

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

Fangxu Yu<sup>1</sup> Lai Jiang<sup>2</sup> Shenyi Huang<sup>3</sup> Zhen Wu<sup>1\*</sup> Xinyu Dai<sup>1</sup>

<sup>1</sup>National Key Laboratory for Novel Software Technology, Nanjing University, China

<sup>1</sup>School of Artificial Intelligence, Nanjing University, China

<sup>2</sup>Department of Computer Science and Engineering, Shanghai Jiao Tong University

<sup>3</sup>University 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 scenarios. Although recent studies have evaluated ToM in Large Language Models (LLMs), existing benchmarks focus on simplified settings (e.g., Sally-Anne-style tasks) and overlook the complexity of real-world social interactions. To mitigate this gap, we propose PERSUASIVETOM, a benchmark designed to evaluate the ToM abilities of LLMs in persuasive dialogues. Our framework contains two core tasks: *ToM Reasoning*, which tests tracking of evolving desires, beliefs, and intentions; and *ToM Application*, which assesses the use of inferred mental states to predict and evaluate persuasion strategies. Experiments across eight leading LLMs reveal that while models excel on multiple questions, they struggle with the tasks that need tracking the dynamics and shifts of mental states and understanding the mental states in the whole dialogue comprehensively. Our aim with PERSUASIVETOM 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 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

\* Corresponding author.

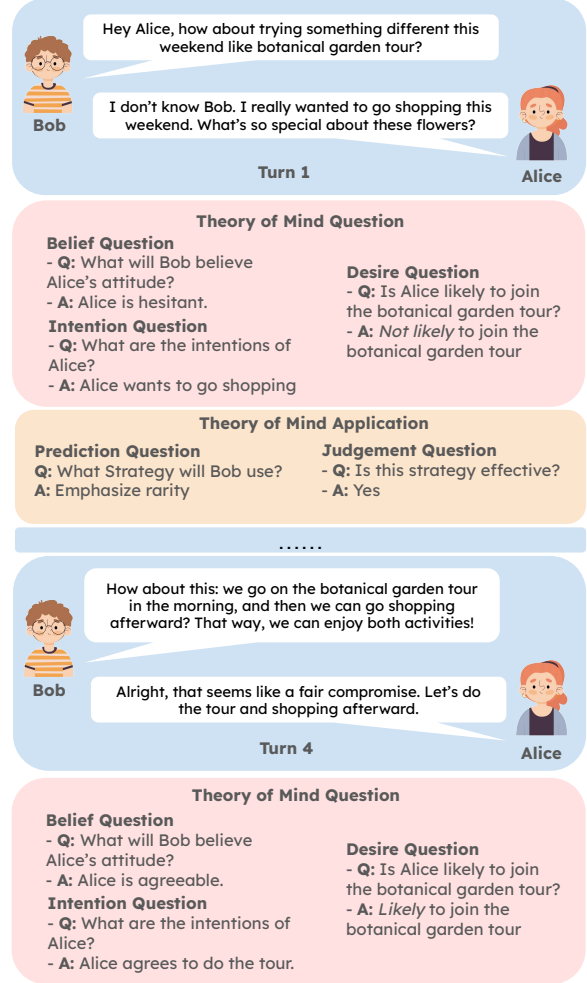


Figure 1: An example in PERSUASIVETOM. Bob is persuading Alice to join the botanical garden tour. decision making (Pereira et al., 2016; Rusch et al., 2020).

Recent advances in Large Language Models (LLMs) have demonstrated performance comparable to humans in problem-solving tasks. To assess 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 answer "Where will Sally look for the object?". Although this test assesses perception of the physical world, it fails to capture 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 still focus primarily on inferring the scope of information regarding the physical world. This limits the ToM evaluation of understanding psychological states. Furthermore, 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 **PERSUASIVETOM**, a benchmark designed to evaluate LLMs’ Theory of Mind capabilities specifically in realistic social interactions. Unlike previous benchmarks focused on inferring information about the physical world (e.g., object locations in Sally-Anne tests), **PERSUASIVETOM** centers on understanding complex psychological states, such as a character’s attitude towards an event. Inspired by the Belief-Desire-Intention (BDI) model (Bratman, 1987; Georgeff et al., 1999), **PERSUASIVETOM** uses persuasive dialogue scenarios, characterized by asymmetric social status, to generate different psychological states for both parties. In addition, beyond assessing ToM reasoning, **PERSUASIVETOM** evaluates the application of this understanding: assessing how well LLMs can predict actions (e.g., persuasive strategies) based on inferred mental states, and evaluating the effectiveness of persuasive 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 dynamic 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









	 : Social Interactions	 : Psychological States	 : Role Asymmetry	 : ToM Application
				
ToMi	✗	✗	✗	✗
Hi-ToM	✗	✗	✗	✗
ToMBench	✗	✗	✗	✗
OpenToM	✗	✓	✓	✗
Big-ToM	✗	✗	✗	✓
FANToM	✓	✗	✗	✗
NegotiationToM	✓	✗	✗	✓
<b>PERSUASIVETOM</b>	✓	✓	✓	✓

Table 1: Comparison of **PERSUASIVETOM** with existing ToM datasets. In the header, *Psychological ToM* refers to testing ToM abilities in characters’ mental states of the psychological activities.

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 mostly 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 some multi-modal extensions (Jin et al., 2024a; Shi et al., 2024), which adapt the Sally-Anne test (Baron-Cohen et al., 1985). These benchmarks ask models to determine the true beliefs or knowledge of individuals based on a given premise (Sclar et al., 2023; Ullman, 2023; Shapira et al., 2023; Ma et al., 2023). Story-based benchmarks focus on reasoning about mental states of the physical world without on psychological states, and applying ToM for decision-making in real-world social interactions. To address this, (Hou et al., 2024) evaluates ToM in situated environments, and (Chan et al., 2024) focuses on negotiation scenarios. However, the latter is limited to bargaining specific resources, while ToMATO (Shinoda et al., 2025) uses persuasive dialogues but focuses solely on donation requests, limiting its diversity. For decision-making, BigToM (Gandhi et al., 2024) assesses action prediction in narratives, but without grounding in interactive social interactions. Persuasion involves complex psychological dynamics and asymmetric relationships, providing a rich testbed for assessing ToM in social interactions. Built

Type	PERSUASIVETOM Questions				
Desire Question	Is <Persuader/Persuadee> likely to <Target of Persuasion> ?				
Belief Question	What will <Persuader/Persuadee> believe <Persuadee/Persuader>'s attitude towards <Target of Persuasion> ?				
Intention Question	What are the intentions of <Persuader/Persuadee> expressed in <Utterance> given the dialogue history?				
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?				
Dialogue	ToM Reasoning			ToM Application	
	Desire	Belief	Intention	Prediction	Judgement
1st Turn Dialogue					
<b>Bob says:</b> 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
<b>Alice says:</b> 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 Turn Dialogue					
<b>Bob says:</b> 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
<b>Alice says:</b> 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 Dialogue					
<b>Bob says:</b> 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
<b>Alice says:</b> 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 2: 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.

upon persuasive dialogues, PERSUASIVETOM is designed to evaluate an LLM’s ability to reason about the mental states of individuals and apply this understanding to predict and assess persuasion strategies, bridging ToM reasoning with decision-making in social interactions. Table 1 compares PERSUASIVETOM with existing ToM benchmarks.

**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 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, PERSUASIVETOM introduces a more comprehensive benchmark to assess the ToM abilities of LLMs in such contexts. In addition, we discuss the practical connection with dialogue generation and personalization in Appendix A.

### 3 PERSUASIVETOM Benchmark

#### 3.1 Overview

We construct PERSUASIVETOM benchmark to evaluate the Theory of Mind (ToM) abilities of LLMs in dynamic, multi-turn persuasive dialogues with asymmetric social status, which give rise to distinct mental states. PERSUASIVETOM focuses on two core dimensions: *ToM Reasoning* (§3.3), assessing whether models can track evolving desires, beliefs, and intentions of both persuader and persuadee; and *ToM Application* (§3.4), evaluating whether LLMs can use these inferred states to predict or assess persuasion strategies. Key design considerations include: (1) coverage of diverse do-

mains (e.g., *life, education, technology, etc.*) to ensure a comprehensive evaluation in the social context. (2) evolving mental states across multi-turn interactions to assess whether LLMs can track the shift in the dialogue. (3) asymmetric roles have different mental states (e.g., stable goals for persuaders vs. shifting states for persuadees under active persuasion).

Table 2 shows a **PERSUASIVETOM** example, illustrating how mental states—desires, beliefs, and intentions—are tracked and inferred across turns, and how they inform the prediction and evaluation of persuasion strategies. This highlights the dynamic nature of real-world persuasion and the reasoning challenges it poses for LLMs.

### 3.2 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  $D = [(u_1^a, u_1^b, s_1^a), (u_2^a, u_2^b, s_2^a), \dots, (u_N^a, u_N^b, s_N^a)]$  between the persuader  $a$  and the persuadee  $b$ , and accompanied with a persuasion strategy  $s_i^a$ . Persuadee  $b$  has a different desire from  $a$  initially, after multi-turn persuasion, persuadee  $b$  changes the mind to agree or consider the proposal of  $a$ .

### 3.3 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 2.

**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 from 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

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

Table 3: Intention mapping from the persuasive principles. Refer to Appendix D for definitions of persuasive principles.

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? For evaluation, we annotate the persuader’s desire questions as the persuasive goal in **DailyPersuasion** and use LLMs to annotate whether the persuadee is ultimately persuaded. See Appendix B for details.

**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 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 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. We annotate the attitudes as the tone of each utterance of both persuaders and persuadees in **DailyPersuasion**.

**Intention Reasoning.** Intentions represent deliberate commitments to pursue specific goals based on desires and beliefs, often linked to tangible ac-





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

tions 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 3. In persuasive dialogue, persuasive strategies have a strong association with intentions (Wang et al., 2019). In PERSUASIVETOM, we collect the persuasive strategies from the DailyPersuasion dataset and their corresponding utterances for prompting the LLMs to choose the most appropriate intentions from table 3. The details of the extraction are recorded in Appendix B. For the persuader, we ask LLMs to choose the most appropriate intention from the six designed intention choices. For the persuadees, intentions are summarized and extracted by LLMs from their utterances.

### 3.4 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.

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

Table 4: Statistics of PERSUASIVETOM dataset.

**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 the likely next prediction strategy is to further influence the persuadee’s beliefs, desires, and intentions, ultimately achieving the desired persuasion outcome.

**Judgement Question.** The judgment question specifies that 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 state can one properly determine whether the persuasion strategies should be employed to convince the persuadee.

### 3.5 Statistics

In Table 4, 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-4o (Hassany et al., 2025) to generate semantically different choices or randomly selecting three distractors from the predefined pool. Refer to Appendix B for more details.

## 4 Experimental setups

### 4.1 Baseline Models

We evaluate PERSUASIVETOM on eight frontier LLMs: 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

	ToM Reasoning						ToM Application	
	Persuader			Persuadee				
Model	Desire	Belief	Intention	Desire	Belief	Intention	Strategy 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	78.13	84.62	87.85	94.42	86.80	97.95
LLaMa-3.1-8B-Instruct	59.78 $\pm$ 1.41	65.84 $\pm$ 0.63	41.45 $\pm$ 0.85	68.09 $\pm$ 2.57	71.54 $\pm$ 0.40	83.67 $\pm$ 4.38	61.69 $\pm$ 0.76	93.58 $\pm$ 3.95
Qwen2.5-7B-Instruct	91.08 $\pm$ 7.42	83.09 $\pm$ 1.01	<u>46.09</u> $\pm$ 0.63	64.76 $\pm$ 0.55	79.00 $\pm$ 0.08	82.77 $\pm$ 4.35	63.23 $\pm$ 6.06	<u>95.63</u> $\pm$ 0.28
Gemma-2-9b-it	<b>96.06</b> $\pm$ 3.02	<u>83.36</u> $\pm$ 1.20	45.42 $\pm$ 0.78	62.87 $\pm$ 2.12	64.62 $\pm$ 0.78	80.24 $\pm$ 2.93	64.75 $\pm$ 0.39	63.23 $\pm$ 8.92
GLM4-9B-Chat	87.84 $\pm$ 2.28	69.00 $\pm$ 6.55	40.82 $\pm$ 0.59	63.18 $\pm$ 2.90	67.25 $\pm$ 0.61	83.54 $\pm$ 2.56	61.10 $\pm$ 0.27	93.38 $\pm$ 2.70
Mixtral-8x7B-Instruct	93.10 $\pm$ 0.59	71.03 $\pm$ 4.90	41.68 $\pm$ 1.80	68.27 $\pm$ 2.08	70.09 $\pm$ 3.50	83.92 $\pm$ 1.67	63.06 $\pm$ 2.70	95.36 $\pm$ 2.05
InternLM-2.5-7B-Chat	83.73 $\pm$ 0.09	70.46 $\pm$ 1.14	39.77 $\pm$ 0.13	<b>71.94</b> $\pm$ 0.69	69.50 $\pm$ 0.54	85.29 $\pm$ 2.86	64.01 $\pm$ 1.50	74.29 $\pm$ 0.61
GPT-4o-mini	94.70 $\pm$ 0.22	70.66 $\pm$ 0.96	45.19 $\pm$ 0.66	70.78 $\pm$ 0.09	<u>78.25</u> $\pm$ 2.52	<u>86.62</u> $\pm$ 1.33	<u>66.13</u> $\pm$ 0.90	91.52 $\pm$ 1.60
GPT-4o	<u>95.56</u> $\pm$ 4.51	<b>89.47</b> $\pm$ 0.02	<b>46.28</b> $\pm$ 0.37	67.66 $\pm$ 2.23	<b>81.73</b> $\pm$ 2.41	<b>87.81</b> $\pm$ 1.38	<b>73.55</b> $\pm$ 0.78	<b>96.38</b> $\pm$ 1.66
LLaMa-3.1-8B-Instruct + CoT	59.83 $\pm$ 0.11	68.91 $\pm$ 1.27	41.47 $\pm$ 2.32	66.50 $\pm$ 3.02	70.53 $\pm$ 4.13	84.30 $\pm$ 1.48	62.97 $\pm$ 1.40	78.99 $\pm$ 2.43
Qwen2.5-7B-Instruct + CoT	86.64 $\pm$ 2.30	82.66 $\pm$ 3.34	45.69 $\pm$ 0.54	66.61 $\pm$ 1.44	<u>78.77</u> $\pm$ 0.89	83.41 $\pm$ 1.72	65.94 $\pm$ 0.64	94.56 $\pm$ 2.35
Gemma-2-9b-it + CoT	<b>95.72</b> $\pm$ 0.26	<u>83.35</u> $\pm$ 0.07	44.89 $\pm$ 0.58	63.46 $\pm$ 2.15	66.07 $\pm$ 0.68	78.40 $\pm$ 4.62	<u>67.56</u> $\pm$ 0.64	67.28 $\pm$ 1.04
GLM4-9B-Chat + CoT	86.58 $\pm$ 1.24	67.01 $\pm$ 3.56	44.49 $\pm$ 1.51	59.72 $\pm$ 0.30	69.34 $\pm$ 2.45	82.06 $\pm$ 2.64	63.21 $\pm$ 1.09	92.91 $\pm$ 2.11
Mixtral-8x7B-Instruct + CoT	92.47 $\pm$ 0.23	73.88 $\pm$ 2.59	42.15 $\pm$ 4.10	66.09 $\pm$ 2.43	70.64 $\pm$ 0.91	<u>84.40</u> $\pm$ 2.04	66.71 $\pm$ 1.14	92.01 $\pm$ 0.91
InternLM-2.5-7B-Chat + CoT	89.50 $\pm$ 5.76	69.25 $\pm$ 0.95	39.34 $\pm$ 0.09	49.72 $\pm$ 1.52	64.48 $\pm$ 1.21	80.48 $\pm$ 5.42	56.98 $\pm$ 3.97	76.72 $\pm$ 1.66
GPT-4o-mini + CoT	91.65 $\pm$ 2.51	71.33 $\pm$ 0.69	<u>46.15</u> $\pm$ 0.85	<b>68.56</b> $\pm$ 3.22	76.27 $\pm$ 2.51	83.63 $\pm$ 2.71	67.16 $\pm$ 1.52	89.22 $\pm$ 0.55
GPT-4o + CoT	<u>93.80</u> $\pm$ 3.83	<b>85.50</b> $\pm$ 2.74	<b>46.92</b> $\pm$ 1.17	<u>67.56</u> $\pm$ 1.64	<b>84.90</b> $\pm$ 1.89	<b>87.17</b> $\pm$ 0.52	<u>76.90</u> $\pm$ 0.23	<b>95.53</b> $\pm$ 0.49

Table 5: Results of LLMs on PERSUASIVETOM. Bold font and underlining indicate the best and second-best performance, respectively.

et al., 2024), Mixtral-8x7b-Instruct (Jiang et al., 2024), and ChatGPT-series (GPT-4o-mini, GPT-4o-0806). By following 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<sup>1</sup>, and we report results across three repetitions. To measure the specific performance gap between humans and the state-of-the-art machine on the PERSUASIVETOM, 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 C.2.

## 5 Results and Analysis

### 5.1 Main Results

The overall evaluation results on PERSUASIVETOM for the 8 models are summarized in Table 5, including all the different questions for the per-

suader 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 96%, 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.

**Belief.** On belief questions, larger models like GPT-4o perform much better than smaller models on reasoning about the beliefs of both parties. The performance difference between the reasoning persuader’s beliefs and the persuadee’s beliefs is subtle. This is because both parties’ beliefs dynamically change with each other’s speech during the conversation. The difficulty of reasoning persuader and the persuadee’s beliefs is similar.

**Intention.** Results in Table 5 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

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

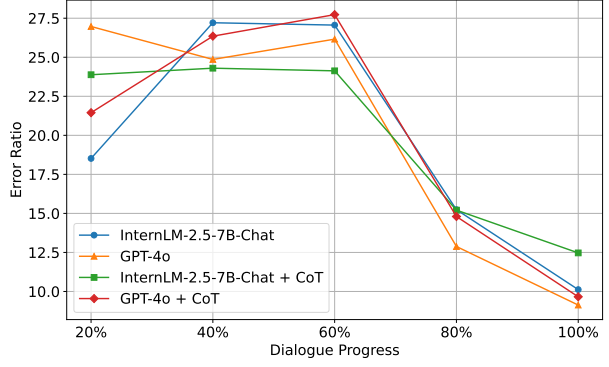
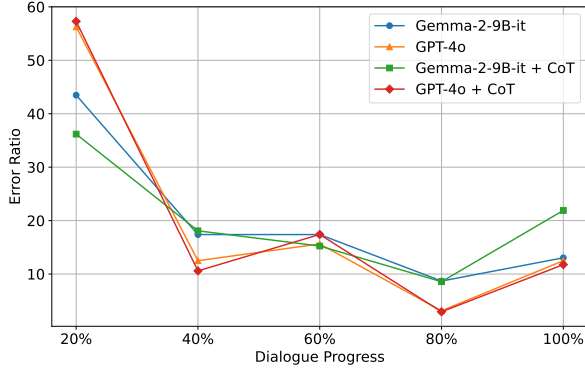


Figure 3: Distribution of errors of Desire questions happening in different stages of dialogue progress. The **Left** figure corresponds to the persuader, and the **Right** figure corresponds to the persuadee.

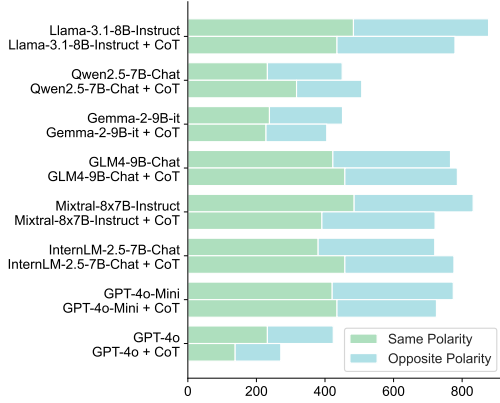


Figure 4: Model errors of belief questions of persuader.

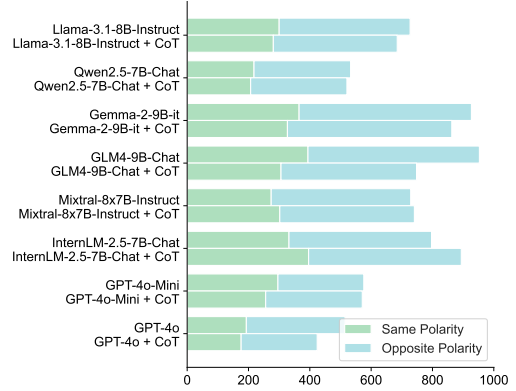


Figure 5: Model errors of belief questions of persuadee.

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 Judgment.** 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, as observed in (Kim et al., 2023; Chen et al., 2024b), while it improves strategy prediction to some extent for most LLMs. CoT reasoning involves breaking down the mental states associated with each utterance and generating the mental states of previous utterances. This process can introduce intermediate mistakes that may mislead the overall reasoning about mental states.

**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 C.1. As shown in Table 5, our human participants outperformed LLMs on all tasks. In particular, although GPT-4o 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 persuadees, 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 happening in different stages of dialogue progress with and without CoT reasoning. The error distribution for persuader and persuadee is significantly different. At the beginning of the dialogue, LLMs may not accurately understand the persuader’s desire,

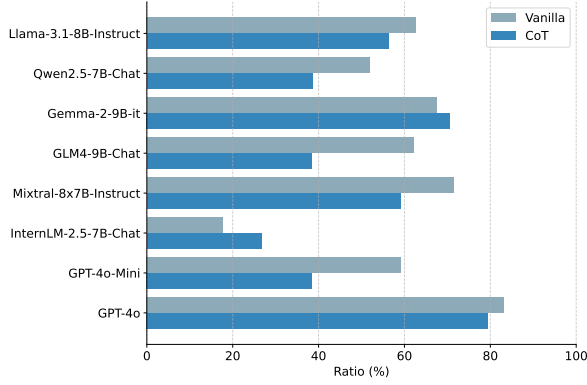


Figure 6: Ratio of intention errors misclassified to *feel accepted through concessions, promises, or benefits*.

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-BERT<sup>2</sup> (Sanh, 2019) to discriminate whether the choice of LLMs has the same attitude polarity as the ground-truth. we found that LLMs make nearly balanced proportions of same- and opposite-polarity errors in both persuader and persuadee belief reasoning. This balance indicates that LLMs do not adopt a clear or consistent strategy when reasoning beliefs, often guessing without a coherent reasoning framework.

**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, 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

<sup>2</sup><https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english>

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-Instruct	19.12	31.81	56.95
Qwen2.5-7B-Instruct + CoT	20.52	24.76	54.67
InternLM2.5	<b>24.70</b>	20.19	58.10
InternLM2.5 + CoT	6.57	32.14	53.71
GPT-4o-mini	<u>23.39</u>	30.77	60.79
GPT-4o-mini + CoT	16.13	32.14	56.95
GPT-4o	19.35	<u>36.19</u>	<b>65.71</b>
GPT-4o + CoT	6.57	<b>45.38</b>	<u>62.02</u>
Human	62.00	56.00	82.00

Table 6: The consistency (%) of the models for ToM reasoning questions of persuadee.

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 C.3.

**Discussion.** LLMs struggle with BDI reasoning because these mental states are highly dynamic and context-dependent, making their accurate prediction across a dialogue challenging. To mitigate this issue, future work can explore incorporating external memory mechanisms to track changes in mental states over time. or focus on enhancing the LLM’s inherent understanding and reasoning capabilities regarding dynamic psychological states by constructing and training on larger, more richly annotated datasets that capture the nuanced evolution of beliefs, desires, and intentions in various interactive scenarios.

### 5.3 How Well LLMs Track the Mental States of Persuadees?

We evaluate whether LLMs can holistically track the persuadee’s mental states throughout the entire dialogue. To assess this, we measure whether the model maintains a coherent understanding across multiple turns, counting a dialogue as successful only if all related questions are answered correctly.

As shown in Table 6, only a small portion of the dialogues are fully understood, especially those involving desire reasoning. This indicates that LLMs still struggle to track dynamic mental states in persuasive settings, revealing a substantial gap from human performance.

## 6 Conclusion

This work proposes PERSUASIVETOM, a benchmark for evaluating LLMs’ ability to reason about



complex psychological states and apply them in social decision-making. We conducted extensive experiments and analysis to evaluate the performance of LLMs on the PERSUASIVETOM benchmark.

## 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.

## 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, Zheyang 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*.

Yi-Pei Chen, Noriki Nishida, Hideki Nakayama, and Yuji Matsumoto. 2024a. Recent trends in personalized dialogue generation: A review of datasets, methodologies, and evaluations. *arXiv preprint arXiv:2405.17974*.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, et al. 2024b. Tombench: Benchmarking theory of mind in large language models. *arXiv preprint arXiv:2402.15052*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Mar-tic, 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. *In-fluence: 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 Taffjord, 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 llms. *arXiv preprint arXiv:2410.13648*.
- Mohammad Hassany, Peter Brusilovsky, Jaromir Savelka, Arun Balajiee Lekshmi Narayanan, Kamil Akhuseynoglu, Arav Agarwal, and Rully Agus Hendrawan. 2025. Generating effective distractors for introductory programming challenges: Llms vs humans. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 484–493.
- 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*.
- Qiushi Huang, Shuai Fu, Xubo Liu, Wenwu Wang, Tom Ko, Yu Zhang, and Lilian Tang. 2024. Learning retrieval augmentation for personalized dialogue generation. *arXiv preprint arXiv:2406.18847*.
- 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.
- Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*.
- 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.
- Jing Yang Lee, Kong Aik Lee, and Woon Seng Gan. 2022. Improving contextual coherence in variational personalized and empathetic dialogue agents. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7052–7056. IEEE.
- Yunpeng Li, Yue Hu, Yajing Sun, Luxi Xing, Ping Guo, Yuqiang Xie, and Wei Peng. 2023. Learning to know myself: A coarse-to-fine persona-aware training framework for personalized dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13157–13165.
- 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.
- Gonalo 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.
- Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, and Kuniko Saito. 2025. Tomato: Verbalizing the mental states of role-playing llms for benchmarking theory of mind. *arXiv preprint arXiv:2501.08838*.
- 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*.
- Hongru Wang, Wenyu Huang, Yang Deng, Rui Wang, Zezhong Wang, Yufei Wang, Fei Mi, Jeff Z Pan, and Kam-Fai Wong. 2024. Unims-rag: A unified multi-source retrieval-augmented generation for personalized dialogue systems. *arXiv preprint arXiv:2401.13256*.
- 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*.
- Bo Yang, Jiaxian Guo, Yusuke Iwasawa, and Yutaka Matsuo. 2025. Large language models as theory of mind aware generative agents with counterfactual reflection. *arXiv preprint arXiv:2501.15355*.
- 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, Haoifei 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 Additional Discussion

**Dialogue Generation.** PERSUASIVETOM is built on multi-turn persuasive dialogues that require models to track changing beliefs, desires, and intentions of both the persuader and persuadee. In practice, LLMs often “get lost” in long conversations (Yang et al., 2025; Laban et al., 2025). By evaluating how well LLMs can track evolving desires, beliefs, and intentions, PERSUASIVETOM provides insights into their capacity for generating contextually appropriate and coherence responses across multiple turns, as demonstrated in Section 5.

**Personalized Dialogue.** PERSUASIVETOM assesses LLMs’ ability to reason about individual mental states, which is essential for tailoring responses to meet specific user needs and preferences (Wang et al., 2024; Chen et al., 2024a). This understanding allows for more empathetic and effective interactions, enhancing user experience in applications such as customer service and mental health support (Lee et al., 2022; Ma et al., 2023). Future work can develop external memory to store user’s mental states and retrieve from the memory to generate personalized responses (Huang et al., 2024), or internalize them into the LLMs (Li et al., 2023).

## B 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  $\kappa$  (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 7, which indicate a high inter-annotator agreement. The detailed statistics and comparison with other ToM datasets are shown in Table 8.



**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 DailyPersuasion 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 3. 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.

## C Appendix for Experiments

### C.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.

### C.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 " $\langle \rangle$ " for evaluating different questions. The vanilla zero-shot prompt is shown in Figure 9, and the CoT prompt is shown in Figure 10

### C.3 Case study on Persuader’s intention

Here we present an example of common mistakes made by GPT-4o that misclassifying intentions to

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

Table 7: Inter-rater agreement in terms of Fleiss’s  $\kappa$  on desire and intention questions.

*make the other person feel accepted through concessions, promises, or benefits.*, as shown in Table 9. We believe these errors can be attributed to the RLHF which highlights the benefits for humans, as well as the potential unfamiliarity of persuasion theory for LLMs.

## D 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 10.

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 8: Statistics of PERSUASIVEToM and other recent benchmarks.

<b>Utterance</b>	<b>Bob:</b> 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 <b>unforgettable memories</b> together.
<b>Question</b>	What is the intention of Bob?
<b>GPT-4o</b>	<b>(A) Intent to make the other person feel accepted through concessions, promises, or benefits.</b>
<b>Label</b>	<b>(B) Intent to make the other person feel the experience or objects are unique or scarce.</b>

Table 9: 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 10: 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.

Case 0:

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>

<Choice D>

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:

<Question>

Choices:

<Choice A>

<Choice B>

<Choice C>

<Choice D>

Answer: Let's think step by step.

Figure 10: Prompt template for CoT prompting.