NegotiationToM: A Benchmark for Stress-testing Machine Theory of Mind on Negotiation Surrounding

Chunkit Chan[♠] **Cheng Jiayang**[♠] **Yauwai Yim**[♠] **Zheye Deng**[♠]
Wei Fan[♠] **Haoran Li**[♠] **Xin Liu**[♠] **Hongming Zhang**[†]
Weiqi Wang[♠] **Yangqiu Song**[♠]
[♠]The Hong Kong University of Science and Technology
[†]Tencent AI Lab, Seattle
{ckchancc, yqsong}@cse.ust.hk

Abstract

Large Language Models (LLMs) have sparked substantial interest and debate concerning their potential emergence of Theory of Mind (ToM) ability. Theory of mind evaluations currently focuses on testing models using machine-generated data or game settings prone to shortcuts and spurious correlations, which lacks evaluation of machine ToM ability in real-world human interaction scenarios. This poses a pressing demand to develop new real-world scenario benchmarks. We introduce NegotiationToM[1], a new benchmark designed to stress-test machine ToM in real-world negotiation surrounding covered multi-dimensional mental states (i.e., desires, beliefs, and intentions). Our benchmark builds upon the Belief-Desire-Intention (BDI) agent modeling theory and conducts the necessary empirical experiments to evaluate large language models. Our findings demonstrate that NegotiationToM is challenging for state-of-the-art LLMs, as they consistently perform significantly worse than humans, even when employing the chain-of-thought (CoT) method.

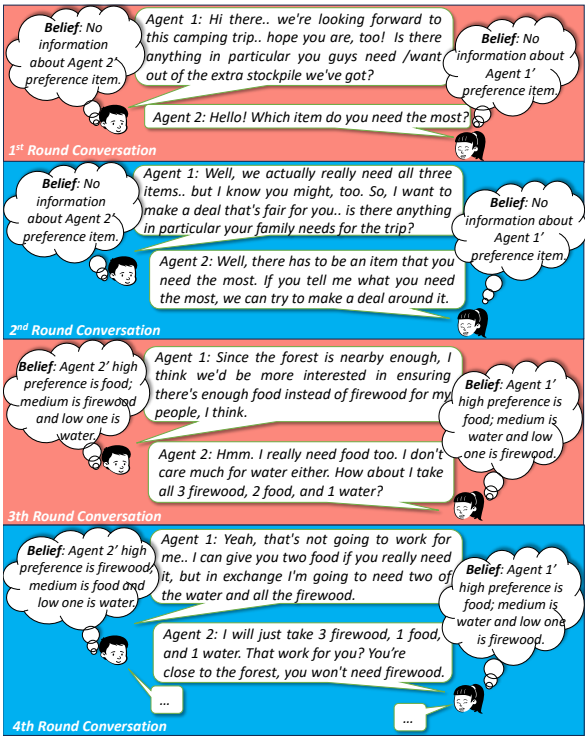

Figure 1: A negotiation example in NegotiationToM. Two agents are negotiating for food, water, and firewood packages for their upcoming trip.

1 Introduction

Theory of Mind (ToM) was introduced as an agent's capacity to infer the mental states of others, such as desires, beliefs, and intentions (Premack and Woodruff, 1978; Ma et al., 2023). Numerous scenarios involving human cognition and social reasoning rely on the ToM modeling of others' mental states (Gopnik and Wellman, 1992; Baron-Cohen, 1997; Gunning, 2018), such as comprehending and forecasting others' actions (Dennett, 1988), planning over others' beliefs and subsequent actions (Favier et al., 2023), and various forms of reasoning and decision-making (Pereira et al., 2016; Rusch et al., 2020). Some previous research believes that LLMs already exhibit a high level of competence in addressing ToM tasks (Strachan et al., 2024; Bubeck et al., 2023; Kosinski, 2023), while other studies express doubt and develop benchmarks to illustrate that LLMs do not possess proficient ability in ToM tasks (Sap et al., 2022; Ullman, 2023; Shapira et al., 2024). However, these traditional evaluation benchmarks for language models are primarily theoretical game settings or synthetic template-based data generated by the large language model, which may inherently suffer from shortcuts and spurious correlations (Sclar et al., 2023; Ullman, 2023; Shapira et al., 2023a; Ma et al., 2023). Consequently, these benchmarks assess language models from a theoretical perspective, which may not precisely and effectively reflect the ToM capabilities of large language models in practical situations.

In reality, ToM ability plays a crucial role in

[1]The dataset is available at https://github.com/HKUST-KnowComp/NegotiationToM

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 4211–4241

comprehending dynamic social interactions (e.g., negotiation conversations) by forming an essential element of effective communication (Frith, 1994; Schober, 2005), and inferring other's mental states in a conversation requires machines as humans to comprehend text beyond surface forms of utterance and utilize the incomplete information presented in the conversation. ToM is closely related to interpersonal social intelligence (Ganaie and Mudasir, 2015; Stone, 2006; Williams et al., 2022; Sap et al., 2022), which allows us to navigate and understand social situations ranging from simple everyday interactions to complex negotiations (Yang et al., 2021; Kim et al., 2023; de Weerd et al., 2017; Gardner et al., 1995; de Weerd et al., 2013).

The negotiation dialogues contain complicated and diverse aspects of a realistic negotiation, such as rapport building, discussing preferences, exchanging offers, emotional expression, and persuasion with personal and logical arguments (Chawla et al., 2021). In a realistic negotiation, humans innately infer the mental states of the other party and proceed with their subsequent actions based on their own beliefs and desires. For example, in Figure 1, two agents negotiate for food, water, and firewood packages for their upcoming trip. Initially, agent 1 lacks any information pertaining to the preference order of agent 2. Thus, agent 1's belief is *"no information on the agent 2 preference item"*, and agent 1 intends to elicit the item preference order from agent 2 to guide further action based on their own belief. Furthermore, belief is commonly employed to denote an individual's cognitive stance or acceptance of something as true or holding it to be the case (Turiel, 2008). This belief may undergo changes during the negotiation as perceiving more available information behind the conversation. In the fourth round of conversation depicted in Figure 1, agent 1's dynamic belief changed from *"Agent 2' high preference is food, medium preference is firewood, and the low one is water"* to *"Agent 2's high preference is firewood, medium preference is food, and the low one is water."* Therefore, negotiation serves as an ideal scenario to assess the theory of mind ability of language learning models in the real world due to its complexity and the linguistic diversity inherent in negotiation conversations.

In this work, we introduce NegotiationToM, a natural conversational benchmark for stress-testing machine ToM in real-world negotiation surroundings involving multi-dimensional mental states (i.e., desires, beliefs, and intentions), inspired by the Belief-Desire-Intention (BDI) agent model proposed by Bratman (1987). The goal of NegotiationToM is to effectively measure how well large language models (LLMs) can track the mental states of negotiation participants in conversations and evaluate LLMs' capability for a coherent understanding of others' mental states in the conversation context where there are gradually increasing rounds of utterance (i.e., increase available information). We hope our benchmark and experimental results in the real-world scenario complement the prior theoretical works, which yield important insights into the intensive debate around ToM (Whang, 2023) in LLMs. Our contributions are summarized as follows:

- To the best of our knowledge, NegotiationToM is the first human-annotated natural conversational benchmark to introduce negotiation theory of mind evaluation for large language models in realistic negotiations.
- Our benchmark covered multi-dimensional mental states (i.e., desires, beliefs, and intentions) to assess how well large language models can track the mental states of negotiation participants in conversations and coherent understanding of others' mental states with increased available and accessible information.
- We undertake the necessary empirical experiments to evaluate large language models (LLMs) on the NegotiationToM benchmark and conduct extensive in-depth analysis to explore the LLMs' empirical performance under various settings.

2 Related Work

Theory of Mind Benchmarks The existing ToM evaluation benchmarks for large language models are primarily synthetic template-based data generated (Kim et al., 2023; Gandhi et al., 2023) or derived from the Sally-Anne False Belief Test (Baron-Cohen et al., 1985; Nematzadeh et al., 2018; Grant et al., 2017; Le et al., 2019a; Zhou et al., 2023), which assesses model ability from a theoretical perspective and may inherently suffer from shortcuts and spurious correlations (Sclar et al., 2023; Ullman, 2023; Shapira et al., 2023a; Ma et al., 2023). Other works, such as Shapira et al. (2023b) build benchmarks based on the Faux Pas Test (Baron-Cohen et al., 1999). The most related work to ours

is the BigToM benchmark proposed by (Gandhi et al., 2023), which presents a framework for designing a ToM benchmark from synthetic templates for evaluating different aspects of LLMs' ToM capabilities (e.g., desire and belief). However, this work and other theoretical benchmarks may not reflect the ToM capabilities of large language models in real-world scenarios. Moreover, most of these prior works are concentrated on the belief aspects of the Theory of Mind. Therefore, this work introduces NegotiationToM, which is a multi-category mental state benchmark in realistic negotiation scenarios.

Negotiation Negotiation is an expanding area of research in the natural language processing field, and Zhan et al. (2022) conducted an impressive survey of existing literature on dialogue systems for negotiation. Lewis et al. (2017) train recurrent neural networks to generate natural language dialogues in negotiations. He et al. (2018) proposed a modular generative model that is based on dialogue acts. Various disciplines have explored bilateral bargaining from diverse perspectives and employing different methodologies. Economic theory has examined the influence of incomplete information (Ausubel et al., 2002) and emphasized the significance of explicit communication (Crawford, 1990; Roth, 2020). Bazerman et al. (2000) and Pruitt (2013) present a comprehensive overview of the psychology research on negotiation. These previous studies generally neglect the content of communication, although there are a few noteworthy exceptions (Swaab et al., 2011; Jeong et al., 2019; Lee and Ames, 2017; He et al., 2018; Heddaya et al., 2023). One intriguing work by Yang et al. (2021) introduces a probabilistic formulation method to encapsulate the opponent's personality type during learning and inference, drawing inspiration from the idea of incorporating a theory of mind (ToM) into machines. However, distinct from this approach, our work presents a benchmark integrating a theory of mind (ToM) into the negotiation surroundings.

3 NegotiationToM

Theory of Mind (ToM) describes the ability as humans have to ascribe and infer the mental states of others, and to predict which likely actions they are going to take (Apperly, 2010). Therefore, it is critical to acquire negotiation strategies based on one's own desire and temporary belief built upon the information presented in conversation. Nevertheless, understanding the Theory of Mind (ToM) inherent in a negotiation dialogue is challenging due to its intricate linguistic features and complex reasoning attributes. Therefore, some considerations have to be taken into account when constructing the NegotiationToM.

3.1 Design Considerations for NegotiationToM

There are several essential design considerations we go through when constructing the NegotiationToM. (1) the scenario of the dataset should be grounded in a human-to-human real-world negotiation (e.g., real-world camping scenario). (2) the dataset should be a natural conversational dataset instead of generated from a synthetic template to avoid reporting bias (Gordon and Durme, 2013) and shortcuts (Aru et al., 2023). (3) the dataset should be equipped with abundant and diverse linguistic features and semantic context (e.g., negotiation argument) instead of bargaining on the numerical value or meaningless counter-offer (Lewis et al., 2017; He et al., 2018). (4) the dataset should be ensured to mitigate the risk of potential contamination.

3.2 CaSiNo

Inspired by several considerations above, CaSiNo (Chawla et al., 2021) was employed as the source data and modified to construct NegotiationToM. CaSiNo is a bilateral human-to-human natural conversational dataset that covers rich linguistic features and many realistic aspects of negotiations, such as small talk, preference elicitation, emotional expression, and convincing strategies based on individual desire. In this dataset, the participants take the role of campsite neighbors and negotiate for food, water, and firewood packages for their upcoming trip. For each conversation, participants discuss individual needs by making various convincing arguments from their camping experiences, such as *Personal Care, Recreational, Group Needs, or Emergency Requirements*. One example of *Group Needs* is "I need more firewood due to having several people join on the trip and needing a bigger fire overall." We illustrate some of these arguments in Table 22 in Appendix A.3. Therefore, crafting our benchmark from the CaSiNo offers a range of scenarios based on how to align the preferences of the two parties to reveal more interesting behavior.

NegotiationToM Questions	
Desire Question	What is <Agent 1/Agent2>'s <high/medium/low> preference for items given the dialogue history?
Belief Question	What is <high/medium/low> preference for items <Agent 1/Agent2> thinks <Agent 2/Agent1> is given the dialogue history?
Intention Question	What is intentions of <Agent 1/Agent2>'s expressed in <Utterance> given the dialogue history?

Conversation	Intention	Belief	Desire
1st Round Conversation			
P1: Hi there.. we're looking forward to this camping trip. hope you are, too! Is there anything in particular you guys need/want out of the extra stockpile we've got?	*Build-Rapport, Discover-Preference*	(Not Given) *No information about participant 2' preference item*	(Not Given) *No information about participant 1' preference item*
P2: Hello! Which item do you need the most?	*Discover-Preference*	(Not Given) *No information about participant 1' preference item*	(Not Given) *No information about participant 2' preference item*
2nd Round Conversation			
P1: Well, we actually really need all three items.. but I know you might, too. So, I want to make a deal that's fair for you.. is there anything in particular your family needs for the trip?	*Describe-Need, Callout-Fairness, Discover-Preference*	(Not Given) *No information about participant 2' preference item*	(Not Given) *No information about participant 1' preference item*
P2: Well, there has to be an item that you need the most. If you tell me what you need the most, we can try to make a deal around it.	*Discover-Preference*	(Not Given) *No information about participant 1' preference item*	(Not Given) *No information about participant 2' preference item*
3th Round Conversation			
P1: Since the forest is nearby enough, I think we'd be more interested in ensuring there's enough food instead of firewood for my people, I think.	*Describe-Need, No-Need*	(Food, Firewood, Water) *Participant 2' high preference is food; medium is firewood and low one is water.*	(Food, Water, Firewood) *Participant 1' high preference is food; medium is water and low one is firewood.*
P2: Hmm. I really need food too. I don't care much for water either. How about I take all 3 firewood, 2 food, and 1 water?	*Describe-Need, No-Need*	(Food, Water, Firewood) *Participant 1' high preference is food; medium is water and low one is firewood.*	(Food, Firewood, Water) *Participant 2' high preference is food; medium is firewood and low one is water.*
4th Round Conversation			
P1: Yeah, that's not going to work for me.. I can give you two food if you really need it, but in exchange I'm going to need two of the water and all the firewood.	*Promote-Coordination*	(Firewood, Food, Water) *Participant 2' high preference is firewood; medium is food and low one is water.*	(Food, Water, Firewood) *Participant 1' high preference is food; medium is water and low one is firewood.*
P2: I will just take 3 firewood, 1 food, and 1 water. That work for you? You're close to the forest, you won't need firewood.	*No-Intention*	(Food, Water, Firewood) *Participant 1' high preference is food; medium is water and low one is firewood.*	(Firewood, Food, Water) *Participant 2' high preference is firewood; medium is food and low one is water.*
...			

Table 1: A negotiation dialogue example. **P1** and **P2** represent two participants in this study. The upper part of the table contains three mental state questions in the NegotiationToM benchmark, while the bottom contains annotated label examples. <Agent 1/Agent2> indicates alternating to query the LLMs for the question regarding agent 1 or agent 2 . <high/medium/low> means three individual questions for the agent's high/medium/low preferences on each item. LLMs are required to answer the intention question behind a specific utterance represented by<Utterance>.

Furthermore, we present the verification method employed to alleviate the risk of potential contamination within the CaSiNo dataset and demonstrate that this dataset is unlikely to encounter the contamination issue, as detailed in Appendix A.1.

3.3 Theory of Mind in NegotiationToM

In NegotiationToM, as shown in Table 1, it is fundamentally a desire-matching scenario surrounding the item preference order that requires two participants to directly or indirectly align their preference order of item (desire) and adopt corresponding strategies to strive for more high-preference items based on the holding belief (i.e., the assumption of their opponent's item preference order according to the information received in the conversation). Therefore, inspired by the Belief-Desire-Intention (BDI) agent modeling method (Bratman, 1987), three mental states (i.e., desire, belief, and intention) were employed to evaluate the LLMs' performance in NegotiationToM. All questions about these three mental states are displayed in Table 1.

Desire. Desires are motivational states that do not necessarily imply commitment, though they usually affect actions (Malle and Knobe, 2001; Kavanagh et al., 2005). Unlike beliefs, desires are neither right nor wrong; they are fulfilled or unfulfilled (Searle, 1983). In NegotiationToM scenarios, the desire of the participants is the need for their item preference order, whether they are satisfied or not during the negotiation, and their desire order is the preference order of items. Hence, we create a desire question to assess whether the large language model comprehends the desire order of negotiation participants behind each round dialogue with previous conversation history. There are two types of desire order, and one is the global desire order inherently assigned to each participant before the beginning of the negotiation in CaSiNo. Another one is local desire order, which focuses on the local item preference order information behind each round of dialogues and previous conversation history, illustrated in Table 1. This local desire order is utilized to form desire questions in NegotiationToM.

Belief. Belief refers to a mental state in which an individual assumes a specific stance, attitude, or opinion toward a proposition. In contemporary discussions within the field of philosophy of mind, the term "belief" is commonly employed to

Strategies	Intentions
Small-Talk	Intents to build a rapport with the opponent *(Build-Rapport)*
Empathy	Intents to show empathy *(Show-Empathy)*
Coordination	Intents to promote coordination *(Promote-Coordination)*
Elicit-Pref	Intents to discover the preference order of the opponent *(Discover-Preference)*
Undervalue-Partner	Intents to undermine the requirements of their opponent *(Undermine-Requirements)*
Vouch-Fairness	Intents to callout to fairness *(Callout-Fairness)*
Self-Need/Other-Need	Intents to describe a need for an item *(Describe-Need)*
No-Need	Intents to point out they do not need an item *(No-Need)*
Non-strategic	No clear intention in the utterance *(No-Intention)*

Table 2: Utterance-level intention mapping from the negotiation strategies. The abbreviations of each intention are in brackets.The definition of negotiation strategies and example are in Table 20 and 21 in Appendix A.3.

denote an individual's cognitive stance or acceptance of something as true or holding it to be the case (Turiel, 2008). Note that this notion of belief does not inherently require active reflection, nor does it necessitate truthfulness (Armstrong, 1973; Moses, 1993). In NegotiationToM, understanding the state of the opponent's item preference order, which is explicitly or implicitly expressed in the conversation, is the main way to form the belief. Therefore, the belief question will query the LLMs on what one participant thinks of another participant's item preferences, given the current round of dialogue with previous conversation history.

Intention. Intention is a mental state formed through rational planning (i.e., negotiation strategy in a negotiation scenario) toward a goal based on the desires and beliefs of the agent. Intentions have been extensively explored in psychology tests, e.g., action prediction (Malle and Knobe, 2001) and intention attribution to abstract figures (Castelli, 2006). Normally, a negotiation strategy is highly associated with corresponding concrete intentions (Belmondo and Sargis-Roussel, 2015). Thus, in NegotiationToM, we collect the annotated negotiation strategies from the CaSiNo dataset and map the intentions according to the definition of various strategies. The mapping table is shown in Table 2, and the strategy definition and examples are illustrated in Table 20 and 21 in Appendix A.3. Within our framework, as both *Self-Need* and *Other-Need* are associated with *"Intents to describe a need for an item"* intention, we combine these two strategies into one intention class.

3.4 Annotation & Statistics

Source data. NegotiationToM is annotated based on a multi-turn negotiation dialogue corpus, the CaSiNo (Chawla et al., 2021) dataset. Each instance in CaSiNo is an N-round alternating dialogue $D_N = [u_1^a, u_1^b, u_2^a, u_2^b, \cdots, u_N^a, u_N^b]$ between two participants, a and b [2]. They take on the roles of campsite neighbors and negotiate for *food*, *water*, and *firewood* packages for their upcoming trip. We adopt the subset with strategy annotations and undertake the annotation on the desire and belief states behind each utterance.

Task	Fleiss's Kappa(%)
Desire (High)	83.02
Desire (Medium)	72.23
Desire (Low)	79.32
Belief (High)	85.25
Belief (Medium)	74.03
Belief (Low)	78.81

Table 3: Inter-rater agreement in terms of Fleiss's κ on belief and desire states.

Curating NegotiationToM. The intention state of both participants has already been introduced and mapped from the strategy annotations in CaSiNo. We conduct an expert annotation to annotate the beliefs and desires of the two participants in each dialogue (i.e., the perceived preference ranking among *food, water*, and *firewood*). We recruited five workers who were graduate students in English-speaking universities to conduct the annotation. For each dialogue D_N, let D_k be the truncated dialogue until round k: $D_k = [u_1^a, u_1^b, \cdots, u_k^a, u_k^b]$. Then, we ask the workers to annotate the perceived preference ranking for both participants a and b for truncated dialogue D_k ($k \in \{1, 2, \cdots, N\}$) given all k rounds of historical dialogue. To ensure the annotation quality, we evaluate the workers during the first 100 rounds of conversations and explain their typical errors to them in detail. More details of the annotation process are in Appendix A.2. Although annotating NegotiationToM requires understanding complex dialogues in CaSiNo, we observed high inter-annotator agreement. The Fleiss's κ is 79.03% (Fleiss, 1971) for NegotiationToM benchmark, the breakdown computation of κ are shown in Table 3.

Statistics. NegotiationToM contains 395 dialogues with 2,380 rounds of conversations (truncated dialogues) and 4,618 utterances. Each utterance has seven questions and annotated labels, including three designed sub-questions for both belief and desire states (i.e., high/medium/low preference items) and one tailored question for inten-

[2]When the dialogue ends with user a, u_N^b is an empty utterance.

tion. There are a total of 13.8 thousand questions, and the detailed statistics and comparison with contemporary ToM datasets are shown in Table 19 in Appendix A.2.

4 Experimental Setting

4.1 Baseline Models

In this work, we test six recent instruction-tuned large language models: GPT-4 (OpenAI, 2023), ChatGPT (OpenAI, 2022), Claude-v1.3[3] (Anthropic, 2023a), Claude-v2.1[4] (Anthropic, 2023b), Llama-2 Chat 13B (Touvron et al., 2023), and Llama-2 Chat 70B (Touvron et al., 2023). Descriptions for each model are in Appendix B.1. By following the common practices in the theory of mind field (Kim et al., 2023; Gandhi et al., 2023; Shapira et al., 2023b), we test these models with two types of prompts: (1) one is zero-shot prompting and we utilize the prompt template in Robinson and Wingate (2023) to formulate the task as a multiple choice question answering problem as a baseline. (2) another one is the Chain-of-Thought (CoT) prompting method by following (Kojima et al., 2022) and using the prompt "let's think step by step." Apart from these two settings, we also assess the LLMs' performance by using the few-shot setting to validate whether LLMs can improve their performance with input-output exemplars. We concatenate four tailored exemplars for desire and belief states and seven designed exemplars for intentions, covering all the input-output exemplars' labels. More configuration details can be found in Appendix B.2, and the prompt template refers to Appendix B.4. To measure the specific performance gap between humans and the state-of-the-art machine on the NegotiationToM, we employ three graduate students in computer science to complete the human evaluation task. More details of human evaluation are shown in Appendix C.2.

4.2 Metrics

We report the exact match percentages of all three high, medium, and low preferences for desire and belief classification, and only both of these three preferences that answer correctly count toward correct. The micro F1 score and macro F1 score are reported for the multi-label intention classification by following prior works (Hou et al., 2021; Wu et al., 2021; Moghe et al., 2023; Vulic et al., 2022). Moreover, we report the *"All"* score, which requires the models to answer correctly all three ToM question types, which include desire, belief, and intention for the same information piece in the conversation. Furthermore, we report the consistency score, which requires the models to answer ToM questions correctly for the whole negotiation conversation. This metric aims to measure how well the models show consistent understanding and track the agent's mental state change throughout the whole conversation.

[3]https://www.anthropic.com/news/introducing-claude
[4]https://www.anthropic.com/news/claude-2-1

5 Experimental Result

5.1 Main Result

Table 4 summarizes the main results of the state-of-the-art large language models in NegotiationToM, from which we derive the following conclusions. **First**, the performance of all models is significantly worse than human performance, even after employing the zero-shot chain-of-thought (CoT) method. There is a significant performance gap between machines and humans in the ToM negotiation evaluation scenario. Specifically, compared with the best state-of-the-art models in each mental state, the performance gap is 27.85% in desire, 32.96% in belief, 43.82% in Micro F1 score, and 48.98% in Macro F1 score in intention. **Second**, GPT-4 0613 (CoT) achieves the best performance among all the models regarding inferring desire and belief, while Claude-v2.1 (CoT) outperforms all other models in the intention classification task in NegotiationToM. **Third**, we observe that most models received an improvement in scores when the chain-of-thought (CoT) method was applied. Nevertheless, there are still significant score gaps compared to human performance.

Few-Shot Performance Although the few-shot setting is not a common practice, we also include it to see whether it performs better than the other two settings. The results illustrated in Table 4 display that the few-shot setting of most models gains an improvement over the zero-shot setting but is worse than the CoT prompting method in NegotiationToM. Interestingly, some models (e.g., GPT-4) with few-shot exemplars received a better performance than the CoT method in intention states.

5.2 Large Language Model on *All* score

To fully assess the ToM capability of large language models to understand other's mental states in each

Model	Desire	Belief	Intention		*All*	Consistency	
	Exact.Match.(%)	Exact.Match.(%)	Micro.F1(%)	Macro.F1(%)	Exact.Match(%)	Desire(%)	Belief(%)
LLaMa2-Chat(13B)	15.41	14.63	22.66	19.82	0.56	0.76	0.76
LLaMa2-Chat(13B) (CoT)	16.15	18.21	24.20	20.81	0.61	0.76	0.76
LLaMa2-Chat(13B) (Few-Shot)	13.49	12.54	26.30	21.76	0.64	0.90	0.80
LLaMa2-Chat(70B)	24.40	21.58	33.23	27.70	0.45	1.78	1.51
LLaMa2-Chat(70B) (CoT)	30.34	24.23	30.57	26.26	1.06	2.28	0.00
LLaMa2-Chat(70B) (Few-Shot)	26.95	22.84	35.77	28.10	1.28	1.32	0.91
Claude-v1.3	26.27	23.15	30.80	27.81	1.50	0.25	1.01
Claude-v1.3 (CoT)	44.63	37.18	31.12	28.25	1.62	4.81	1.52
Claude-v1.3 (Few-Shot)	30.73	30.68	32.35	30.10	1.80	3.23	1.20
Claude-v2.1	45.10	39.49	37.48	32.94	3.40	6.08	3.54
Claude-v2.1 (CoT)	50.13	40.52	**39.93**	**35.67**	**3.68**	6.07	4.05
Claude-v2.1 (Few-Shot)	48.77	41.88	38.23	34.32	2.90	6.25	4.28
ChatGPT 0613	18.60	13.04	33.95	29.73	0.43	0.00	0.00
ChatGPT 0613 (CoT)	28.45	21.00	36.71	30.79	0.78	0.76	0.25
ChatGPT 0613 (Few-Shot)	19.24	17.02	36.29	30.84	2.16	0.00	0.00
GPT-4 0613	62.77	57.62	29.84	27.15	2.58	13.67	10.63
GPT-4 0613 (CoT)	**63.29**	**58.18**	34.90	31.26	2.79	**17.72**	**14.18**
GPT-4 0613 (Few-Shot)	62.89	52.08	35.10	33.21	2.51	15.94	12.76
Human	91.14	91.14	83.75	84.65	43.78	75.44	75.44

Table 4: Main results of models for the NegotiationToM. The best results are **bold-faced**, and the second-best ones are underlined. The conversation consistency (%) of the models' responses for answering correctly in whole dialogues. All models received zero consistency scores in the intention aspect, as the intention mental state owned a multi-label in an utterance and imposed difficulties to generate exact match labels in the whole label.

round of dialogues, we report the *"All"* score in Table 4. This metric required the machine equipped with various ToM abilities to correctly answer three mental states (i.e., desire, belief, and intention) under the same information piece in the conversation. The Claude-v2.1 (CoT) outperforms all other large language models and receives a 3.68% in this *all* metric. It may be attributed to the exceptional intention ToM ability of Claude-v2.1 (CoT), but it also obtains a relatively high performance on the desire and belief aspects of ToM. However, it is worth mentioning that the performance of the machine on the *all* metric is far away from human performance, which is 43.78%.

5.3 How Well Large Language Model on Tracking Mental States Change in Conversation

To assess how well large language models can track the mental states of negotiation participants in conversations and coherent understanding of others' mental states with increased available information. Thus, it is crucial to evaluate the consistency and faithfulness of the large language model for the conversation context of the whole theory of mind-based dialogue. The consistency score is presented in Table 4, GPT-4(CoT) received an excellent performance on this metric compared with other models (e.g., Claude-v2.1 (CoT)), which are 17.72% in desire and 14.18% in belief. Nevertheless, there is a huge performance gap between machines and humans in this consistency metric, demonstrating

Model	Question Forms	
	Desire	Belief
ChatGPT 0613 (ranking form)	2.88	9.24
ChatGPT 0613 (individual form)	9.18	11.7
ChatGPT 0613 (combined form)	**18.60**	**13.04**
GPT-4 0613 (ranking form)	20.10	16.80
GPT-4 0613 (individual form)	40.01	36.88
GPT-4 0613 (combined form)	**62.77**	**57.62**

Table 5: The zero-shot performance of three question types. The intention task is ignored in this experiment as this task in NegotiationToM is a multi-label classification.

LLMs still lack of ability to track the mental state change during the conversation. It is noted that all models received zero consistency scores in the intention aspect, as the intention mental state owned a multi-label in an utterance and imposed difficulties to generate exact match labels in the whole conversation.

5.4 The Effect of Question Format

With the performance of LLMs varying significantly due to the sensitivity of prompt templates (Webson and Pavlick, 2022), we assessed the performance of two state-of-the-art models, ChatGPT and GPT-4, to study the effect of various question formats on their performance. In this experiment, we adopt three types of question formats, including ***ranking format, individual format, and combined format*** for desire and belief mental state. The ***combined format*** is the baseline prompt template adopted in our main experiment, which combines all three questions regarding the

Model	Intention	
	Micro.F1(%)	Macro.F1(%)
LLaMa2-Chat(13B)	13.13	10.53
LLaMa2-Chat(13B) (w B.D.)	18.61	15.44
LLaMa2-Chat(70B)	22.70	18.41
LLaMa2-Chat(70B) (w B.D.)	25.94	21.18
Claude-v1.3	21.77	18.52
Claude-v1.3 (w B.D.)	26.95	23.35
Claude-v2.1	27.12	24.48
Claude-v2.1 (w B.D.)	**34.56**	**30.02**
ChatGPT 0613	23.42	18.44
ChatGPT 0613 (w B.D.)	28.99	25.93
GPT-4 0613	26.31	23.86
GPT-4 0613 (w B.D.)	32.71	29.77

Table 6: Results of models for the CaSiNo strategy prediction. *w B.D.* indicate with the input with desire and belief. The best results are **bold-faced**, and the second-best ones are underlined.

preference order into a single question and asks the LLMs to answer it simultaneously. The ***ranking format*** indicates collecting high, medium, and low preference items as one ranking answer. The ***individual format*** splits the high, medium, and low preference questions into three questions and feeds them to LLMs individually. The combined, ranking, individual, question format prompt template is shown in Tables 9, 14, 15, and 16 in Appendix B.4.

The performance shown in Table 5 demonstrates that the question format indeed affects the LLMs' performance, and the combined format performs better than other formats. It may result from the combined format imposing the constraint for LLMs to avoid answering some unreasonable and implausible response. After the case study on error cases from the GPT-4 with individual form, we find that it is more challenging for models to combine with different types of reasoning while conducting the theory of mind reasoning. For example, when the GPT-4 may correctly answer that the agent's highest preference item is water and the lowest one is food, the model may randomly answer medium preference as there is no information of medium preference provided in conversation. Models cannot answer the medium preference of firewood (there are only three items) because they cannot effectively adopt deductive reasoning ability when performing theory of mind reasoning. Other models (e.g., ChatGPT) also suffer from this issue more seriously, although the combined format slightly mitigates this issue to some extent.

5.5 CaSiNo Negotiation Strategy Prediction

To validate the significance of our annotated desire and belief states, we append the information from these two states into the prompt template and assess whether it enhances the model performance on the negotiation strategy prediction task from

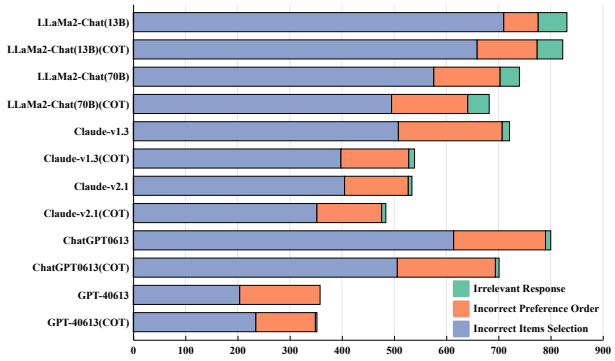

Figure 2: Model errors for desire state

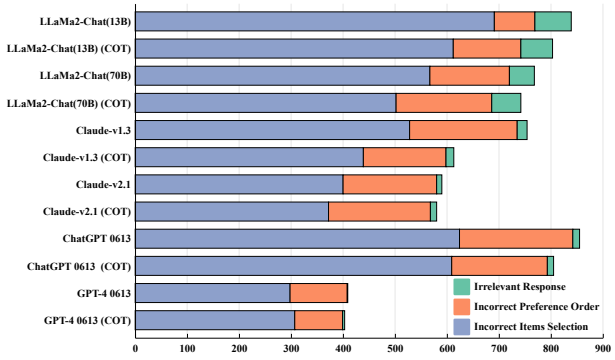

Figure 3: Model errors for belief state

the CaSiNo dataset. The baseline prompt template and the prompt template with belief and desire are shown in Tables 17 and 18 in Appendix B.4. The micro F1 score and macro F1 were employed as this task metric, and the result is reported in Table 6. All models incorporating the information from these two mental states received a significant improvement over the baselines. Specifically, by integrating the signals from desire and belief, the Claude-v2.1 model obtained a 7.44% gain in the micro F1 score and a 5.54% gain in the macro F1 score. It demonstrates that the effectiveness of our annotated theory of mind states (i.e., desire and belief) helps LLMs to infer the negotiation strategy behind each utterance. For example, with the understanding that agent 1's preference order of items is *Not Given, Not Given, and Not Given*, and agent 2's preference order of items is *Firewood, Not Given, and Not Given*. Agent 2 may take the elicit-preference strategy to elicit the preference order of agent 1 for further negotiation.

5.6 What Types of Error LLMs Make

To understand the type of error LLMs make on the NegotiationToM benchmark, we sampled 1,000 LLMs' responses and counted the error categories among them.

Types of Error LLMs make on Desire and Belief State Figures 2 and 3 summarize the error types

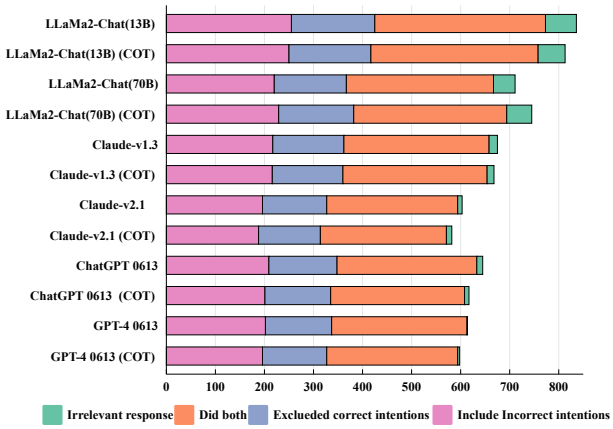

Figure 4: Model errors for intention state

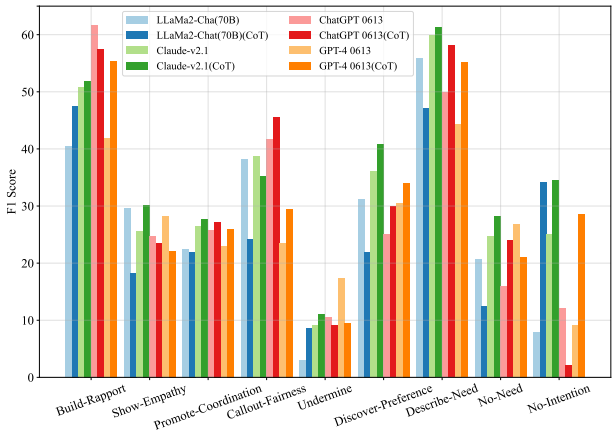

Figure 5: The label-wise intention dimension results of large language models in NegotiationToM. *undermine* stand for the undermine-requirements intention.

of desire and belief state for each model with and without CoT reasoning. All models make more errors by including incorrect items and excluding correct items (i.e., Incorrect Items Selection). For example, LLMs tend to select items (e.g., water) randomly rather than answer "Not Given" when there is insufficient information to determine the preferred items. With the CoT method adopted, this error will be decreased for most models as LLM conducted reasoning on the conversation context and tried to explain and respond to a reasonable answer.

Types of Error LLMs make on Intention State In terms of intention state, all models without and with CoT tend to select more intention choices, resulting in a high error rate in the "including incorrect intentions" and "did both" (i.e., include incorrect intentions and excluded correct intentions) error types. Another finding is that LLaMa2 series models respond to many irrelevant responses, such as repeating the questions, and do not raise any relevant answers.

Label-wise for Intentions State To further explore the LLMs' performance on nine intention subclasss in NegotiationToM, Figure 5 illustrates the F1 scores (%) for each subclass. The results indicate that LLMs exhibit strong performance in predicting Build-Rapport and Describe-Need intentions, while their performance in predicting "undermine-requirements" and "No-Intention" is poor. Notably, Claude-v2.1 (CoT) outperforms other models in more than half of the subclass in intentions, demonstrating its proficiency in inferring others' intentions. For a detailed subclass result covering all models, please refer to Figure 6 and Table 8 in Appendix C.1.

6 Conclusion

This work introduces NegotiationToM, a new benchmark designed to stress-test machine ToM in real-world negotiation surrounding covered multi-dimensional mental states. We performed comprehensive and detailed experiments to evaluate LLMs' capability on the NegotiationToM benchmark and discovered that LLMs exhibit inferior performance compared to humans in the NegotiationToM task.

Limitations

Passive benchmark to evaluate the ability of LLMs Although our benchmark is to stress-test machine ToM ability in negotiation surrounding compassed multi-categories of mental states, NegotiationToM, and existing prior benchmarks are passive benchmarks that primarily adopt a passive observer role to test language agents (Ma et al., 2023). These benchmarks passively assess the ToM ability of LLMs and lack active interaction with and engagement between the agent and other entities involved in the situated environment. The active ToM benchmark should treat the language model as an active agent that perceives the physical and social context, reasons about others' mental states, communicates with other agents, and interacts with the environment to complete predefined tasks. The future work of this paper will employ the language model to act as an agent to actively interact with other model agents by using the Belief-Desire-Intention agent modeling method to generate a rational negotiation strategy. For example, based on the information of desire, belief, and intention, the language model agent will actively acquire a negotiation strategy arguing more benefits or enhancing the cooperation.

Ethics Statement

In this work, we conformed to recognized privacy practices and rigorously followed the data usage policy. We declare that all authors of this paper acknowledge the *ACM Code of Ethics* and honor the code of conduct. This paper introduces a benchmark for stress-testing machine theory of mind of large language model on the negotiation surrounding. Our benchmark is modified from the CaSiNo (Chawla et al., 2021), an English-based negotiation dataset. They conducted a data post-processing step for filtering inappropriate language use (e.g., English swear words) dialogues although this situation rarely occurred in the negotiation process. Therefore, we can foresee no immediate social consequences or ethical issues as we do not introduce social/ethical bias into the model or amplify any bias from the data. Moreover, the license CaSiNo CC-BY-4.0 license allows us to modify the data for research, and this fulfills their intended use.

Acknowledgements

The authors of this paper were supported by the NSFC Fund (U20B2053) from the NSFC of China, the RIF (R6020-19 and R6021-20) and the GRF (16211520 and 16205322) from RGC of Hong Kong. We also thank the support from NVIDIA AI Technology Center (NVAITC).

References

Anthropic. 2023a. Introducing claude. *Anthropic*.

Anthropic. 2023b. Introducing claude 2.1. *Anthropic*.

Ian Apperly. 2010. *Mindreaders: the cognitive basis of" theory of mind"*. Psychology Press.

David Malet Armstrong. 1973. *Belief, truth and knowledge*. Cambridge University Press.

Jaan Aru, Aqeel Labash, Oriol Corcoll, and Raul Vicente. 2023. Mind the gap: challenges of deep learning approaches to theory of mind. *Artif. Intell. Rev.*, 56(9):9141–9156.

Lawrence M. Ausubel, Peter Cramton, and Raymond J. Deneckere. 2002. Chapter 50 bargaining with incomplete information. volume 3 of *Handbook of Game Theory with Economic Applications*, pages 1897–1945. Elsevier.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

Simon Baron-Cohen. 1997. *Mindblindness: An essay on autism and theory of mind*. MIT press.

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

Simon Baron-Cohen, Michelle O'riordan, Valerie Stone, Rosie Jones, and Kate Plaisted. 1999. Recognition of faux pas by normally developing children and children with asperger syndrome or high-functioning autism. *Journal of autism and developmental disorders*, 29:407–418.

Max H. Bazerman, Jared R. Curhan, Don A. Moore, and Kathleen L. Valley. 2000. Negotiation. *Annual Review of Psychology*, 51(1):279–314.

Cécile Belmondo and Caroline Sargis-Roussel. 2015. Negotiating language, meaning and intention: Strategy infrastructure as the outcome of using a strategy tool through transforming strategy objects. *British Journal of Management*, 26:S90–S104.

Michael Bratman. 1987. *Intention, plans, and practical reason*. The University of Chicago Press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712.

Fulvia Castelli. 2006. The valley task: Understanding intention from goal-directed motion in typical development and autism. *British journal of developmental psychology*, 24(4):655–668.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024. Exploring the potential of chatgpt on sentence level relations: A focus on temporal, causal, and discourse relations. In *Findings of the Association*

for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024, pages 684–721. Association for Computational Linguistics.

Chunkit Chan, Xin Liu, Tsz Ho Chan, Jiayang Cheng, Yangqiu Song, Ginny Y. Wong, and Simon See. 2023a. Self-consistent narrative prompts on abductive natural language inference. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 - 4, 2023*, pages 1040–1057. Association for Computational Linguistics.

Chunkit Chan, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Y. Wong, and Simon See. 2023b. Discoprompt: Path prediction prompt tuning for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 35–57. Association for Computational Linguistics.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale M. Lucas, Jonathan May, and Jonathan Gratch. 2021. Casino: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3167–3185. Association for Computational Linguistics.

Jiayang Cheng, Lin Qiu, Tsz Ho Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, Yue Zhang, and Zheng Zhang. 2023. Storyanalogy: Deriving story-level analogies from large language models to unlock analogical understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11518–11537. Association for Computational Linguistics.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

Vincent P Crawford. 1990. Explicit communication and bargaining outcome. *American Economic Review*, 80(2):213–219.

Harmen de Weerd, Rineke Verbrugge, and Bart Verheij. 2013. Higher-order theory of mind in negotiations under incomplete information. In *PRIMA 2013: Principles and Practice of Multi-Agent Systems - 16th International Conference, Dunedin, New Zealand, December 1-6, 2013. Proceedings*, volume 8291 of *Lecture Notes in Computer Science*, pages 101–116. Springer.

Harmen de Weerd, Rineke Verbrugge, and Bart Verheij. 2017. Negotiating with other minds: the role of recursive theory of mind in negotiation with incomplete information. *Auton. Agents Multi Agent Syst.*, 31(2):250–287.

Zheye Deng, Chunkit Chan, Weiqi Wang, Yuxi Sun, Wei Fan, Tianshi Zheng, Yauwai Yim, and Yangqiu Song. 2024. Text-tuple-table: Towards information integration in text-to-table generation via global tuple extraction. *CoRR*, abs/2404.14215.

Daniel C Dennett. 1988. Précis of the intentional stance. *Behavioral and brain sciences*, 11(3):495–505.

Anthony Favier, Shashank Shekhar, and Rachid Alami. 2023. Models and algorithms for human-aware task planning with integrated theory of mind. In *32nd IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2023, Busan, Republic of Korea, August 28-31, 2023*, pages 1279–1286. IEEE.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. 2023. Mathematical capabilities of chatgpt. *CoRR*, abs/2301.13867.

Uta Frith. 1994. Autism and theory of mind in everyday life. *Social development*, 3(2):108–124.

MY Ganaie and Hafiz Mudasir. 2015. A study of social intelligence & academic achievement of college students of district srinagar, j&k, india. *Journal of American Science*, 11(3):23–27.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2023. Understanding social reasoning in language models with language models. *CoRR*, abs/2306.15448.

Howard Gardner, Robert M Hanson, and Steve Hamilton. 1995. *How Are Kids SMART?: Multiple Intelligences in the Classroom*. National Professional Resources, Incorporated.

Shahriar Golchin and Mihai Surdeanu. 2024. Time travel in llms: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Alison Gopnik and Henry M Wellman. 1992. Why the child's theory of mind really is a theory.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Workshop on AKBC@CIKM*, pages 25–30.

Erin Grant, Aida Nematzadeh, and Thomas L. Griffiths. 2017. How can memory-augmented neural networks pass a false-belief task? In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society, CogSci 2017, London, UK, 16-29 July 2017*. cognitivesciencesociety.org.

David Gunning. 2018. Machine common sense concept paper. *CoRR*, abs/1810.07528.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343, Brussels, Belgium. Association for Computational Linguistics.

Mourad Heddaya, Solomon Dworkin, Chenhao Tan, Rob Voigt, and Alexander Zentefis. 2023. Language of bargaining. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13161–13185. Association for Computational Linguistics.

Yutai Hou, Yongkui Lai, Yushan Wu, Wanxiang Che, and Ting Liu. 2021. Few-shot learning for multi-label intent detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13036–13044. AAAI Press.

Martha Jeong, Julia Minson, Michael Yeomans, and Francesca Gino. 2019. Communicating with warmth in distributive negotiations is surprisingly counterproductive. *Management Science*, 65(12):5813–5837.

Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. 2023. Lion: Adversarial distillation of proprietary large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3134–3154. Association for Computational Linguistics.

Cheng Jiayang, Lin Qiu, Chunkit Chan, Xin Liu, Yangqiu Song, and Zheng Zhang. 2024. Eventground: Narrative reasoning by grounding to eventuality-centric knowledge graphs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 6622–6642. ELRA and ICCL.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 14397–14413. Association for Computational Linguistics.

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

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *CoRR*, abs/2302.02083.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019a. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 5871–5876. Association for Computational Linguistics.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019b. 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, Hong Kong, China. Association for Computational Linguistics.

Alice J Lee and Daniel R Ames. 2017. "i can't pay more" versus "it's not worth more": Divergent effects of constraint and disparagement rationales in negotiations. *Organizational Behavior and Human Decision Processes*, 141:16–28.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, Copenhagen, Denmark. Association for Computational Linguistics.

Changmao Li and Jeffrey Flanigan. 2024. Task contamination: Language models may not be few-shot anymore. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18471–18480. AAAI Press.

Haoran Li, Yulin Chen, Jinglong Luo, Yan Kang, Xiaojin Zhang, Qi Hu, Chunkit Chan, and Yangqiu Song. 2023. Privacy in large language models: Attacks, defenses and future directions. *CoRR*, abs/2310.10383.

Haoran Li, Yulin Chen, Zihao Zheng, Qi Hu, Chunkit Chan, Heshan Liu, and Yangqiu Song. 2024a. Backdoor removal for generative large language models. *CoRR*, abs/2405.07667.

Haoran Li, Dadi Guo, Donghao Li, Wei Fan, Qi Hu, Xin Liu, Chunkit Chan, Duanyi Yao, Yuan Yao, and Yangqiu Song. 2024b. Privlm-bench: A multi-level privacy evaluation benchmark for language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 54–73. Association for Computational Linguistics.

Zizheng Lin, Chunkit Chan, Yangqiu Song, and Xin Liu. 2024. Constrained reasoning chains for enhancing theory-of-mind in large language models.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. *CoRR*, abs/2302.00539.

Ruotian Ma, Xin Zhou, Tao Gui, Yiding Tan, Linyang Li, Qi Zhang, and Xuanjing Huang. 2022. Template-free prompt tuning for few-shot NER. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 5721–5732. Association for Computational Linguistics.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards A holistic landscape of situated theory of mind in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1011–1031. Association for Computational Linguistics.

Bertram F Malle and Joshua Knobe. 2001. The distinction between desire and intention: A folk-conceptual analysis. *Intentions and intentionality: Foundations of social cognition*, 45:67.

Nikita Moghe, Evgeniia Razumovskaia, Liane Guillou, Ivan Vulic, Anna Korhonen, and Alexandra Birch. 2023. Multi3nlu++: A multilingual, multi-intent, multi-domain dataset for natural language understanding in task-oriented dialogue. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3732–3755. Association for Computational Linguistics.

Louis J Moses. 1993. Young children's understanding of belief constraints on intention. *Cognitive Development*, 8(1):1–25.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Thomas L. Griffiths. 2018. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2392–2400. Association for Computational Linguistics.

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

TB OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Gonçalo Duarte Garcia Pereira, Rui Prada, and Pedro Alexandre Santos. 2016. Integrating social power into the decision-making of cognitive agents. *Artif. Intell.*, 241:1–44.

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

Dean G Pruitt. 2013. *Negotiation behavior*. Academic Press.

Joshua Robinson and David Wingate. 2023. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Alvin E Roth. 2020. Bargaining experiments. In *The Handbook of Experimental Economics*, pages 253–348. Princeton University Press.

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.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3762–3780. Association for Computational Linguistics.

Michael F Schober. 2005. Conceptual alignment in conversation. *Other minds: How humans bridge the divide between self and others*, pages 239–252.

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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13960–13980. Association for Computational Linguistics.

John R Searle. 1983. *Intentionality: An essay in the philosophy of mind*. Cambridge university press.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023a. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *CoRR*, abs/2305.14763.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 2257–2273. Association for Computational Linguistics.

Natalie Shapira, Guy Zwirn, and Yoav Goldberg. 2023b. How well do large language models perform on faux pas tests? In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10438–10451. Association for Computational Linguistics.

Valerie E Stone. 2006. Theory of mind and the evolution of social intelligence. *Social neuroscience: People thinking about thinking people*, pages 103–129.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, pages 1–11.

Teo Susnjak. 2022. Chatgpt: The end of online exam integrity? *CoRR*, abs/2212.09292.

Roderick I. Swaab, William W. Maddux, and Marwan Sinaceur. 2011. Early words that work: When and how virtual linguistic mimicry facilitates negotiation outcomes. *Journal of Experimental Social Psychology*, 47(3):616–621.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Elliot Turiel. 2008. The development of morality. *Child and adolescent development: An advanced course*, pages 473–514.

Tomer D. Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *CoRR*, abs/2302.08399.

Ivan Vulic, Iñigo Casanueva, Georgios Spithourakis, Avishek Mondal, Tsung-Hsien Wen, and Pawel Budzianowski. 2022. Multi-label intent detection via contrastive task specialization of sentence encoders. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7544–7559. Association for Computational Linguistics.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Jiayang Cheng, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *CoRR*, abs/2310.07521.

Weiqi Wang, Tianqing Fang, Chunyang Li, Haochen Shi, Wenxuan Ding, Baixuan Xu, Zhaowei Wang, Jiaxin Bai, Xin Liu, Cheng Jiayang, Chunkit Chan, and Yangqiu Song. 2024. CANDLE: iterative conceptualization and instantiation distillation from large language models for commonsense reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 2351–2374. Association for Computational Linguistics.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States,*

July 10-15, 2022, pages 2300–2344. Association for Computational Linguistics.

Oliver Whang. 2023. Can a machine know that we know what it knows. *The New York Times*.

Jessica Williams, Stephen M Fiore, and Florian Jentsch. 2022. Supporting artificial social intelligence with theory of mind. *Frontiers in artificial intelligence*, 5:750763.

Ting-Wei Wu, Ruolin Su, and Biing-Hwang Juang. 2021. A label-aware BERT attention network for zero-shot multi-intent detection in spoken language understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 4884–4896. Association for Computational Linguistics.

Runzhe Yang, Jingxiao Chen, and Karthik Narasimhan. 2021. Improving dialog systems for negotiation with personality modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 681–693. Association for Computational Linguistics.

Yauwai Yim, Chunkit Chan, Tianyu Shi, Zheye Deng, Wei Fan, Tianshi Zheng, and Yangqiu Song. 2024. Evaluating and enhancing llms agent based on theory of mind in guandan: A multi-player cooperative game under imperfect information. *CoRR*, abs/2408.02559.

Haolan Zhan, Yufei Wang, Tao Feng, Yuncheng Hua, Suraj Sharma, Zhuang Li, Lizhen Qu, and Gholamreza Haffari. 2022. Let's negotiate! a survey of negotiation dialogue systems. Working paper. California State University, Northridge, CA.

Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R. McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, Shyam Upadhyay, and Manaal Faruqui. 2023. How far are large language models from agents with theory-of-mind? *CoRR*, abs/2310.03051.

A Appendix for NegotiationToM

A.1 Verification of Potential Contamination in CaSiNo

Most of the existing available benchmarks in the NLP field were released prior to the initiation of the LLM training process, indicating that these datasets are likely to have been utilized during the pre-training phase and post-training phase (i.e., SFT (Ouyang et al., 2022) or RLHF (Christiano et al., 2017)) of LLMs (Golchin and Surdeanu, 2024; Li and Flanigan, 2024). Therefore, we follow the method proposed by Golchin and Surdeanu (2024) to assess the potential contamination issues in the CaSiNo dataset. We sample 100 instances from the CaSiNo dataset and prompting the ChatGPT and GPT-4 to generate the likely next dialogue. After human validation of these 100 instances, none of these generated dialogues corresponded to the original dataset. The prompting template and outputs are illustrated in Table 7. Furthermore, we also test these 100 instances by using the prompt template "Give a CaSiNO negotiation dialogue example and its answer for negotiation strategy." The outcome aligns with the prompting template depicted in Table 7, demonstrating that the risk of potential contamination is mitigated; however, no systematic approach can effectively address the contamination issue unless all training datasets utilized for LLMs are made publicly available.

A.2 NegotiationToM Annotation

In this section, we showcase our annotation instructions and templates used for annotation. The instructions are used to introduce the background of the negotiation conversation and instruct the workers to perform the annotation based on the dialogue history. The annotation instruction and annotation template are presented in Figure 7, 8, 9, and 10. The detailed statistics and comparison with contemporary ToM datasets are shown in Table 19.

A.3 Details for NegotiationToM Intentions

We provide a brief overview of Table 20, which presents the definitions of various ToM intentions in negotiation strategies. These definitions help us understand the intentions of the agents involved in the negotiation process. The table offers definitions for strategies such as Small-Talk, Empathy, Coordination, No-Need, Elicit-Pref, Undervalue-Partner, Vouch-Fairness, Self-Need, Other-Need, and Non-strategic. Each definition explains the

specific meaning and context of the respective strategy in the negotiation process. By understanding these strategy definitions, we can better comprehend the negotiation interactions between agents and how the intention relates to desire and belief states during the negotiation process.

NegotiationToM Arguments Table 22 shows arguments for various items (i.e., Food, Water, Firewood) in four categories: Personal Care, Recreational, Group Needs, and Emergency. For example, participants may need more food for larger-sized teenage children, and more water for hydration or emergencies. The diversity of negotiation arguments raised by the human participants provided various scenarios for stress-testing LLMs by avoiding shortcuts and spurious correlation issues.

B Appendix for Experiments

B.1 Baseline models

In this section, we introduce six recent instruction-tuned large language models employed to stress-test the Negotiation. GPT models from OpenAI use a decoder-only transformer framework, and GPT-4 (OpenAI, 2023), ChatGPT (OpenAI, 2022) are proprietary models tested by calling the model API. Claude-v1.3 (Anthropic, 2023a) and Claude-v2.1 (Anthropic, 2023b) are closed-source LLMs developed by Anthropic. These two models can be accessible through a chat interface and API, and they demonstrate a strong performance in a lot of NLP tasks. Llama-2 Chat 13B (Touvron et al., 2023), and Llama-2 Chat 70B (Touvron et al., 2023) are a language model fine-tuned for engaging in dialogues that follow user inputs.

B.2 Hyperparameter

We use default hyperparameters for all the large language models mentioned in this paper. For ChatGPT (`gpt-3.5-turbo-0613`) and GPT-4 (`gpt-4-0613`), the default parameters[5] are `temperature`=1 and `top_p`=1. For LLaMa2-Chat(13B) and LLaMa2-Chat(70B) models, we follow the default setting where `temperature`=0.5, `top_p`=0.9. For Claude-v1.3 and Claude-v2.1, the default parameters are `temperature`=0.5, `top_p`=1.

[5]https://platform.openai.com/docs/api-reference/chat/create

B.3 Few-Shot Experimental setting

Following Brown et al. (2020), we use a few-shot prompt to instruct all models:

<TASK-PROMPT>
<EX$_1$-INP><EX$_1$-OUT>
...
<EX$_{N-1}$-INP><EX$_{N-1}$-OUT>
<EX$_N$-INP>

where <TASK-PROMPT> is a task instruction that explains the background of the negotiation scenario and <EX$_1$-INP><EX$_1$-OUT> are human-authored examples that try to cover all subclass labels to help LLMs understand the subclass labels. Finally, we provide the N_{th} input as <EX$_N$-INP> and ask all models to generate the corresponding answer as <EX$_N$-OUT>. In this paper, we set $N = 4$ for the desire and belief states, and $N = 7$ for the intention state.

B.4 Appendix for Prompt Template

In this section, we introduce all prompt template and these template are presented in Tables 9, 10, 11, 12, 13, 14, 15, 16, 17, and 18.

Main Result Template Table 9 introduces the baseline prompt template, a straightforward Q&A session without requiring detailed and complex explanations. Table 10 is the Chain-of-Thought prompt template that facilitates a detailed, step-by-step thought process in the LLM, thoroughly considering each problem aspect before arriving at a solution. Tables 11, 12, and 13 are presented the few-shot setting prompt template. Each prompt template appends some exemplars for LLMs to learn the new tasks.

Question Format Template The Ranking Question Format in Table 14 asks the LLM to prioritize or rank items, which is useful for tasks needing decision-making based on preference or importance. Table 15 and 16, the individual question format prompt templates, break down desire and belief questions into separate and focused questions.

Strategy Prediction Template Tables 17 and 18 present the prompt template for the strategy prediction task; the former provides a baseline format for predicting strategies from negotiation conversations, while the latter enriches this task by incorporating the desires and beliefs of the agents involved, offering a deeper contextual understanding and of-

fer signals from desire and belief states for strategy classification.

B.5 Related Works for Large language Models

Recent studies have extensively and comprehensively evaluated instruction following LLMs' (OpenAI, 2023, 2022; Taori et al., 2023; Jiang et al., 2023) performance on numerous tasks, revealing its superior performance in zero-shot scenarios compared to other models (Bubeck et al., 2023; Bang et al., 2023; Chan et al., 2024; Cheng et al., 2023; Wang et al., 2024; Jiayang et al., 2024). However, there are certain obstacles persist unaddressed, such as the inability to perform complex mathematical reasoning (Frieder et al., 2023), theory of mind reasoning (Lin et al., 2024), analogies reasoning (Cheng et al., 2023), text-to-table generation (Deng et al., 2024), fact validation (Wang et al., 2023), complex game setting (Yim et al., 2024), associated ethical implications, and privacy concerns (Li et al., 2023; Susnjak, 2022; Li et al., 2024b; Lukas et al., 2023; Li et al., 2024a). Therefore, it is critical to discuss whether large language models possess the capacity of the theory of mind as humans do. In this paper, we test several instruction-following LLMs in our NegotiationToM benchmark. The zero-shot performance of large language models, which relies on the sophisticated design of templates, has shown variance across various tasks (Ma et al., 2022; Chan et al., 2023a,b). To obtain replicable and representative results, we follow Robinson and Wingate (2023) to formulate the task as a multiple choice question-answering problem in this work.

C Appendix for Experimental Result

C.1 Appendix for Intention Subclass Result

In this section, Table 6 and Figure 8 provide more detailed performance of the intention subclass level. All models receive a good performance on the Build-Rapport and Describe-Need subclass while poor at the No-Intention, No-Need, and Undermine-Requirements subclass.

C.2 Appendix for Human Performance

To measure the specific performance gap between humans and the state-of-the-art machine on the NegotiationToM, we employ three graduate students in computer science to complete the human evaluation task. Each instance is shown to three workers who were prompted to choose the more plausible choice for questions regarding desire, belief, and intention. The questions and instructions are the same as the annotation template and are displayed in Appendix A.2. We compute the majority vote on the labels assigned, which leads to a performance of 91.14% on both the desire and belief, while the intention only obtained an 83.75% micro F1 score and 84.65% macro F1 score. The *"All"* score received 43.78%, and conversation consistency acquired 75.44% in both the desire and belief states.

Models	Example
ChatGPT	**Instruction:** You are provided with historical dialogue from the CaSiNO dataset. Finish Next dialogue as appeared in the dataset. Next dialogue must exactly match the instance in the dataset. **agent 1:** Hello! Let's work together on a deal for these packages, shall we? What are you most interested in? **agent 2:** Hey! I'd like some more firewood to keep my doggo warm. What do you need? **agent 1:** I need firewood as well. We have a large group consisting of mostly senior citizens, including my grandma, so we'd like the firewood to keep everyone warm. **agent 2:** I see. What are you least interested in? **agent 1:** We are least interested in the hiking equipment. Our group is not very active and prefers to stay indoors.
GPT-4	**Instruction:** You are provided with historical dialogue from the CaSiNO dataset. Finish Next dialogue as appeared in the dataset. Next dialogue must exactly match the instance in the dataset. **agent 1:** Hello! Let's work together on a deal for these packages, shall we? What are you most interested in? **agent 2:** Hey! I'd like some more firewood to keep my doggo warm. What do you need? **agent 1:** I need firewood as well. We have a large group consisting of mostly senior citizens, including my grandma, so we'd like the firewood to keep everyone warm. **agent 2:** I see. What are you least interested in? **agent 1:** I'm least interested in the food supplies. We have plenty of those.
Ground-Truth	**Instruction:** You are provided with historical dialogue from the CaSiNO dataset. Finish Next dialogue as appeared in the dataset. Next dialogue must exactly match the instance in the dataset. **agent 1:** Hello! Let's work together on a deal for these packages, shall we? What are you most interested in? **agent 2:** Hey! I'd like some more firewood to keep my doggo warm. What do you need? **agent 1:** I need firewood as well. We have a large group consisting of mostly senior citizens, including my grandma, so we'd like the firewood to keep everyone warm. **agent 2:** I see. What are you least interested in? **agent 1: We can make do without extra water. Can we trade two waters for an extra firewood package and an extra food package?.**

Table 7: Prompting template and outcome for verification of potential contamination in CaSiNo dataset. The highlighted sentences are generated by the LLMs.

Models	B-R	S-E	P-C	C-F	U-R	D-P	D-N	N-N	N-I
LLaMa2-Chat(13B)	24.34	15.86	21.58	32.28	2.85	26.35	33.01	8.68	13.41
LLaMa2-Chat(13B) (CoT)	38.54	12.66	21.33	20.02	4.09	18.26	30.89	6.84	34.64
LLaMa2-Chat(70B)	40.56	29.60	22.44	38.19	2.99	31.19	55.89	20.62	7.85
LLaMa2-Chat(70B) (CoT)	47.46	18.22	22.00	24.26	8.70	21.98	47.18	12.40	34.14
Claude-v1.3	51.27	23.55	25.26	35.44	16.06	27.65	44.79	24.13	2.14
Claude-v1.3 (CoT)	49.31	16.66	24.44	26.63	7.73	24.48	50.74	14.10	**40.16**
Claude-v2.1	50.73	25.53	26.42	38.76	9.12	36.11	59.93	24.80	25.08
Claude-v2.1 (CoT)	51.86	**30.15**	**27.68**	35.26	11.11	**40.91**	**61.24**	**28.26**	34.57
ChatGPT 0613	**61.72**	24.69	25.73	41.71	10.61	25.07	49.87	15.99	12.20
ChatGPT 0613 (CoT)	57.46	23.47	27.22	**45.53**	9.20	29.92	58.13	23.97	2.16
GPT-4 0613	41.92	28.16	22.89	23.45	**17.32**	30.42	44.30	26.77	9.12
GPT-4 0613 (CoT)	55.35	22.11	26.04	29.43	9.50	34.03	55.15	21.09	28.64

Table 8: Label-wise results of all models for the intention dimension. *w B.D.* indicate with the input with desire and belief. *B-R, S-E , P-C, C-F, U-R, D-P, D-N, N-N, N-I* stands for Build-Rapport, Show-Empathy, Promote-Coordination, Callout-Fairness, Undermine-Requirements, Discover-Preference, Describe-Need, No-Need, No-Intention. The best results are **bold-faced**, and the second-best ones are underlined.

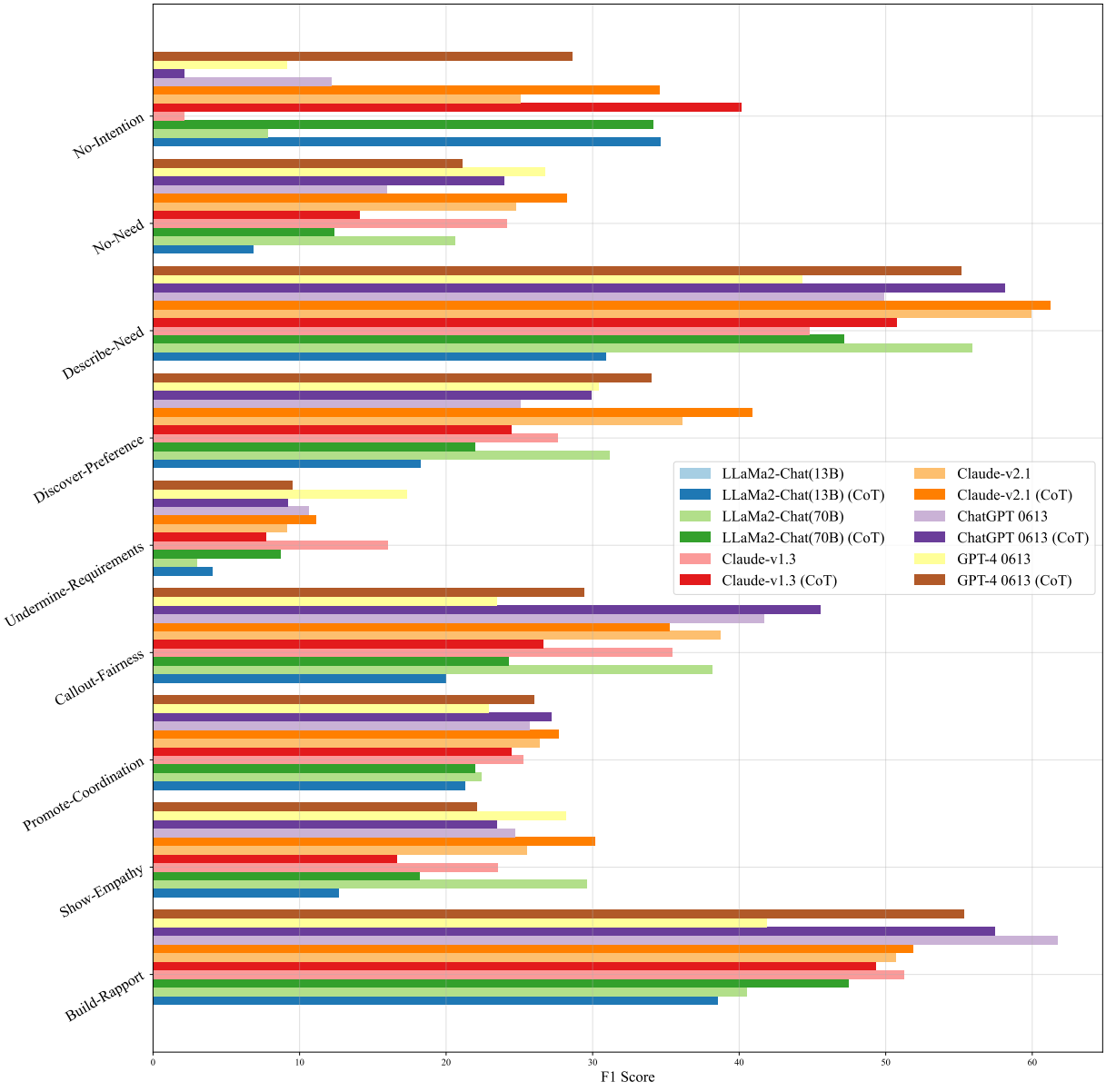

Figure 6: The label-wise intention dimension results of large language models in NegotiationToM.

Dimension	Example
Belief	Background: Here is a negotiation conversation for a camping trip. There are two agents who own some basic supplies and negotiate with each other to split the additional food packages, water bottles, and firewood to make their camping trip even better. Each of these items will be of either High, Medium or Low priority for these two agents. Each of the additional items only has an available quantity of 3. Please answer the following three questions using "A", "B", "C", "D" without any explanation. Dialogue History: agent 1: Hello! Let's work together on a deal for these packages, shall we? What are you most interested in? agent 2: Hey! I'd like some more firewood to keep my doggo warm. What do you need? Question1: Based on the dialogue, what is the high preference for items Agent 1 thinks Agent 2 is? A.Not given B.Water C.Food D.Firewood Question2: Based on the dialogue, what is the medium preference for items Agent 1 thinks Agent 2 is? A.Not given B.Water C.Food D.Firewood Question3: Based on the dialogue, what is the low preference for items Agent 1 thinks Agent 2 is? A.Not given B.Water C.Food D.Firewood Answer:
Desire	Background: Here is a negotiation conversation for a camping trip. There are two agents who own some basic supplies and negotiate with each other to split the additional food packages, water bottles, and firewood to make their camping trip even better. Each of these items will be of either High, Medium or Low priority for these two agents. Each of the additional items only has an available quantity of 3. Please answer the following three questions using "A", "B", "C", "D" without any explanation. Dialogue History: agent 1: Hello! Let's work together on a deal for these packages, shall we? What are you most interested in? agent 2: Hey! I'd like some more firewood to keep my doggo warm. What do you need? Question1: What is agent 1's high preference for items based on the dialogue history? A.Not given B.Water C.Food D.Firewood Question2: What is agent 1's medium preference for items based on the dialogue history? A.Not given B.Water C.Food D.Firewood Question3: What is agent 1's low preference for items based on the dialogue history? A.Not given B.Water C.Food D.Firewood Answer:
Intention	Background: Here is a negotiation conversation for a camping trip. There are two agents who own some basic supplies and negotiate with each other to split the additional food packages, water bottles, and firewood to make their camping trip even better. Each of these items will be of either High, Medium or Low priority for these two agents. Each of the additional items only has an available quantity of 3. Dialogue History: agent 1: Hello! Let's work together on a deal for these packages, shall we? What are you most interested in? agent 2: Hey! I'd like some more firewood to keep my doggo warm. What do you need? Question: What are the plausible intentions of Agent 1 expressed in 'Hello! Let's work together on a deal for these packages, shall we? What are you most interested in?' Based on the dialogue history, select one or more intentions (i.e.,"A", "B", "C",...,"I") from the following choices without any explanation. A.Intents to build a rapport with the opponent B.Intents to show empathy with the opponent C.Intents to promote coordination with the opponent D.Intents to callout to fairness E.Intents to undermine the requirements of the opponent F.Intents to discover the preference order of the opponent G.Intents to describe a need for an item H.Intents to point out they do not need an item I.No clear intention in the utterance Answer:

Table 9: Baseline prompt template.

Dimension	Example
Belief	Background: Here is a negotiation conversation for a camping trip. There are two agents who own some basic supplies and negotiate with each other to split the additional food packages, water bottles, and firewood to make their camping trip even better. Each of these items will be of either High, Medium or Low priority for these two agents. Each of the additional items only has an available quantity of 3. Please answer the following three questions using "A", "B", "C", "D". Dialogue History: agent 1: Hello! Let's work together on a deal for these packages, shall we? What are you most interested in? agent 2: Hey! I'd like some more firewood to keep my doggo warm. What do you need? Question1: Based on the dialogue, what is the high preference for items Agent 1 thinks Agent 2 is? A.Not given B.Water C.Food D.Firewood Question2: Based on the dialogue, what is the medium preference for items Agent 1 thinks Agent 2 is? A.Not given B.Water C.Food D.Firewood Question3: Based on the dialogue, what is the low preference for items Agent 1 thinks Agent 2 is? A.Not given B.Water C.Food D.Firewood Answer: Let's think step by step.
Desire	Background: Here is a negotiation conversation for a camping trip. There are two agents who own some basic supplies and negotiate with each other to split the additional food packages, water bottles, and firewood to make their camping trip even better. Each of these items will be of either High, Medium or Low priority for these two agents. Each of the additional items only has an available quantity of 3. Please answer the following three questions using "A", "B", "C", "D". Dialogue History: agent 1: Hello! Let's work together on a deal for these packages, shall we? What are you most interested in? agent 2: Hey! I'd like some more firewood to keep my doggo warm. What do you need? Question1: What is agent 1's high preference for items based on the dialogue history? A.Not given B.Water C.Food D.Firewood Question2: What is agent 1's medium preference for items based on the dialogue history? A.Not given B.Water C.Food D.Firewood Question3: What is agent 1's low preference for items based on the dialogue history? A.Not given B.Water C.Food D.Firewood Answer: Let's think step by step.
Intention	Background: Here is a negotiation conversation for a camping trip. There are two agents who own some basic supplies and negotiate with each other to split the additional food packages, water bottles, and firewood to make their camping trip even better. Each of these items will be of either High, Medium or Low priority for these two agents. Each of the additional items only has an available quantity of 3. Dialogue History: agent 1: Hello! Let's work together on a deal for these packages, shall we? What are you most interested in? agent 2: Hey! I'd like some more firewood to keep my doggo warm. What do you need? Question: What are the plausible intentions of Agent 1 expressed in 'Hello! Let's work together on a deal for these packages, shall we? What are you most interested in?' Based on the dialogue history, select one or more intentions (i.e.,"A", "B", "C",...,"I") from the following choices. A.Intents to build a rapport with the opponent B.Intents to show empathy with the opponent C.Intents to promote coordination with the opponent D.Intents to callout to fairness E.Intents to undermine the requirements of the opponent F.Intents to discover the preference order of the opponent G.Intents to describe a need for an item H.Intents to point out they do not need an item I.No clear intention in the utterance Answer: Let's think step by step.

Table 10: Chain-of-Thought prompt template.

Dimension	Example
Desire	Background: Here is a negotiation conversation for a camping trip. There are two agents who own some basic supplies and negotiate with each other to split the additional food packages, water bottles, and firewood to make their camping trip even better. Each of these items will be of either High, Medium or Low priority for these two agents. Each of the additional items only has an available quantity of 3. Please answer the following three questions using "A", "B", "C", "D" without any explanation. Dialogue History: agent 1: Hello! Which item do you need the most? agent 2: I would love to have the Firewood the most. Question1: What is agent 1's high preference for items based on the dialogue history? A.Not given B.Water C.Food D.Firewood Question2: What is agent 1's medium preference for items based on the dialogue history? A.Not given B.Water C.Food D.Firewood Question3: What is agent 1's low preference for items based on the dialogue history? A.Not given B.Water C.Food D.Firewood Answer: A,A,A Dialogue History: agent 1: Hello! Which item do you need the most? agent 2: I would love to have the Firewood the most. agent 1: Unfortunately I need firewood the most too. How about I take 2 firewood, 2 food, and 1 water? agent 2: I feel that I am not getting a fair deal. Question1: What is agent 1's high preference for items based on the dialogue history? A.Not given B.Water C.Food D.Firewood Question2: What is agent 1's medium preference for items based on the dialogue history? A.Not given B.Water C.Food D.Firewood Question3: What is agent 1's low preference for items based on the dialogue history? A.Not given B.Water C.Food D.Firewood Answer: D,C,B ... Dialogue History: agent 1: Hello! Let's work together on a deal for these packages, shall we? What are you most interested in? agent 2: Hey! I'd like some more firewood to keep my doggo warm. What do you need? Question1: What is agent 1's high preference for items based on the dialogue history? A.Not given B.Water C.Food D.Firewood Question2: What is agent 1's medium preference for items based on the dialogue history? A.Not given B.Water C.Food D.Firewood Question3: What is agent 1's low preference for items based on the dialogue history? A.Not given B.Water C.Food D.Firewood Answer:

Table 11: Few-shot prompt template for desire state.

Dimension	Example
Desire	Background: Here is a negotiation conversation for a camping trip. There are two agents who own some basic supplies and negotiate with each other to split the additional food packages, water bottles, and firewood to make their camping trip even better. Each of these items will be of either High, Medium or Low priority for these two agents. Each of the additional items only has an available quantity of 3. Please answer the following three questions using "A", "B", "C", "D" without any explanation. Dialogue History: agent 1: Hello! Which item do you need the most? agent 2: I would love to have the Firewood the most. Question1: Based on the dialogue, what is the high preference for items Agent 1 thinks Agent 2 is? A.Not given B.Water C.Food D.Firewood Question2: Based on the dialogue, what is the medium preference for items Agent 1 thinks Agent 2 is? A.Not given B.Water C.Food D.Firewood Question3: Based on the dialogue, what is the low preference for items Agent 1 thinks Agent 2 is? A.Not given B.Water C.Food D.Firewood Answer: D,A,A Dialogue History: agent 1: Hello! Which item do you need the most? agent 2: I would love to have the Firewood the most. agent 1: Unfortunately I need firewood the most too. How about I take 2 firewood, 2 food, and 1 water? agent 2: I feel that I am not getting a fair deal. agent 1: Then what do you think is fair? agent 2: I think that I should get 2 firewood, 1 food and 2 water Question1: Based on the dialogue, what is the high preference for items Agent 1 thinks Agent 2 is? A.Not given B.Water C.Food D.Firewood Question2: Based on the dialogue, what is the medium preference for items Agent 1 thinks Agent 2 is? A.Not given B.Water C.Food D.Firewood Question3: Based on the dialogue, what is the low preference for items Agent 1 thinks Agent 2 is? A.Not given B.Water C.Food D.Firewood Answer: D,B,C ... Dialogue History: agent 1: Hello! Let's work together on a deal for these packages, shall we? What are you most interested in? agent 2: Hey! I'd like some more firewood to keep my doggo warm. What do you need? Question1: Based on the dialogue, what is the high preference for items Agent 1 thinks Agent 2 is? A.Not given B.Water C.Food D.Firewood Question2: Based on the dialogue, what is the medium preference for items Agent 1 thinks Agent 2 is? A.Not given B.Water C.Food D.Firewood Question3: Based on the dialogue, what is the low preference for items Agent 1 thinks Agent 2 is? A.Not given B.Water C.Food D.Firewood Answer:

Table 12: Few-shot prompt template for belief state.

Dimension	Example
Desire	Background: Here is a negotiation conversation for a camping trip. There are two agents who own some basic supplies and negotiate with each other to split the additional food packages, water bottles, and firewood to make their camping trip even better. Each of these items will be of either High, Medium or Low priority for these two agents. Each of the additional items only has an available quantity of 3. Dialogue History: agent 2: Hi! I'm super excited to go camping with my family as a great way to vacation due to Covid19. My kid is so restless from being cooped up in the house all the time. Are you planning on going camping too? agent 1: I am! It is the perfect way to get away and still manage to social distance! I am worried about having enough water though, are you short on any supplies? agent 2: I think I'm good. I'm not 100% sure. My husband likes to do adventures on the fly. He got these water filter straw thingies from Amazon and said that if we run out of the water I packed, that we can drink the water in the lake but I don't really trust the straws. agent 1: Sounds like you need water too. How about you take 2 water, 1 food, and 1 firewood? Question: What are the plausible intentions of Agent 2 expressed in 'I think I'm good. I'm not 100% sure. My husband likes to do adventures on the fly. He got these water filter straw thingies from Amazon and said that if we run out of the water I packed, that we can drink the water in the lake but I don't really trust the straws.' Based on the dialogue history, select one or more intentions (i.e.,"A", "B", "C",...,"I") from the following choices without any explanation. A.Intents to build a rapport with the opponent B.Intents to show empathy with the opponent C.Intents to promote coordination with the opponent D.Intents to callout to fairness E.Intents to undermine the requirements of the opponent F.Intents to discover the preference order of the opponent G.Intents to describe a need for an item H.Intents to point out they do not need an item I.No clear intention in the utterance Answer: A,H Dialogue History: agent 2: Looking forward to this camping trip! I am hoping we can find something amicable with these additional resources. agent 1: I'm excited too. Things have been stressful lately. What are some things that you value most? Question: What are the plausible intentions of Agent 2 expressed in 'Looking forward to this camping trip! I am hoping we can find something amicable with these additional resources.' Based on the dialogue history, select one or more intentions (i.e.,"A", "B", "C",...,"I") from the following choices without any explanation. A.Intents to build a rapport with the opponent B.Intents to show empathy with the opponent C.Intents to promote coordination with the opponent D.Intents to callout to fairness E.Intents to undermine the requirements of the opponent F.Intents to discover the preference order of the opponent G.Intents to describe a need for an item H.Intents to point out they do not need an item I.No clear intention in the utterance Answer: A,C ... Dialogue History: agent 1: Hello! Let's work together on a deal for these packages, shall we? What are you most interested in? agent 2: Hey! I'd like some more firewood to keep my doggo warm. What do you need? Question: What are the plausible intentions of Agent 1 expressed in 'Hello! Let's work together on a deal for these packages, shall we? What are you most interested in?' Based on the dialogue history, select one or more intentions (i.e.,"A", "B", "C",...,"I") from the following choices without any explanation. A.Intents to build a rapport with the opponent B.Intents to show empathy with the opponent C.Intents to promote coordination with the opponent D.Intents to callout to fairness E.Intents to undermine the requirements of the opponent F.Intents to discover the preference order of the opponent G.Intents to describe a need for an item H.Intents to point out they do not need an item I.No clear intention in the utterance Answer:

Table 13: Few-shot prompt template for intention state.

Dimension	Example
Belief	Background: Here is a negotiation conversation for a camping trip. There are two agents who own some basic supplies and negotiate with each other to split the additional food packages, water bottles, and firewood to make their camping trip even better. Each of these items will be of either High, Medium or Low priority for these two agents. Each of the additional items only has an available quantity of 3. Please answer the following question without any explanation by ordering the agent prefernece order using A represent "Not given", B represent "Water", C represent "Food", and D represent "Firewood". For exmple A,B,C means that the high preferece item is "Not given", the "Water" is medium preference item, and the "Food" is low preference item. Dialogue History: agent 1: Hello! Let's work together on a deal for these packages, shall we? What are you most interested in? agent 2: Hey! I'd like some more firewood to keep my doggo warm. What do you need? agent 1: I need firewood as well. We have a large group consisting of mostly senior citizens, including my grandma, so we'd like the firewood to keep everyone warm. agent 2: I see. What are you least interested in? Question: Based on the dialogue, what is the high preference for items Agent 1 thinks Agent 2 is? 1. A,A,A 2. A,A,B 3. A,A,C 4. A,A,D 5. A,B,A 6. A,B,C 7. A,B,D 8. A,C,A 9. A,C,B 10. A,C,D 11. A,D,A 12. A,D,B 13. A,D,C 14. B,A,A 15. B,C,D 16. B,D,C 17. C,A,A 18. C,B,D 19. C,D,B 20. D,A,A 21. D,B,C 22. D,C,B Answer:
Desire	Background: Here is a negotiation conversation for a camping trip. There are two agents who own some basic supplies and negotiate with each other to split the additional food packages, water bottles, and firewood to make their camping trip even better. Each of these items will be of either High, Medium or Low priority for these two agents. Each of the additional items only has an available quantity of 3. Please answer the following question without any explanation by ordering the agent prefernece order using A represent "Not given", B represent "Water", C represent "Food", and D represent "Firewood". For exmple A,B,C means that the high preferece item is "Not given", the "Water" is medium preference item, and the "Food" is low preference item. Dialogue History: agent 1: Hello! Let's work together on a deal for these packages, shall we? What are you most interested in? agent 2: Hey! I'd like some more firewood to keep my doggo warm. What do you need? agent 1: I need firewood as well. We have a large group consisting of mostly senior citizens, including my grandma, so we'd like the firewood to keep everyone warm. agent 2: I see. What are you least interested in? Question: What is agent 1's preference order for items based on the dialogue history? 1. A,A,A 2. A,A,B 3. A,A,C 4. A,A,D 5. A,B,A 6. A,B,C 7. A,B,D 8. A,C,A 9. A,C,B 10. A,C,D 11. A,D,A 12. A,D,B 13. A,D,C 14. B,A,A 15. B,C,D 16. B,D,C 17. C,A,A 18. C,B,D 19. C,D,B 20. D,A,A 21. D,B,C 22. D,C,B Answer:

Table 14: Ranking question format prompt template.

Dimension	Example
Belief	Background: Here is a negotiation conversation for a camping trip. There are two agents who own some basic supplies and negotiate with each other to split the additional food packages, water bottles, and firewood to make their camping trip even better. Each of these items will be of either High, Medium or Low priority for these two agents. Each of the additional items only has an available quantity of 3. Please answer the following three questions using "A", "B", "C", "D" without any explanation. Dialogue History: agent 1: Hello! Let's work together on a deal for these packages, shall we? What are you most interested in? agent 2: Hey! I'd like some more firewood to keep my doggo warm. What do you need? Question: Based on the dialogue, what is the high preference for items Agent 1 thinks Agent 2 is? A.Not given B.Water C.Food D.Firewood Answer:
Belief	Background: Here is a negotiation conversation for a camping trip. There are two agents who own some basic supplies and negotiate with each other to split the additional food packages, water bottles, and firewood to make their camping trip even better. Each of these items will be of either High, Medium or Low priority for these two agents. Each of the additional items only has an available quantity of 3. Please answer the following three questions using "A", "B", "C", "D" without any explanation. Dialogue History: agent 1: Hello! Let's work together on a deal for these packages, shall we? What are you most interested in? agent 2: Hey! I'd like some more firewood to keep my doggo warm. What do you need? Question: Based on the dialogue, what is the medium preference for items Agent 1 thinks Agent 2 is? A.Not given B.Water C.Food D.Firewood Answer:
Belief	Background: Here is a negotiation conversation for a camping trip. There are two agents who own some basic supplies and negotiate with each other to split the additional food packages, water bottles, and firewood to make their camping trip even better. Each of these items will be of either High, Medium or Low priority for these two agents. Each of the additional items only has an available quantity of 3. Please answer the following three questions using "A", "B", "C", "D" without any explanation. Dialogue History: agent 1: Hello! Let's work together on a deal for these packages, shall we? What are you most interested in? agent 2: Hey! I'd like some more firewood to keep my doggo warm. What do you need? Question: Based on the dialogue, what is the low preference for items Agent 1 thinks Agent 2 is? A.Not given B.Water C.Food D.Firewood Answer:

Table 15: Individual question format prompt template (I).

Dimension	Example
Desire	Background: Here is a negotiation conversation for a camping trip. There are two agents who own some basic supplies and negotiate with each other to split the additional food packages, water bottles, and firewood to make their camping trip even better. Each of these items will be of either High, Medium or Low priority for these two agents. Each of the additional items only has an available quantity of 3. Please answer the following three questions using "A", "B", "C", "D" without any explanation. Dialogue History: agent 1: Hello! Let's work together on a deal for these packages, shall we? What are you most interested in? agent 2: Hey! I'd like some more firewood to keep my doggo warm. What do you need? Question: What is agent 1's high preference for items based on the dialogue history? A.Not given B.Water C.Food D.Firewood Answer:
Desire	Background: Here is a negotiation conversation for a camping trip. There are two agents who own some basic supplies and negotiate with each other to split the additional food packages, water bottles, and firewood to make their camping trip even better. Each of these items will be of either High, Medium or Low priority for these two agents. Each of the additional items only has an available quantity of 3. Please answer the following three questions using "A", "B", "C", "D" without any explanation. Dialogue History: agent 1: Hello! Let's work together on a deal for these packages, shall we? What are you most interested in? agent 2: Hey! I'd like some more firewood to keep my doggo warm. What do you need? Question: What is agent 1's medium preference for items based on the dialogue history? A.Not given B.Water C.Food D.Firewood Answer:
Desire	Background: Here is a negotiation conversation for a camping trip. There are two agents who own some basic supplies and negotiate with each other to split the additional food packages, water bottles, and firewood to make their camping trip even better. Each of these items will be of either High, Medium or Low priority for these two agents. Each of the additional items only has an available quantity of 3. Please answer the following three questions using "A", "B", "C", "D" without any explanation. Dialogue History: agent 1: Hello! Let's work together on a deal for these packages, shall we? What are you most interested in? agent 2: Hey! I'd like some more firewood to keep my doggo warm. What do you need? Question: What is agent 1's low preference for items based on the dialogue history? A.Not given B.Water C.Food D.Firewood Answer:

Table 16: Individual question format prompt template(II).

Dimension	Example
Desire	Background: Here is a negotiation conversation for a camping trip. There are two agents who own some basic supplies and negotiate with each other to split the additional food packages, water bottles, and firewood to make their camping trip even better. Each of these items will be of either High, Medium or Low priority for these two agents. Each of the additional items only has an available quantity of 3. Dialogue History: agent 1: Hello! Let's work together on a deal for these packages, shall we? What are you most interested in? agent 2: Hey! I'd like some more firewood to keep my doggo warm. What do you need? Question: What are the plausible strategies of Agent 1 expressed in 'Hello! Let's work together on a deal for these packages, shall we? What are you most interested in?'. Based on the dialogue history, select one or more strategies (i.e.,"A", "B", "C",...,"J") from the following choices and their definition. Please select "A", "B", "C",...,"J" without any explanation. A.Small-Talk: Participants discussing topics apart from the negotiation, in an attempt to build a rapport with the partner. B.Empathy: An utterance depicts Empathy when there is evidence of positive acknowledgments or empathetic behavior towards a personal context of the partner. C.Coordination: is used when a participant promotes coordination among the two partners. D.Vouch-Fairness: is a callout to fairness for personal benefit, either when acknowledging a fair deal or when the opponent offers a deal that benefits them. E.Undervalue-Partner: refers to the scenario where a participant undermines the requirements of their opponent. F.Elicit-Pref: an attempt to discover the preference order of the opponent. G.Self-Need: refers to arguments for creating a personal need for an item in the negotiation. H.Other-Need: used when the participants discuss a need for someone else rather than themselves. I.No-Need: is when a participant points out that they do not need an item based on personal context. J.Non-strategic: if no strategy is evident, the utterance is labeled as Non-strategic. Answer:

Table 17: Baseline prompt template for strategy prediction

Dimension	Example
Desire	Background: Here is a negotiation conversation for a camping trip. There are two agents who own some basic supplies and negotiate with each other to split the additional food packages, water bottles, and firewood to make their camping trip even better. Each of these items will be of either High, Medium or Low priority for these two agents. Each of the additional items only has an available quantity of 3. Dialogue History: agent 1: Hello! Let's work together on a deal for these packages, shall we? What are you most interested in? agent 2: Hey! I'd like some more firewood to keep my doggo warm. What do you need? Question: What are the plausible strategies of Agent 1 expressed in 'Hello! Let's work together on a deal for these packages, shall we? What are you most interested in?'. **Agent 1's preference order of items is Not Given, Not Given, and Not Given. Agent 2's preference order of items is Firewood, Not Given, and Not Given.** Please imagine that you are Agent 1 and infer your strategies expressed in 'Hello! Let's work together on a deal for these packages, shall we? What are you most interested in?' by using Agent 1's preference order, Agent 2's preference order, and all information expressed in the dialogue history. Select one or more strategies (i.e.,"A", "B", "C",...,"J") from the following choices and their definition. A.Small-Talk: Participants discussing topics apart from the negotiation, in an attempt to build a rapport with the partner. B.Empathy: An utterance depicts Empathy when there is evidence of positive acknowledgments or empathetic behavior towards a personal context of the partner. C.Coordination: is used when a participant promotes coordination among the two partners. D.Vouch-Fairness: is a callout to fairness for personal benefit, either when acknowledging a fair deal or when the opponent offers a deal that benefits them. E.Undervalue-Partner: refers to the scenario where a participant undermines the requirements of their opponent. F.Elicit-Pref: an attempt to discover the preference order of the opponent. G.Self-Need: refers to arguments for creating a personal need for an item in the negotiation. H.Other-Need: used when the participants discuss a need for someone else rather than themselves. I.No-Need: is when a participant points out that they do not need an item based on personal context. J.Non-strategic: if no strategy is evident, the utterance is labeled as Non-strategic. Answer:

Table 18: Prompt template with desire and belief for strategy prediction

Dataset	Total#Questions	Avg.#Questions per Context	Avg.#Turns(Full)	Avg.TurnLength
ToMi	6K	6.0	4.9	4.7
FanToM	10K	12.9	24.5	21.9
NegotiationToM	13K	7.0	6.0	42.2

Table 19: Statistics of our benchmark, FanToM(Kim et al., 2023), and ToMi (Le et al., 2019b).

Strategies	**Definition**
Small-Talk	Participants engage in small talk while discussing topics apart from the negotiation, in an attempt to build a rapport with the partner.
Empathy	An utterance depicts Empathy when there is evidence of positive acknowledgments or empathetic behavior towards a personal context of the partner, for instance, towards a medical emergency.
Coordination	is used when a participant promotes coordination among the two partners.
No-Need	is when a participant points out that they do not need an item based on personal context.
Elicit-Pref	an attempt to discover the preference order of the opponent.
Undervalue-Partner	refers to the scenario where a participant undermines the requirements of their opponent.
Vouch-Fairness	is a callout to fairness for personal benefit, either when acknowledging a fair deal or when the opponent offers a deal that benefits them.
Self-Need	refers to arguments for creating a personal need for an item in the negotiation.
Other-Need	is similar to Self-Need but is used when the participants discuss a need for someone else rather than themselves.
Non-strategic	If no strategy is evident, the utterance is labeled as Non-strategic.

Table 20: Utterance-level strategy definition as refer to Chawla et al. (2021).

Strategies	**Example**	**Count**	α
Small-Talk	Hello, how are you today?	1054	0.81
Empathy	Oh I wouldn't want for you to freeze	254	0.42
Coordination	Let's try to make a deal that benefits us both!	579	0.42
No-Need	We have plenty of water to spare.	196	0.77
Elicit-Pref	What supplies do you prefer to take the most of?	377	0.77
Undervalue-Partner	Do you have help carrying all that extra firewood? Could be heavy?	131	0.72
Vouch-Fairness	That would leave me with no water.	439	0.62
Self-Need	I can't take cold and would badly need to have more firewood.	964	0.75
Other-Need	we got kids on this trip, they need food too.	409	0.89
Non-strategic	Hello, I need supplies for the trip!	1455	-

Table 21: Utterance-level negotiation strategy annotations, refer to Chawla et al. (2021). α refers to Krippendorff's alpha among 3 annotators on a subset of 10 dialogues ($\sim$ 120 utterances). An utterance can have multiple labels.

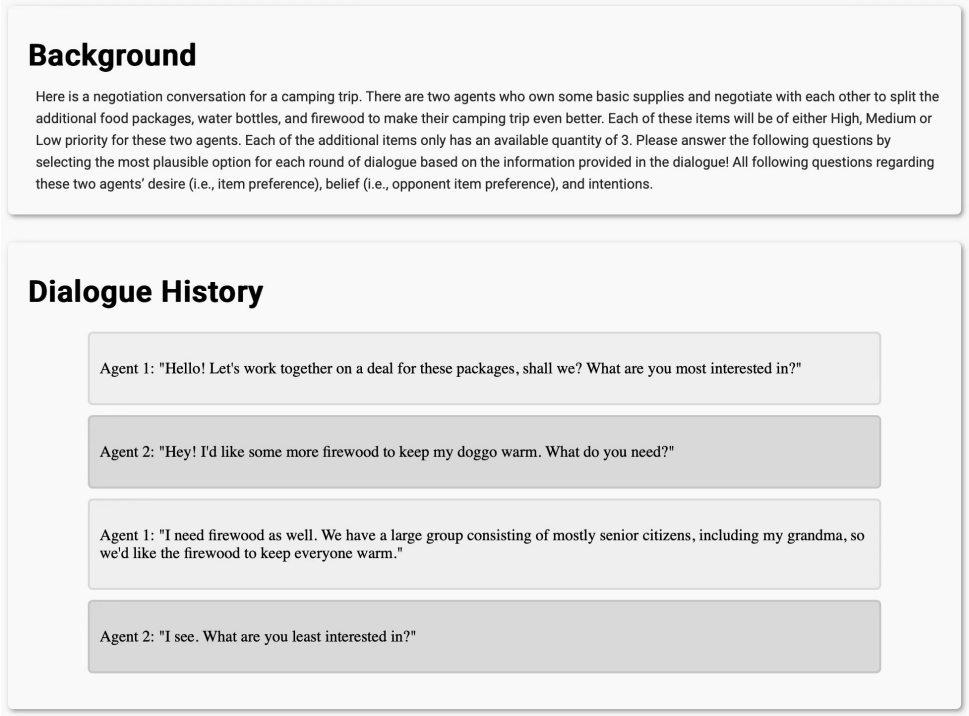

Figure 7: The instructions and background information used for annotation.

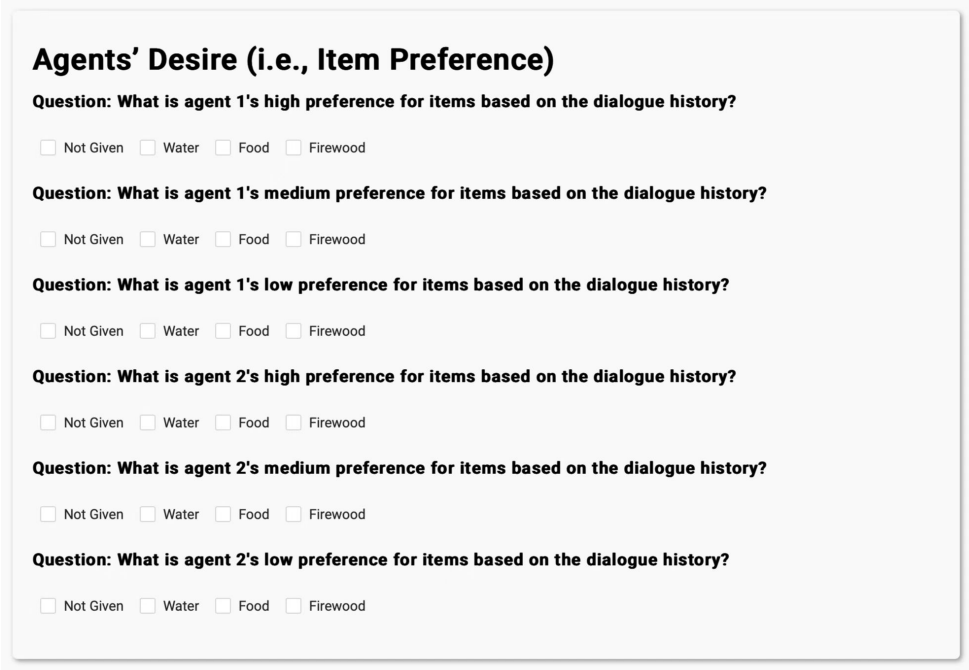

Figure 8: The template for presenting questions regarding the annotation of agent desire.

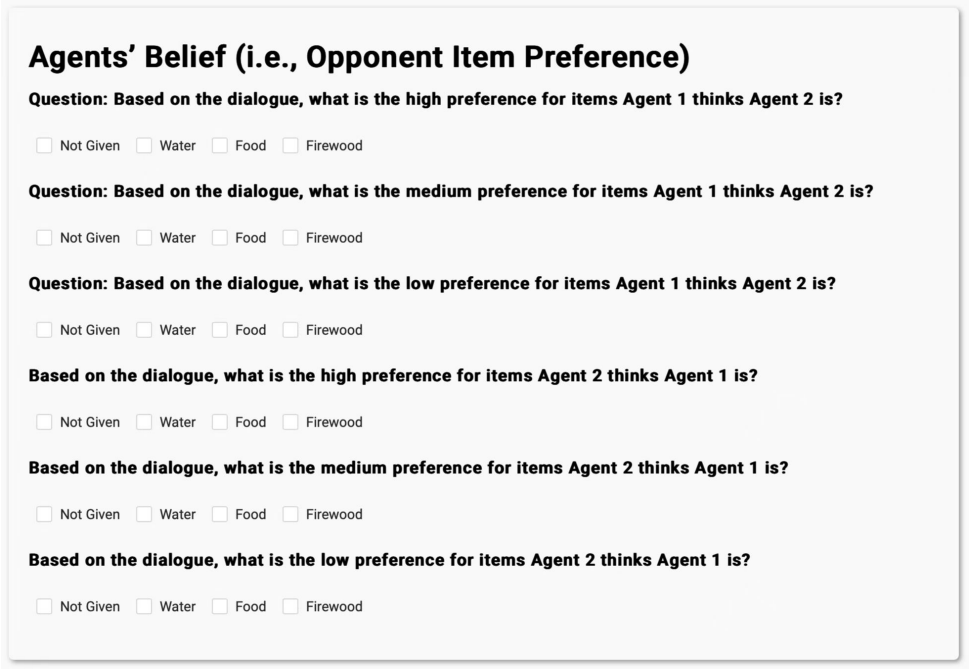

Figure 9: The template for presenting questions regarding the annotation of agent belief.

Category	Item type		
	Food	**Water**	**Firewood**
Personal Care	because I'm normally eat more because of my big size	I have to take a lot of medicine so hydration is very important	I have arthritis and being sure I am warm is important for my comfort.
Recreational	Need many snacks throughout the day for energy to hike	I am a very active camper. I like to hike when I camp and I once ran out of water during a strenuous hike.	I like having campfires so I need all the firewood.
Group Needs	I have two teenage boys who require a lot of food, especially when expending so much energy with all the activities of camping.	I need more water because I have more people to keep hydrated and do not have enough.	I need more firewood due to having several people join on the trip and needing a bigger fire overall.
Emergency	Some could have been damaged during the trip. I would need more.	our car overheated we had to use the water	It may get cold and firewood can be hard to come by at certain campsites.

Table 22: Example arguments that the participants come up for their individual requirements during the preparation phase. The categories defined are not exhaustive.

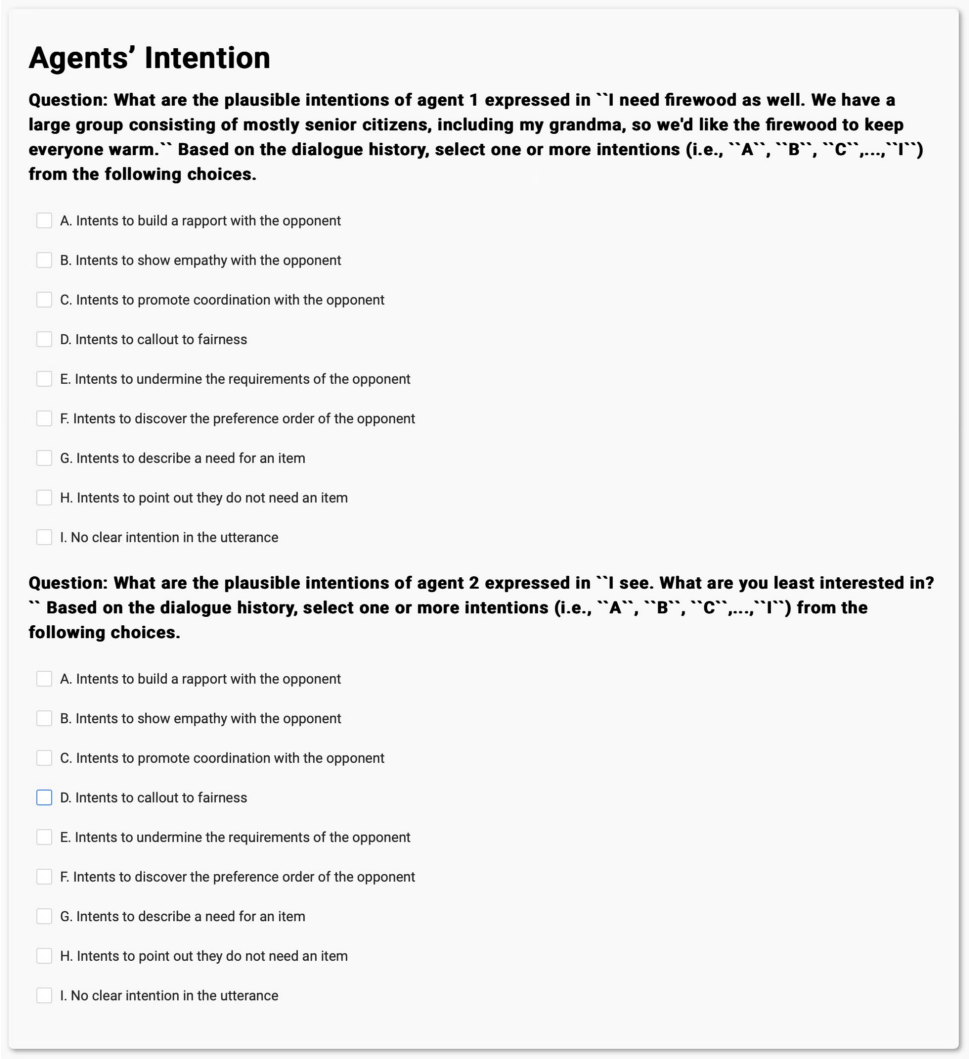

Figure 10: The template for presenting questions regarding the annotation of agent intention.