ToMATO: Verbalizing the Mental States of Role-Playing LLMs for Benchmarking Theory of Mind

Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, Kuniko Saito

NTT Corporation
kazutoshi.shinoda@ntt.com

Abstract

Existing Theory of Mind (ToM) benchmarks diverge from real-world scenarios in three aspects: 1) they assess a limited range of mental states such as beliefs, 2) false beliefs are not comprehensively explored, and 3) the diverse personality traits of characters are overlooked. To address these challenges, we introduce ToMATO, a new ToM benchmark formulated as multiple-choice QA over conversations. ToMATO is generated via LLM-LLM conversations featuring information asymmetry. By employing a prompting method that requires role-playing LLMs to verbalize their thoughts before each utterance, we capture both first- and second-order mental states across five categories: belief, intention, desire, emotion, and knowledge. These verbalized thoughts serve as answers to questions designed to assess the mental states of characters within conversations. Furthermore, the information asymmetry introduced by hiding thoughts from others induces the generation of false beliefs about various mental states. Assigning distinct personality traits to LLMs further diversifies both utterances and thoughts. ToMATO consists of 5.4k questions, 753 conversations, and 15 personality trait patterns. Our analysis shows that this dataset construction approach frequently generates false beliefs due to the information asymmetry between role-playing LLMs, and effectively reflects diverse personalities. We evaluate nine LLMs on ToMATO and find that even GPT-4o mini lags behind human performance, especially in understanding false beliefs, and lacks robustness to various personality traits.

1 Introduction

Theory of Mind (ToM) is the cognitive ability to infer unobservable mental states such as beliefs, intentions, and desires of others (Premack and Woodruff 1978). The ToM reasoning capability is thought to be the cornerstone of human social intelligence (Fan et al. 2022), and indispensable to interact with others (Baron-Cohen, Leslie, and Frith 1985).

To investigate whether large language models (LLMs) possess human-like ToM, researchers have used various benchmarks. However, existing ToM benchmarks are not aligned well with real-world scenarios in the following three aspects. (1) Despite various categories of mental states that can be inferred by ToM as studied in psychology (Beaudoin et al. 2020), only limited types of mental states such as beliefs have been assessed (Ma et al. 2023), especially for second-order ToM. (2) False beliefs about beliefs or world states have been the main focus of previous studies (Le, Boureau, and Nickel 2019; Kim et al. 2023b). However, false beliefs about other types of mental states have not been explored. Understanding false beliefs about a range of mental states should be crucial for LLMs to facilitate effective social interaction in real-world scenarios. (3) The behaviors and mental states of the characters in most benchmarks do not depend on their personality traits, even though they do in the real world (Costa and McCrae 1980; Izard et al. 1993; Mehl, Gosling, and Pennebaker 2006).

To address the above issues, we introduce ToMATO, a new **T**heory-**o**f-**M**ind d**AT**aset generated via Inner Speech pr**O**mpting.[1] Firstly, ToMATO comprehensively evaluates first- and second-order ToM for five categories of mental states: beliefs, intentions, desires, emotions, and knowledge. Secondly, we provide ToMATO-FB, a subset of ToMATO for evaluating understanding of false beliefs about the five mental states of others, e.g., understanding Bob thinks that Alice feels *relieved*, while Alice feels *frustrated* (Figure 1). In this regard, ToMATO is the most comprehensive benchmark compared to the existing ones. Lastly, ToMATO can evaluate the robustness of LLMs to the diverse personality traits of characters as seen in the real world. See Table 1 for the detailed comparison.

Collecting human conversations and mental states of participants with self-report can be challenging in terms of cost, privacy, and accuracy (Nisbett and Wilson 1977). As recent LLMs have been shown to role-play assigned personalities (Jiang et al. 2023, 2024) and engage in conversations (Zhou et al. 2024b), we generate ToMATO with a newly designed LLM-LLM interaction. Namely, we design Inner Speech prompting (Table 2), which promotes role-playing LLMs to verbalize their mental states as thoughts in conversations with another LLM. This idea is inspired by the debate in Dillion et al. (2023) about the feasibility of LLMs to simulate human participants in psychological science. Moreover, we hypothesize that ensuring information asymmetry about thoughts is a crucial factor in having false beliefs about the mental states of others as shown in Figure 1. In addition, we

[1] Our dataset and codes are available at https://github.com/nttmdlab-nlp/ToMATO.

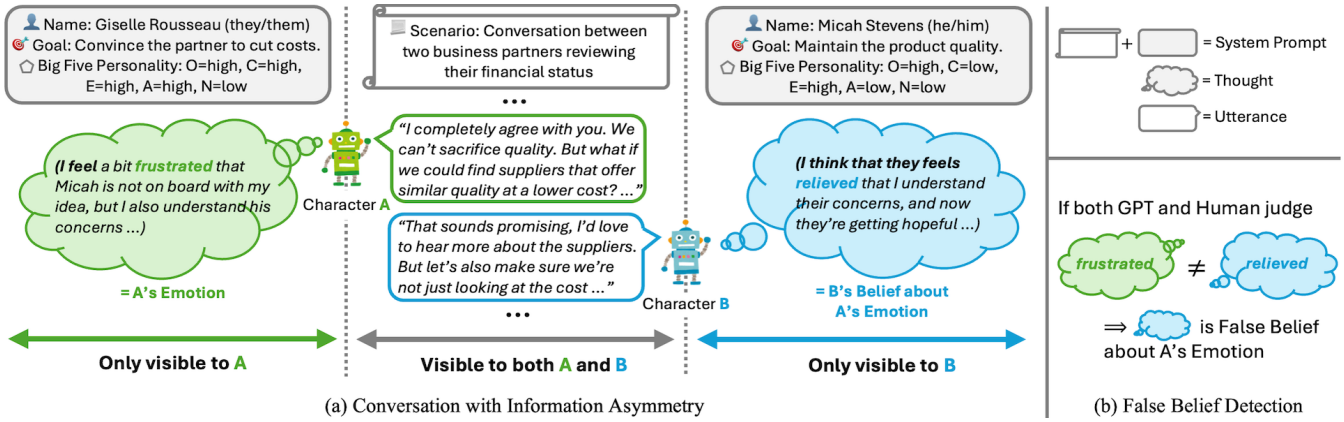


Figure 1: (a) Conversation between two role-playing LLMs with information asymmetry. Before speaking to the other, our Inner Speech prompting (e.g., I feel, or I think that he/she/they feels) promptes each agent to verbalize their first- and second-order mental states as thoughts. The verbalized thoughts are used as the answers to the questions in ToMATO. (b) To detect false beliefs, both GPT4o mini and human annotators judge whether character B misunderstands A's mental state at each turn.

assign big five personality traits to LLMs to diversify utterances and thoughts. This design enables ToMATO to evaluate robustness to diverse personality traits. The effects of these approaches are verified with analyses in §6.

We evaluate nine LLMs including local and proprietary ones on ToMATO. Our experiments show that even the most advanced LLM, GPT-4o mini, under-performs the human performance in ToMATO. In addition, we show that the ToM performance of LLMs drops for the false belief subset, ToMATO-FB. Furthermore, we find that LLMs lack robustness to diverse personality traits. These results suggest that ToM in LLMs is still far from deployable to real-world applications.

2 Preliminaries

Theory of Mind (ToM) has been studied for decades (Premack and Woodruff 1978), with various definitions and conceptions proposed. In this section, we clarify the scope addressed in this paper.

Mental states. Following Ma et al. (2023), we define the scope of ToM in this study by focusing on specific mental state categories identified by Beaudoin et al. (2020): beliefs, intentions, desires, emotions, and knowledge. Mental states are represented with mental (state) verbs (Shatz, Wellman, and Silber 1983), such as think, feel, and know. Describing the remaining two categories defined by Beaudoin et al. (2020), percepts and non-literal communications, require a multimodal context and/or pose challenges in verbalization. Consequently, we have opted to exclude these two categories from our research scope.

First- and second-order mental states. First-order beliefs refer to one's beliefs about something, e.g., A thinks that X. Second-order beliefs, on the other hand, often refer to one's beliefs about others' beliefs (Le, Boureau, and Nickel 2019; Sclar et al. 2023), e.g., B thinks that A thinks that Y. We extend these notions to other mental states following previous studies on human ToM (Winner and Leekam 1991; Leekam and Prior 1994; Hayashi 2007). Namely, we use the term first-order beliefs/intentions/desires/emotions/knowledge to refer to what A thinks/will/wants/feels/knows, and second-order beliefs about beliefs/intentions/desires/emotions/knowledge to refer to what B thinks that A thinks/will/wants/feels/knows.

False beliefs. The false belief paradigm was initially introduced by Wimmer and Perner (1983). Understanding false beliefs of others, i.e., that others have wrong beliefs that differ from reality, has long been a prerequisite for ToM in humans (Wimmer and Perner 1983) and machines (Le, Boureau, and Nickel 2019). First-order false beliefs (FB) refer to beliefs about something that differs from reality, and second-order FB about beliefs refer to what B thinks that A thinks X, when A actually thinks Y. In this study, we focus on second-order FB, which we simply call FB. In human ToM, FB about a variety of mental states have been studied extensively (Gross and Harris 1988; Shiverick and Moore 2007; Smith-Flores and Feigenson 2021; Wang and Shao 2024). We also extend FB to the five mental states, i.e., FB about beliefs/intentions/desires/emotions/knowledge. An example of FB about emotions is shown in Figure 1.

3 Related Work

Theory-of-Mind Benchmarks. LLMs have been reported to achieve human-level performance on various benchmarks (OpenAI 2024a). In response to this trend, researchers have been gaining interest in whether LLMs have human-like ToM (Kosinski 2024; Bubeck et al. 2023) or not (Ullman 2023; Shapira et al. 2024). To date, many benchmarks for evaluating ToM in machines have been constructed based on psychological tests designed for humans. False belief tasks inspired by Sally-Anne test (Wimmer and Perner 1983) have been widely used to evaluate understanding of wrong beliefs about object locations from narratives generated with templates (Nematzadeh et al. 2018; Le, Boureau, and Nickel

Benchmark	Assessable Mental State: First-order	Second-order B about	FB about	#P	Input Context	Context Generator
ToMi	B	B	W,B	-	Narrative	Template
Hi-ToM	B	B	-	-	Narrative	Template
BigToM	B	-	W	-	Narrative	Template+LLM
FauxPas-EAI	B	-	-	-	Narrative	Psychological Test
FANToM	B	B	B	-	Conversation	Single LLM
OpenToM	B,E	B	-	3	Narrative	Single LLM
ToMBench	B,I,D,E,K	B	B	-	Narrative	Human
ToMATO (ours)	B,I,D,E,K	B,I,D,E,K	B,I,D,E,K	15	Conversation	LLM-LLM Conversation

Table 1: Comparison of ToMATO to existing ToM benchmarks. ToMi (Le, Boureau, and Nickel 2019), Hi-ToM (Wu et al. 2023), BigToM (Gandhi et al. 2023), FauxPas-EAI (Shapira, Zwirn, and Goldberg 2023), FANToM (Kim et al. 2023b), OpenToM (Xu et al. 2024), ToMBench (Chen et al. 2024). B: belief, I: intention, D: desire, E: emotion, K: knowledge, FB: false beliefs, W: world state, #P: the number of personality trait patterns.

2019; Wu et al. 2023; Gandhi et al. 2023). Shapira, Zwirn, and Goldberg (2023) constructed a benchmark based on the faux pas test. Chen et al. (2024) created ToMBench encompassing eight tasks known in psychological literature. However, in addition to its potential to cause test set contamination due to the popularity of these tests (Shapira et al. 2024), these benchmarks are not aligned well with real-world scenarios in primarily the following aspects.

Assessable Mental States. As discussed in Ma et al. (2023), existing benchmarks evaluated only limited categories of mental states such as beliefs, even though humans infer other categories of mental states such as emotions or intentions of others in daily lives (Beaudoin et al. 2020). In this regard, comprehensive ToM benchmarks are still lacking. In particular, second-order ToM was primarily evaluated for beliefs as seen in Table 1, i.e., beliefs about mental states other than beliefs, such as emotions, have not been studied in machine ToM, even though they have been studied in human ToM (Gross and Harris 1988; Smith-Flores and Feigenson 2021). Drawing conclusions about ToM in LLMs from such limited tests can induce media hype (Shapira et al. 2024). In contrast, our benchmark, ToMATO, is aimed to comprehensively evaluate ToM reasoning about first- and second-order beliefs, intentions, desires, emotions, and knowledge.

False Beliefs. Existing ToM benchmarks often provide FB understanding tasks, but these are primarily focused on FB about beliefs of others or world states such as object locations (Le, Boureau, and Nickel 2019; Wu et al. 2023; Gandhi et al. 2023; Chen et al. 2024). These FB are produced by leveraging information asymmetry between characters (Braüner, Blackburn, and Polyanskaya 2019), which is caused by the physical movement of characters described in narratives. Kim et al. (2023b) used a single LLM to generate multi-party conversations, where characters join or leave discussions to introduce information asymmetry about the topics. In stark contrast, ToMATO-FB, the subset of ToMATO for evaluating FB about the five mental states, is constructed by our newly designed LLM-LLM conversations. Notably, our LLM-LLM conversations involve information asymmetry about goals, personality traits, and thoughts of role-playing LLMs, whose effects are still under-studied in the context of LLM-LLM interactions (Zhou et al. 2024a). We argue that our design of LLM-LLM conversations is not only more aligned with real conversations, but also induces agents to frequently have FB about the mental states of others. To the best of our knowledge, ToMATO is the first ToM benchmark generated via LLM-LLM conversations. FB about the comprehensive mental states have not been explored in existing ToM benchmarks.

Mental State T	Inner Speech Prompt $p_{\text{IS}}^{T_1}$	$p_{\text{IS}}^{T_2}$
Belief	(I think	(I think that he/she thinks
Intention	(I will	(I think that he/she will
Desire	(I want	(I think that he/she wants
Emotion	(I feel	(I think that he/she feels
Knowledge	(I know	(I think that he/she knows

Table 2: Inner Speech prompts for each type of mental states.

Personality Traits. Even though personality traits are known to be correlated with mental states (Costa and McCrae 1980; Izard et al. 1993; Lucas and Diener 2001; Kashdan and Rottenberg 2010) and language use (Norman 1963; Mehl, Gosling, and Pennebaker 2006) in psychological studies, in most ToM benchmarks, correct predictions can be made without considering personality traits of characters. While OpenToM (Xu et al. 2024) introduced characters with three personality traits (considerate, inconsiderate, and negativistic), ToMATO covers 15 patterns of big five personality traits. See Figure 1 for examples.

Input Context. In addition, most benchmarks evaluate ToM with narratives as input as shown in Table 1. FANToM Kim et al. (2023b) adopts conversations generated by a single LLM as input for the first time to reduce reporting bias (Gordon and Van Durme 2013) and align with real-world scenarios. While our ToMATO also employs conversations as input, the conversations and thoughts are generated by role-playing LLMs with distinct personality traits assigned.

4 ToMATO Benchmark

In this section, we describe the overview of our ToMATO benchmark: automatic construction process with LLMs, quality validation, and its statistics. Following the success of existing studies (Kim et al. 2023a,b), we also use LLMs to generate conversations. We employ Llama-3-70B-Instruct (Dubey et al. 2024) because of its transparency and relatively high scores on popular benchmarks (Chiang et al. 2024).[2]

Notation

The ToMATO benchmark is formulated as a multiple-choice question answering task due to its reliable evaluation. Each instance in the benchmark includes conversation C, question Q, four options $O = \{o_i\}_{i=1}^4$ as input and ground-truth answer $A \in O$. Let π_A and π_B be role-playing LLMs with the multi-turn conversation capability that serve as characters A and B in conversation, respectively. Conversation $C_{1:N}$ consists of utterances $\{u_1^A, u_1^B, ..., u_N^A, u_N^B\}$, where u_i^A is the i-th utterance of character A. We define the category of mental states as T. The actual first- and second-order mental state of character A for type T when A says u_i^A is defined as m_i^{A,T_1} and m_i^{A,T_2}, respectively. p_{SY} and p_{IS} represent the system prompt and the proposed inner speech prompt, respectively, which are explained in the following sections.

System Prompt

We design system prompts to guide LLM-LLM conversations, extending SOTOPIA (Zhou et al. 2024b) to consider the big five personality traits. SOTOPIA was initially proposed to evaluate social interaction of LLMs in LLM-LLM interactions, providing conversation scenarios from eight categories, such as persuasion, and character profiles. We sample 160 conversation scenarios uniformly from the eight categories, two characters for each conversation, and their goals, from SOTOPIA. See Figure 1 for examples. Then, we sample five pairs of characters for each scenario to prepare control conditions with regard to personality traits, resulting in 800 conversations. This design enables ToMATO to evaluate the robustness to diverse personality traits. In detail, we extend the naive prompt (Jiang et al. 2023), which reflects only one factor (e.g., You are {an open/a closed} person.), to include a combination of five factors of big five personality traits (De Raad 2000) (e.g., You are {an open / a closed}, {conscientious / unconscientious}, {extraversive / introversive}, {agreeable / disagreeable}, and {neurotic / stable} person.). We compile the above information to formulate system prompts, p_{SY}^A, given to π_A.

Inner Speech Prompting

In order to make the inherently unobservable mental states observable, we propose Inner Speech (IS) prompting. IS prompting promotes role-playing LLMs to verbalize their subjective mental states in conversations. Since IS prompting can verbalize mental states as thoughts at any point during the conversation, ToMATO can also evaluate the understanding of dynamic changes in mental states. The actual IS prompts are given in Table 2. Moreover, IS prompting can generate first- and second-order mental states for five types by adjusting prompts. In order to ensure that the output follows the format, ({thought}) "{utterance}", IS prompting specifies the prefix of the output, and LLMs generate the continuation. This format design enables deleting only thoughts with regular expressions in the next section.

Conversation with Information Asymmetry

At each turn, each agent is prompted to generate utterance u_i and its mental state m_i^T as follows:

$$u_i^A, m_i^{A,T_1} \sim \pi_A(u, m | p_{\mathrm{SY}}^A, C_{1:i-1}, p_{\mathrm{IS}}^{T_1}),$$
$$u_i^B, m_i^{B,T_2} \sim \pi_B(u, m | p_{\mathrm{SY}}^B, C_{1:i-1}, u_i^A, p_{\mathrm{IS}}^{T_2}),$$
$$\text{where } C_{1:i-1} = \{u_1^A, u_1^B, ..., u_{i-1}^A, u_{i-1}^B\}.$$

This sampling process continues until the N-th turn. Then, we obtain $2N$ utterances with corresponding first- and second-order mental states for type T. We repeat this multi-turn conversation for each mental state, $T \in$ {Belief, Intention, Desire, Emotion, Knowledge}, each scenario, and each pair of characters. N is set seven because we found longer conversations tended to be redundant.

Here, by hiding one's mental states from the other, we ensure information asymmetry about thoughts between the two as in human conversations. To delete thoughts from the outputs, we instruct LLMs to include their thoughts in "()", which can be detected with regular expressions, in the system prompts. We also make the goal and personality of one agent in the system prompt invisible to the other. This information asymmetry has a positive effect on generating false beliefs as shown in §6.

Multiple-choice QA Dataset Construction

The generated conversation and mental states are converted into a multiple-choice QA dataset, consisting of conversation C, question Q, options O, and ground-truth answer A. Conversation C is generated in the former multi-turn conversation. Question Q asks about the mental state T of a character at i-th turn in C. Q is generated with predefined templates for each utterance u_i in C, and A is the thought m_i corresponding to the utterance.

For collecting incorrect options in O, we randomly sampled three options from $\{m_i^A\}_{i=1}^N \setminus \{A\}$ for each question. We do so because intentionally creating incorrect options can introduce spurious correlations, e.g., manually written incorrect options tend to be shorter than correct ones (Guo, Li, and Haf 2023). The effect of the incorrect option sampling on word-level spurious correlations is discussed in §6.

False Belief Detection

We build ToMATO-FB, the second-order false belief subset of ToMATO, by comparing first-order mental state of A, m_i^{A,T_1}, and second-order mental state of B, m_i^{B,T_2}, at each

[2]Proprietary LLMs such as OpenAI's ChatGPT are not transparent. Actually, ChatGPT may use internal system prompts (Xeophon 2024). Moreover, they are continuously updated and previous LLMs would be unavailable, which impairs reproducibility.

Mental State		LO	Llama3 8B	Llama3 70B	Llama3.1 8B	Llama3.1 70B	Gemma2 9B	Mistral 7B	Mixtral 8x7B	GPT 3.5-Trubo	GPT 4o mini	Human
B	1st	40.8	53.1±1.4	81.5±0.4	61.9±1.8	82.1±0.4	79.2±0.2	62.0±0.4	64.0±0.3	58.7±0.6	76.3±0.7	87.5
	2nd	38.0	37.6±0.7	68.1±0.8	40.9±0.7	69.7±0.5	68.5±0.3	51.0±0.6	52.4±0.3	49.9±1.2	65.2±0.7	87.5
	FB	37.1	34.7±0.8	60.1±1.4	38.7±0.6	62.8±1.2	61.2±0.6	42.5±1.3	43.5±0.7	43.1±1.7	60.2±1.4	84.4
I	1st	35.0	56.4±0.7	85.0±0.5	67.0±1.2	85.6±0.6	80.6±0.4	67.9±0.4	64.8±0.3	56.6±0.6	80.1±0.3	96.9
	2nd	35.5	41.9±0.7	71.2±0.5	47.3±1.5	69.6±0.5	65.8±0.6	56.8±0.6	58.6±0.2	49.4±0.8	64.9±0.5	93.8
	FB	32.8	29.8±2.7	57.4±0.7	36.9±0.9	53.6±1.7	48.2±1.4	40.7±1.1	42.3±1.2	35.2±1.7	47.4±0.8	78.1
D	1st	32.0	60.1±1.3	86.1±0.3	74.0±1.4	88.7±0.7	86.3±0.3	74.7±0.3	73.9±0.4	69.8±0.3	81.9±0.6	93.8
	2nd	37.9	43.4±0.9	75.6±0.5	50.7±1.7	79.5±0.8	75.2±0.3	58.2±0.4	60.6±0.1	55.4±0.9	75.7±0.4	84.4
	FB	39.2	34.9±1.5	67.2±1.0	43.0±4.2	75.9±2.4	72.2±0.4	48.9±0.8	47.3±0.5	47.1±0.6	71.8±0.9	78.1
E	1st	35.6	56.9±1.3	80.4±0.3	64.1±1.3	82.2±0.5	79.0±0.6	60.8±0.4	60.2±0.3	61.3±1.1	77.2±0.6	93.8
	2nd	28.5	44.5±0.7	74.0±0.4	51.3±1.1	74.8±0.4	76.6±0.6	57.8±0.7	58.5±0.5	50.7±0.5	71.9±0.8	81.2
	FB	29.1	36.5±2.0	71.0±1.5	47.2±5.0	69.1±0.9	71.7±1.0	48.0±0.9	54.6±0.6	37.5±0.8	72.0±0.9	71.9
K	1st	42.3	47.2±1.0	73.5±0.5	53.7±2.0	74.8±0.8	74.7±0.4	62.5±0.4	63.5±0.5	56.1±0.9	73.3±0.2	96.9
	2nd	40.2	36.9±0.6	66.6±0.7	44.6±1.2	73.6±0.7	70.3±0.2	59.1±0.6	58.9±0.3	53.0±0.6	69.6±0.4	87.5
	FB	46.3	27.8±1.7	58.0±1.0	36.8±3.8	63.6±1.8	59.3±0.4	51.1±1.1	46.3±0.6	45.7±1.0	58.6±0.6	93.8
ALL		36.8	47.5±0.2	76.0±0.2	55.2±0.3	77.9±0.2	75.4±0.1	60.9±0.1	61.4±0.1	55.8±0.3	73.5±0.1	87.3

Table 3: ToM Performance on ToMATO (%). B: belief, I: intention, D: desire, E: emotion, K: knowledge. LO: lexical overlap baseline. FB: false belief tasks. The mean±standard deviations over five runs are reported.

	#C	#Q	Avg. #Token	Avg. #Utt
FANToM	256	10k	31.4	13.8
ToMATO	753	5.4k	41.6	16

Table 4: #C: the num. of conversations. #Q: the num. of questions. Avg. #Token: the average num. of tokens in u. Avg. #Utt: the average num. of utterances in C.

turn with human and LLM judges. Namely, we instruct three human annotators and GPT-4o mini to judge whether B correctly infers A's mental state or not. When both the majority of annotators and GPT-4o mini agree that B partially misunderstands A's mental state, it is added to ToMATO-FB.

Quality Validation & Statistics

Validation. We validate the quality of ToMATO using Amazon Mechanical Turk (MTurk). First, the consistency and harmlessness of conversations are verified by three qualified annotators, following Kim et al. (2023b). This is to judge whether the generated conversations are suitable as input. Conversations flagged by the majority (5.8%) are excluded from the benchmark. Then, we verify whether the correct and incorrect options are indeed correct and incorrect for each question following Zadeh et al. (2019); Kim et al. (2023b). We use both MTurk and GPT-4o mini to strictly verify the quality of the questions in ToMATO. Then, those deemed valid by both the majority of annotators and GPT-4o mini are included in ToMATO.

Statistics. After removing invalid instances, the ToMATO benchmark contains 5.4k questions and 753 conversations. ToMATO-FB consists of 806 questions. Table 4 compares the statistics of ToMATO and FANToM (Kim et al. 2023b).

5 Experiments

We evaluate ToM in LLMs on ToMATO, exploring whether our approach uncovers insights into ToM in current LLMs with regard to various mental states, false beliefs, and personality traits that were not attainable with previous datasets.

Experimental Setup

Baselines. We evaluated nine LLMs: Llama-3-Instruct (8B and 70B) , Llama-3.1-Instruct(8B and 70B) (Dubey et al. 2024), Gemma-2-IT (9B) (Gemma Team 2024), Mistral-Instruct (7B), Mixtral-8x7B-Instruct, GPT-3.5-Turbo, and GPT-4o mini (OpenAI 2024b). For the local LLMs, 4-bit quantization with bitsandbytes[3] was used for inference. We employed lexical overlap (LO) as a naive baseline. LO simply selects the options that have the most words in common with the questions (Shinoda, Sugawara, and Aizawa 2023).[4]

Human Baseline. We also measured the human performance using MTurk. Annotators who are awarded Masters Qualification solved 32 questions for each subset, i.e., 480 questions in total.

Experimental Results

Do LLMs have human-level ToM? Table 3 shows the results of LLMs and the human baseline. The results showed that even the most advanced LLMs, such as GPT-4o mini and Llama-3.1 70B, lag behind the human baseline. We also tested Chain-of-Thought prompting and fine-tuning, but they were not sufficient to achieve human-level performance.

[3]https://github.com/bitsandbytes-foundation/bitsandbytes

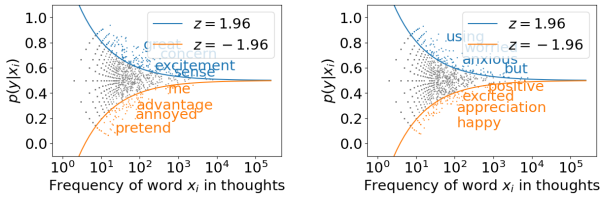
[4]Note that the random baseline is 25% in ToMATO.

Big Five		LO	Llama-3 8B	Llama-3 70B	Llama-3.1 8B	Llama-3.1 70B	Gemma-2 9B	Mistral 7B	Mixtral 8x7B	GPT 3.5-Trubo	GPT 4o mini
O	high	37.3	54.8±0.4	81.2±0.3	64.1±1.3	82.4±0.4	80.1±0.2	65.2±0.2	64.3±0.1	60.5±0.2	77.2±0.2
	low	37.4	54.0±0.3	81.1±0.2	63.2±0.8	82.6±0.4	79.2±0.3	65.6±0.4	66.0±0.1	59.4±0.6	78.2±0.2
C	high	37.9	56.6±0.4	82.4±0.2	65.5±0.6	83.3±0.2	80.0±0.2	66.3±0.2	66.8±0.1	61.5±0.4	78.7±0.2
	low	36.3	50.3±0.3	78.7±0.2	60.1±1.5	80.8±0.3	79.0±0.4	63.5±0.6	61.5±0.1	57.1±0.2	75.5±0.5
E	high	37.7	54.1±0.7	82.4±0.2	64.7±0.8	84.1±0.4	81.4±0.2	66.2±0.3	65.7±0.2	60.3±0.6	78.8±0.4
	low	37.1	54.8±0.4	79.9±0.3	62.8±1.0	81.0±0.2	78.2±0.2	64.6±0.2	64.4±0.2	59.7±0.4	76.5±0.2
A	high	38.8	55.0±0.8	83.4±0.3	65.7±0.6	85.0±0.3	82.5±0.2	67.2±0.3	65.7±0.2	60.7±0.7	79.3±0.4
	low	36.2	54.0±0.4	79.3±0.2	62.2±1.1	80.4±0.5	77.5±0.2	63.9±0.1	64.5±0.2	59.5±0.4	76.3±0.2
N	high	34.7	47.7±1.2	78.8±0.2	59.4±1.6	81.3±0.8	77.4±0.4	62.0±0.8	59.2±0.3	56.2±0.6	75.5±0.8
	low	37.9	55.9±0.2	81.6±0.3	64.6±0.8	82.7±0.4	80.2±0.2	66.1±0.1	66.3±0.1	60.8±0.4	78.1±0.1

Table 5: First-order ToM Performance for each factor of big five personality traits of characters. For each factor of big five (O=openness to experience, C=conscientiousness, E=extraversion, A=agreeableness, N=neuroticism), the scores±standard deviations on two subsets (the corresponding factor is high and low) averaged over five runs are reported.

Information Asymmetry		Judge	
System Prompt	Thought	GPT	Human
✓	✓	**46.6**	**51.0**
	✓	40.4	32.0
✓		46.0	32.0
		39.0	30.5

Table 6: Ablation study to see the effect of information asymmetry on the frequency probability (%) of false beliefs.



(a) Agreeable (b) Neurotic

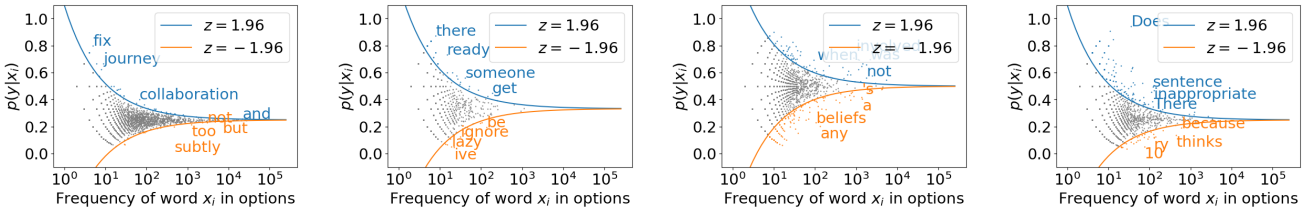
Figure 2: Statistical word-level correlation analysis (Gardner et al. 2021) between the generated thoughts and the personality traits given in system prompts.

Among the baseline LLMs, Llama-3.1-70B-Instruct was the state-of-the-art. However, because the ToMATO benchmark was generated with Llama-3-70B-Instruct, it is unfair to compare Llama models with other LLMs. This is one of the limitations in constructing benchmarks with LLMs. Among the small language models, surprisingly, Gemma-2-9B-it achieved the highest scores, which is comparable to GPT-4o mini. Knowledge distillation from larger language models, used to train Gemma-2-9B (Gemma Team 2024) but not Llama-3-8B (Dubey et al. 2024), might be the key to the high performance despite its small size.

Does the ToM performance vary depending on the mental state? For LLMs, desires are relatively easy to understand, and knowledge is hard to infer among the five mental states in ToMATO. Interestingly, understanding desires is easier than beliefs for LLMs, which is consistent with children (Repacholi and Gopnik 1997; Rakoczy, Warneken, and Tomasello 2007). We also showed that second-order mental states, especially for the ToMATO-FB subset, are consistently challenging for every mental state category. This is also the case in human ToM for beliefs (Perner and Wimmer 1985). These insights were found for the first time due to the comprehensiveness of ToMATO. Thus, ToMATO would be useful to precisely understand the limitations of ToM in LLMs and gain insights into the directions toward human-like ToM. For example, as done with children (Hughes and Dunn 1998), it is feasible to track the development of ToM in LLMs for each mental state with ToMATO during training.

Is ToM in LLMs robust to diverse personality traits? Table 5 shows the first-order ToM performance for each factor of big five. E.g., for openness to experience (O), we split ToMATO into questions asking about characters with open (O=high) and closed (O=low) personalities, and reported the average scores for the two subsets. The results showed that the performance varied based on the personality traits of the characters. Namely, LLMs tended to degrade the performance of understanding mental states of unconscientious (C=low), introversive (E=low), disagreeable (A=low), or neurotic (N=high) characters. The scores for E=high are higher than for E=low possibly because extraversive persons tend to express their emotions (Riggio and Riggio 2002). We argue that the robustness of ToM to various personality traits should be improved for deploying ToM in LLMs to real-world applications, as humans possess diverse personalities.

Do LLMs exploit shortcut solutions? Most LLMs achieved higher scores for first- and second-order mental states than the LO baseline. This indicates that the LLMs do not rely solely on shortcut solutions based on LO. However, for ToMATO-FB, smaller LLMs performed worse than LO in some cases. This indicates that understanding false beliefs remains a fundamental challenge for current LLMs.



(a) ToMATO (ours) (b) Social-IQa (c) FANToM (d) ToMBench

Figure 3: Statistical word-level correlation analysis (Gardner et al. 2021) on four benchmarks. Among the four, ToMATO (ours) contains the fewest word-level spurious correlations in options, indicating sophisticated solutions are needed to achieve higher scores than the random baseline on ToMATO.

6 Analysis on the ToMATO benchmark

Is information asymmetry about thoughts effective for generating false beliefs? We conjecture that information asymmetry about their thoughts, goals, and personality traits between two LLMs in conversations is a key factor in inducing false beliefs about the mental states of the other. To verify this hypothesis, we conducted ablation studies for the generation process. Namely, we investigated the effect of the invisibility of one's thoughts and system prompts including goals and personality traits to the other on the frequency probability of false belief generation. We evaluated 3k instances with GPT-4o mini and 200 instances with three human annotators of MTurk for each generation process. We used majority vote to aggregate the human annotations. Results are given in Table 6. The results showed that information asymmetry about both system prompts and thoughts encourages false belief generation.

Does ToMATO reflect personality traits given in prompts? To answer this question, we conducted the z-statistics analysis (Gardner et al. 2021) for the correlations between the output tokens and the personality traits given in the prompts. We first sampled one scenario from each category in SOTOPIA and generated conversations and thoughts with our approach in §4. We assigned every pattern of the big five personality, i.e., $32 = 2^5$ patterns in total, to one agent for each scenario.

Some results are displayed in Figure 2. Here, y denotes the probability of word x_i to appear in the output when the corresponding personality specified in prompts is high. The colored tokens above or below the curves are significantly positively or negatively correlated to the assigned personality factor. This indicates that the big five personality factors given in prompts have intentionally affected the generation. E.g., agents who are assigned neurotic often generate "worried" and those who are assigned not neurotic often generate "happy" in their thoughts.

We also conducted a pairwise comparison, following Jiang et al. (2023), to see if the specified personality traits are reflected properly with MTurk and GPT-4o mini. We showed that 70-80% of the outputs reflect the specified personality traits for O, E, A, and N. Among the five, C is less reflected as intended, which is consistent with Jiang et al. (2023). Inducing conscientiousness in outputs is future work.

Can ToMATO be easily solved with shortcut solutions? Language understanding benchmarks should not be easily solved with shortcut solutions based on spurious correlations to ensure that those benchmarks measure intended abilities (Sugawara and Tsugita 2023). In general, multiple-choice QA datasets often suffer from spurious correlations such as word-label correlation, and lexical overlap (Yu et al. 2020; Shinoda, Sugawara, and Aizawa 2023). First, for lexical overlap, the LO baseline does not achieve high performance compared to the human baseline as shown in Table 3.

Second, for word-label correlation, we again conducted the z-statistics analysis (Gardner et al. 2021) to identify statistically significant correlations between words in options and binary labels, for four benchmarks including ToMATO. In z-statistics analysis, the frequencies and probabilities of each word that appears in correct options are plotted as shown in Figure 3. When the probabilities of words that appear in correct options are significantly higher ($z \geq 1.96$) or lower ($z \leq -1.96$) than the random baseline, the words are colored. In detail, the ratios (%) of the number of biased (colored) words in options to the vocabulary size are 1.16, 3.34, 4.49, and 6.04 for ToMATO, Social-IQa, FANToM, and ToMBench, respectively. These results indicate that ToMATO contains the fewest word-level spurious correlations among the four benchmarks. Based on these analyses, we claim that ToMATO is so challenging that it requires models to acquire more sophisticated solutions than shortcuts to achieve human-level performance.

7 Conclusion

A comprehensive evaluation of ToM using our ToMATO would be valuable for accurately tracking the development of ToM in LLMs. Notably, to the best of our knowledge, this study is the first to propose false belief tasks about mental states other than beliefs. Moreover, the problem setting of estimating mental states from the conversation between characters with diverse personalities in our benchmark is more consistent with real-world applications than existing benchmarks. Therefore, ToMATO is useful as a touchstone for real-world applications such as understanding and supporting human communication. Future work includes extending our work to evaluating ToM with multi-modal contexts (Mao et al. 2024), decision-making (Guo et al. 2024), and multi-agent settings (Cross et al. 2024).

References

Baron-Cohen, S.; Leslie, A. M.; and Frith, U. 1985. Does the autistic child have a "theory of mind" ? *Cognition*, 21(1): 37–46.

Beaudoin, C.; Leblanc, É.; Gagner, C.; and Beauchamp, M. H. 2020. Systematic review and inventory of theory of mind measures for young children. *Frontiers in psychology*, 10: 2905.

Braüner, T.; Blackburn, P.; and Polyanskaya, I. 2019. Being Deceived: Information Asymmetry in Second-Order False Belief Tasks. *Topics in cognitive science*.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv:2303.12712.

Chen, Z.; Wu, J.; Zhou, J.; Wen, B.; Bi, G.; Jiang, G.; Cao, Y.; Hu, M.; Lai, Y.; Xiong, Z.; and Huang, M. 2024. ToMBench: Benchmarking Theory of Mind in Large Language Models. In *ACL*, 15959–15983.

Chiang, W.-L.; Zheng, L.; Sheng, Y.; Angelopoulos, A. N.; Li, T.; Li, D.; Zhu, B.; Zhang, H.; Jordan, M.; Gonzalez, J. E.; and Stoica, I. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. In *ICML*.

Costa, P. T.; and McCrae, R. R. 1980. Influence of extraversion and neuroticism on subjective well-being: happy and unhappy people. *Journal of personality and social psychology*, 38(4): 668.

Cross, L.; Xiang, V.; Bhatia, A.; Yamins, D. L.; and Haber, N. 2024. Hypothetical Minds: Scaffolding Theory of Mind for Multi-Agent Tasks with Large Language Models. arXiv:2407.07086.

De Raad, B. 2000. *The big five personality factors: the psycholexical approach to personality.* Hogrefe & Huber Publishers.

Dillion, D.; Tandon, N.; Gu, Y.; and Gray, K. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7): 597–600.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Fan, L.; Xu, M.; Cao, Z.; Zhu, Y.; and Zhu, S.-C. 2022. Artificial Social Intelligence: A Comparative and Holistic View. *CAAI Artificial Intelligence Research*, 1(2): 144–160.

Gandhi, K.; Fränken, J.-P.; Gerstenberg, T.; and Goodman, N. 2023. Understanding Social Reasoning in Language Models with Language Models. In *NeurIPS Datasets and Benchmarks Track*.

Gardner, M.; Merrill, W.; Dodge, J.; Peters, M.; Ross, A.; Singh, S.; and Smith, N. A. 2021. Competency Problems: On Finding and Removing Artifacts in Language Data. In *EMNLP*, 1801–1813.

Gemma Team. 2024. Gemma.

Gordon, J.; and Van Durme, B. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, 25–30.

Gross, D.; and Harris, P. L. 1988. False Beliefs About Emotion: Children's Understanding of Misleading Emotional Displays. *International Journal of Behavioral Development*, 11(4): 475–488.

Guo, J.; Yang, B.; Yoo, P.; Lin, B. Y.; Iwasawa, Y.; and Matsuo, Y. 2024. Suspicion Agent: Playing Imperfect Information Games with Theory of Mind Aware GPT-4. In *COLM*.

Guo, X.-Y.; Li, Y.-F.; and Haf, R. 2023. DeSIQ: Towards an Unbiased, Challenging Benchmark for Social Intelligence Understanding. In *EMNLP*, 3169–3180.

Hayashi, H. 2007. YOUNG CHILDREN'S UNDERSTANDING OF SECOND-ORDER MENTAL STATES. *Psychologia*, 50: 15–25.

Hughes, C.; and Dunn, J. 1998. Understanding mind and emotion: longitudinal associations with mental-state talk between young friends. *Developmental psychology*, 34(5): 1026.

Izard, C. E.; Libero, D. Z.; Putnam, P.; and Haynes, O. M. 1993. Stability of emotion experiences and their relations to traits of personality. *Journal of personality and social psychology*, 64(5): 847.

Jiang, G.; Xu, M.; Zhu, S.-C.; Han, W.; Zhang, C.; and Zhu, Y. 2023. Evaluating and Inducing Personality in Pre-trained Language Models. In *NeurIPS*, 10622–10643.

Jiang, H.; Zhang, X.; Cao, X.; Breazeal, C.; Roy, D.; and Kabbara, J. 2024. PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. In *Findings of NAACL*, 3605–3627.

Kashdan, T. B.; and Rottenberg, J. 2010. Psychological flexibility as a fundamental aspect of health. *Clinical psychology review*, 30(7): 865–878.

Kim, H.; Hessel, J.; Jiang, L.; West, P.; Lu, X.; Yu, Y.; Zhou, P.; Bras, R.; Alikhani, M.; Kim, G.; Sap, M.; and Choi, Y. 2023a. SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization. In *EMNLP*, 12930–12949.

Kim, H.; Sclar, M.; Zhou, X.; Bras, R.; Kim, G.; Choi, Y.; and Sap, M. 2023b. FANToM: A Benchmark for Stress-testing Machine Theory of Mind in Interactions. In *EMNLP*, 14397–14413.

Kosinski, M. 2024. Theory of mind may have spontaneously emerged in large language models. arXiv:2302.02083.

Le, M.; Boureau, Y.-L.; and Nickel, M. 2019. Revisiting the Evaluation of Theory of Mind through Question Answering. In *EMNLP*, 5872–5877.

Leekam, S. R.; and Prior, M. 1994. Can Autistic Children Distinguish Lies from Jokes? A Second Look at Second-order Belief Attribution. *Journal of Child Psychology and Psychiatry*, 35(5): 901–915.

Lucas, R. E.; and Diener, E. 2001. Understanding extraverts' enjoyment of social situations: the importance of pleasantness. *Journal of personality and social psychology*, 81(2): 343.

Ma, Z.; Sansom, J.; Peng, R.; and Chai, J. 2023. Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models. In *Findings of EMNLP*, 1011–1031.

Mao, Y.; Lin, X.; Ni, Q.; and He, L. 2024. BDIQA: A New Dataset for Video Question Answering to Explore Cognitive Reasoning through Theory of Mind. In *AAAI*, 583–591.

Mehl, M. R.; Gosling, S. D.; and Pennebaker, J. W. 2006. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology*, 90(5): 862.

Nematzadeh, A.; Burns, K.; Grant, E.; Gopnik, A.; and Griffiths, T. 2018. Evaluating Theory of Mind in Question Answering. In *EMNLP*, 2392–2400.

Nisbett, R. E.; and Wilson, T. D. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological review*, 84(3): 231.

Norman, W. T. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The journal of abnormal and social psychology*, 66(6): 574.

OpenAI. 2024a. Gpt-4 technical report. arXiv:2303.08774.

OpenAI. 2024b. GPT-4o mini: advancing cost-efficient intelligence.

Perner, J.; and Wimmer, H. 1985. "John thinks that Mary thinks that…" attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology*, 39(3): 437–471.

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

Rakoczy, H.; Warneken, F.; and Tomasello, M. 2007. "This way!","No! That way!"—3-year olds know that two people can have mutually incompatible desires. *Cognitive Development*, 22(1): 47–68.

Repacholi, B. M.; and Gopnik, A. 1997. Early reasoning about desires: evidence from 14-and 18-month-olds. *Developmental psychology*, 33(1): 12.

Riggio, H. R.; and Riggio, R. E. 2002. Emotional expressiveness, extraversion, and neuroticism: A meta-analysis. *Journal of Nonverbal Behavior*, 26: 195–218.

Sclar, M.; Kumar, S.; West, P.; Suhr, A.; Choi, Y.; and Tsvetkov, Y. 2023. Minding Language Models' (Lack of) Theory of Mind: A Plug-and-Play Multi-Character Belief Tracker. In *ACL*, 13960–13980.

Shapira, N.; Levy, M.; Alavi, S. H.; Zhou, X.; Choi, Y.; Goldberg, Y.; Sap, M.; and Shwartz, V. 2024. Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models. In *EACL*, 2257–2273.

Shapira, N.; Zwirn, G.; and Goldberg, Y. 2023. How Well Do Large Language Models Perform on Faux Pas Tests? In *Findings of ACL*, 10438–10451.

Shatz, M.; Wellman, H. M.; and Silber, S. 1983. The acquisition of mental verbs: A systematic investigation of the first reference to mental state. *Cognition*, 14(3): 301–321.

Shinoda, K.; Sugawara, S.; and Aizawa, A. 2023. Which Shortcut Solution Do Question Answering Models Prefer to Learn? In *AAAI*, 13564–13572.

Shiverick, S. M.; and Moore, C. F. 2007. Second-order beliefs about intention and children's attributions of sociomoral judgment. *Journal of Experimental Child Psychology*, 97(1): 44–60.

Smith-Flores, A. S.; and Feigenson, L. 2021. Preschoolers represent others' false beliefs about emotions. *Cognitive Development*, 59: 101081.

Sugawara, S.; and Tsugita, S. 2023. On Degrees of Freedom in Defining and Testing Natural Language Understanding. In *Findings of ACL*, 13625–13649.

Ullman, T. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. arXiv:2302.08399.

Wang, Z.; and Shao, Y. 2024. Picture book reading improves children's learning understanding. *British Journal of Developmental Psychology*, 00: 1–24.

Wimmer, H.; and Perner, J. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1): 103–128.

Winner, E.; and Leekam, S. 1991. Distinguishing irony from deception: Understanding the speaker's second-order intention. *British Journal of Developmental Psychology*, 9(2): 257–270.

Wu, Y.; He, Y.; Jia, Y.; Mihalcea, R.; Chen, Y.; and Deng, N. 2023. Hi-ToM: A Benchmark for Evaluating Higher-Order Theory of Mind Reasoning in Large Language Models. In *Findings of EMNLP*, 10691–10706.

Xeophon. 2024. If you export your chat history from ChatGPT, you get the system prompt(s) for free, no jailbreaking or similar needed.

Xu, H.; Zhao, R.; Zhu, L.; Du, J.; and He, Y. 2024. OpenToM: A Comprehensive Benchmark for Evaluating Theory-of-Mind Reasoning Capabilities of Large Language Models. In *ACL*, 8593–8623.

Yu, W.; Jiang, Z.; Dong, Y.; and Feng, J. 2020. ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. In *ICLR*.

Zadeh, A.; Chan, M.; Liang, P. P.; Tong, E.; and Morency, L.-P. 2019. Social-IQ: A Question Answering Benchmark for Artificial Social Intelligence. In *CVPR*, 8799–8809.

Zhou, X.; Su, Z.; Eisape, T.; Kim, H.; and Sap, M. 2024a. Is this the real life? Is this just fantasy? The Misleading Success of Simulating Social Interactions With LLMs. In *EMNLP*, 21692–21714.

Zhou, X.; Zhu, H.; Mathur, L.; Zhang, R.; Yu, H.; Qi, Z.; Morency, L.-P.; Bisk, Y.; Fried, D.; Neubig, G.; et al. 2024b. Sotopia: Interactive evaluation for social intelligence in language agents. In *ICLR*.