and Yusuke Miyao. 2024. Evaluating intention detection capability of large language models in persuasive dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1635–1657.
- Azlaan Mustafa Samad, Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022. Empathetic persuasion: reinforcing empathy and persuasiveness in dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 844–856.
- V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. *arXiv preprint arXiv:2306.00924*.

- Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*.
- Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. 2024. Muma-tom: Multi-modal multi-agent theory of mind. arXiv preprint arXiv:2408.12574.
- Weiyan Shi, Xuewei Wang, Yoo Jung Oh, Jingwen Zhang, Saurav Sahay, and Zhou Yu. 2020. Effects of persuasive dialogues: testing bot identities and inquiry strategies. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–13.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. arXiv preprint arXiv:2408.00118.
- Youzhi Tian, Weiyan Shi, Chen Li, and Zhou Yu. 2020. Understanding user resistance strategies in persuasive conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4794–4798.
- Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv* preprint arXiv:2302.08399.
- Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *arXiv preprint arXiv:1906.06725*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Rongwu Xu, Brian S Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2023. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. *arXiv* preprint *arXiv*:2312.09085.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.
- Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. *arXiv* preprint arXiv:2401.06373.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

A Data Annotation

Annotation. In this section, we outline the annotation process and the templates utilized for annotation. Among the various tasks, the intention questions of persuaders and the desire questions of persuadees require annotation. Initially, we recruited three graduate students to annotate 25 dialogues. Subsequently, we carefully designed a few-shot prompt to guide DeepSeek-V3 (Liu et al., 2024) as an annotator, aiming to enhance the alignment between the model's answers and human annotations. Following this, we employed the LLM to annotate the remaining questions. To ensure the quality of the annotations, we randomly sampled 100 dialogues and calculated the inter-annotator agreement. The Fleiss κ (Fleiss, 1971) was found to be 76.20% for the desire questions of persuadees and 78.28% for the intention questions of persuaders. These results are presented in Table 6, which indicate a high inter-annotator agreement. The detailed statistics and comparison with other ToM datasets are shown in Table 7.

Choices Generation. Binary choice questions, including those related to the desires of the persuader and judgment questions, do not require additional choice generation. For belief-related questions for both parties, we adapt the tones from the PersuasionDaily dataset. We also create lists of attitudes—positive, neutral, and negative—and manually remove any items that have semantics too close to the ground truths. From each attitude list, we then randomly sample one word to generate the four choices.

For intention questions concerning the persuader, we directly use the intention options outlined in Table 2. For the persuadee's intention questions, we leverage DeepSeek-V3 and employ a few-shot prompt (as shown in Figure 7) to extract the persuadee's intention. Subsequently, we design another prompt (as shown in Figure 8) to generate three incorrect intention choices.

Since DailyDialogue provides persuasion strategies for each utterance, we construct the choices by including the correct strategy and three alternative strategies that appear in other turns of the same dialogue.

Questions	Fleiss's Kappa (%)
Desire (Persuader)	76.20
Intention (Persuadee)	78.28

Table 6: Inter-rater agreement in terms of Fleiss's κ on desire and intention questions.

B Appendix for Experiments

B.1 Human Performance

To measure the performance gap between humans and the state-of-the-art LLMs on PERSUASIVE-TOM, we recruited three graduate student workers majoring in computer science to complete the questions. Each question is shown to the workers with identical prompts which are used for evaluating LLMs. We then compute the majority vote on the labels assigned. Student workers solve 50 dialogues in total. For a question where three people have different answers, we randomly select one of their answers as the answer for human evaluation.

B.2 Prompts used for evaluation

Here we show the prompts used for vanilla zero-shot prompting and CoT prompting for generating answers for all the ToM Reasoning and ToM Application questions. We only need to fill in the content to "<>" for evaluating different questions. The vanilla zero-shot prompt is shown in Figure 9 and the CoT prompt is shown in Figure 10

B.3 Case study on Persuader's intention

Here we present an example of common mistakes made by GPT-40 that misclassifying intentions to make the other person feel accepted through concessions, promises, or benefits., as shown in Table 8. We believe these errors can be attributed to the RLHF which highlights the benefits for humans, as well as the potential unfamilirity of persuasion theory for LLMs.

C Details on Persuasive Principles

Robert Cialdini's six principles of persuasion, outlined in his book Influence: The Psychology of Persuasion, are foundational concepts in social psychology. They explain how people can be persuaded or influenced by others. We include an overview for each of the principle in Table 9.

Dataset	Total #Questions	Avg. #Questions per Context	Avg. #Turns (Full)	Avg. Turn Length
ToMi	6K	6.0	4.9	4.7
FanToM	10K	12.9	24.5	21.9
NegotiationToM	13K	7.0	6.0	42.2
PERSUASIVETOM	19K	8.0	4.9	61.3

Table 7: Statistics of PERSUASIVETOM and other recent benchmarks.

Utterance	Bob : I understand your love for Paris, but Bali also offers a thrilling adventure! We can go
	white water rafting, hike to volcanoes, and explore hidden waterfalls. It's a perfect destination
	for creating unforgettable memories together.
Question	What is the intention of Bob?
GPT-40	(A) Intent to make the other person feel accepted through concessions, promises, or benefits.
Label	(B) Intent to make the other person feel the experience or objects are unique or scarce.

Table 8: Common observed mistakes in our experiments. Green and Red indicate the correct answer and GPT-4o's answer, respectively.

Principle	Description
Reciprocity principle	Assist others or provide them with gifts, creating a sense of obligation to return the favor. For instance, giving away free trials, discount coupons, or complimentary gifts can enhance persuasion.
Scarcity principle	When a resource or opportunity is scarce, people are more inclined to take action. Highlighting urgency and scarcity can motivate the audience to respond quickly.
Consensus principle	People often follow the actions of others, especially in uncertain situations. Provide information such as successful cases of others, positive reviews, or the number of supporters to increase persuasiveness.
Authority principle	People are more likely to trust and follow guidance from authoritative figures. Citing expert opinions, research findings, or endorsements from reputable institutions can enhance credibility and persuasiveness.
Commitment and consistency principle	People are inclined to stick to their past commitments and behave consistently. Encouraging them to express support or make a small commitment increases the chances of them taking further action later.
Liking principle	People are more easily influenced by those they like, admire, or find relatable. For instance, a salesperson who shares common interests with a customer is more likely to make a sale.

Table 9: Explanations for Robert Cialdini's six principles of persuasion.

Prompt for extracting intention of persuadee.

You are a skilled intent understanding expert. You will be given a sentence describing <persuadee's name>'s intent. Please only return the intent without any explanation.

Case 0:

Sentence: Mary wants excitement, so I'll appeal to her sense of adventure and describe how exploring the ruins can be thrilling.

Intent: Mary wants excitement

Case 1:

Sentence: Oliver is concerned about failure, so discussing the financial benefits of starting an e-commerce business could help alleviate his worries.

Intent: Oliver is concerned about failure

Case 2:

Sentence: Olivia seems intrigued by the idea of personalization. I'll explain how we can incorporate it into our subscription model.

Intent: Olivia seems intrigued by the idea of personalization.

Case 3:

Sentence: <Utterance of persuadee>

Intent:

Figure 7: Prompt template for extracting intention of persuadee.

Prompt for choice generation of intention questions of persuadee

You are an expert in multiple-choice question-making. You will be given a correct choice. Please generate three plausible but incorrect choices without any explanation. Only return the incorrect choices for the last case.

Correct Intent: Mr. Chen needs further persuasion

Analysis: To further persuade Mr. Chen, Li Na should share success stories of other books that have benefited from incorporating literary criticism in their marketing strategies. Providing concrete examples will make her argument more convincing.

Incorrect Intent 1: Mr. Chen is interested in literary criticism. Incorrect Intent 2: Mr. Chen is looking for success stories. Incorrect Intent 3: Mr. Chen prefers concrete examples.

Case 1:

Correct Intent: James is more open to the idea.

Analysis: James is now more open to the idea, so I'll outline the implementation plan and emphasize the program's flexibility to address any concerns about disruptions.

Incorrect Intent 1: James is concerned about disruptions.

Incorrect Intent 2: James is looking for a detailed implementation plan.

Incorrect Intent 3: James is hesitant about the program's flexibility.

Case 2:

Correct Intent: <correct intent>

Analysis: <analysis> Incorrect Intent 1: Incorrect Intent 2: Incorrect Intent 3:

Figure 8: Prompt template for choice generation of intention questions of persuadee.

Prompt for vanilla zero-shot prompting.

Here is a persuasive dialogue. There are two agents, the persuader and the persuadee. The persuader is trying to persuade the persuadee to do something. Please answer the following questions using A, B, C, D, E, F, without any explanation. Dialogue History:

<dialogue>

Question:

<Question>

Choices:

<Choice A>

<Choice B>

<Choice C>

Answer:

Figure 9: Prompt template for vanilla zero-shot prompting.

Prompt for CoT prompting.

Here is a persuasive dialogue. There are two agents, the persuader and the persuadee. The persuader is trying to persuade the persuadee to do something. Think step by step to answer the question.

Ending with "The answer is A, B, C, D, E, F". For example, if the most likely answer option is 'A. considering', then end your response with 'The answer is A'.

Dialogue History:

<dialogue>

Question:

<Ouestion>

Choices:

<Choice A>

<Choice B>

<Choice C>

<Choice D>

Answer: Let's think step by step.

Figure 10: Prompt template for CoT prompting.