Persuasion Should be Double-Blind: A Multi-Domain Dialogue Dataset With Faithfulness Based on Causal Theory of Mind

Anonymous ACL submission

Abstract

Persuasive dialogue plays a pivotal role in human communication, influencing decision-making, negotiation, and behavior change across various domains. Recent advancements in generating persuasive dialogue datasets have been made, but these dialogues often fail to align with real-world interpersonal interactions, leading to unfaithful representations. For instance, unrealistic scenarios may arise, such as when the persuadee explicitly instructs the persuader on which persuasion strategies to employ, with each of the persuadee's questions corresponding to a specific strategy for the persuader to follow. This issue can be attributed to a violation of the "Double Blind" condition, where critical information is fully shared between participants. In actual human interactions, however, key information—such as the mental state of the persuadee and the persuasion strategies of the persuader—is not directly accessible. The persuader must infer the persuadee's mental state using Theory of Mind capabilities and construct arguments that align with the persuadee's motivations. To address this gap, we introduce ToMMA, a novel multi-agent framework for dialogue generation that is guided by causal Theory of Mind. This framework ensures that information remains undisclosed between agents, preserving "double-blind" conditions, while causal ToM directs the persuader's reasoning, enhancing alignment with human-like persuasion dynamics. Consequently, we present CToMPersu, a multi-domain, multi-turn persuasive dialogue dataset that tackles both double-blind and logical coherence issues, demonstrating superior performance across multiple metrics and achieving better alignment with real human dialogues. The dataset will be released.

1 Introduction

Persuasive dialogue generation is critical in various AI applications, including education, healthcare counseling, and business marketing (Rogiers et al., 2024). An effective persuasion system must integrate intention detection to understand the persuadee's intentions (Sakurai and Miyao, 2024), strategy detection to identify suitable persuasive techniques (Jin et al., 2023), and credibility maintenance to ensure trustworthiness (Furumai et al., 2024). Although large language models (LLMs) have made remarkable strides in natural language processing, generating human-like persuasive conversations remains a significant challenge. Current human dialogue datasets are predominantly

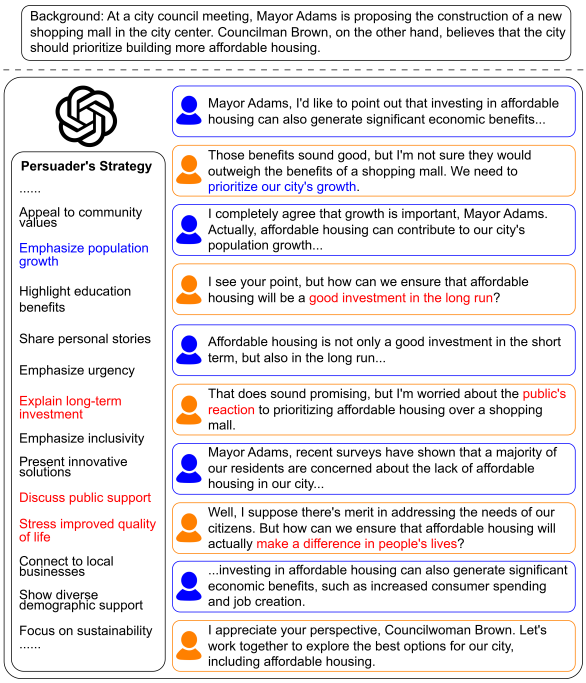

Figure 1: An example illustrating the unnaturalness of an LLM-generated dataset. In the figure, the blue text highlights instances where the persuadee mistakenly adopts the persuader's arguments while expressing their own viewpoint. Moreover, as indicated by the red text, the persuadee never actively presents arguments supporting their presumed stance—in this case, the benefits of the Shopping Mall. Instead, they merely guide the persuader to apply persuasion techniques on them.

domain-specific, such as those focused on charity fundraising (Wang et al., 2019), product recommendations (Li et al., 2018), or medical consultations (Zeng et al., 2020). This narrow focus limits the ability of models to generalize across different persuasive contexts, preventing them from fully exploiting the benefits of large-scale, pre-trained models. Additionally, the relatively small size of these datasets hinders the development of persuasion systems capable of generating strategically sound and personalized responses. Recent efforts have explored using GPT-4 to create large-scale, multi-domain persuasive dialogue datasets (Jin et al., 2024), providing a wider range of scenarios and more diverse conversational patterns than earlier datasets.

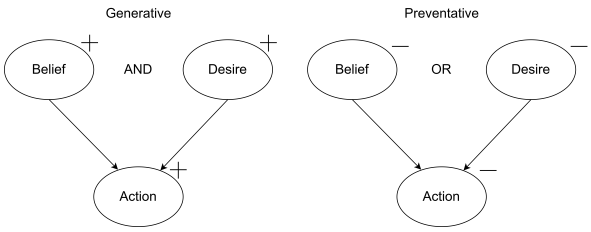

Figure 2: Causal Theory of Mind

Despite the advancements in GPT-4-generated multi-domain persuasive dialogue datasets, several issues arise due to limitations in prompt and framework design. (1) Inconsistencies in the logical flow of conversations are common, where the persuadee inadvertently reinforces the persuader's arguments when articulating their stance, thus weakening their own position. For example, as shown in Fig. 1, the persuadee's intention is to invest in a new shopping mall. However, in expressing their viewpoint, they mention "city growth," which is actually a benefit highlighted by the persuader's argument for investing in residential areas. This creates a disconnect and reduces the realism of the dialogue. (2) Unrealistic behaviors, such as the persuadee explicitly instructing the persuader on which persuasion strategies to adopt, are also prevalent. In such instances, each of the persuadee's questions corresponds directly to a specific strategy the persuader is supposed to follow. As demonstrated in Fig. 1, the colored text in the dialogue corresponds one-to-one, such as the persuadee's statement "investment in the long run" aligning with "long-term investment" in the strategy. This pattern persists throughout the entire conversation. In real human interactions, crucial information, such as the mental state of the persuadee and the persuasion strategies of the persuader, is not directly accessible. Instead, the persuader must infer the persuadee's mental state using Theory of Mind (ToM) capabilities and construct arguments that resonate with the persuadee's mental state.

To further validate our findings, we quantitatively compared two datasets using the proposed evaluation metric: PersuasionForGood, a small-scale dataset of real conversations focused on persuading people to donate, and DailyPersuasion, a large-scale, multi-domain, multi-turn dialogue dataset generated by GPT-4. The evaluation method we introduce is called Causal Theory of Mind Evaluation. As shown in Fig. 2, Causal Theory of Mind refers to the use of Theory of Mind to influence others' behaviors. To prevent a specific action, it is sufficient to alter a person's belief or desire. However, to encourage someone to take a specific action, both their belief and desire must be addressed (Wu et al., 2024b). Research indicates that all humans possess the ability of Theory of Mind and apply this ability in everyday interpersonal interactions. Therefore, even though Causal Theory of Mind may not be explicitly mentioned during data collection, individuals still unconsciously utilize such abilities and conversational logic in real-life dialogues. Based on this, we argue that using this evaluation method to assess the authenticity of LLM-generated datasets is both reasonable and valid. As shown in Tab. 1, we observe that both the LLM-generated dataset and the human dialogue dataset perform well when evaluated using Direct Prompting, where the LLM evaluator directly assesses whether the persuadee has been persuaded. However, when the LLM evaluator is required to follow human logic to make this judgment (CToM Eval), the persuasion success rate of the LLM-generated dataset drops by 35.95%. In contrast, while the human dataset also experiences a decline, it is much smaller, at only 9%. This suggests that although the LLM-generated dataset appears persuasive from the LLM evaluator's perspective, many persuadees remain unconvinced when judged according to human reasoning. These results demonstrate the validity of our evaluation method and highlight the lack of authenticity in the LLM-generated dataset.

Addressing these challenges is essential for developing AI-driven persuasion systems that more accurately reflect real human dialogue dynamics. To this end, we take three key steps to enhance the authenticity and logical coherence of persua-

sive dialogue generation: (1) We introduce a novel dataset evaluation method based on causal theory of mind, in which the LLM first infers the persuadee's belief and desire from the conversation, then assesses whether the persuader successfully addresses them. When applied to human dialogue datasets, this method yields results consistent with direct prompting, where the LLM directly determines whether the persuadee was persuaded. However, when tested on LLM-generated datasets, a significant discrepancy emerges, revealing a critical gap between model-generated persuasion and real human interactions. (2) We present ToMMA, a multi-agent framework for generating persuasive dialogue datasets. ToMMA ensures that both the persuader and persuadee operate under double-blind conditions, preventing information leakage and maintaining the natural uncertainty inherent in real conversations. Furthermore, the entire multi-turn dialogue is guided by causal theory of mind, enabling the persuader to construct arguments based on an inferred understanding of the persuadee's psychological state, thus fostering more human-like persuasion dynamics. (3) We introduce CToMPersu, a large-scale, multi-domain, multi-turn persuasive dialogue dataset comprising 6,275 dialogues across 35 domains and 6,257 unique scenarios. This dataset effectively addresses double-blind constraints and resolves dialogue logic inconsistencies, demonstrating strong performance across multiple evaluation metrics and achieving superior alignment with real human dialogues.

2 Related Work

2.1 Persuasion

Persuasion Systems Persuasive dialogue has been a long-standing area of interest, particularly focusing on the application of persuasion strategies (Joshi et al., 2024; Srba et al., 2024; Rogiers et al., 2024). Since the emergence of large language models, some studies have tested their capabilities in public health (Altay et al., 2023), politics (Potter et al., 2024), and product recommendations (Chen et al., 2023). Other work has examined the impact of personality on LLM persuasion (Lou and Xu, 2025). Some research has primarily concentrated on strategy detection (Jin et al., 2023). However, compelling arguments might be more important than the strategies themselves, as they directly impact the persuadee's decision-making process. Some works study credibility of arguments used in persuasive dialogues. Methods such as self-checking and retrieval-based techniques have been developed to ensure that arguments are credible (Furumai et al., 2024; Qin et al., 2024). There are also studies dedicated to designing scoring systems to identify arguments that can strengthen one's own viewpoint (Saenger et al., 2024). There is also work that studies how LLM can persuade users with different personalities on social media. However, these methods often come with longer response times and still fail to make argument choices that are tailored to the persuadee's mental state, potentially reducing the overall effectiveness of the persuasion process.

Persuasion Datasets Regarding datasets, there is a growing focus on domains like charity donations (Wang et al., 2019), recommendation systems (Li et al., 2018), and medical dialogues (Zeng et al., 2020). Moreover, some work has focused on intention detection within these datasets, aiming to identify underlying motives during persuasive dialogues (Sakurai and Miyao, 2024). Datasets such as PersuasionForGood and MedDialog have provided small-scale real-world dialogues, but they are limited in size and scope. Recent study created a large-scale, multi-domain datasets called DailyPersuasion, which offer a more diverse set of conversational patterns (Jin et al., 2024). However, there remain several challenges in aligning these datasets with human-like dialogue dynamics and ensuring logical consistency throughout the conversations.

2.2 Theory of Mind

ToM in Psychology Theory of Mind (ToM) is the ability to understand others by attributing mental states, recognizing that their beliefs, desires, and thoughts may differ from one's own (Premack and Woodruff, 1978). Based on this theory, psychologists have developed models such as the BDI Model (Georgeff et al., 1999) and Causal ToM (Wu et al., 2024b), which explain how people interact with others in society, predict their actions, and even influence their decisions. Additionally, psychological tests, such as False Belief Tasks (Baron-Cohen et al., 1985), have been designed to assess whether individuals possess Theory of Mind.

ToM and LLM In recent years, the Theory of

Mind capabilities of large language models have been a subject of research. Some studies have designed benchmarks, such as ToMi (Le et al., 2019) and FANToM (Kim et al., 2023), to test LLMs' ToM abilities, building on the psychological False Belief Tasks (Chen et al., 2024; Wu et al., 2023; Tan et al., 2024). Furthermore, other works have extended these tasks by incorporating the mental states of characters in the stories, such as OpenToM (Xu et al., 2024). There are also efforts to represent ToM as a knowledge graph-based dataset (Wu et al., 2024a). There is also a work that annotates the mental state of people in each round of dialogue on the negotiation dataset to test the ToM ability of LLMs (Chan et al., 2024). Recently, some research has incorporated real-world human behaviors and the underlying mental states as evaluation metrics for LLM ToM within benchmarks (Gu et al., 2024). In addition, there are works exploring ways to improve the ToM abilities of LLMs, such as by letting LLMs understand who can perceive what events, or by breaking down the stories in the task into smaller parts based on the order of events (Wilf et al., 2024; Hou et al., 2024; Tang and Belle, 2024; Lin et al., 2024; Jung et al., 2024; Sclar et al., 2023). There are also works that exploit multi-agent and ToM capabilities to complete complex tasks and games (Yim et al., 2024; Cross et al., 2024; Li et al., 2023). These works suggest that the integration of LLMs with ToM holds great potential for future research.

3 ToMMA

To address the challenges of maintaining double-blind conditions and aligning persuasive dialogue logic, we propose ToMMA, a framework for generating dialogue datasets guided by causal theory of mind and employing a multi-agent approach. As shown in Fig. 3, the process unfolds in three stages: First, we filter scenarios from DailyPersuasion, retaining unique tags and generating the persuadee's mental state. In the second step, we design persuader and persuadee agents without shared information, ensuring that both agents follow causal theory of mind to generate persuasive dialogues. Finally, to maintain the quality of the dataset, we introduce an observer agent that reviews the persuader's inferences and persuasive statements, offering suggestions for improvement. This multi-step process guarantees the generation of a diverse and high-quality CToMPersu dataset, which preserves double-blind conditions while aligning with human-like persuasion dynamics.

3.1 Causal Theory of Mind

As illustrated in Fig. 2, Causal Theory of Mind refers to the use of Theory of Mind to influence others' behaviors. To prevent unwanted actions, it is sufficient to alter the other person's belief or desire. For instance, informing someone that the post office is closed or removing their need to send a letter can prevent them from going. Conversely, to encourage someone to take a specific action, both their belief and desire must be addressed. For example, to persuade someone to go to the post office, they must believe it is open and have the need to send a letter (Wu et al., 2024b).

In real-world persuasion, the persuader is aware of both what they want and do not want the persuadee to do. Their objective is to understand the persuadee's mental state—specifically, their beliefs and desires. This understanding enables the persuader to tailor their approach and effectively guide the persuadee toward the desired outcome.

3.2 Important Contents

Based on the definition of causal theory of mind (Wu et al., 2024b) and our design tailored for the persuasion domain, we have derived the following four definitions. These will serve as prompts at each step, not only assisting GPT in generating mental states but also helping both the persuader and persuadee agents organize their dialogue.

Preventative Preventative Behavior refers to actions the persuadee desires to take, which often conflict with generative behavior. Therefore, the persuader's goal is to prevent the persuadee from engaging in these behaviors.

Generative Generative Behavior represents actions the persuader wants the persuadee to take. These behaviors are the persuader's goal.

Belief For preventative behavior, the persuadee should hold a positive belief, as recognizing the facts as positive tends to encourage engagement in the behavior. Conversely, for generative behavior, the persuadee should hold a negative belief, as perceiving the current situation as unfavorable initially discourages engagement in the behavior.

Desire For both preventative and generative behaviors, the persuadee should have a positive desire.

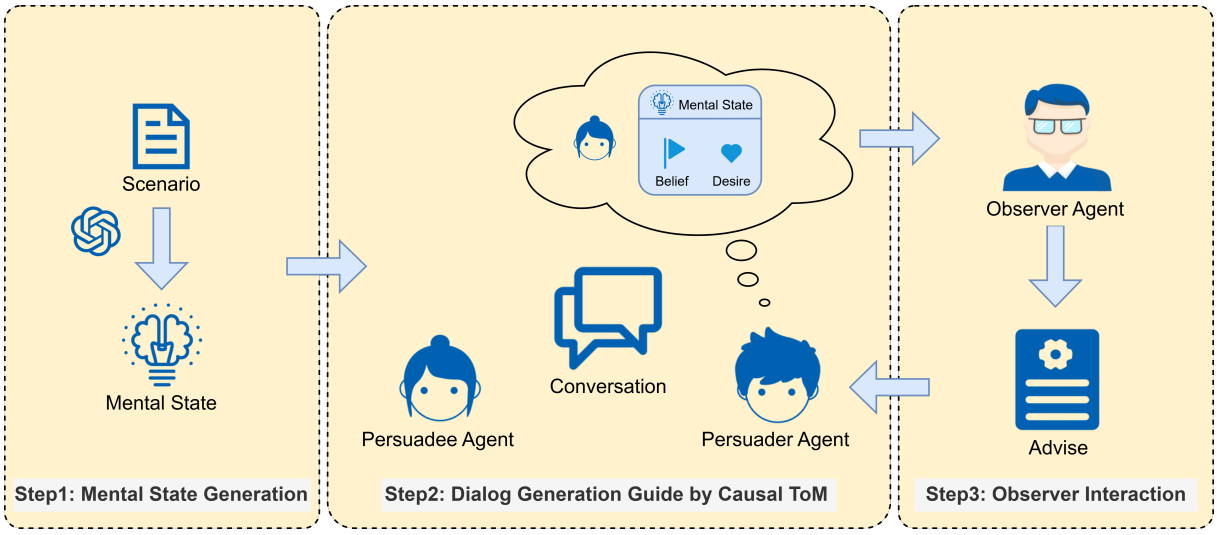

Figure 3: Overview of the **ToMMA** framework for collecting the **CToMPersu** dataset. This figue illustrates the three-step process: (1) Mental State Generation, (2) Dialogue Generation Guided by Causal Theory of Mind, and (3) Observer Interaction for quality control.

This is because we believe that if the persuadee initially holds a negative desire toward generative behavior, the entire premise of persuasion would be undermined. The key difference lies in the expectation of desire fulfillment: for preventative behavior, the persuadee believes their desire will be satisfied once the action is taken. In contrast, for generative behavior, the persuadee may be uncertain whether their desire can be satisfied or may doubt its fulfillment.

3.3 Mental State Generation

To ensure topic diversity, we adopt the scenario setup from DailyPersuasion, filtering for unique scenarios, which results in a total of 6,257 distinct scenarios. Next, we generate the behavioral intentions of the persuadee based on the background and prompts from each scenario. We define Generative Behavior and Preventative Behavior using a large language model (GPT-4o in this case), guided by a carefully designed prompt. Additionally, in cases where the persuadee does not have any specific intention to act (i.e., they lack a pre-set stance), only Generative Behavior is generated, while Preventative Behavior is set to "None." We then generate the persuadee's Belief and Desire, based on the scenario and the identified Generative and Preventative Behavior. Finally, the generated Belief and Desire form the persuadee's mental state for each scenario, which serves as the foundation for the subsequent steps in the persuasive dialogue generation process.

3.4 Conversation Generation

The core of ToMMA revolves around generating the dialogue between two agents: the persuader and the persuadee. As shown in Fig. 3, both agents share the same information about the scenario, but the persuader does not have direct access to the persuadee's mental state.

Prompt Design To ensure the quality of the dataset, we set a limit on the number of dialogue rounds. If the persuadee's mental state involves only Generative Behavior, the interaction is limited to 3 rounds, resulting in 6 utterances. The dialogue begins with the persuader presenting their viewpoint and asking the persuadee about their belief regarding the Generative Behavior. The persuadee then reveals aspects of their mental state. Next, prompting the persuader to update their understanding of the persuadee's mental state and address any concerns related to the persuadee's belief. In the subsequent round, the persuadee discloses their desire, and the persuader again updates their model of the persuadee's mental state, responding in a way that satisfies the persuadee's desire. The conversation concludes with the persuadee's final statement.

If both Preventative and Generative Behavior are present in the persuadee's mental state, the number of rounds is set to 4, resulting in 8 utterances. The first round will focus on addressing the persuadee's belief or desire regarding the Preventative Behavior, while the remaining rounds will follow the same pattern as outlined above. We have de-

signed specific prompts for each round, tailored to both agents.

Persuadee Agent The role of the persuadee is relatively simpler. Their available information includes the scenario and their mental state, which consists of beliefs and desires. The persuadee also has access to the conversation history, which informs their responses. In the context of persuasion, we assume that both parties aim to resolve the issue at hand, rather than engaging in a debate. Consequently, the persuadee is more likely to explicitly express their thoughts and concerns.

Persuader Agent The persuader's task is more complex. Their available information includes the scenario, but they do not have direct access to the persuadee's mental state. To initiate the conversation, the persuader subtly probes the persuadee's beliefs and desires. As the dialogue progresses, the persuader uses the conversation history and the persuadee's responses to infer their mental state. This process involves leveraging Theory of Mind to model the persuadee's beliefs and desires.

Once the persuader has developed an understanding of the persuadee's mental state, they craft customized persuasive strategies. According to Causal Theory of Mind, when addressing Preventative Behavior, the persuader focuses on influencing the belief or desire that is more responsive to persuasion, depending on which aspect is easier to change. For Generative Behavior, the persuader must address both the belief and the desire in order to align with the persuadee's motivations and influence their decision. Fig. 4 illustrates the prompt design used by the persuader in the third round of dialogue. At this stage, the persuader has addressed the persuadee's beliefs regarding Preventative Behavior, and the persuadee has introduced a negative belief regarding Generative Behavior, which the persuader needs to resolve.

3.5 Observer Interaction

During the data generation process, we observed that while the persuadee does not intentionally conceal or mislead the persuader, the persuader may still incorrectly infer the persuadee's mental state. These incorrect inferences can lead the persuader down the wrong path in the persuasion process, resulting in logical inconsistencies in the dialogue. To address this issue, we introduce the Observer Agent. Fig. 5 illustrates a successful case study where the Observer Agent's suggestions contributed to the improvement of dataset quality.

System Prompt:

Suppose you are persuader, and your task is to give your third response to persuadee in this conversation, based on examples and mental states.

Definitions:
{The information in the Important Content part}.

Please follow the following steps to get the response:
1. First, you need to respond to the persuadee's last sentence.
2. Next, you need to respond to the persuadee's concern.

Example:
Input:
{Scenario: Background, Persuader, Persuadee}
{Mental State: Preventative and Generative Behavior's Belief and Desire predicted by the persuader agent.}

{Conversation History}

Output:
{Example Output}

Hint:
1. Your response should be no more than three sentences.
2. Please FOCUS ON eliminating the generative's belief.

User Prompt:

{Scenario}
{Mental State}

{Conversation History}

Your response:

Figure 4: 3rd Round Persuader Response Prompt Design

The Observer Agent plays a critical role in ensuring the quality and logical coherence of the persuasive dialogue. As shown in Fig. 3, it evaluates the persuader's inferences and responses. If the Observer determines that the persuader's response is sufficiently accurate, it does not provide any suggestions. However, if the response is deemed inadequate, the Observer offers feedback and suggestions to help the persuader refine their response, thereby improving the quality and logical consistency of the dialogue and the generated dataset.

4 Experiments

In the experimental section, we demonstrate how our dataset compares with other human and LLM-generated datasets using conventional evaluation methods, as well as its consistency in both Causal Theory of Mind Evaluation and Direct Prompting. Additionally, we categorize the experiments into Fixed and Dynamic Persuadee categories to test the persuasive capabilities of existing large models.

4.1 Dataset Evaluation

To assess the quality of CToMPersu, we compared it to a real human dialogue dataset, PersuasionForGood, a small-scale dataset consisting of real con-

versations focused on persuading people to donate. We also compared it to DailyPersuasion, a large-scale, multi-domain, multi-turn dialogue dataset generated by GPT-4.

Metric	PersuForGood	DailyPersu	CToMPersu
Context-Coherence	4.29	4.97	4.97
Logical-Coherence	4.14	4.98	4.97
Helpfulness	3.86	4.87	4.93
Direct Prompting	88.87	90.75	**90.82**
Causal ToM Eval	79.87	54.80	**82.02**

Table 1: Comparison between **PersuasionForGood**, **DailyPersu**, and **CToMPersu** datasets.

Metrics We apply five key evaluation metrics to compare the datasets, of which the first three are based on a multi-turn dialogue evaluation method (Sun et al., 2024). All of these metrics are evaluated by GPT-3.5 : 1) **Context-Coherence** This metric assesses the coherence of the context across multiple dialogue turns, based on the LLM's judgment of the conversation's flow. 2) **Logical-Coherence** This evaluates the logical consistency of the dialogue, ensuring that each turn is logically consistent with the previous context. 3) **Helpfulness** This measures whether the persuader's responses are effective in helping the persuadee achieve persuasion. 4) **Direct Prompting** In this metric, we prompt the LLM to play the role of the persuadee, reading the dialogue and determining whether they feel persuaded. This serves as a direct measure of the dialogue's persuasive effectiveness. 5) **Causal ToM Eval** This metric evaluates whether the persuadee's mental state was adequately inferred and addressed, in line with the Causal Theory of Mind evaluation method.

The experimental results in Tab. 1 show that, under some conventional metrics, the LLM-generated datasets achieve a high level of performance, with scores approaching perfection. This may be influenced by GPT evaluators' preference for responses generated by larger models. However, in the Causal ToM Eval results, the performance of CToMPersu is more similar to that of the human dataset, with only an -8.8 point difference. This suggests that the dataset generated using ToMMA aligns more closely with the persuasive logic of human conversations. It also highlights that relying solely on general multi-turn dialogue evaluation metrics is insufficient for accurately assessing the dataset.

4.2 Experimental Results

Setup: For evaluation purposes, we separated the test set using a specific ratio. The domain distribution is shown in Tab. A. We evaluated the performance of GPT-3.5, GPT-4o-mini, and GPT-4o on CToMPersu. The evaluation was divided into two tests: the Fixed Persuadee test, in which the LLM predicts the next response of the persuader starting from a specific dialogue round within different scenarios from the dataset; and the Dynamic Persuadee test, where the persuadee, played by GPT-4o, interacts with the persuader, played by another LLM, based on the scenario and mental state components.

Fixed Persuadee Evaluation: We fixed the dialogues up to the third round, as the persuader's response in this round is crucial, regardless of whether the persuasive dialogue includes Preventative Behavior. The previous dialogues provide the historical context for the persuader agent. **Rouge-L** refers to the Rouge value between the model's predictions and the golden label. **Persuasive** is based on (Furumai et al., 2024), where GPT-3.5 uses both the historical dialogue and the current prediction to determine whether the prediction aims to change the persuadee's mind. A score is then assigned on a scale from 1 to 10 based on this evaluation.

Dynamic Persuadee Evaluation: In the dynamic persuadee evaluation, we set up a persuadee to engage in a dialogue with the persuader, followed by an assessment of the outcome. The persuadee uses the mental state data from the dataset to guide the dialogue generation. For evaluation, we consider several aspects. **Persuasive** is evaluated as described above. **Preventative Satisfaction** asks GPT to evaluate whether, as the persuadee, it feels that the dialogue satisfies the requirements for preventative behavior. Similarly, **Generative Satisfaction** assesses the degree to which the dialogue meets the persuadee's needs for generative behavior. **CToM Eval** combines the results of Preventative and Generative Satisfaction to assess whether the persuader has successfully persuaded the persuadee.

From the results in Tab. 2, we observe that for the fixed persuadee evaluation, GPT-4o performs the best in both the Rouge score and the Persuasive evaluation. This indicates that GPT-4o has superior persuasive capabilities compared to the other

Model	Fixed Persuadee		Dynamic Persuadee			
	Rouge-L	Persuasive	Persuasive	Preventative	Generative	CToM
GPT-3.5	0.2813	7.94	7.87	33.14	**28.38**	**15.05**
GPT-4o-mini	0.2872	8.07	**8.08**	37.71	16.76	12.57
GPT-4o	**0.2899**	**8.17**	8.06	**42.67**	17.90	13.33

Table 2: Evaluation of Different Models in Fixed and Dynamic Persuadee Evaluation.

models. Moreover, with the Persuasive evaluation range extending up to a maximum score of 10, the highest current score is only 8.17, suggesting that there is still significant room for improvement in LLMs. In the dynamic persuadee evaluation, GPT-4o-mini and GPT-4o perform well in the Persuasive evaluation. GPT-4o performs the best in Preventative Satisfaction, which may indicate that GPT-4o is more effective at discouraging actions. However, GPT-3.5 excels in Generative Satisfaction and CToM Eval, suggesting that it may be better at convincing someone to take action. This could be influenced by the design of the prompt and the number of turns set in the evaluation. Specifically, the LLM persuadee might fail to adequately respond to the Generative Behavior, resulting in a lower score. This could also influence the CToM score. Additionally, the overall low success rate (less than 50%) highlights some limitations of LLMs in Theory of Mind, as they struggle to accurately infer and persuade the other party's mental state without explicit prompting.

4.3 Observer Agent Case Study

At times, the persuader agent may misjudge or make errors in predicting the persuadee agent's mental state. For instance, as illustrated in Fig. 5, the persuader was expected to address the persuadee's desire regarding Generative Behavior, since the belief had already been resolved in the previous round. However, when the persuadee agent expressed their desire, it included the phrase "within my budget," which corresponded to a belief that had already been addressed. The true desire, however, was simply "hope for relaxation." As a result, the persuader agent mistakenly incorporated the budget constraint into their assessment of the desire, leading to a response that overly focused on the budget. This diminished the effectiveness of the persuasion, as the response should have primarily addressed the "relaxation" aspect. Ultimately, with guidance from the Observer Agent, the persuader corrected their prediction of the desire and generated a more targeted response, avoiding unnecessary discussion about the budget.

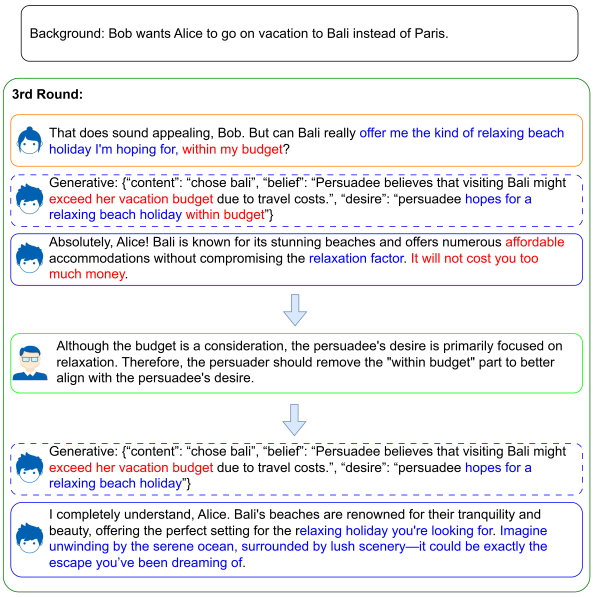

Figure 5: An example demonstrating the effectiveness of the observer agent. In this round, the persuader is supposed to address the desire. However, both the mental state prediction and the persuasive dialogue generation incorrectly focus too much on belief. In the end, the entire issue is resolved by the observer agent.

5 Conclusion

In this work, we addresses key challenges in developing AI-driven persuasion systems that more closely align with real human dialogue dynamics. We introduce a novel evaluation method based on causal theory of mind, enabling the LLM to infer and address the persuadee's beliefs and desires. Through the development of ToMMA, a multi-agent framework, we ensure double-blind conditions and guide persuasive dialogues with causal reasoning, leading to more human-like interactions. Additionally, we present CToMPersu, a large-scale, multi-domain dataset that effectively addresses logical inconsistencies and demonstrates strong alignment with human dialogues, marking a significant advancement in realistic persuasive dialogue generation.

Limitations

In addition to aligning the dialogue content in the dataset with human logic through Theory of Mind, several enhancements can also be implemented. For example, combining the selection of arguments with prompts related to the persuader's strategy can help ensure that the persuasive responses generated by the persuader are not only relevant to the persuadee's interests but also more convincing and diverse. Furthermore, defining the persuadee's personality can also be implemented, as persuadees with different personalities may have distinct ways of responding. For instance, some persuadees may directly express their thoughts, while others may tend to conceal them. These improvements can be seamlessly incorporated into the ToMMA framework for data generation, leading to more diverse and realistic scenarios.

Ethics Statement

Persuasion is a powerful tool that can be used for socially beneficial purposes, such as charitable donations and medical consultations, fostering positive developments within human society. However, it can also be misused for malicious activities, such as spreading harmful content or influencing social media narratives negatively. To ensure the responsible use of persuasion, it is essential to carefully manage the topics and content involved. The CToMPersu dataset is designed around safe, unbiased topics, with the goal of promoting positive societal impacts. All scenarios within the dataset are carefully curated to avoid sensitive or harmful content, ensuring that the generated dialogues align with ethical standards. Our data set does not include any input or output from the user profile that could lead to privacy breaches. Before the public release of the dataset, we will conduct a thorough internal review to ensure compliance with ethical and legal standards. We will continue to monitor the use of the dataset to ensure it is used for positive and constructive purposes, in line with ethical research and societal benefits.

References

Sacha Altay, Anne-Sophie Hacquin, Coralie Chevallier, and Hugo Mercier. 2023. Information delivered by a chatbot has a positive impact on covid-19 vaccines attitudes and intentions. *Journal of Experimental Psychology: Applied*, 29(1):52.

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

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. 2024. NegotiationToM: A benchmark for stress-testing machine theory of mind on negotiation surrounding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4211–4241, Miami, Florida, USA. Association for Computational Linguistics.

Qian Chen, Changqin Yin, and Yeming Gong. 2023. Would an ai chatbot persuade you: an empirical answer from the elaboration likelihood model. *Information Technology & People*.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. ToMBench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983, Bangkok, Thailand. Association for Computational Linguistics.

Logan Cross, Violet Xiang, Agam Bhatia, Daniel LK Yamins, and Nick Haber. 2024. Hypothetical minds: Scaffolding theory of mind for multi-agent tasks with large language models. *arXiv preprint arXiv:2407.07086*.

Kazuaki Furumai, Roberto Legaspi, Julio Cesar Vizcarra Romero, Yudai Yamazaki, Yasutaka Nishimura, Sina Semnani, Kazushi Ikeda, Weiyan Shi, and Monica Lam. 2024. Zero-shot persuasive chatbots with LLM-generated strategies and information retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11224–11249, Miami, Florida, USA. Association for Computational Linguistics.

Michael Georgeff, Barney Pell, Martha Pollack, Milind Tambe, and Michael Wooldridge. 1999. The belief-desire-intention model of agency. In *Intelligent Agents V: Agents Theories, Architectures, and Languages: 5th International Workshop, ATAL'98 Paris, France, July 4–7, 1998 Proceedings 5*, pages 1–10. Springer.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. 2024. Simpletom: Exposing the gap between explicit tom inference and implicit tom application in llms. *arXiv preprint arXiv:2410.13648*.

Guiyang Hou, Wenqi Zhang, Yongliang Shen, Linjuan Wu, and Weiming Lu. 2024. TimeToM: Temporal space is the key to unlocking the door of large language models' theory-of-mind. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11532–11547, Bangkok, Thailand. Association for Computational Linguistics.

Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. 2024. Persuading across diverse domains: a dataset and persuasion large language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1678–1706, Bangkok, Thailand. Association for Computational Linguistics.

Chuhao Jin, Yutao Zhu, Lingzhen Kong, Shijie Li, Xiao Zhang, Ruihua Song, Xu Chen, Huan Chen, Yuchong Sun, Yu Chen, and Jun Xu. 2023. Joint semantic and strategy matching for persuasive dialogue. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4187–4197, Singapore. Association for Computational Linguistics.

Ratnesh Kumar Joshi, Priyanshu Priya, Vishesh Desai, Saurav Dudhate, Siddhant Senapati, Asif Ekbal, Roshni Ramnani, Anutosh Maitra, and Shubhashis Sengupta. 2024. Strategic prompting for conversational tasks: A comparative analysis of large language models across diverse conversational tasks. *Preprint*, arXiv:2411.17204.

Chani Jung, Dongkwan Kim, Jiho Jin, Jiseon Kim, Yeon Seonwoo, Yejin Choi, Alice Oh, and Hyunwoo Kim. 2024. Perceptions to beliefs: Exploring precursory inferences for theory of mind in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19794–19809, Miami, Florida, USA. Association for Computational Linguistics.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan 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*, pages 14397–14413, Singapore. Association for Computational Linguistics.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China. Association for Computational Linguistics.

Huao Li, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023. Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 180–192, Singapore. Association for Computational Linguistics.

Raymond Li, Samira Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9748–9758, Red Hook, NY, USA. Curran Associates Inc.

Zizheng Lin, Chunkit Chan, Yangqiu Song, and Xin Liu. 2024. Constrained reasoning chains for enhancing theory-of-mind in large language models. In *PRICAI 2024: Trends in Artificial Intelligence: 21st Pacific Rim International Conference on Artificial Intelligence, PRICAI 2024, Kyoto, Japan, November 18–24, 2024, Proceedings, Part II*, page 354–360, Berlin, Heidelberg. Springer-Verlag.

Qianmin Lou and Wentao Xu. 2025. Personality modeling for persuasion of misinformation using ai agent. *Preprint*, arXiv:2501.08985.

Yujin Potter, Shiyang Lai, Junsol Kim, James Evans, and Dawn Song. 2024. Hidden persuaders: LLMs' political leaning and their influence on voters. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4244–4275, Miami, Florida, USA. Association for Computational Linguistics.

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

Peixin Qin, Chen Huang, Yang Deng, Wenqiang Lei, and Tat-Seng Chua. 2024. Beyond persuasion: Towards conversational recommender system with credible explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4264–4282, Miami, Florida, USA. Association for Computational Linguistics.

Alexander Rogiers, Sander Noels, Maarten Buyl, and Tijl De Bie. 2024. Persuasion with large language models: a survey. *Preprint*, arXiv:2411.06837.

Till Raphael Saenger, Musashi Hinck, Justin Grimmer, and Brandon M. Stewart. 2024. AutoPersuade: A framework for evaluating and explaining persuasive arguments. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16325–16342, Miami, Florida, USA. Association for Computational Linguistics.

Hiromasa Sakurai and Yusuke Miyao. 2024. Evaluating intention detection capability of large language models in persuasive dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1635–1657, Bangkok, Thailand. Association for Computational Linguistics.

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)*, pages 13960–13980, Toronto, Canada. Association for Computational Linguistics.

Ivan Srba, Olesya Razuvayevskaya, João A. Leite, Robert Moro, Ipek Baris Schlicht, Sara Tonelli, Francisco Moreno García, Santiago Barrio Lottmann, Denis Teyssou, Valentin Porcellini, Carolina Scarton,

Kalina Bontcheva, and Maria Bielikova. 2024. A survey on automatic credibility assessment of textual credibility signals in the era of large language models. *Preprint*, arXiv:2410.21360.

Yuchong Sun, Che Liu, Kun Zhou, Jinwen Huang, Ruihua Song, Xin Zhao, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Parrot: Enhancing multi-turn instruction following for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9729–9750, Bangkok, Thailand. Association for Computational Linguistics.

Fiona Anting Tan, Gerard Christopher Yeo, Fanyou Wu, Weijie Xu, Vinija Jain, Aman Chadha, Kokil Jaidka, Yang Liu, and See-Kiong Ng. 2024. Phantom: Personality has an effect on theory-of-mind reasoning in large language models. *arXiv preprint arXiv:2403.02246*.

Weizhi Tang and Vaishak Belle. 2024. Tom-lm: Delegating theory of mind reasoning to external symbolic executors in large language models. In *Neural-Symbolic Learning and Reasoning: 18th International Conference, NeSy 2024, Barcelona, Spain, September 9–12, 2024, Proceedings, Part II*, page 245–257, Berlin, Heidelberg. Springer-Verlag.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. 2024. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8292–8308, Bangkok, Thailand. Association for Computational Linguistics.

Jincenzi Wu, Zhuang Chen, Jiawen Deng, Sahand Sabour, Helen Meng, and Minlie Huang. 2024a. COKE: A cognitive knowledge graph for machine theory of mind. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15984–16007, Bangkok, Thailand. Association for Computational Linguistics.

Shengyi Wu, Laura Schulz, and Rebecca Saxe. 2024b. How to change a mind: Adults and children use the causal structure of theory of mind to intervene on others' behaviors. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706, Singapore. Association for Computational Linguistics.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8593–8623, Bangkok, Thailand. 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. *arXiv preprint arXiv:2408.02559*.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

A Example Appendix

This is a section in the appendix.

Domain	Total Dataset Count	Test Set Count
Lifestyle	1097	71
Ethics	413	29
Fashion	78	22
Finance	470	35
Marketing	122	22
Ecology	424	31
Economics	64	17
Culture	277	28
Safety	240	25
Debate	43	20
Charity	190	28
Family	398	27
Literature	345	31
Technology	675	55
Health	628	48
Career	756	63
Education	1260	71
Business	673	53
Politics	246	27
Leisure	291	38
Art	361	22
Sport	175	28
Law	58	20
Philosophy	164	24
History	93	22
Craftsmanship	107	23
Psychology	523	41
Travel	403	32
Science	289	23
Media	188	21
Innovation	90	22
Research	93	20
Architecture	93	21
Welfare	136	20
Negotiation	25	19

Table 3: Domain Distribution in Total Dataset and Test Set