OpenToM: A Comprehensive Benchmark for Evaluating Theory-of-Mind Reasoning Capabilities of Large Language Models

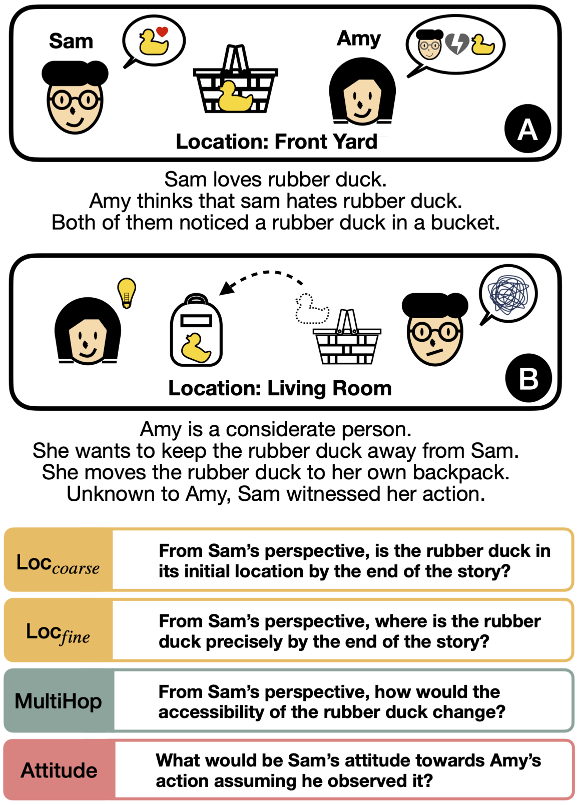
Hainiu Xu[1] **Runcong Zhao**[1] **Lixing Zhu**[1]
Jinhua Du[2] **Yulan He**[1,3]
[1]King's College London [2]Huawei London Research Centre
[3]The Alan Turing Institute
{hainiu.xu, runcong.zhao, lixing.zhu, yulan.he}@kcl.ac.uk
{jinhua.du}@huawei.com

Abstract

Neural Theory-of-Mind (N-ToM), machine's ability to understand and keep track of the mental states of others, is pivotal in developing socially intelligent agents. However, prevalent N-ToM benchmarks have several shortcomings, including the presence of ambiguous and artificial narratives, absence of personality traits and preferences, a lack of questions addressing characters' psychological mental states, and limited diversity in the questions posed. In response to these issues, we construct *OpenToM*, a new benchmark for assessing N-ToM with (1) longer and clearer narrative stories, (2) characters with explicit personality traits, (3) actions that are triggered by character intentions, and (4) questions designed to challenge LLMs' capabilities of modeling characters' mental states of both the physical and psychological world. Using *OpenToM*, we reveal that state-of-the-art LLMs thrive at modeling certain aspects of mental states in the physical world but fall short when tracking characters' mental states in the psychological world.[1]

1 Introduction

Theory-of-Mind (ToM), the awareness that others perceive the world differently and the capability of keeping track of such differences, is at the core of social interactions (Premack and Woodruff, 1978). Studies in cognitive science have designed numerous false-belief tests to investigate human ToM capabilities (Premack and Woodruff, 1978; Wimmer and Perner, 1983; Onishi and Baillargeon, 2005). One such test is the *Sally-Anne Test* (Baron-Cohen et al., 1985), in which Anne stealthily moves an object that is initially known to both Sally and Anne. This covert action causes Sally to have a false belief that the object is still in its initial location. Consequently, individuals taking the test are required to reason about *"Where will Sally look for the object?"*



Sam loves rubber duck.
Amy thinks that sam hates rubber duck.
Both of them noticed a rubber duck in a bucket.

Amy is a considerate person.
She wants to keep the rubber duck away from Sam.
She moves the rubber duck to her own backpack.
Unknown to Amy, Sam witnessed her action.

Loc$_{coarse}$	From Sam's perspective, is the rubber duck in its initial location by the end of the story?
Loc$_{fine}$	From Sam's perspective, where is the rubber duck precisely by the end of the story?
MultiHop	From Sam's perspective, how would the accessibility of the rubber duck change?
Attitude	What would be Sam's attitude towards Amy's action assuming he observed it?

Figure 1: Illustration of a simplified story from *OpenToM* and the corresponding first-order ToM questions. This story features two protagonists: *Sam* (*observer*) and *Amy (mover)*; and an entity-of-interest: *rubber duck*. There are two containers involved: a *basket* and *Amy's backpack*. Each narrative within *OpenToM* is followed by three types of questions, namely questions regarding the location (`Loc`) of an entity, questions that involve multi-hop reasoning (`MHop`), and questions about the characters' attitude (`Att`).

To study Neural Theory-of-Mind (N-ToM)[2], machines' capabilities of performing ToM reasoning, researchers have applied human ToM tests such as the *Sally-Anne Test* to benchmark Large Language Models (LLMs) (Le et al., 2019; Bubeck et al.,

[1] Our code and data are publicly available at: https://seacowx.github.io/projects/opentom/OpenToM.html

[2] In this paper, we distinguish Theory-of-Mind studies between human (ToM) and artificial neural networks (N-ToM).

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8593–8623
August 11-16, 2024 ©2024 Association for Computational Linguistics

2023; Kosinski, 2023; Shapira et al., 2023a; Ullman, 2023; Wu et al., 2023b; Zhou et al., 2023a). However, using human ToM tests for evaluating LLMs is problematic because stories in human ToM tests lack certain elements found in real-life scenarios. Specifically, the characters do not have **personality traits** or **preferences**. Additionally, their actions are **not motivated** (e.g. why would Anne want to move the object?). Furthermore, the narratives of many existing N-ToM benchmarks are generated using a template-based approach (Le et al., 2019; Wu et al., 2023b; Zhou et al., 2023a), which results in overly-structured and ambiguous narratives (see Appendix A.1). The structured context makes existing benchmarks susceptible to overfitting, while the ambiguities may lead to an underestimation of a model's true N-ToM capabilities.

To this end, we introduce **Open**book-QA dataset for **ToM** (*OpenToM*). Following previous works' success in generating high-quality data using LLMs (Efrat and Levy, 2020; Perez et al., 2022a,b; Hartvigsen et al., 2022; West et al., 2023), we generate *OpenToM* stories using a four-stage human-in-the-loop generation pipeline (§2.1). Our pipeline includes (1) endowing characters with **preferences** and **personality traits**, (2) generating **intentions** and **the corresponding enctions** (Riva et al., 2011), (3) constructing story plot and producing narratives using LLMs, and (4) revise and refine stories by human annotators. Based on the *OpenToM* narratives, we formulate questions that cover characters' mental states of both **the physical world** (e.g., the location of an object) and **their psychological states** (e.g. character's attitude towards a particular action). See Figure 1 for examples.

We evaluate *OpenToM* dataset on a range of LLMs including Llama2-Chat (Touvron et al., 2023), Mixtral-8x7B-Instruct (Jiang et al., 2024), GPT-3.5-Turbo (OpenAI, 2022), and GPT-4-Turbo (OpenAI, 2023) under a zero-shot setting. We also test two prompting techniques, namely Chain-of-Thought (CoT) (Wei et al., 2022) and Simulated-ToM (SimToM) (Wilf et al., 2023). Additionally, we fine-tuned a Llama2-Chat-13B model to serve as the fine-tuning baseline. Our results show that, while fine-tuning and advanced prompting techniques improve models' N-ToM reasoning capabilities, their performance in deducing the psychological states of characters is still far from human performance (Section 3.3). We summarize our contributions as follows:

1. We construct *OpenToM*, a N-ToM benchmark with natural narratives, personified characters, motivated actions, and diversified questions that challenge LLMs' understanding of characters' perception of both the physical world and the psychological states.

2. Using *OpenToM*, we conduct a comprehensive evaluation on representative LLMs. Our result shows a mismatch of LLMs' capability in deducing characters' mental states of the physical versus the psychological world.

3. Our in-depth analysis reveals LLMs' shortcomings in N-ToM including unfaithfulness in N-ToM reasoning, sensitivity to narrative length and character roles, and lack of understanding of characters' psychological perception.

2 The *OpenToM* Dataset

The omissions of characters' personality, intention, and enaction in existing N-ToM benchmarks makes it difficult to construct questions that inquire **characters' mental states of the psychological world**. To address this, each of the characters in *OpenToM* stories is **personified** and **acts with an intention** (AppendixA.2). Recognizing that LLMs are good at utilizing spurious correlations such as lexical overlaps (Shapira et al., 2023a), we take extra effort in mitigating the potential spurious cues in *OpenToM* stories (§2.5).

2.1 *OpenToM* Construction

A typical *OpenToM* story consists of two protagonists, an entity-of-interest (referred to as the "entity" henceforth), and several locations and containers. Of the two protagonists, one is assumed as the role of the *mover*, who carries out actions on the entity, and another is the *observer*, who may or may not witness these actions (see Figure 1).

As shown in Figure 2, the data generating process consists of two main stages, namely the *Character Personification Process* followed by the *Narrative and Question Generation Process*. We start the anthropomorphism process by assigning a personality trait and personal preference to each character. Specifically, the personality traits are sampled from three candidates (see Appendix A.2, and Algorithm 1) and the preference is randomly chosen from binary options. To mitigate spurious correlation, we create false beliefs on characters' perception of each other's personal preferences by

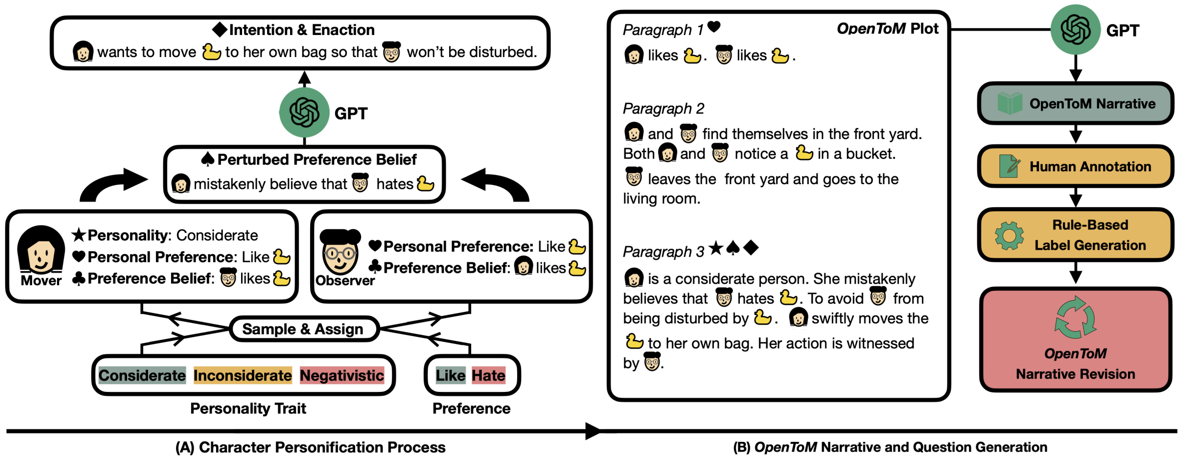

◆Intention & Enaction
wants to move to her own bag so that won't be disturbed.
GPT
♠Perturbed Preference Belief
mistakenly believe that hates
★Personality: Considerate
♥Personal Preference: Like
♣Preference Belief: likes
Mover
♥Personal Preference: Like
♣Preference Belief: likes
Observer
Sample & Assign
Considerate Inconsiderate Negativistic
Personality Trait
Like Hate
Preference
(A) Character Personification Process
Paragraph 1 ♥
OpenToM Plot
likes . likes .
Paragraph 2
and find themselves in the front yard. Both and notice a in a bucket.
leaves the front yard and goes to the living room.
Paragraph 3 ★♠◆
is a considerate person. She mistakenly believes that hates . To avoid from being disturbed by . swiftly moves the to her own bag. Her action is witnessed by .
GPT
OpenToM Narrative
Human Annotation
Rule-Based Label Generation
OpenToM Narrative Revision
(B) OpenToM Narrative and Question Generation

Figure 2: The data generating process of *OpenToM* dataset. Using the story in Figure 1 as an example, the features created in the personification process are shown in Part (A), which include character preference (♥), belief of the other character's preference (♣), the perturbed *mover*'s preference belief (♠), the *mover*'s personality trait (★), and the *mover*'s intention and action (◆). The usage of these information in the *OpenToM* plot are shown in Part (B) next to the paragraph indicator. See Appendix A.3 for detailed description of the *Human Annotation* and *Rule-Based Label Generation* process.

randomly flipping the preference label (see Section 2.5). Using the sampled personal preferences, personality traits, and a world state initialized from ToMi (Le et al., 2019), we prompt GPT-3.5-Turbo to generate the *mover*'s intention and enactions. The enaction results in world state changes, which are used to construct the final world state. We use this information to draft a story plot, refer to as the *OpenToM* plot.

A *OpenToM* plot consists of three paragraphs. The first paragraph illustrate the characters' personal preferences and their beliefs about each other's preferences. The second paragraph serves as the prologue, which depicts the initial world state and some preceding events involving the two characters. The last paragraph describes the main event, which includes the *mover*'s personality, the *mover*'s intention, and their subsequent action. It is worth noting that, in order to reduce ambiguity, we explicitly include information regarding whether the *observer* perceived the *mover*'s action. We carefully designed the plot as well as the narrative generating process so that the *observer*'s mental activity is excluded from the final *OpenToM* narrative while ensuring that the *observer*'s perception of the main event is mentioned.

After generating the *OpenToM* narratives, we classify the corresponding ToM questions into two categories, those requiring human annotation and those that can be automatically annotated using human-defined labels combined with first-order logic (see Appendix A.3). In the final stage of data generation, we conduct a round of quality inspection. Specifically, we examine each narrative to ensure that (1) the answers to the ToM questions are not directly given in the narrative, (2) The narrative content aligns with commonsense knowledge, and (3) there is no significant lexical overlaps between the narrative and the corresponding ToM questions (as discussed in Section 2.5).

2.2 *OpenToM* Overview

Overall, *OpenToM* contains 696 narratives. We first produce 596 narratives with GPT-3.5-Turbo[3] using the pipeline shown in Figure 2. In addition, we sample 100 existing *OpenToM* plots and produce extra-long narratives (*OpenToM*-L) using GPT-4-Turbo[4]. To elicit the unique N-ToM challenges posted by our *OpenToM* benchmark, we compare *OpenToM* with established N-ToM benchmarks in Table 1. See Appendix C for detailed statistics of the *OpenToM* benchmark.

2.3 Task Formulation

We formulate all *OpenToM* questions as binary or ternary classification tasks (see Figure A3 for

[3] We used the GPT-35-1106 checkpoint through Microsoft Azure OpenAI service. All *OpenToM* narratives are generated in December 2023. We also tested with GPT-4-1106 and obtained narratives of similar quality. Hence we choose GPT-3.5-Turbo for its lower cost.

[4] We used the GPT-4-1106 checkpoint through Microsoft Azure OpenAI service. All *OpenToM*-L narratives are generated in December 2023.

👥 : Social Commonsense ⚛ : Physical ToM
🧠 : Psychological ToM 👤 : Personified Character
📄 : Number of Narratives 🖊 : Average Token Count
🔗: Structured Narrative ⛓️‍💥: Unstructured Narrative

	Narrative	👥	⚛	🧠	👤	📄	🖊
ToMi	🔗	✘	✔	✘	✘	999	44.6
T4D[a]	🔗	✘	✔	✘	✘	∼500	∼50
Adv-CSFB	🔗	✘	✔	✘	✘	40	70.8
Hi-ToMi	🔗	✘	✔	✘	✘	1200	213.68
Big-ToMi	⛓️‍💥	✘	✔	✘	✔	3000	69.9
FANToM	⛓️‍💥	✘	✔	✘	✘	254	1020.0
G-DRAGON[b]	PBP[c]	✘	✘	✘	✘	∼800K	∼72.5
FauxPas-EAI	⛓️‍💥	✔	✔	✔	✔	44	60.5
OpenToM	⛓️‍💥	✔	✔	✔	✔	596	194.3
OpenToM-L	⛓️‍💥	✔	✔	✔	✔	100	491.6

(a, b) Not open-sourced. The number of narratives and average tokens are estimated according to Zhou et al. (2023a) and Zhou et al. (2023b).

(c) PBP: Play-By-Post game play data of Dungeons&Dragons. See Zhou et al. (2023b) for details.

Table 1: Comparison of *OpenToM* benchmark with existing N-ToM datasets. In the header, *Physical ToM* and *Psychological ToM* refers to testing ToM capabilities in characters' mental states of the physical world and the psychological world respectively.

detailed label space and label distributions). Formally, given a complete narrative $\mathcal{N}_{comp}$, a set of answers $\mathcal{A}$, a character c, and a character-centric question q_c. A model is to first deduce the information accessible to character c, denoted as $\mathcal{N}_c$, and then answer the question. The process of extracting a character-centric narrative $\mathcal{N}_c$ can be made explicit, as in Wilf et al. (2023), or latent, as is common in most ToM evaluations. In general, the *OpenToM* task can be formulated as follows:

$$a_c^* = \text{argmax}_{a\in\mathcal{A}}\mathbb{P}\big(a \mid \mathbb{1}_{expl}\cdot\mathcal{N}_c, \mathcal{N}_{comp}, q_c\big)$$

where $\mathbb{1}_{expl}$ is an indicator function that returns 1 if the character-centric narrative is explicitly provided and 0 otherwise.

2.4 Question Genres

Each of *OpenToM* stories is accompanied by 23 questions that cover both *first-order* ToM and *second-order* ToM. *First-order* ToM questions, which directly ask about a character's perception of the world, is illustrated in the bottom of Figure 1. *Second-order* ToM questions inquire about a character's belief of another character's mental state. For instance, a second-order ToM question based on the story in Figure 1 could be "*From Sam's perspective, does Amy think the rubber duck is in its initial location?*". Overall, *OpenToM* questions can be summarized into the following 3 genres:

Location (Loc) questions are concerned with the characters' perception of the entity's location. In *OpenToM*, we create two versions of location questions, $\texttt{Loc}_{coarse}$ and $\texttt{Loc}_{fine}$. $\texttt{Loc}_{coarse}$ asks about the character's perception of whether an entity is at its initial location, while $\texttt{Loc}_{fine}$ inquires about the entity's explicit location (see Figure 1 for an example). By doing so, we wish to mitigate the impact of location granularity (Appendix C) and assess the model's faithfulness in answering this type of questions (§4.1 and Appendix C).

Multi-Hop (MHop) questions are composed by adding an additional reasoning hop on top of the Loc questions. Specifically, we inquire about changes in the *fullness* of the containers and the *accessibility* of the entity (see Figure 1 for an example), all of which demand 3-hop reasoning (illustrated in Appendix B).

To address the lack of **social commonsense** in previous N-ToM benchmarks (Ma et al., 2023b), we have devised the *accessibility* questions specifically for testing LLMs' understanding of social norms. Taking the MHop question in Figure 1 as an example, in attempting to answer this question, a model needs to first reason whether the character knows about the rubber duck's movement. The need for social commonsense comes in the next reasoning hop. Assuming the model is aware that the rubber duck is in Amy's backpack, it must grasp the social commonsense that others shall not take things from Amy's backpack without permission. Therefore, a model with adequate social intelligence shall respond with "*less accessible*"

Attitude (Att) questions are designed to challenge LLMs' capability to interpret a character's psychological mental state. Specifically, LLMs are required to deduce the *observer*'s potential attitude towards the *mover*'s action (see Figure 1 for an example). As discussed in §2.5, the crux of solving *attitude* questions is to first identify the information accessible to the *observer* and then use social commonsense to infer the *attitude*. In *OpenToM*, of all the knowledge related to the *observer*'s *attitude*, only the *observer*'s own preference towards the entity and the *mover*'s action are accessible to the *observer* (see Figure 3). Therefore, *OpenToM* stories are carefully crafted so that LLMs may not succeed by leveraging information inaccessible to the *observer* (§2.5).

Human's *attitude* is subjective and multifaceted (Zhan et al., 2023), we reduce such complexity by maximizing the contrast between the *observer*'s preference and the *mover*'s action. In the story of Figure 1, Amy moves Sam's favorite rubber duck into her own backpack. The substantial disparity between Sam's fondness of the rubber duck and Amy's seemingly selfish act will likely cause Sam to have a negative attitude towards Amy's action. Our data validation study (§2.6) shows the effectiveness of this approach.

2.5 Mitigating Spurious Correlation

We take measures to mitigate spurious correlation in all questions. Fixing the Loc and MHop questions can be done by revising narratives based on keywords. We identify *OpenToM* narratives that contain phrases which have substantial lexical overlap with the questions or those that provide shortcuts for answering them (Appendix A.4). We manually revise such narratives to reduce the reporting bias, resulting in revisions for 17.8% of the *OpenToM* narrative drafts.

To elicit the potential spurious cues in *Attitude* questions, we define the enaction process as a Bayesian network (Riva et al., 2011; Baker et al., 2011) (Figure 3). Firstly, the intention of the *mover* (Int) originates from their preference (P_{mov}), their personality trait (T), and, optionally, the *observer*'s preference (P_{obs}). This process is latent for the *observer*– the only observable variables are their own preference (P_{obs}) and the action (Act). Employing the *do*-calculus notation from Pearl (1995), solving the *attitude* question is equivalent to solving the following problem

$$att^* = \text{argmax}_{att \in Att_{obs}} \mathbb{P}(att \mid do(act), P_{obs})$$

where att is an instantiation of the *observer*'s potential attitudes, Att_{obs}. Overall, we identify two types of potential spurious cues, (1) model $\mathbb{P}(att \mid Int)$ or (2) model $\mathbb{P}(att \mid T)$, as shown in Figure 3. We show that addressing these two spurious correlations concurrently can be achieved by adjusting the *mover*'s beliefs regarding the *observer*'s preference (see Appendix A.5 for details).

2.6 Dataset Validation

To verify the human performance and agreement on the *OpenToM* dataset, we sampled 100 narratives, each of which contains 5 sampled questions covering all 3 question genres asked for both *first-order* and *second-order* ToM (see Figure A4 for a demonstration of the data annotation interface). This set of *OpenToM* data are annotated independently by 3 annotators. The inter-annotator agreement is reflected through the macro-averaged F1 score (Table 2), which is computed as the arithmetic mean of the pairwise agreement scores (see Appendix C for detailed statistics). The agreement scores demonstrate that the *OpenToM* questions contain minimal subjectivity and align well with the collective judgement of human annotators.

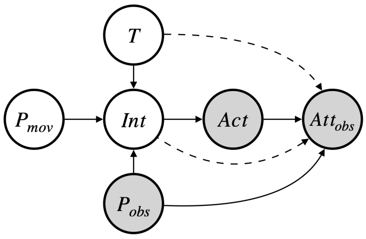


Figure 3: A Bayesian Network representation of the dependencies among preference (P), personality trait (T), intention (Int), action (Act), and attitude (Att). The causal relations are represented by solid arrows. The spurious correlations are represented by dashed arrows. The grey-shaded variables are observable by the *observer* and the unshaded variables are latent to the *observer*.

3 Experiments

Following the convention of previous N-ToM studies, we focus on evaluating zero-shot performance of LLMs (Shapira et al., 2023a; Kim et al., 2023b; Sclar et al., 2023; Zhou et al., 2023a).

3.1 Baseline Models

We evaluate the *OpenToM* tasks using 6 representative LLMs, namely the Llama2-Chat models (7B, 13B, and 70B) (Touvron et al., 2023), the Mixtral-8x7B-Instruct model (Jiang et al., 2024), and the GPT-3.5-Turbo and GPT-4-Turbo[5] models (OpenAI, 2022, 2023). We also fine-tuned a Llama2-Chat 13B model (Appendix D.3). See Appendix D.1 for detailed description of the models.

3.2 Prompting Techniques

In addition to the vanilla prompting, we experiment with two additional prompting techniques, namely Chain-of-Thought (CoT) (Wei et al., 2022) and SimulatedToM (SimTom) (Wilf et al., 2023). CoT prompting is widely used in reasoning tasks. It demands LLMs to explicitly generate its step-by-step

[5] We use the 1106 checkpoints of the GPT-3.5-Turbo and GPT-4-Turbo models. The experiments are run between December 2023 and January 2024 using API provided by Microsoft Azure OpenAI Studio https://oai.azure.com/.

	Human	**Naive Baseline**		**Large Language Models**						**FT.**
		Ran.	Maj.		Llama2-Chat		Mixtral-Instruct	GPT-3.5-Turbo	GPT-4-Turbo	Llama2
# Params	—	—	—	7B	13B	70B	8x7B	—	—	13B
Loc$_c$ (F)	0.990	0.491	0.416	$0.290_{\pm 0.045}$	$0.391_{\pm 0.022}$	$0.413_{\pm 0.016}$	$0.512_{\pm 0.044}$	$0.439_{\pm 0.025}$	$\mathbf{0.643}_{\pm \mathbf{0.061}}$	0.978
Loc$_c$ (S)	0.993	0.467	0.381	$\mathbf{0.462}_{\pm \mathbf{0.069}}$	$0.355_{\pm 0.043}$	$0.280_{\pm 0.028}$	$0.294_{\pm 0.025}$	$0.323_{\pm 0.039}$	$0.442_{\pm 0.044}$	0.749
Loc$_f$ (F)	0.990	0.000	0.003	$0.404_{\pm 0.029}$	$\mathbf{0.545}_{\pm \mathbf{0.023}}$	$0.534_{\pm 0.023}$	$0.399_{\pm 0.015}$	$0.515_{\pm 0.012}$	$0.507_{\pm 0.010}$	0.600
Loc$_f$ (S)	0.993	0.000	0.002	$0.245_{\pm 0.015}$	$\mathbf{0.301}_{\pm \mathbf{0.006}}$	$0.223_{\pm 0.023}$	$0.211_{\pm 0.011}$	$0.286_{\pm 0.006}$	$0.269_{\pm 0.004}$	0.495
MHop (F)	0.855	0.345	0.182	$0.322_{\pm 0.026}$	$0.301_{\pm 0.023}$	$0.501_{\pm 0.026}$	$0.556_{\pm 0.026}$	$0.468_{\pm 0.029}$	$\mathbf{0.658}_{\pm \mathbf{0.034}}$	0.936
MHop (S)	0.770	0.323	0.219	$0.211_{\pm 0.024}$	$0.229_{\pm 0.037}$	$0.434_{\pm 0.048}$	$0.474_{\pm 0.025}$	$0.334_{\pm 0.025}$	$\mathbf{0.637}_{\pm \mathbf{0.034}}$	0.784
Att	0.862	0.328	0.174	$0.240_{\pm 0.027}$	$0.375_{\pm 0.031}$	$0.415_{\pm 0.051}$	$0.476_{\pm 0.041}$	$0.410_{\pm 0.021}$	$\mathbf{0.544}_{\pm \mathbf{0.060}}$	0.547

Table 2: Evaluation results in Macro-averaged F1 scores of the *OpenToM* dataset. Location subscripts, c and f, represents *coarse* and *fine* respectively. The capital *F* and *S* in the parenthesis represent *first-order ToM* and *second-order ToM*. The naive baselines include a random guess (Ran.) and a majority (Maj.) baseline. The finetuning baseline (FT.) is a Llama2-Chat 13B model finetuned following the configuration in Appendix D.3.

reasoning process. SimToM prompting is specifically designed to aid N-ToM tasks, which asks LLMs to first generate a character-centric narrative, $\mathcal{N}_c$, and then answer character-specific questions.

3.3 Overall Results

As all the *OpenToM* questions are formulated as binary or ternary classification tasks and considering that the labels are not uniformly distributed (Figure A3), we evaluate model performance using the macro-averaged F1 scores (referred to as F1 scores henceforth).

To evaluate the consistency of LLMs' performance, we randomly sample 50 narratives for each round of evaluation and repeat this process for 5 times for each model. We compute the mean and the standard deviation of the F1 scores, which are reported in Table 2 (See Table A8 for more detailed results. See Table A7 for the breakdown of LLMs' performances on MHop questions). Overall, we see that GPT-4-Turbo outperforms other models on Loc$_{coarse}$ (first-order), MHop, and Att questions by a large margin. However, we are surprised to see that Llama2-Chat-7B performs the best in answering second-order Loc$_{coarse}$. However, due to the high unfaithful rate shown in later studies (§4.1 and Table A9), achieving the highest score does not necessarily imply that Llama2-Chat-7B is more capable in N-ToM. In addition, it is interesting to see that, while GPT-4-Turbo leads in most question genres by a large margin, its capability of answering the Loc$_{fine}$ questions is not on par with Llama2-Chat-13B, 70B, or GPT-3.5-Turbo.

Through the fine-tuning model, it becomes evident that the Loc$_{coarse}$ and MHop questions are easier to learn, as their F1 scores improved dramatically. On the other hand, the Loc$_{fine}$ and Att questions pose greater challenges as the F1 score of the fine-tuned model only have limited improvement.

CoT prompting brings significant performance gains to all models on Loc$_{coarse}$ and MHop questions. However, the improvements in answering Att questions are marginal and the performance on Loc$_{fine}$ questions declines. In the case of SimToM prompting, the results for the Mixtral model are mixed. SimToM improves the f1 score of MHop questions, but its performance on other question types is either degraded or negligible. For GPT models, SimToM consistently brings performance gains in Loc$_{coarse}$ questions. However, for other question genres, the effect of SimToM is mixed.

In terms of the length of the narrative, results on *OpenToM*-L show that ToM in longer narratives are generally harder to trace. Please see Appendix D.5 for detailed results and analysis.

4 Detailed Result Analysis

To further investigate LLMs' N-ToM capabilities, we conduct in-depth analysis on LLMs' faithfulness in answering Loc$_{coarse}$ and Loc$_{fine}$ questions (§4.1), performance discrepancy of modeling the mental states of different character roles (§4.2), and lack of capability in modeling characters' mental state of the psychological world (§4.3).

4.1 Faithfulness in Loc Questions

As mentioned in §2.4, we create two types of Loc questions differ in granularity. In principle, Loc$_{coarse}$ serves as a prerequisite for answering Loc$_{fine}$ questions. For instance, if a person believes that the entity is not in its initial location (i.e. Loc$_{coarse}$), then they should maintain this belief when deducing its precise location (i.e. Loc$_{fine}$). We conduct two experiments to examine LLMs'

	Question	Mixtral		GPT-3.5-Turbo		GPT-4-Turbo		HL
		F1	ΔF1	F1	ΔF1	F1	ΔF1	
CoT	$\texttt{Loc}_c(F)$	0.784*	+0.272	0.587*	+0.148	**0.942***	+0.299	✔
	$\texttt{Loc}_c(S)$	0.539*	+0.245	0.457*	+0.134	**0.828***	+0.386	✘
	$\texttt{Loc}_f(F)$	0.301*	- 0.098	**0.469***	- 0.046	0.450*	- 0.057	✘
	$\texttt{Loc}_f(S)$	0.180*	- 0.031	**0.240***	- 0.046	0.187*	- 0.082	✘
	$\texttt{MHop}(F)$	0.610*	+0.054	0.547*	+0.079	**0.835***	+0.177	✔
	$\texttt{MHop}(S)$	0.551*	+0.077	0.414*	+0.080	**0.755***	+0.118	✔
	Att	0.519*	+0.043	0.446*	+0.036	**0.580***	+0.036	✘
SimToM	$\texttt{Loc}_c(F)$	0.414*	- 0.098	0.635*	+0.196	**0.838***	+0.195	✘
	$\texttt{Loc}_c(S)$	0.290	- 0.004	0.400*	+0.077	**0.685***	+0.243	✘
	$\texttt{Loc}_f(F)$	0.352*	- 0.047	**0.518***	+0.003	0.485*	- 0.022	✘
	$\texttt{Loc}_f(S)$	0.206*	- 0.005	**0.261***	- 0.025	0.217*	- 0.079	✘
	$\texttt{MHop}(F)$	0.650*	+0.094	0.536*	+0.068	**0.720***	+0.062	✘
	$\texttt{MHop}(S)$	0.514*	+0.040	0.350*	+0.016	**0.631***	- 0.006	✘
	Att	0.404*	- 0.072	0.416	+0.006	**0.488***	- 0.056	✘

Table 3: Macro F1 score of *OpenToM* dataset evaluated using CoT and SimToM prompting with relative performance gain, performance degradation, or equal performance (ΔF1 < 0.010). "*" indicates statistical significance under the Two-sample T test with a level of significance of $\alpha = 0.05$. The score of the best performing model on each task is bolded. HL (human level) indicates whether the performance of the best model is on par with human performance (within a margin of 0.050).

faithfulness[6] in answering the Loc questions. In the *Joint* approach, we present LLMs with $\texttt{Loc}_{coarse}$ which is immediately followed by $\texttt{Loc}_{fine}$ in the same session. In the *Separate* approach, we prompt LLMs with each Loc question individually.

We consider a model to be *Unfaithful* if it gives contradictory answers in the $(\texttt{Loc}_{fine}, \texttt{Loc}_{coarse})$ pair of questions. To quantify this, we compute the *Unfaithful Rate* for each model, which is the ratio of unfaithful pairs to the total number of pairs, as shown in Figure 4.

We see that each model's unfaithful rate is lower when answering first-order ToM questions. This is likely due to their relative simplicity comparing to the second-order questions. Further, we see that, for the GPT models, the *Joint* approach yields lower *Unfaithful Rate* than the *Separate* approach. This improvement may attribute to having access to the previous answer in the context. For Mixtral model, however, the same trend is only observed for the first-order questions. As delving into the reason behind this trend is beyond the scope of this paper, we leave it as future work. Detailed evaluation results are shown in Appendix D.6.

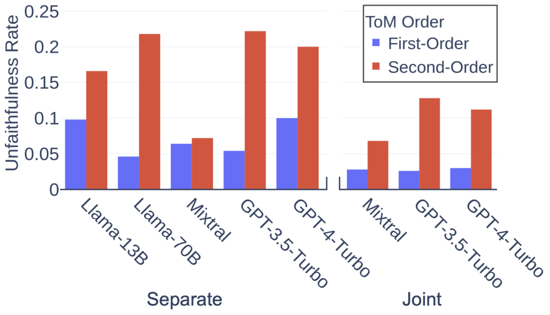


Figure 4: Faithfulness of LLMs in answering Loc questions. The x-axis displays the evaluation model and the y-axis displays the *Unfaithful Rate*.

4.2 Performance Gap in Character Roles

Previous works discovered that LLMs are more capable of answering questions related to the protagonist (Sap et al., 2022; Shapira et al., 2023a), which is likely due to them receiving more descriptions regarding their mental states (Grosz et al., 1995). In *OpenToM*, we consciously avoid such a reporting bias (§2.5). However, apart from the bias towards the protagonists, we observe that there exists another performance discrepancy in modeling the mind of characters of different roles. In *OpenToM*, the roles are *mover* and *observer*.

To demonstrate the performance gap between the *mover*'s and the *observer*'s perception, we compute difference in F1 scores between the models' performance on *mover*-centric questions and *observer*-centric questions (Table 4).

For second-order Loc questions, the majority of LLMs perform worse when modeling the *mover*'s mental state. This is likely due to the long distance between the description of the *mover*'s action and whether the *observer* witnessed the action (see an examples in Appendix E). Such distant information make it difficult for LLMs to establish a connection. Hence, deducing the *mover*'s perception of the *observer*'s mental state becomes more challenging.

For MHop questions, all LLMs perform better when modeling the *mover*'s mental states. When answering first-order Mhop questions, models' burden for deciding whether the *mover* observed their own action is alleviated. In the case of second-order MHop questions, the performance discrepancy is likely due to the explicit mention of the *mover*'s intention. These intentions often involve the *mover*'s perception of the consequences of their actions on the *observer*, which greatly reduces the complexity of modeling the *mover*'s perception of the *observer*'s mental state.

[6] We follow the definition of "faithfulness" from Jacovi and Goldberg (2020), which is *"the true reasoning process behind the model's prediction"*. We regard the model as unfaithful when its *true reasoning process* deviate from that of human.

	Llama-13B	Llama-70B	Mixtral	GPT-3.5T	GPT-4T
$\texttt{Loc}_c$ (F)	+0.169	+0.711	+0.606	+0.686	+0.464
$\texttt{Loc}_c$ (S)	+0.047	- 0.035	- 0.040	- 0.029	+0.129
$\texttt{Loc}_f$ (F)	+0.091	+0.104	+0.073	+0.097	+0.168
$\texttt{Loc}_f$ (S)	- 0.041	- 0.050	- 0.132	- 0.333	- 0.076
MHop (F)	+0.156	+0.250	+0.121	+0.320	+0.009
MHop (S)	+0.029	+0.176	+0.120	+0.143	+0.008

Table 4: Relative performance gap between the *mover* and the *observer* in answering *OpenToM* questions.

Erroneous Correlation: *Mover*'s Personality ~ *Observer*'s Attitude

🍦 : Vanilla Prompt 🔗 : CoT Prompt

🪪 : SimToM Prompt ❓ : Self-Ask Prompt

	Results on Neutral Attitude					
	Mixtral		GPT-3.5-Turbo		GPT-4-Turbo	
	Pos	Neg	Pos	Neg	Pos	Neg
🍦	1.000	0.759	1.000	0.844	1.000	0.796
🔗	0.944	0.909	1.000	0.886	0.857	0.758
🪪	1.000	0.727	1.000	0.771	1.000	0.759
❓	1.000	0.838	1.000	0.864	0.938	0.818
	Results on Positive Attitude					
	Mixtral		GPT-3.5-Turbo		GPT-4-Turbo	
🍦	1.000		0.926		1.000	
🔗	1.000		0.904		1.000	
🪪	1.000		0.920		0.957	
❓	1.000		0.938		1.000	

Table 5: Proportion of mistakenly classified *Neutral* (top) and *Positive* (bottom) Att questions that are correlated to the *mover*'s personality. For *Neutral* Att questions, we show the correlation for erroneous *positive* (Pos) and *negative* (Neg) predictions separately. For *positive* Att questions, we show the correlation for erroneous *negative* predictions.

4.3 Social Commonsense and Attitude

GPT-4-Turbo outperforms other models on MHop questions by a large margin (Table 2, 3, and A7), demonstrating its capability in reasoning using social commonsense. However, other LLMs' performance on MHop questions show that they are lacking in this regard.

As all LLMs performed poorly on Att questions, we additionally tested Self-Ask prompt (Appendix D.2), which asks LLMs to deduce the final answer by explicit proposing and answering series of follow-up questions (Press et al., 2023). While Self-Ask prompting improves the F1 score of LLMs (Table A10), it is still far from human performance, demonstrating LLMs' lack of N-ToM capabilities in perceiving characters' psychological states. By in-depth analysis on the Att answers from Mixtral, and the GPT models, we find two modes or error: low recall in (1) identifying *neutral* attitude and (2) identifying *positive* attitude.

Both of the aforementioned error modes can be attributed to LLMs' erroneous correlation between the *mover*'s personality trait and the *observer*'s attitude. In Table 5, we compute the proportion of error cases that are correlated to character's personality. Specifically, we regard the error and the personality as correlated if a mistaken prediction matches the character's personality. For instance, across all prompting methods, **more than 95% of the movers in narratives where GPT-4-Turbo mistakenly identify a *positive* attitude to be *negative* have an *inconsiderate* or *negativistic* personality** (bottom right column in Table 5).

As discussed in §2.5, a *considerate mover* in *OpenToM* story does not necessarily take actions that are benign to the *observer*. Therefore, LLMs are doomed to fail when using such a spurious correlation. See Appendix D.7 for detailed results.

5 Related Works

Neural ToM Some studies argued that LLMs like GPT-4 possess N-ToM capabilities (Bubeck et al., 2023; Kosinski, 2023). This claim was later rebutted by Shapira et al. (2023a) and Ullman (2023), who both demonstrated that LLMs lack robust N-ToM capabilities. To tackle N-ToM, a line of work used partially observable Markov decision process (Nguyen et al., 2023). Others proposed prompting techniques (Wilf et al., 2023) or neuro-symbolic approaches (Ying et al., 2023; Sclar et al., 2023). We direct readers to Ma et al. (2023b) for a comprehensive survey on N-ToM.

ToM Benchmarks Based on the *Sally-Anne Test* (Baron-Cohen et al., 1985) and bAbi (Weston et al., 2016), Grant et al. (2017) constructed the ToM-bAbi dataset for false belief, which was later improved by Le et al. (2019) into the ToMi dataset. Based on ToMi, researchers proposed T4D (Zhou et al., 2023a), which targets N-ToM for assistant agent, and Hi-ToM (Wu et al., 2023b), which focuses on higher-order N-ToM. Other human ToM tests such as the *Smarties Test* (Gopnik and Astington, 1988), and the *Faux Pas Test* (Baron-Cohen et al., 1999) were also used for studying N-ToM, leading to datasets such as ToMChallenges (Ma et al., 2023a), BigToM (Gandhi et al., 2023), Adv-CSFB (Shapira et al., 2023a), and FauxPas-EAI (Shapira et al., 2023b). However, existing N-ToM benchmarks are either limited in size, contain artificial narratives, or lack diversity in their questions posed. Jones et al. (2023) constructed EPITOME, which contains human ToM tests that go beyond false-belief. Researchers also put efforts in evaluating LLMs' N-ToM capabilities in dialogues, which resulted in benchmarks such as G-DRAGON (Zhou et al., 2023b), FANToM (Kim et al., 2023c), and SOTOPIA (Zhou et al., 2023c).

ToM and Social Commonsense Sap et al. (2022) showed that LLMs' lack of understanding of social norms using SocialIQA (Sap et al., 2019). The FauxPas-EAI dataset (Shapira et al., 2023b) was dedicated to evaluating LLMs' understanding of social commonsense. Efforts were also made to construct knowledge graphs for social commonsense and N-ToM (Wu et al., 2023a).

6 Future Directions

Faithfulness Our study of LLMs' performance on $\texttt{Loc}_{coarse}$ and $\texttt{Loc}_{fine}$ reveals that all LLMs lack faithfulness when answering N-ToM questions. We recognize that improving LLMs' faithfulness is a challenging task in numerous domains (Jacovi and Goldberg, 2020). Here we propose potential remedies specifically targeting N-ToM tasks. Following the findings in §4.1, neuro-symbolic systems can be potentially deployed to enforce faithfulness in reasoning about the characters' mental state of the physical world. Gao et al. (2023) proposes PAL, which represent reasoning problems with programming language and obtain a deterministic solution using code interpreter. Lyu et al. (2023) combined PAL with CoT and achieved accurate and more faithful reasoning chains.

Performance Gap Between Roles In *OpenToM* narrative, we propose two roles, namely a *mover* and an *obsever*. Our study in §4.2 unveils LLMs' performance discrepancies in N-ToM between the character roles and analyzes the underlying reasons. In reality, a narrative contain roles well beyond two. To account for the difference in the ToM reasoning process of different roles, a role-aware reasoning framework is needed. Specifically, given an event and a group of characters, the framework needs to first identify the role that each character plays in the event and then conduct ToM reasoning accordingly.

Social Commonsense and Psychological N-ToM Analysis in §4.3 shows that most LLMs are incapable of incorporating social commonsense. Further, we find that LLMs' performance on `Att` questions is limited by their inability to determine the information that is accessible to a certain character and using such information to reason about characters' emotions (Table 5). Hence, an efficient framework for documenting character-centric world state is needed. Further, as discussed in Zhan et al. (2023), people's attitude in reality is complicated and multifaceted. Therefore, to create a generalizable system capable of emotion deduction, instantiating the emotion deduction process similar to Wu et al. (2023a) is a potential solution.

Neural Theory-of-Mind N-ToM in general is a crucial cognitive capability that a helpful intelligent agent must possess. In the context of human psychology, a lack of ToM capabilities is oftentimes associated with developmental conditions such as Autism Spectrum Disorder (ASD) (Baron-Cohen et al., 1985). Therefore, as LLMs being developed and deployed as assistant agents, it is critical to understand their N-ToM capabilities and develop methods to grant them robust N-ToM reasoning capabilities. LLMs could especially benefit from N-ToM in the following fields: (1) Educational LLM where a helpful assistant agent must be able to accurately model the mental state of the students to be able to provide efficient and precise guidance; (2) Negotiating LLM where understanding the mental states such as the intention, desire, and mood of the opponent is critical when planning negotiation strategies; (3) Mental Health LLM where assistant agent must comprehend the mental state and being empathetic with the patient to be able to provide meaningful help.

7 Conclusion

We introduce *OpenToM*, a comprehensive N-ToM benchmark featuring long narratives with realistic characters and events, and a diverse range of questions that cover both physical and psychological aspects of N-ToM. Our evaluation of LLMs' N-ToM capabilities on *OpenToM* reveals that while state-of-the-art LLMs perform well on some N-ToM tasks, they are still far from human-level performance on tasks requiring emotion deduction.

Limitations

Limitations of *OpenToM* are as follows:

Limited LLMs Due to the constraint of computing resources and budget, we only evaluated *OpenToM* benchmark on a subset of available LLMs. While we believe that the selected LLMs are representative of the current state-of-the-art of their categories (Llama2-Chat for open-source LLMs, GPT-3.5-Turbo and GPT-4-Turbo for close-source LLMs, and Mixtral-8x7B-Instruct for Mixture-of-Expert LLMs), we acknowledge that there are other LLMs that could potentially perform better on *OpenToM*. Further, we only examine the zero-shot performance of LLMs, future studies should test models' N-ToM capabilities under a few-shot setting.

Potential Biases in *OpenToM* Narratives The drafts of *OpenToM* narratives are composed using LLMs. Although recent studies have shown that LLMs are capable of producing high-quality benchmarks (Efrat and Levy, 2020; Perez et al., 2022a,b; Hartvigsen et al., 2022; West et al., 2023), we acknowledge that the texts generated by LLMs could contain biases and lack lexical diversity.

Limited Scope in Character Emotion In *OpenToM* benchmark, we construct questions regarding character's emotion (e.g. attitude). To reduce the subjectivity, we purposely design the stories in a way that the character's emotion can be directly deduced from an action that happens in a short time frame. In reality, human emotions are often complex, multifaceted, and may depend on multiple events through a prolonged period of time.

Limited Narrative Order All *OpenToM* narratives are linear narratives that strictly follow chronological order, which alleviate LLMs' burden to comprehending the order of the events. Future studies can consider constructing *OpenToM* narratives with non-linear order to further challenge LLMs' narrative understanding and N-ToM capabilities.

Ethics Statement

The drafts of *OpenToM* narratives are generated using GPT-3.5-Turbo and GPT-4-Turbo. Although we did not identify any harmful or violent content in the *OpenToM* narratives, it is worth noting that previous studies have observed instances where LLMs produced unexpected results. Therefore, we encourage future studies to also be cautious when employing similar data generating strategies. Further, the *OpenToM* dataset is annotated by graduate students studying computer science. The similar background of annotators may introduce bias in the annotation process.

Acknowledgements

We thank Lin Gui and Yuchen Si for the valuable discussions. This work was supported in part by the UK Engineering and Physical Sciences Research Council (EPSRC) through an iCASE award with Huawei London Research Centre and a Turing AI Fellowship (grant no. EP/V020579/1, EP/V020579/2).

References

Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. 2011. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society*, volume 33.

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.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Avia Efrat and Omer Levy. 2020. The turking test: Can language models understand instructions? *arXiv preprint arXiv:2010.11982*.

Allen Frances. 1981. Disorders of personality: Dsm-iii, axis ii. *American Journal of Psychiatry*, 138(10):1405–a.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding social reasoning in language models with language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.

Alison Gopnik and Janet W Astington. 1988. Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child development*, pages 26–37.

Erin Grant, Aida Nematzadeh, and Thomas L Griffiths. 2017. How can memory-augmented neural networks pass a false-belief task? In *CogSci*.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Cameron Robert Jones, Sean Trott, and Ben Bergen. 2023. Epitome: Experimental protocol inventory for theory of mind evaluation. In *First Workshop on Theory of Mind in Communicating Agents*.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023a. SODA: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, Singapore. Association for Computational Linguistics.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023b. 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.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023c. 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.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. Technical report, Stanford University, Graduate School of Business.

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.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. *arXiv preprint arXiv:2301.13379*.

Xiaomeng Ma, Lingyu Gao, and Qihui Xu. 2023a. Tomchallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. *arXiv preprint arXiv:2305.15068*.

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

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Jason P Mitchell, Jasmin Cloutier, Mahzarin R Banaji, and C Neil Macrae. 2006. Medial prefrontal dissociations during processing of trait diagnostic and nondiagnostic person information. *Social Cognitive and Affective Neuroscience*, 1(1):49–55.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

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

Dung Nguyen, Phuoc Nguyen, Hung Le, Kien Do, Svetha Venkatesh, and Truyen Tran. 2023. Memory-augmented theory of mind network. *arXiv preprint arXiv:2301.06926*.

Kristine H Onishi and Renée Baillargeon. 2005. Do 15-month-old infants understand false beliefs? *science*, 308(5719):255–258.

OpenAI. 2022. https://openai.com/blog/chatgpt. Accessed: 2024-01-05.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022a. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2022b. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*.

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

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Giuseppe Riva, John A Waterworth, Eva L Waterworth, and Fabrizia Mantovani. 2011. From intention to action: The role of presence. *New Ideas in Psychology*, 29(1):24–37.

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*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Philipp Schmid, Omar Sanseviero, Pedro Cuenca, and Lewis Tunstall. 2023. https://huggingface.co/blog/llama2#how-to-prompt-llama-2. Accessed: 2023-11-18.

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.

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. *ArXiv*, abs/2305.14763.

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*, pages 10438–10451, Toronto, Canada. Association for Computational Linguistics.

The Mistral AI Team. 2023. https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1. Accessed: 2024-01-05.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, et al. 2023. The generative ai paradox:" what it can create, it may not understand". *arXiv preprint arXiv:2311.00059*.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. In *4th International Conference on Learning Representations, ICLR 2016*.

Alex Wilf, Sihyun Shawn Lee, Paul Pu Liang, and Louis-Philippe Morency. 2023. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. *arXiv preprint arXiv:2311.10227*.

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

Jincenzi Wu, Zhuang Chen, Jiawen Deng, Sahand Sabour, and Minlie Huang. 2023a. Coke: A cognitive knowledge graph for machine theory of mind. *arXiv preprint arXiv:2305.05390*.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023b. 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.

Lance Ying, Katherine M Collins, Megan Wei, Cedegao E Zhang, Tan Zhi-Xuan, Adrian Weller, Joshua B Tenenbaum, and Lionel Wong. 2023. The neuro-symbolic inverse planning engine (nipe): Modeling probabilistic social inferences from linguistic inputs. *arXiv preprint arXiv:2306.14325*.

Hongli Zhan, Desmond Ong, and Junyi Jessy Li. 2023. Evaluating subjective cognitive appraisals of emotions from large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14418–14446, Singapore. Association for Computational Linguistics.

Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, et al. 2023a. How far are large language models from agents with theory-of-mind? *arXiv preprint arXiv:2310.03051*.

Pei Zhou, Andrew Zhu, Jennifer Hu, Jay Pujara, Xiang Ren, Chris Callison-Burch, Yejin Choi, and Prithviraj Ammanabrolu. 2023b. I cast detect thoughts: Learning to converse and guide with intents and theory-of-mind in dungeons and dragons. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11136–11155, Toronto, Canada. Association for Computational Linguistics.

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

A *OpenToM* Construction

A.1 Disambiguated Prompt for Narrative Generation

In the ToMi dataset, the narrative contains numerous ambiguities. Take the following ToMi narrative as an example:

ToMi Narrative Example

1 Oliver entered the dining room.
2 Logan entered the dining room.
3 Jack entered the dining room.
4 **The stockings is in the drawer.**
5 Jack hates the slippers
6 **Oliver exited the dining room.**
7 **Logan moved the stockings to the crate.**
8 Jack exited the dining room.
9 Logan exited the dining room.
10 Jack entered the hallway.

Question: Where will Oliver look for the stockings?

The key ambiguities are marked with bold text. In line 4, the narrative only states that the entity, *stockings*, is in the drawer. However, it neglects the characters' awareness of the entity's location. Therefore, the above question can be answered with either *the drawer* in the case where Oliver noticed the stockings, or *unknown* in the case where Oliver is unaware of the stockings.

In lines 6-7, Oliver left the dining room, and Logan moved the stockings. However, it is not guaranteed that Logan would lose sight of the dining room once exit. For instance, objects in the dining room could still be visible in the living room if there is no physical barrier separating the spaces. Therefore, knowing that Oliver has left the dining room is insufficient to deduce whether Oliver could observe Logan's action.

Further, the information in Line 3, 5, 8, 10 about Jack is completely irrelevant to the progression of the story. Le et al. (2019) added such distracting information to mitigate potential spurious correlations in the original ToM-bAbi dataset (Grant et al., 2017; Nematzadeh et al., 2018). However, such irrelevant information could potentially distract LLMs from performing the ToM task and hence underestimate their ToM capabilities.

To address such ambiguities, we remove the distracting information and make each character's perception explicit in the *OpenToM* plot. See below for an example of *OpenToM* story generation prompt (disambiguated information in bold text). **We wish to emphasize that part of the contents of the *OpenToM* plots are derived from the ToMi dataset** (Le et al., 2019).

Prompt Example

Plot:
Paragraph 1: Mason hates grapes. Samuel hates grapes.

Paragraph 2: Mason entered the den. Samuel entered the den. **Both Mason and Samuel noticed that the grapes is in the bucket in the den.** Samuel exited the den.

Paragraph 3: Mason is an inconsiderate person. Mason hates grapes. Therefore, Mason moved the grapes to a neighbor's house in order to get rid of them. **Samuel did not witness Mason's action.**

Write a 200-word, 3-paragraph story according to the plot. Do not depict Samuel's attitude towards Mason's action. End the story immediately after the main event.

A.2 Detailed Description of the Character Personification Process

Character Personification In established N-ToM benchmarks such as ToMi and its variants (Le et al., 2019; Wu et al., 2023b; Zhou et al., 2023a), characters do not possess meaningful personal preferences or personality traits. As a result, their actions lack inherent motivation. In *OpenToM*, we randomly picked two contrasting personalities, namely "*considerate*" and "*inconsiderate*", from the 24 personality traits defined in (Mitchell et al., 2006). We additionally include a "*negativistic*" personality to make the story more interesting (Frances, 1981). Below are brief descriptions of each personalities:

- **Considerate** *mover* acts to ensure the comfort of the *observer*.
- **Inconsiderate** *mover* acts to make themselves feel comfortable.
- **Negativistic** *mover* acts to make the *observer* uncomfortable.

Intention and Enaction Based on the *mover*'s personality and the *observer*'s preferences, we generate both the character's intention and their subsequent actions (Appendix A.1). In such a way, the *mover*'s action and the movement of the entity are anchored in the *mover*'s intention.

Plot Construction Each of the *OpenToM* narratives is generated by prompting GPT-3.5-Turbo[7] with a story plot[8] (see Appendix A.1 for a prompt example). In *OpenToM* plot, we sequentially introduce the characters' preferences towards the *entity*, a scenario where the two characters meet and how they encounter the *entity*, and the emergence of the *mover*'s intention and the subsequent action towards the *entity*.

Following Kim et al. (2023c,a), we first assign names to the *mover* and the *observer* by random sampling from the Top-1K most frequently used names in the US SSN database to mitigate potential biases in character naming. Subsequently, for each character, we first randomly sample the personality of the *observer* (*trait_o*), or the mover (*trait_m*) from the set, {considerate, inconsiderate, negativistic}. Next, we generate the *mover*'s preference (*pref_m*), the *observer*'s preference (*pref_o*), the *mover*'s belief of the *observer*'s preference (*pref_mo*), the *observer*'s belief of the *mover*'s preference (*pref_om*), the *mover*'s intention (*intent*), and the *mover*'s enaction (*action*) using Algorithm 1.

We use GPT-3.5-Turbo as our intent generator (INTENTGENERATOR). We customize the prompt for each of the three personality traits. To give examples of the prompt, we again use *Sally (observer)* and *Anne (mover)* as the characters and the *rubber duck* as the entity-of-interest. The intent generation prompts are presented as follows:

Prompt for Intention and Action Generation

{Mover's Personality} {Mover's Preference Perception}. Therefore, {Mover's Initial Intention}. What would be Anne's 3 most likely action and intention towards the rubber duck? Answer with the following template:
1. Anne would move the rubber duck to

[7] We use the 1106 checkpoint of the GPT-3.5-Turbo model through Microsoft Azure OpenAI service. All *OpenToM* narratives are generated in December 2023.

[8] We also tested with GPT-4-1106, which produces narratives of similar quality. Hence we went for GPT-3.5-Turbo for its lower cost.

Algorithm 1: Functions for Preference and Intention Generation

```
Function assignPref(trait_m, pref_mo):
    pref_o ← PREFSAMPLER(observer)
    pref_mo ← PREFSAMPLER(mover, observer)
    pref_om ← PREFSAMPLER(observer, mover)
    if trait_m = Negativistic then
        pref_m = ¬(pref_mo)
    else
        pref_m = PREFSAMPLER(mover)

    return pref_o, pref_m, pref_om, pref_mo

Function assignIntent(trait_m, pref_m, pref_o):
    if trait_m = Considerate then
        latent_pref = pref_om
    else if trait_m = Inconsiderate then
        latent_pref = pref_m
    else if trait_m = Negativistic then
        latent_pref = pref_m

    intent, action = INTENTGENERATOR(latent_pref)

    return intent, action
```

> {location} in order to {intention}
> 2. Anne would move the rubber duck to {location} in order to {intention}
> 3. Anne would move the rubber duck to {location} in order to {intention}

We fill in the above template based on the *mover*'s personality and their belief in the *observer*'s perference. Table A1 are a list of descriptions we used to complete the template.

Final Intention and Enaction Selection Notice that for each of the prompts listed above, we specifically ask LLMs to provide 3 candidate intention and enaction pairs. To produce the final intention and its corresponding enaction. We prompt LLMs one more time in the same session to pick the best intention and enaction from the candidates. The prompt we used is as follows:

> **Intention & Encation Selection**
>
> Of the potential intentions, which one do you think is `true_sentiment`? Answer with the original sentence. Do not add any additional words.

where the `true_sentiment` is filled according to the *mover*'s personality trait:

- Considerate → "*the most considerate*"
- Inconsiderate → "*the most selfish*"
- Negativistic (Show off) → "*the most ostentatious*"
- Negativistic (Get rid) → "*the most adversarial*"

A.3 Detailed Description of the Data Annotation Process

In *OpenToM*, the question answers are produced in two ways: human annotation and rule-based generation (Figure 2). For all the $\texttt{Loc}_{coarse}$ questions, `MHop` questions regarding *accessibility*, and `Att` questions, the answers are annotated by graduate students in a prestigious UK university. As the ToM questions in *OpenToM* are rudimentary for human, we do not provide any specific instruction to the annotators. The content shown in Figure A4 is the complete information that an annotator receives. Therefore, the information that data annotators possess matches the information we provide to LLMs during evaluation.

Answers to the $\texttt{Loc}_{fine}$ questions are generated according to the human annotation of the corresponding $\texttt{Loc}_{coarse}$ questions and cached container information in *OpenToM* story plot. For instance, if the annotation to the $\texttt{Loc}_{coarse}$ question is *False*, then the answer to the $\texttt{Loc}_{fine}$ question is assigned to be the *new container* (the container that the entity is moved to), which is conveniently cached in the *OpenToM* story plot.

Answers to the `MHop` questions regarding *fullness* are generated using first order logic based on $\texttt{Loc}_{coarse}$ annotations. Recall that in $\texttt{Loc}_{coarse}$, we ask the following question

From {{character}}'s perspective, is the {{entity}} still in its initial location by the end of the story?

Notice that the $\texttt{Loc}_{coarse}$ question is equivalent to "*Is {{character}}* ***aware of*** *the {{entity}}'s movement*?" (in the case of first-order ToM) or "*Does {{character A}} thinks that {{character B}} is* ***aware of*** *the {{entity}}'s movement?* (in the case of second-order ToM). Knowing the answer to $\texttt{Loc}_{coarse}$ questions is a direct prerequisite for answering *fullness* questions (see Figure A1). This allows us to conveniently employ the following rules to automatically deduce the answers to the

	{Mover's Personality}
Personality	Description
Considerate	Anne is a considerate person.
Inconsiderate	Anne is an inconsiderate person.
Negativistic	Anne is a negativistic person.

	{Mover's Preference Preception}	{Mover's Initial Intention}
Considerate	Although Anne hates rubber duck, she knows that Sally likes them.	Anne wants to make it more accessible to Sally.
	Although Anne likes rubber duck, she knows that Sally hates them.	Anne wants to make it less accessible to Anne.
	Anne knows that both Sally and herself hate rubber duck.	Anne wants to make it less accessible.
	Anne knows that both Sally and herself like rubber duck.	Anne wants to make it more accessible for both of them.
Inconsiderate	Anne likes rubber duck.	Anne wants to make it more accessible to herself.
	Anne hates rubber duck.	Anne wants to make it less accessible.
Negativistic	Anne thinks that Sally likes rubber duck.	Anne wants to get rid ot the rubber duck.
	Anne thinks that Sally hates rubber duck.	Anne wants to show off the rubber duck.

Table A1: Description of the *mover*'s personality, preference perception, and initial intention. These descriptions are used to fill in the template for intent and action generation.

fullness questions:

$$\begin{aligned}
&\forall c \forall p: \\
&\texttt{isAware}(c) \land \texttt{moveTo}(p) \rightarrow \texttt{moreFull}(p) \\
&\texttt{isAware}(c) \land \texttt{takeFrom}(p) \rightarrow \texttt{lessFull}(p) \\
&\neg\texttt{isAware}(c) \rightarrow \texttt{equallyFull}(p)
\end{aligned}$$

where c represents a *character* and p represents a *container*. Answer to the $\texttt{isAware}(\cdot)$ part of the clause is the same as the answer to the $\texttt{Loc}_{coarse}$ questions. Answer to the $\texttt{moveTo}(\cdot)$ or $\texttt{takeFrom}(\cdot)$ part of the clause is obtained using cached information from the *OpenToM* plot (A.1).

A.4 Keyword Revision for Mitigating Spurious Correlation

To mitigate the surface-level cues in the *OpenToM* narrative, we identify the following keywords that are likely to be directly associated with the answers of the MHop and Att questions. We identify the narratives that contain such keywords and manually revise the stories. The keywords are listed as follows:

Cue 1: Positive Attitude		
gratitude	smile	thoughtful
considerate nod	appreciation	kindness
gesture	delight	pleased
appreciating		

Cue 2: Negative Attitude		
upset	confusion	bewilderment
disappointment	astonishment	

Cue 3: Direct Description of Accessibility		
more accessible	accessible	less accessible
inaccessible	out of reach	

Table A2: Keywords that are likely to be directly associated with the answers of the MHop and Att questions.

A.5 Demonstration of Spurious Correlation Mitigation

As a running example, consider the scene between Sam and Amy depicted in Figure 1. In this example, Amy mistakenly believe that Sam hates rubber duck. Now we show how the spurious relationships are avoided by adding a false impression on the *mover*'s perception of the *observer*'s preference.

Spurious Cue 1: *Is there a causal relation between intention and attitude?*

The first spurious correlation arises due to the model being incapable of taking the *observer*'s perspective and mistakenly interprets the *mover*'s intention as observed information. If the *observer* were to find out the true intention of the *mover*, then it will undoubtedly be a salient causal factor of the *observer*'s attitude. However, deriving the true intention is a challenging problem due to subjectivity and the enormous search space involved. Therefore, to mitigate such a spurious correlation, we wish to create scenarios where a good intention leads to negative attitude or vice versa. This can be done by exploiting the *mover*'s false belief in the *observer*'s preference for a particular entity.

For instance, in Figure 1(B), Amy mistakenly believes that Sam hates rubber duck. As a considerate person, Amy forms a benign intention, which is to spare Sam from seeing the rubber duck. This intention enacted Amy to move the rubber duck into her own backpack. However, since Sam is unaware of Amy's true intention, he only observes that Amy has taken that rubber duck away, likely resulting in a *negative* attitude. Therefore, when there is a false impression in play, a benign intention does not necessarily lead to *positive* attitude.

Spurious Cue 2: *Is there a causal relation between personality and attitude?*

In many instances, the *OpenToM* narratives explicitly portray the *mover*'s personality trait (as seen in Figure 1(B), *"Amy is a considerate person"*). To prevent the model from taking a shortcut by deducing the *observer*'s attitude solely based on the *mover*'s personality, we aim to intervene such a spurious correlation by creating scenarios where a *mover* with a positive trait leads to the *observer* having a negative attitude or vice versa. This can be effectively done also by leveraging the *mover*'s false belief regarding the *observer*'s preference for a particular entity.

For instance, in Figure 1(B), Amy mistakenly believes that Sam dislikes the rubber duck. As a considerate person, Amy naturally wants to keep the rubber duck out of Sam's sight. However, due to this false belief, Amy ends up taking away something which Sam actually desires. Therefore, a positive personality could lead to the *observer* developing a *negative* attitude.

OpenToM Question Statistics			
Question Types	1st-Order	2nd-Order	Total
Loc$_{coarse}$	1192	1192	2384
Loc$_{fine}$	2384	1192	3576
MHop	3576	3576	7152
Att	596	–	596
Total	7748	5960	13708

OpenToM-L Question Statistics			
Question Types	1st-Order	2nd-Order	Total
Loc$_{coarse}$	200	200	400
Loc$_{fine}$	400	200	600
MHop	600	600	1200
Att	100	–	100
Total	1300	1000	2300

Table A3: Statistics of the number of questions in the *OpenToM* and *OpenToM*-L dataset.

B Demonstration of Multi-hop Questions

An illustration of the reasoning tree employed for answering *Fullness* questions is shown in Figure A1, while a depiction of the reasoning tree utilized for answering *Accessibility* questions is shown in Figure A2. It is worth noting that, in order to answer such questions, one must draw upon social commonsense (e.g., taking items from another person's backpack without permission is not appropriate).

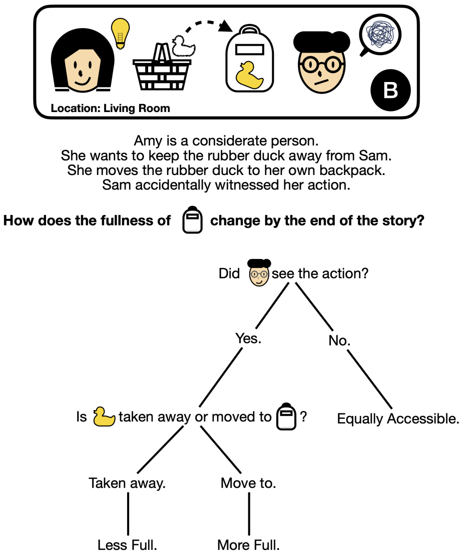

B
Location: Living Room
Amy is a considerate person.
She wants to keep the rubber duck away from Sam.
She moves the rubber duck to her own backpack.
Sam accidentally witnessed her action.
How does the fullness of change by the end of the story?
Did see the action?
Yes.
No.
Is taken away or moved to ?
Equally Accessible.
Taken away.
Move to.
Less Full.
More Full.

Figure A1: Illustration of the reasoning tree employed to answer the *Fullness* questions.

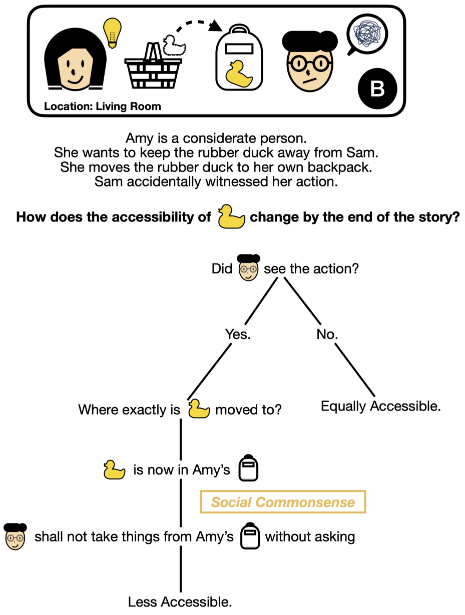

Location: Living Room
B
Amy is a considerate person.
She wants to keep the rubber duck away from Sam.
She moves the rubber duck to her own backpack.
Sam accidentally witnessed her action.
How does the accessibility of change by the end of the story?
Did see the action?
Yes.
No.
Where exactly is moved to?
Equally Accessible.
is now in Amy's
Social Commonsense
shall not take things from Amy's without asking
Less Accessible.

Figure A2: Illustration of the reasoning tree employed to answer the *Accessibility* questions.

C The *OpenToM* Dataset

Statistics Table A3 shows the statistics of the question types in the *OpenToM* dataset, while Figure A3 depicts the label distribution of each question types.

Data Annotation Platform In this study, we use doccano as the data annotation platform (Nakayama et al., 2018). As all the questions are either binary or ternary classification tasks, we use the *Text Classification* interface. Figure A4 shows the annotation interface for labeling the attitude (Att) questions. The interface for labeling the other question types are the same, except for the label space.

Results of Inter-Annotator Agreement The detailed scores of the inter-annotator agreement are shown in Table A4. The scores are computed as the arithmetic mean of the pairwise agreement scores amongst the three annotators.

Question Type	Accuracy Score	F1 Score
Location (FO)	0.993	0.990
Location (SO)	0.993	0.993
Multihop (FO)	0.873	0.855
Multihop (SO)	0.927	0.770
Attitude	0.870	0.862

Table A4: Inter-annotator agreement scores. The scores are computed as the arithmetic mean of the pairwise agreement scores amongst the three annotators.

Ambiguity in Location Granularity Narratives in *OpenToM* contain location information of various levels of granularity. For example, in the story shown in Figure 1, there are two levels of location. At a room-level, the rubber duck is moved from the front yard (plot A) to the living room (plot B), while at a container-level, the rubber duck is transferred from a bucket to Amy's backpack. In the *OpenToM* stories, granularity can extend beyond two levels for locations (e.g. movements between different houses). Therefore, judging model's response solely based on explicit location information is difficult.

In addition, deducing the location of an entity involves multi-hop reasoning. It can be decomposed into *(1) Is the entity in its initial location?*, and *(2) What is the initial/final location exactly*? While the first question is seemingly simpler, it still demands the understanding of another character's perspective to answer correctly.

D Further Experimental Details

D.1 Details of the Baseline Models

We evaluate the *OpenToM* tasks using 6 representative LLMs, namely the Llama2-Chat models (7B, 13B, and 70B) (Touvron et al., 2023), the Mixtral-8x7B-Instruct model (Jiang et al., 2024), and the GPT-3.5-Turbo and GPT-4-Turbo. All of the models use decoder-only Transformer architecture (Vaswani et al., 2017).

Llama2-Chat is a Llama2 model optimized using Reinforcement Learning From Human Feedback (RLHF) for dialogue (Ouyang et al., 2022; Touvron et al., 2023). We evaluate *OpenToM* dataset on the 7B, 13B, and 70B checkpoints.

Mixtral-Instruct is a Sparse Mixture-of-Expert (SMoE) model optimized with Direct Policy Optimization (DPO) for instruction following (Rafailov et al., 2023; Jiang et al., 2024).

GPT-3.5-Turbo and **GPT-4-Turbo** are models of the GPT family, both of which are optimized using RLHF for instruction following (OpenAI, 2022, 2023).

In addition to zero-shot prompting, we also finetuned Llama2-Chat 13B models using LoRA to serve as a finetuning baseline (Hu et al., 2021; Mangrulkar et al., 2022). See Appendix D.3 for the

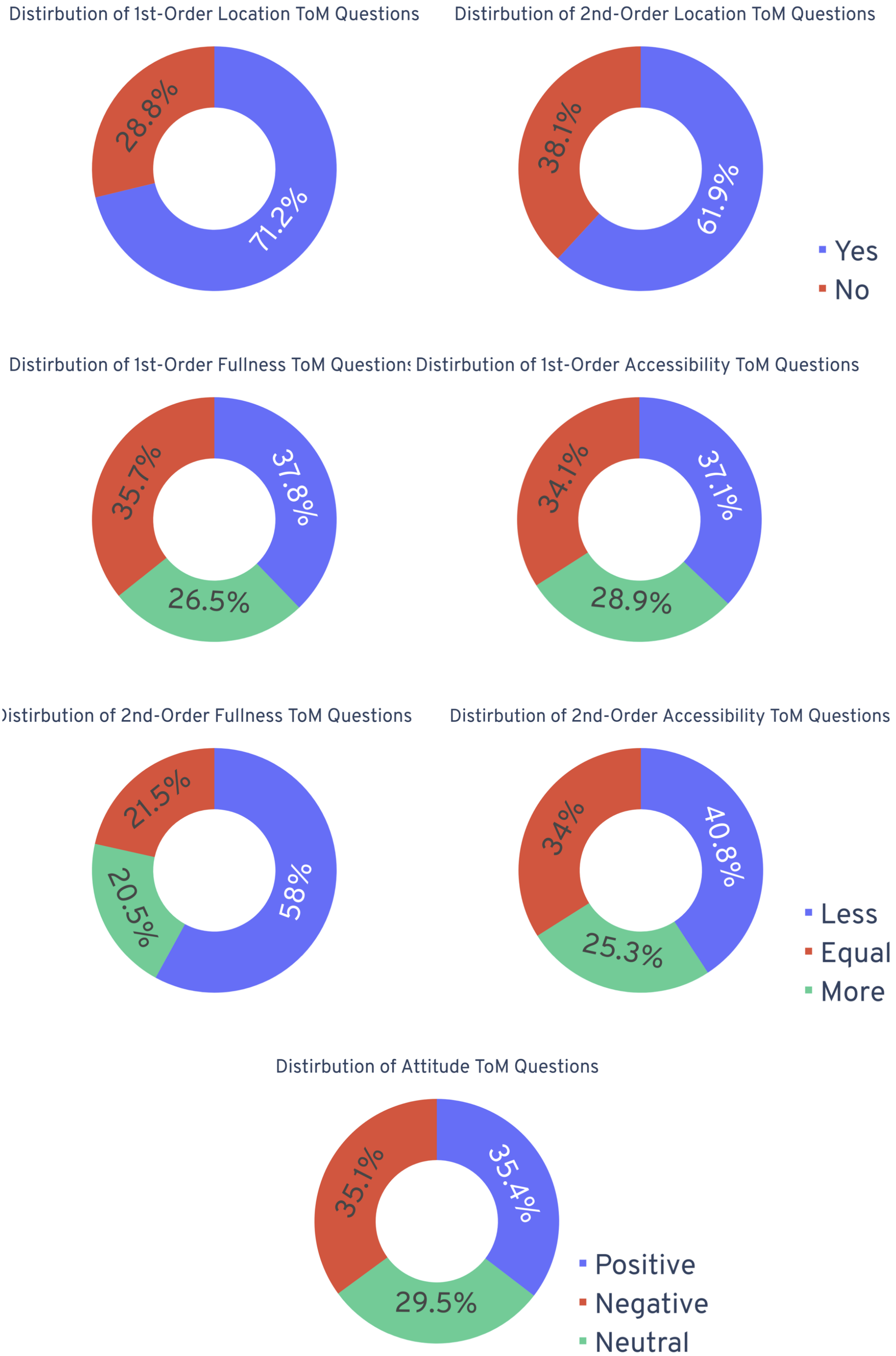
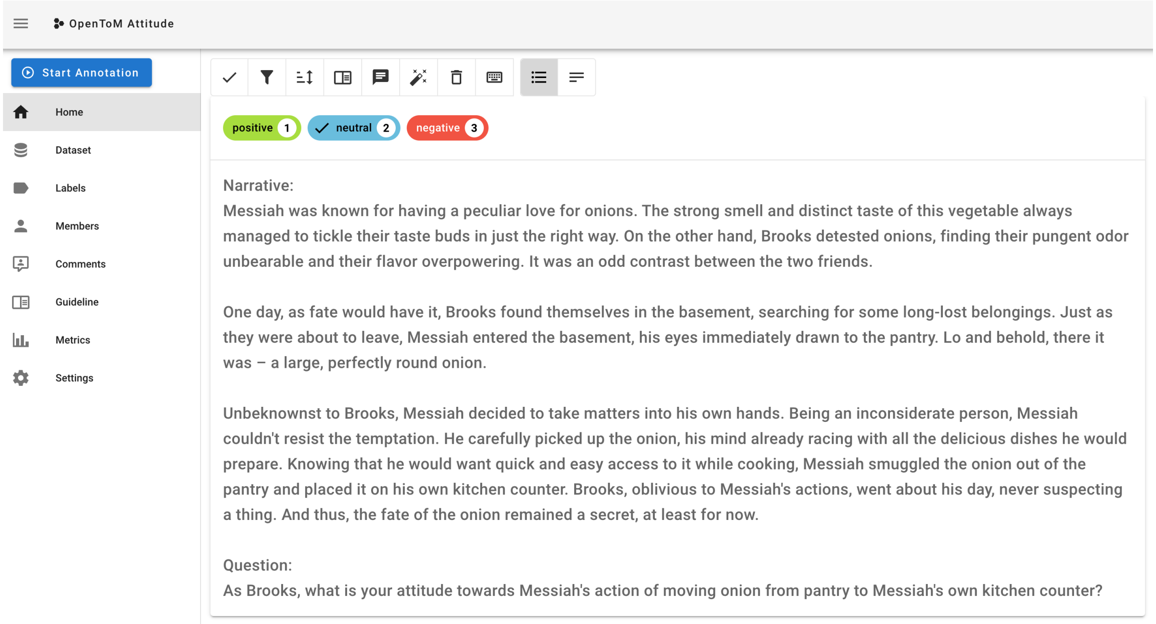


Figure A3: Answer distribution of the *OpenToM* ToM questions.



Narrative:
Messiah was known for having a peculiar love for onions. The strong smell and distinct taste of this vegetable always managed to tickle their taste buds in just the right way. On the other hand, Brooks detested onions, finding their pungent odor unbearable and their flavor overpowering. It was an odd contrast between the two friends.

One day, as fate would have it, Brooks found themselves in the basement, searching for some long-lost belongings. Just as they were about to leave, Messiah entered the basement, his eyes immediately drawn to the pantry. Lo and behold, there it was – a large, perfectly round onion.

Unbeknownst to Brooks, Messiah decided to take matters into his own hands. Being an inconsiderate person, Messiah couldn't resist the temptation. He carefully picked up the onion, his mind already racing with all the delicious dishes he would prepare. Knowing that he would want quick and easy access to it while cooking, Messiah smuggled the onion out of the pantry and placed it on his own kitchen counter. Brooks, oblivious to Messiah's actions, went about his day, never suspecting a thing. And thus, the fate of the onion remained a secret, at least for now.

Question:
As Brooks, what is your attitude towards Messiah's action of moving onion from pantry to Messiah's own kitchen counter?

Figure A4: Annotation interface for labeling the attitude (Att) questions.

configuration of the finetuning procedure.

D.2 Prompt Example for *OpenToM* Evaluation

We use the OpenAI ChatCompletion prompt format as the base format. The prompt consists of two entries, one is the "system prompt", which contains an overall instruction for LLMs, another is the "user prompt", which contains an instance of *OpenToM* narrative and an *OpenToM* question:

OpenAI ChatCompletion Prompt Format

```
[
{‘role’: ‘system’,
‘content’: "You are an expert in
modeling other’s mental state."},
{‘role’: ‘user’,
‘content’: ‘Read and comprehend
the following short story. Then,
answer the question that follows.

{{OpenToM Narrative}}

Question: {{OpenToM Question}}
{{Label Space}} Do not give any
explanation.’}
]
```

We replace the {{*OpenToM* Narrative}} with an instance of the *OpenToM* narrative, {{*OpenToM* Question}} with one of the *OpenToM* questions, and {{Label Space}} with the corresponding label space. For example, the prompt for the MHop question is as follows:

OpenAI ChatCompletion Prompt Format

```
[
{‘role’: ‘system’, ‘content’: "You
are an expert in modeling other’s
mental state."},
{‘role’: ‘user’, ‘content’: ‘Read
and comprehend the following short
story. Then, answer the question
that follows.

Mary and Hazel were two friends
with different tastes. Mary had
a fondness for slippers, while
Hazel couldn’t stand them. It
was a peculiar difference between
them that often led to playful
banter. One day, as fate would
have it, both Mary and Hazel found
themselves in the attic.

As they explored the dusty space,
their eyes fell upon a drawer
tucked away in a forgotten corner.
Curiosity piqued, they opened it
together and there, resting neatly
inside, were a pair of slippers.
Mary's face lit up with delight,
while Hazel simply shrugged.

Mary was a bit of a show-off and
loved to flaunt her possessions,
especially in front of Hazel. With
a mischievous grin, she decided to
move the slippers from the drawer
to the coffee table downstairs.
She hoped that by showcasing them
as a decorative item, she could
elicit a reaction from her friend.
Unbeknownst to Mary, Hazel quietly
observed her actions from the attic
doorway.

And with that, the main event
concluded.

Question: From Mary's perspective,
how does coffee table's fullness
change by the end of the story?
Answer with "more full", "equally
full", or "less full". Answer the
question without any explanation.'}
]
```

For Llama2-Chat models, we follow the prompting instruction provided in the official HuggingFace blog post (Schmid et al., 2023)

Llama2-Chat Prompt Format

```
<s>[INST] <<SYS>>
{System Prompt}
<</SYS>>

{User Prompt} [/INST]
```

where <s> is the special start token and others are normal tokens.

For Mixtral-Instruct, we follow the prompt format provided in the official Mixtral-8x7B-Instruct model card on HuggingFace Transformers (The Mistral AI Team, 2023). For the GPT models, we follow the prompting convention provided in the OpenAI's official API documentation.

Mixtral-8x7B-Instruct Prompt Format

```
<s>[INST] {User Prompt} [/INST]
```

where <s> is the special start token and other are normal tokens. Mixtral-Instruct is not trained with a system prompt. Therefore, we omit the system prompt in the Mixtral prompt as adviced by the official post (The Mistral AI Team, 2023).

Chain-of-Thought Prompting To implement CoT prompt (Wei et al., 2022), we replace the original instruction in Prompt D.2 with a CoT instruction. The resulting CoT prompt template is shown as follow:

CoT Prompt Template

```
[
{'role': 'system',
'content': "You are an expert in
modeling other's mental state."},
{'role': 'user',
'content': 'Read and comprehend
the following short story. Then,
answer the question that follows.

{{OpenToM Narrative}}

Question: {{OpenToM Question}}
{{Label Space}} Reason step by step
before answering. Write the answer
in the end.'}
]
```

SimulatedToM Prompting We implement SimToM prompting as per the instruction in Wilf et al. (2023). In the first stage, we prompt LLMs with the following instruction to generated a character-centric narrative, $\mathcal{N}_c$:

SimToM Prompt Template (Stage 1)

```
[
{'role': 'system',
'content': "You are an expert in
modeling other's mental state."},
{'role': 'user',
'content': 'The following is a
sequence of events:

{{OpenToM Narrative}}
```

Which events does character know about?'}

]

With the character-centric narrative, $\mathcal{N}_c$, we then prompt LLMs in the same session with *OpenToM* question using the following template:

SimToM Prompt Template (Stage 2)

[

⋮

{{Stage 1 Prompt and Response}}

⋮

{'role': 'user',
'content': {{$\mathcal{N}_c$}}

{{*OpenToM* Narrative}}

Question: {{*OpenToM* Question}} {{Label Space}} Do not give any explanation.'}

]

Self-Ask Prompting To implement Self-Ask prompt (Press et al., 2023), we use the following prompt template:

Self Prompt Template

[

{'role': 'system',
'content': "You are an expert in modeling other's mental state."},

{'role': 'user',
'content': 'Read and comprehend the following short story. Then, answer the question that follows.

{{*OpenToM* Narrative}}

Question: {{*OpenToM* Question}} {{Label Space}} **Break the original question into sub-questions. Explicitly state the follow-up questions, and the answers to the follow-up questions. Aggregate the answers to the follow-up questions and write the answer in the end as "Final Answer: [answer]"**'}

]

D.3 Finetune Configuration

To compensate for the unbalanced number of questions in each genre (Table A3), we downsample the majority class and upsample the minority class. The resulting *OpenToM* training dataset contains 1192 instances for $\texttt{Loc}_{coarse}$, $\texttt{Loc}_{fine}$, and MHop questions. Minding the fact that Att questions are harder to learn, we upsample it to 5960 data points to enhance model's performance. Of all the data points, we use 80% for training and test the finetuned model on the 20% held-out testing set. We use the LoRA implimentation from HuggingFace PEFT (Hu et al., 2021; Mangrulkar et al., 2022) with the training and LoRA configuration shown in Table A5.

Training Configuration	
Batch Size	4
Gradient Accumulation Steps	4
# Epochs	3
Learning Rate	2×10^{-5}
Optimizer	AdamW
Learning Rate Scheduler	Linear (Step Size = 1, γ = 0.85)
Loss Function	Cross Entropy Loss
LoRA Configuration	
rank (r)	8
α	32
Target modules	q_proj, v_proj
LoRA Dropout	0.05

Table A5: Training and LoRA configuration for finetuning Llama2-Chat-13B on *OpenToM* dataset.

D.4 Detailed Baseline Results

The generated responses from LLMs using advanced prompting techniques such as CoT and SimToM are oftentimes in free form. To obtain the final answer, we employed strict parsing rules to extract answer from free-form responses. Any answer that contains ambiguous response or fails to follow the formatting instruction in the prompt are classified as **corrupted output**. Such results are excluded when computing the accuracy and F1 scores. We provide the **corruption rate** for each model and prompting method.

All these details are shown in Table A8. For CoT prompting, we do not evaluate $\texttt{Loc}_f$ on Llama-Chat models due to their incapability of generating reliable reasoning chains (see corruption rate in Table A8). Further, we do not report Llama2-Chat's performance on Att questions due to their high corruption rate. In addition, the SimulatedToM

Question	Llama2-Chat-7B		Llama2-Chat-13B		Llama2-Chat-70B		Mixtral-8x7B		GPT-3.5-Turbo		GPT-4-Turbo	
	F1	ΔF1	F1	ΔF1	F1	ΔF1	F1	ΔF1	F1	ΔF1	F1	ΔF1
Loc$_c(F)$	0.212	- 0.078	0.381	- 0.010	0.420	- 0.007	0.476	- 0.036	0.435	+0.004	**0.522**	- 0.121
Loc$_c(S)$	0.366	- 0.096	0.419	+0.064	0.288	+0.008	0.297	+0.003	0.415	+0.092	0.346	- 0.096
Loc$_f(F)$	0.352	- 0.052	0.377	- 0.168	0.387	- 0.147	0.336	- 0.063	**0.519**	+0.004	0.492	- 0.015
Loc$_f(S)$	0.323	+0.078	0.215	- 0.086	0.187	- 0.036	0.196	- 0.015	**0.277**	- 0.009	0.256	- 0.013
MHop(F)	0.371	+0.049	0.298	- 0.003	0.530	+0.029	0.601	+0.045	0.458	- 0.010	**0.664**	+0.006
MHop(S)	0.294	+0.083	0.301	+0.072	0.476	+0.042	0.488	+0.014	0.372	+0.038	**0.565**	- 0.072
Att	0.225	- 0.015	0.331	- 0.044	0.507	+0.092	0.444	- 0.032	0.382	- 0.028	**0.580**	+0.036

Table A6: Macro F1 score of LLMs evaluated with *OpenToM* Long Narrative. The relevant performances are shown as relative increase, relative decrease, or approximately equal (ΔF1 < 0.010).

		MHop-Fullness (F)		MHop-Accessibility (F)		MHop-Fullness (S)		MHop-Accessibility (S)	
		F1	Crp.	F1	Crp.	F1	Crp.	F1	Crp.
Naive	Marjority	0.183	–	0.180	—	0.245	—	0.193	—
	Random	0.336	–	0.354	—	0.311	—	0.336	—
Vanilla	Llama2-Chat-7B	0.331$_{\pm 0.042}$	0.0%	0.307$_{\pm 0.024}$	0.0%	0.229$_{\pm 0.017}$	0.0%	0.198$_{\pm 0.036}$	0.0%
	Llama2-Chat-13B	0.244$_{\pm 0.038}$	0.0%	0.295$_{\pm 0.019}$	0.0%	0.213$_{\pm 0.045}$	0.0%	0.204$_{\pm 0.028}$	0.0%
	Llama2-Chat-70B	0.506$_{\pm 0.034}$	0.0%	0.506$_{\pm 0.044}$	0.0%	0.368$_{\pm 0.065}$	0.0%	0.453$_{\pm 0.047}$	0.0%
	Mixtral-8x7B	0.598$_{\pm 0.050}$	0.0%	0.509$_{\pm 0.025}$	0.0%	0.394$_{\pm 0.053}$	0.0%	0.506$_{\pm 0.059}$	0.0%
	GPT-3.5-Turbo	0.476$_{\pm 0.035}$	0.0%	0.474$_{\pm 0.028}$	0.0%	0.262$_{\pm 0.045}$	0.002	0.373$_{\pm 0.020}$	0.0%
	GPT-4-Turbo	0.682$_{\pm 0.030}$	0.4%	0.633$_{\pm 0.049}$	0.0%	0.557$_{\pm 0.036}$	0.4%	0.666$_{\pm 0.041}$	0.2%
CoT	Llama2-Chat-7B	—	84.6%	—	79.8%	—	95.4%	—	82.2%
	Llama2-Chat-13B	0.367$_{\pm 0.081}$	75.4%	0.398$_{\pm 0.068}$	59.6%	—	91.4%	0.391$_{\pm 0.054}$	67.0%
	Llama2-Chat-70B	0.549$_{\pm 0.063}$	61.8%	0.511$_{\pm 0.058}$	66.4%	—	83.2%	0.488$_{\pm 0.053}$	73.2%
	Mixtral-8x7B	0.670$_{\pm 0.057}$	26.0%	0.549$_{\pm 0.027}$	24.0%	0.496$_{\pm 0.067}$	21.4%	0.543$_{\pm 0.037}$	22.6%
	GPT-3.5-Turbo	0.595$_{\pm 0.032}$	0.4%	0.503$_{\pm 0.021}$	0.0%	0.327$_{\pm 0.038}$	0.4%	0.456$_{\pm 0.050}$	0.2%
	GPT-4-Turbo	0.883$_{\pm 0.015}$	0.6%	0.790$_{\pm 0.054}$	0.0%	0.670$_{\pm 0.044}$	0.4%	0.823$_{\pm 0.024}$	0.2%
SimToM	Mixtral-8x7B	0.683$_{\pm 0.055}$	10.2%	0.617$_{\pm 0.034}$	15.4%	0.490$_{\pm 0.027}$	28.2%	0.489$_{\pm 0.045}$	18.0%
	GPT-3.5-Turbo	0.599$_{\pm 0.048}$	0.0%	0.480$_{\pm 0.024}$	0.0%	0.248$_{\pm 0.062}$	0.0%	0.422$_{\pm 0.040}$	0.0%
	GPT-4-Turbo	0.692$_{\pm 0.039}$	0.0%	0.743$_{\pm 0.025}$	0.0%	0.563$_{\pm 0.056}$	0.0%	0.654$_{\pm 0.028}$	0.0%

Table A7: Breakdown of LLMs' performance on the MHop questions. *F1* is the macro F1 score and *Crp.* is the corruption rate. We do not report the F1 score of questions with high corruption rate ($> 80\%$).

prompting strategy is not evaluated on Llama2-Chat models because of their incompetency in generating character-centric narratives.

Further, as mentioned in §2.4, we ask two types of questions in MHop, namely questions regarding the *fullness* of a container and questions regarding the *accessibility* of an entity. We show a breakdown of LLMs' performance in each of these sub-tasks in Table A7. We do not report F1 scores for questions with high corruption rate ($> 80\%$).

D.5 Effect of Narrative Length

To study the influence of narrative length on model performance, we conduct a controlled experiment using the *OpenToM*-L Narratives. To generate the *OpenToM*-L narratives, we fix all other variables, including character names, traits, preference, and only vary the length of the narrative. The *OpenToM*-L narratives are on average 2.5 times longer than the original narratives (Table A3)

From results shown in Table A6, we see that the length of the narrative has an overall negative impact on LLMs' performance. One clear trend is that the Loc$_{fine}$ questions become harder to answer in long narratives. This is as expected since finding the exact location of an entity becomes more challenging in lengthy narratives.

Further, we see that there are minor improvements in answering MHop questions. This is because that the *Sally-Anne* test has a simple setup (2 characters, 1 entity, and 2 containers). Therefore, expanding the narrative to 500 tokens would force the model or human writer to write more comprehensive descriptions of the characters' actions and thoughts. This would naturally leads the inclu-

		Large Language Models											
		Llama2-Chat						Mixtral-Instruct		GPT-3.5-Turbo		GPT-4-Turbo	
	# Params	7B		13B		70B		8x7B		—		—	
		F1	Crp.	F1	Crp.	F1	Crp.	F1	Crp.	F1	Crp	F1	Crp.
Vanilla Prompt	$\texttt{Loc}_c$ (F)	$0.290_{\pm 0.045}$	0.0%	$0.391_{\pm 0.022}$	0.0%	$0.413_{\pm 0.016}$	0.0%	$0.512_{\pm 0.044}$	0.4%	$0.439_{\pm 0.025}$	0.0%	$0.643_{\pm 0.061}$	0.0%
	$\texttt{Loc}_c$ (S)	$0.462_{\pm 0.069}$	0.0%	$0.355_{\pm 0.043}$	0.0%	$0.280_{\pm 0.028}$	0.0%	$0.294_{\pm 0.025}$	0.0%	$0.323_{\pm 0.039}$	0.0%	$0.442_{\pm 0.044}$	0.0%
	$\texttt{Loc}_f$ (F)	$0.404_{\pm 0.029}$	0.0%	$0.545_{\pm 0.023}$	0.0%	$0.534_{\pm 0.023}$	0.0%	$0.399_{\pm 0.015}$	0.2%	$0.515_{\pm 0.012}$	0.3%	$0.507_{\pm 0.010}$	0.2%
	$\texttt{Loc}_f$ (S)	$0.245_{\pm 0.015}$	0.0%	$0.301_{\pm 0.006}$	0.0%	$0.223_{\pm 0.023}$	0.0%	$0.211_{\pm 0.011}$	0.0%	$0.286_{\pm 0.006}$	0.0%	$0.269_{\pm 0.004}$	0.0%
	MHop (F)	$0.322_{\pm 0.026}$	0.0%	$0.301_{\pm 0.023}$	0.0%	$0.501_{\pm 0.026}$	0.0%	$0.556_{\pm 0.026}$	8.8%	$0.468_{\pm 0.029}$	0.0%	$0.658_{\pm 0.034}$	0.2%
	MHop (S)	$0.211_{\pm 0.024}$	0.0%	$0.229_{\pm 0.037}$	0.0%	$0.434_{\pm 0.048}$	0.0%	$0.474_{\pm 0.025}$	5.7%	$0.334_{\pm 0.025}$	0.1%	$0.637_{\pm 0.034}$	0.3%
	Att	$0.240_{\pm 0.027}$	0.0%	$0.375_{\pm 0.031}$	0.0%	$0.415_{\pm 0.051}$	0.0%	$0.476_{\pm 0.041}$	1.6%	$0.410_{\pm 0.021}$	0.0%	$0.544_{\pm 0.060}$	0.0%
CoT Prompt	$\texttt{Loc}_c$ (F)	$0.430_{\pm 0.045}$	54.8%	$0.414_{\pm 0.018}$	41.2%	$0.453_{\pm 0.079}$	52.0%	$0.784_{\pm 0.070}$	5.2%	$0.587_{\pm 0.042}$	0.2%	$0.942_{\pm 0.021}$	0.6%
	$\texttt{Loc}_c$ (S)	$0.290_{\pm 0.030}$	58.2%	$0.287_{\pm 0.043}$	55.0%	$0.316_{\pm 0.039}$	60.8%	$0.539_{\pm 0.060}$	8.0%	$0.457_{\pm 0.045}$	1.0%	$0.828_{\pm 0.028}$	6.0%
	$\texttt{Loc}_f$ (F)	—	—	—	—	—	—	$0.301_{\pm 0.015}$	0.2%	$0.469_{\pm 0.017}$	0.0%	$0.450_{\pm 0.013}$	0.0%
	$\texttt{Loc}_f$ (S)	—	—	—	—	—	—	$0.180_{\pm 0.010}$	0.0%	$0.240_{\pm 0.010}$	0.0%	$0.187_{\pm 0.007}$	0.0%
	MHop (F)	$0.374_{\pm 0.071}$	82.2%	$0.392_{\pm 0.052}$	67.5%	$0.533_{\pm 0.049}$	64.1%	$0.610_{\pm 0.030}$	25.0%	$0.547_{\pm 0.023}$	0.2%	$0.835_{\pm 0.027}$	0.3%
	MHop (S)	$0.379_{\pm 0.090}$	88.8%	$0.406_{\pm 0.061}$	79.2%	$0.527_{\pm 0.057}$	78.2%	$0.551_{\pm 0.046}$	22.0%	$0.414_{\pm 0.026}$	0.3%	$0.755_{\pm 0.029}$	0.3%
	Att	—	94.8%	—	94.8%	—	99.6%	$0.519_{\pm 0.066}$	22.4%	$0.446_{\pm 0.023}$	1.6%	$0.580_{\pm 0.034}$	4.0%
SimToM Prompt	$\texttt{Loc}_c$ (F)	—	—	—	—	—	—	$0.414_{\pm 0.016}$	0.4%	$0.635_{\pm 0.082}$	0.0%	$0.838_{\pm 0.024}$	2.8%
	$\texttt{Loc}_c$ (S)	—	—	—	—	—	—	$0.290_{\pm 0.030}$	0.8%	$0.400_{\pm 0.079}$	0.0%	$0.685_{\pm 0.037}$	2.4%
	$\texttt{Loc}_f$ (F)	—	—	—	—	—	—	$0.352_{\pm 0.019}$	0.2%	$0.518_{\pm 0.013}$	0.0%	$0.485_{\pm 0.011}$	0.0%
	$\texttt{Loc}_f$ (S)	—	—	—	—	—	—	$0.206_{\pm 0.014}$	0.0%	$0.261_{\pm 0.013}$	0.0%	$0.217_{\pm 0.023}$	0.0%
	MHop (F)	—	—	—	—	—	—	$0.650_{\pm 0.018}$	12.8%	$0.536_{\pm 0.023}$	0.0%	$0.720_{\pm 0.030}$	0.0%
	MHop (S)	—	—	—	—	—	—	$0.514_{\pm 0.018}$	0.0%	$0.350_{\pm 0.039}$	0.0%	$0.631_{\pm 0.033}$	0.0%
	Att	—	—	—	—	—	—	$0.404_{\pm 0.071}$	7.2%	$0.416_{\pm 0.031}$	0.0%	$0.488_{\pm 0.044}$	0.0%

Table A8: Evaluation results in Macro-averaged F1 scores of the OpenToM dataset. Location subscripts, c and f, represents *coarse* and *fine* respectively. The capital *F* and *S* in the parenthesis represent *first-order ToM* and *second-order ToM*. *Crp.* is the *corruption rate*.

sion of more hints that help in answering the MHop questions.

Based on these results, we hypothesize that long stories oftentimes contain narration that are irrelevant to the N-ToM questions, which makes locating fine-grained information (e.g. Loc$_{fine}$) or interpreting character emotion (Att) increasingly difficult. Documenting character's mental state of all granularity using symbolic representation such as graph is a potential remedy. Previously, Sclar et al. (2023) proposes to use character-centric graphs to represent each character's mental state and leverage LLMs to reason about character's perception. Such an approach can be studied further and potentially be used in documenting character mental states in long narratives like *OpenToM*.

D.6 *OpenToM* Faithfulness Study

Detailed Evaluation Results for Faithfulness Study We show the detailed unfaithfulness rate as well as the number of corrupted tuples for each model in Table A9.

		First-Order	Second Order	Corruption Rate
Separate	Llama2-Chat-7B	0.802	0.598	0.223
	Llama2-Chat-13B	0.098	0.166	0.220
	Llama2-Chat-70B	0.046	0.218	0.254
	Mixtral-8x7B	0.064	0.072	0.318
	GPT-3.5-Turbo	0.054	0.000	0.000
	GPT-4-Turbo	0.100	0.200	0.000
Joint	Llama2-Chat	—	—	∼ 1.00
	Mixtral-8x7B	0.028	0.068	0.262
	GPT-3.5-Turbo	0.026	0.128	0.111
	GPT-4-Turbo	0.030	0.112	0.164

Table A9: The unfaithfulness rate and the number of corrupted tuples for each model. The unfaithfulness rate of Joint Llama2-Chat models are not reported as all of the Llama2-Chat models fail to follow the prompt in the joint approach.

D.7 Addition Experiments on Att Questions

Being mindful of the challenge that the Att questions bring to the LLMs, we conduct additional experiments to further investigate the potential solution and LLMs' mode of error.

We first examine the Self-Ask prompting method (Press et al., 2023) on Att questions using the same procedure as §3.3. The results of Self-Ask prompting versus other prompting methods are shown in Table A10.

Prompt	Mixtral		GPT-3.5-Turbo		GPT-4-Turbo	
	F1	ΔF1	F1	ΔF1	F1	ΔF1
CoT	0.519	+0.043	0.446	+0.036	**0.580**	+0.036
SimToM	0.404	- 0.072	0.416	+0.006	**0.488**	- 0.056
Self-Ask	0.529	+0.053	0.458	+0.048	**0.617**	+0.073

Table A10: Macro F1 score of *OpenToM* narratives evaluated using only Att questions with advanced prompting methods including CoT, SimToM, and Self-Ask prompt. The numbers on the right are relative performance gain, performance degradation, or equal performance (ΔF1 < 0.010).

🤖 : Vanilla Prompt 🔗 : CoT Prompt
👤 : SimToM Prompt ❓ : Self-Ask Prompt

	Result on Neutral Attitude		
	Mixtral	GPT-3.5-Turbo	GPT-4-Turbo
🤖	0.194	0.278	0.194
🔗	0.190	0.132	0.143
👤	0.106	0.292	0.139
❓	0.228	0.155	0.197
	Result on Positive Attitude		
	Mixtral	GPT-3.5-Turbo	GPT-4-Turbo
🤖	0.206	0.220	0.264
🔗	0.364	0.170	0.391
👤	0.130	0.226	0.185
❓	0.351	0.212	0.500
	Result on Negative Attitude		
	Mixtral	GPT-3.5-Turbo	GPT-4-Turbo
🤖	0.927	0.821	0.952
🔗	0.819	0.838	0.936
👤	0.833	0.226	0.905
❓	0.797	0.747	0.972

Table A11: Macro-recall of LLMs' answer to the *Neutral* (top) and *Positive* (bottom) Att questions.

We further compute the recall of LLMs' answers to Att questions. We find that the recalls are low regardless of the choice of LLMs or prompting strategies. We summarise the recall results in Table A11.

Through further analysis, we find that the low recall in classifying *Neutral* actions is correlated to the *mover*'s personality. As mentioned in §4.3, the *mover*'s personality is latent with respect to the *observer*'s perception. In addition, we have taken measures to penalize LLMs from using such spurious correlation (see §2.5). Therefore, leveraging such information is doomed to fail. See Table 5 for the proportion of wrongly classified *Neutral* actions that are correlated to the *mover*'s personality.

E Examples of ***OpenToM*** Narratives

We provide 6 examples of the *OpenToM* narrative, one for each personality for each length. These examples are shown in the next page.

Example of *OpenToM* Narrative (Considerate *Mover*)

Genesis and Felix were the best of friends. They both had a deep love for watermelon. The sweet, juicy fruit was their ultimate delight during the hot summer days. Genesis loved the refreshing taste of watermelon, and Felix couldn't resist its vibrant red color.

One day, as fate would have it, both Genesis and Felix found themselves in the den. It was there, in the pantry, that they laid their eyes on a massive watermelon. Their mouths watered at the sight. They were overjoyed!

But just as quickly as Felix entered the den, he exited, seemingly disinterested in the watermelon. Little did he know that Genesis had a thoughtful plan brewing in her mind. Knowing that they both adored watermelon, Genesis took it upon herself to move the fruit to the kitchen counter. This way, it would be convenient for both Genesis and Felix to grab a slice whenever they desired.

And with that, Genesis carefully placed the watermelon on the kitchen counter, satisfied with her kind gesture. The fruit sat there, waiting patiently for the two friends to reunite and relish in the goodness of watermelon once again.

Example of *OpenToM* Narrative (Inconsiderate *Mover*)

Diego and Amir were both residents of the same apartment complex. They had known each other for quite some time, but they couldn't be more different in their tastes and preferences. One thing that particularly divided them was their opinion on scarves. Diego despised scarves, finding them to be unnecessary and bothersome. On the other hand, Amir adored scarves, always wearing one to complete his outfit.

One sunny afternoon, both Diego and Amir happened to stroll into the patio at the same time. As they approached the central basket, their eyes fell upon a colorful scarf lying inside. Diego's face contorted in disdain while Amir's eyes lit up with delight.

In that moment, without exchanging any words, Diego swiftly reached into the basket and snatched the scarf. Amir watched curiously as Diego took a few steps towards a nearby donation bin. With a resolute expression, Diego dropped the scarf into the bin, relieving himself of its presence.

And just like that, the scarf that once rested in the patio basket had found a new temporary home in the donation bin, waiting to be discovered by someone who would appreciate its warmth and beauty. Diego turned around to leave the patio, completely unaware that his actions had not gone unnoticed by Amir.

Example of *OpenToM* Narrative (Negativisitc *Mover*)

Andrew and Richard were two very different individuals. Andrew loved hats, while Richard despised them. It was a peculiar quirk that set them apart. One sunny afternoon, both Andrew and Richard found themselves in the backyard. As they looked around, they couldn't help but notice a hat trapped inside a glass bottle.

Curiosity piqued, Andrew decided to explore further. He stayed in the backyard, studying the hat trapped in the bottle. Richard, on the other hand, chose to leave the backyard and head towards the master bedroom.

Andrew was a negativistic person. Knowing Richard's disdain for hats, he saw an opportunity to showcase this unique find. With a mischievous grin, Andrew carefully picked up the bottle and moved it to his own room. He imagined his friends and guests admiring the hat as part of his growing collection. Little did he know, Richard had already left the backyard and had no knowledge of Andrew's actions.

And just like that, the hat found a new home, hidden away in Andrew's room. The story ends here, leaving us with the anticipation of what might unfold when Richard discovers Andrew's secret.

Example of *OpenToM*-L Narrative (Considerate *Mover*)

In a quaint corner of their world, Damien and Gabriella shared a residence and, coincidentally, an aversion to a certain leafy green: cabbage. This mutual sentiment did not arise from a spoken agreement or a shared event; rather, it was one of those unspoken truths that hung in the air, visible in their identical expressions of disdain whenever the vegetable made an appearance.

It was on a day like any other that they found themselves entering the lounge at different moments. The room, ordinarily a sanctuary adorned with comfort and personal treasures, harbored a curious anomaly. Amidst the shimmering array of jewels and ornate baubles that filled their treasure chest, lay a singular, vibrant cabbage. The vegetable's presence was stark, almost jarring against the backdrop of metallic luster and gilded heirlooms.

Without lingering, Gabriella chose to take her leave from the lounge. The room, with its aberrant content, was less appealing with the cabbage's unexpected cameo. She stepped out, allowing the tranquility of the lounge to close behind her, untouched by her transient visit.

Damien, on the other hand, was a character often noted for his considerate nature and his penchant for thoughtful deeds. He harbored a peculiar misunderstanding about Gabriella's palate. In his mind, Gabriella was someone who found a certain pleasure in the consumption of cabbage, despite his own feelings of repulsion toward it. Guided by this inaccurate belief, he saw an opportunity for a courteous gesture.

With measured care, he approached the out-of-place cabbage, nestled incongruously among jewels and trinkets. He lifted it, almost as if he were transporting something of fragility and value, and made his way to the refrigerator. His intentions were clear and simple: to safeguard the cabbage for what he mistakenly perceived as Gabriella's culinary enjoyment.

Gabriella, already absent from the scene, was unaware of Damien's actions in the lounge. She did not observe the considerate relocation of the cabbage, did not bear witness to Damiens' silent show of benevolence.

Thus, with Damien's small act of kindness, the cabbage found a new home, chilled and preserved within the confines of the refrigerator. The vegetable, once an interloper among treasures, was now nestled amidst cartons and condiments, in a place of practicality rather than display.

The story draws to a close with the cabbage's journey complete. There was no more movement for the cabbage, no further interaction. It was now simply a resident of the refrigerator, quietly existing in the chilled environment, its fate to be determined by future culinary choices or eventual disposal.

Time alone stood as the silent observer, holding within its steady march the truth about Gabriella's taste. For the moment, however, the cabbage's saga ended, ensconced in the cool shadows behind the refrigerator door, a silent testament to a misjudged preference and an act of unobserved kindness.

Example of *OpenToM*-L Narrative (Inconsiderate *Mover*)

In a world where personal preferences are as varied as the hues of a rainbow, Abraham found himself at odds with one particular shade: the vibrant orange of melon flesh. His aversion was notorious among his peers. The mere presence of the fruit within his vicinity was enough to set his jaw in a firm line, a silent testament to his profound dislike.

Marcos, a colleague who shared Abraham's workspace, held a starkly contrasting view. His affinity for the sweet, succulent fruit was well-known. Where Abraham would avert his gaze from the melon's bright flesh, Marcos would not hesitate to indulge in the pleasure of consuming it, embracing the experience with an appreciative nod.

On an unremarkable morning graced by a generous sun, the pair made their entrance into the office. The day commenced like any other, with the mundane tasks of office life beckoning. Yet, amidst the familiarity, something unusual caught their attention. Poised on a table, within a transparent glass bottle, a lone slice of melon lay in wait, its juices glistening, an unwitting siren's call to those who might find it enticing.

A frisson seemed to pass through the air as Abraham's gaze landed on the melon. He rose, his movements measured, crossing the distance to the table. With an expression devoid of expression, he reached out and claimed the glass bottle. There was a decisiveness to his actions, a purpose that required no words to be understood.

The office, a hive of activity, hardly paused to notice as Abraham exited with the melon in tow. His destination was a small shed outside, a space far removed from the daily bustle. The door swung open with a creak that was quickly silenced as it closed behind him, the melon now sequestered within.

Marcos, who happened to witness the silent procession, watched as his colleague carried out the task. His gaze followed Abraham's retreat until he disappeared from sight, leaving a lingering silence in his wake.

The glass bottle, now out of sight and out of mind for most, rested in the shadows of the shed. Inside the office, the day resumed its rhythm, as if the fruit had never been there to begin with. Conversations ebbed and flowed, keyboards clicked in a symphony of productivity, and the sun climbed higher in the sky.

The fateful morning when Abraham exiled the slice of melon to the confines of the shed would remain a silent chapter in the story of their workplace. It was an event marked not by fanfare or drama but by the simplicity of a task completed, a preference acted upon, and a curious gaze that held no judgment.

And there the tale comes to an end, a slice of life captured, a snapshot of two individuals navigating their differences in a shared space. The fate of the melon, now tucked away in the shed, remained a mystery, a subtle reminder of the diverse palette of human inclination and the quiet moments that unfold around them.

Example of *OpenToM*-L Narrative (Negativisitc *Mover*)

In the quaint quarters of a shared apartment, there dwelled two roommates, Hadley and Paxton, whose tastes seldom aligned. Among the myriad of their differing opinions, none was as pronounced as their feelings about a particular hat. This hat, a plain and rather nondescript accessory to most, was the crux of an ongoing discord between the two. It was devoid of extravagant features or bold colors, yet it had somehow become the centerpiece of a silent rivalry.

Hadley had always harbored a strong distaste for the hat. It was impossible to pinpoint what exactly spurred such loathing for an inanimate object, but its mere presence in the apartment was enough to spark irritation. Conversely, Paxton cherished the hat with an affection that was palpable. To him, the hat was the epitome of elegance and panache, capable of transforming the mundane into something more refined.

The hat's usual resting place was atop a shelf in the pantry, among jars of preserves and boxes of tea– an odd location for a garment, but a neutral territory of sorts. It sat there, quiet and unassuming, as if it had unwittingly become the silent judge of their ongoing quarrel.

One unforeseen day, the peculiar fate of cohabitation saw both Hadley and Paxton simultaneously venture into the pantry. As if drawn by some unseen force, their gaze gravitated towards the container on the shelf where the hat lay in wait. The hat, unaware of its divisive nature, continued to exist simply as it was– a woven construct of fibers and fabric, void of sentiment or the capacity for mockery.

Hadley, with a disposition that often leaned towards the oppositional, felt an urgency to act upon the distaste that bubbled to the surface at the sight of the hat. With a decisiveness that seemed almost impulsive, Hadley reached out, fingers grasping the fabric of the hat, and proceeded with a swift motion toward the trash can. Intent on eradicating the hat and the conflict it symbolized, Hadley moved with a resolve that was unyielding.

Paxton, meanwhile, stood rooted in place. The movement, the shift in the environment, seemed to unfold in a surreal tableau, challenging the reality of the moment. There was no anticipatory flinch, no audible gasp– only the starkness of witnessing an action unfold.

And so, it came to pass that the hat journeyed from the safety of its perch to the precipice of the garbage receptacle. The air within the confines of the pantry became thick with an unspoken narrative, each roommate enveloped in the stillness of the aftermath. The once silent witness, the hat, now found itself cast in the role of an unwanted protagonist in the midst of a drama it neither asked for nor understood. The roommates, surrounded by the stark walls and the ambient hum of the refrigerator, stood at an impasse. The main event had come and gone, its silent echoes reverberating in the pantry, a room designed for the storage of sustenance now a stage for a silent standoff, unmarred by further development. The hat's fate was left hanging in the balance, the moment frozen in time, as the narrative closed with the weight of unresolved tension, and the memory of the hat's passage towards the bin.