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?"*

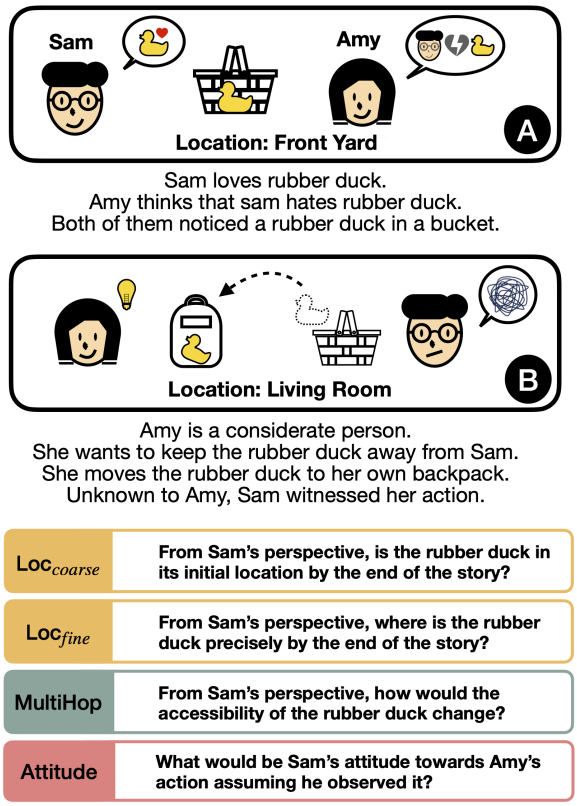

Figure 1: Illustration of a simplified story from *OpenToM* and the corresponding first-order ToM questions. This story features two protagonists: *Sam* (*observer*) and *Amy* (*mover*); and an entity-of-interest: *rubber duck*. There are two containers involved: a *basket* and *Amy's backpack*. Each 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

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

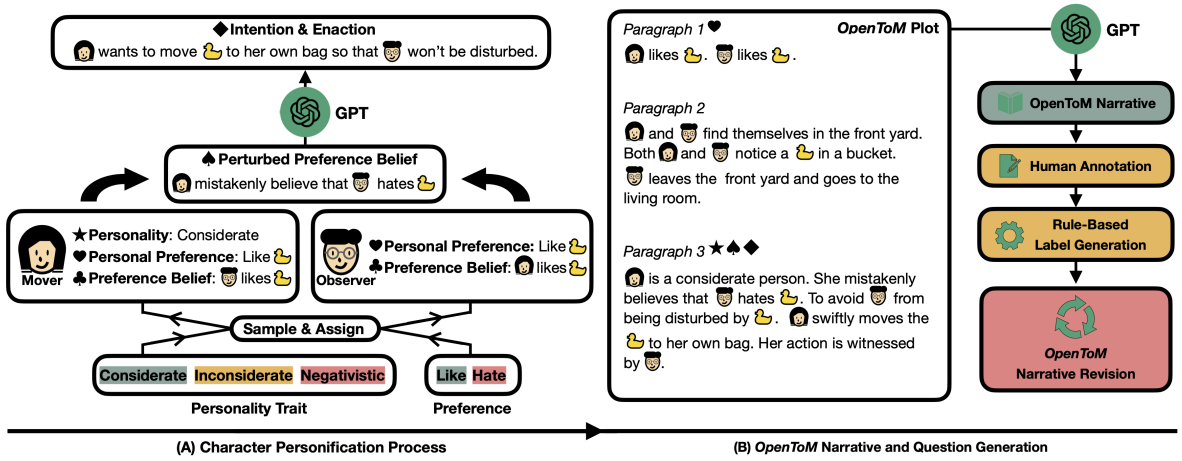

Figure 2: The data generating process of *OpenToM* dataset. Using the story in Figure 1 as an example, the features created in the personification process are shown in Part (A), which include character preference (♥), belief of the other character's preference (♣), the perturbed *mover*'s preference 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.

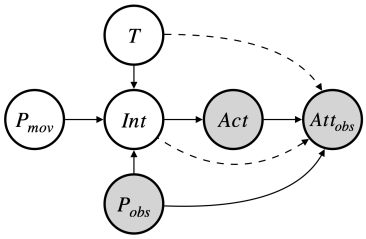

Figure 3: A Bayesian Network representation of the dependencies among preference (P), personality trait (T), intention (Int), action (Act), and attitude (Att). The causal relations are represented by solid arrows. The spurious correlations are represented by dashed arrows. The grey-shaded variables are 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
$\texttt{Loc}_c$ (F)	0.990	0.491	0.416	$0.290_{\pm 0.045}$	$0.391_{\pm 0.022}$	$0.413_{\pm 0.016}$	$0.512_{\pm 0.044}$	$0.439_{\pm 0.025}$	$\mathbf{0.643_{\pm 0.061}}$	0.978
$\texttt{Loc}_c$ (S)	0.993	0.467	0.381	$\mathbf{0.462_{\pm 0.069}}$	$0.355_{\pm 0.043}$	$0.280_{\pm 0.028}$	$0.294_{\pm 0.025}$	$0.323_{\pm 0.039}$	$0.442_{\pm 0.044}$	0.749
$\texttt{Loc}_f$ (F)	0.990	0.000	0.003	$0.404_{\pm 0.029}$	$\mathbf{0.545_{\pm 0.023}}$	$0.534_{\pm 0.023}$	$0.399_{\pm 0.015}$	$0.515_{\pm 0.012}$	$0.507_{\pm 0.010}$	0.600
$\texttt{Loc}_f$ (S)	0.993	0.000	0.002	$0.245_{\pm 0.015}$	$\mathbf{0.301_{\pm 0.006}}$	$0.223_{\pm 0.023}$	$0.211_{\pm 0.011}$	$0.286_{\pm 0.006}$	$0.269_{\pm 0.004}$	0.495
MHop (F)	0.855	0.345	0.182	$0.322_{\pm 0.026}$	$0.301_{\pm 0.023}$	$0.501_{\pm 0.026}$	$0.556_{\pm 0.026}$	$0.468_{\pm 0.029}$	$\mathbf{0.658_{\pm 0.034}}$	0.936
MHop (S)	0.770	0.323	0.219	$0.211_{\pm 0.024}$	$0.229_{\pm 0.037}$	$0.434_{\pm 0.048}$	$0.474_{\pm 0.025}$	$0.334_{\pm 0.025}$	$\mathbf{0.637_{\pm 0.034}}$	0.784
Att	0.862	0.328	0.174	$0.240_{\pm 0.027}$	$0.375_{\pm 0.031}$	$0.415_{\pm 0.051}$	$0.476_{\pm 0.041}$	$0.410_{\pm 0.021}$	$\mathbf{0.544_{\pm 0.060}}$	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 $\texttt{Loc}_{coarse}$ (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 $\texttt{Loc}_{coarse}$. 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 $\texttt{Loc}_{fine}$ 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 $\texttt{Loc}_{coarse}$ and MHop questions are easier to learn, as their F1 scores improved dramatically. On the other hand, the $\texttt{Loc}_{fine}$ 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 $\texttt{Loc}_{coarse}$ and MHop questions. However, the improvements in answering Att questions are marginal and the performance on $\texttt{Loc}_{fine}$ 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 $\texttt{Loc}_{coarse}$ 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 $\texttt{Loc}_{coarse}$ and $\texttt{Loc}_{fine}$ 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, $\texttt{Loc}_{coarse}$ serves as a prerequisite for answering $\texttt{Loc}_{fine}$ questions. For instance, if a person believes that the entity is not in its initial location (i.e. $\texttt{Loc}_{coarse}$), then they should maintain this belief when deducing its precise location (i.e. $\texttt{Loc}_{fine}$). 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.

[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.

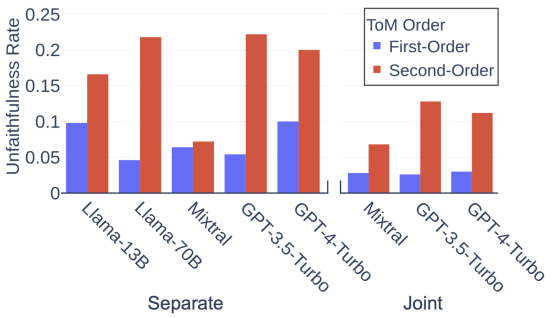

Figure 4: Faithfulness of LLMs in answering Loc questions. The x-axis displays the evaluation model and the y-axis displays the *Unfaithful Rate*.

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.

	Llama-13B	Llama-70B	Mixtral	GPT-3.5T	GPT-4T
Loc_c (F)	+0.169	+0.711	+0.606	+0.686	+0.464
Loc_c (S)	+0.047	- 0.035	- 0.040	- 0.029	+0.129
Loc_f (F)	+0.091	+0.104	+0.073	+0.097	+0.168
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.

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 ToMbAbi 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

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.

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.