

# Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models

Natalie Shapira<sup>1</sup> Mosh Levy\*<sup>1</sup> Seyed Hossein Alavi\*<sup>2,3</sup> Xuhui Zhou\*<sup>4</sup>

Yejin Choi<sup>5,6</sup> Yoav Goldberg<sup>1,5</sup> Maarten Sap<sup>4,5</sup> Vered Shwartz<sup>2,3</sup>

<sup>1</sup> Bar-Ilan University <sup>2</sup> University of British Columbia

<sup>3</sup> Vector Institute for AI <sup>4</sup> Carnegie Mellon University

<sup>5</sup> Allen Institute for Artificial Intelligence <sup>6</sup> University of Washington

nd1234@gmail.com

## Abstract

The escalating debate on AI’s capabilities warrants developing reliable metrics to assess machine “intelligence.” Recently, many anecdotal examples were used to suggest that newer large language models (LLMs) like ChatGPT and GPT-4 exhibit Neural Theory-of-Mind (N-ToM); however, prior work reached conflicting conclusions regarding those abilities. We investigate the extent of LLMs’ N-ToM through an extensive evaluation of 6 tasks and find that while LLMs exhibit certain N-ToM abilities, this behavior is far from being robust. We further examine the factors impacting performance on N-ToM tasks and discover that LLMs struggle with adversarial examples, indicating reliance on shallow heuristics rather than robust ToM abilities. We caution against drawing conclusions from anecdotal examples, limited benchmark testing, and using human-designed psychological tests to evaluate models.

## 1 Introduction

Theory of Mind (ToM) is the ability to understand that other people have thoughts, beliefs, and emotions that differ from one’s own (Wimmer and Perner, 1983). As ToM is inherently linked to human cognition, imbuing machines with capabilities that mimic or resemble ToM has the potential to lead to the “ELIZA effect” (Weizenbaum, 1976), wherein human-like intelligence or even sentience and consciousness is incorrectly ascribed to the machine (e.g., Kosinski, 2023; Bubeck et al., 2023).

In light of these possibly illusory ToM abilities, there is a pressing need to develop robust metrics for assessing Neural-ToM (N-ToM) in machines. This is particularly crucial given the escalating stakes of the debate on the extent to which machines possess ToM-like abilities and the potential ramifications of overblown claims in AI.<sup>1,2</sup>

\* These authors contributed equally to this work.

<sup>1</sup><https://futureoflife.org/open-letter/pause-giant-ai-experiments/>

<sup>2</sup><https://amcs-community.org/open-letters/>

Two recent papers addressed whether Large Language Models (LLMs; Brown et al., 2020; Bommasani et al., 2021; Zhao et al., 2023) have a ToM, and came to opposite conclusions: Sap et al. (2022) shows they lack this ability and Kosinski (2023) claims this ability has emerged in the newer models spontaneously. The latter was criticized for its flawed methodology (Marcus and Davis, 2023). Ullman (2023) further showed that simple changes to the ToM questions break LLMs. But to paraphrase the saying, hype gets halfway around the world before rigorous experiments put on their boots; other researchers continue to spread the word about N-ToM, claiming that GPT-4 “has a very advanced level of theory of mind” based on a few anecdotal examples (Bubeck et al., 2023).

**Do LLMs have robust N-ToM?** This paper aims to address the discrepancy and limited scope of previous work (that each tested 2 tasks) by performing an extensive evaluation on 6 tasks targeting various aspects of ToM. We also experiment with different probing methods (i.e., generative QA format vs. probability of answer choices). We find that contemporary LLMs demonstrate certain N-ToM abilities, but these abilities are not robust (§4).

**ToM or Spurious Correlations?** We investigate through a series of experiments the factors influencing performance on N-ToM tasks. We show that LLMs perform worse on datasets that were designed to prevent annotation artifacts. We also enhanced the dataset originally proposed by Kosinski (2023) to incorporate adversarial examples inspired by Ullman (2023). We find that the performance of LLMs decreases for adversarial examples, suggesting that LLMs don’t have robust ToM abilities but rather rely on shallow heuristics (§5).

We summarize these findings and additional insights in §6. In particular, we warn against drawing conclusions from anecdotal examples, testing on a few benchmarks, and using psychological tests

designed for humans to test models.

Our contribution: (1) to contribute to the ongoing discussion regarding the capabilities of LLMs. Our work presents the largest and most comprehensive quantitative analyses of ToM to date and it reproduces existing experiments as well as presents many novel results. (2) we also created two new test sets. Those test sets are created by experts and focus on high-quality adversarial examples that point out the use of shallow heuristics (“Clever Hans”) by LLMs and that they are not negligible. (3) We caution against drawing conclusions from anecdotal examples, limited benchmark testing, and using human-designed psychological tests to evaluate models. The consequences of the success of psychological tests do not straightforwardly transfer from humans to models.

## 2 Background: ToM and Clinical Tests

ToM has a long history starting in philosophy (Lewis, 1966) and later in psychology and cognitive science (Premack and Woodruff, 1978). ToM involves understanding mental states, beliefs, desires, intentions, and emotions of the self and of others. Clinical psychology tests were developed to test ToM abilities in humans, such as the false belief and faux pas tests detailed here.

### 2.1 False Belief Test

In a false belief test (Wimmer and Perner, 1983) the examinee is told a story in which a character in the story is exposed to partial information and therefore mistakenly believes in something that is not true (“false belief”) in contrast to the listener who is exposed to the full story.

A widely used clinical psychology task to assess false belief understanding is the *Sally–Anne Test* (Baron-Cohen et al., 1985) or *unexpected transfer*. In this test, Sally has a basket, and Anne has a box. Sally puts a marble in her basket and leaves the room. Anne takes the marble out of the basket and puts it in her box. The examinee is asked about **first order belief**, i.e. *where will Sally look for her marble?*; about the **reality**, i.e. *where is the marble?*; and about their **memory**, i.e. *where was the marble in the beginning?*

The answers are that Sally will look in the basket, where she left the marble. Sally’s belief is false because she is unaware of the marble’s relocation to the box. However, a listener exposed to the entire story knows that the marble is no longer in Sally’s basket and that Sally will look in the wrong place.

In more complex versions, **Second Order Belief** question would be, *where does Anne think Sally will look for her marble?*

In a different version of a false belief task, known as the *Smarties Test* (Perner et al., 1987), the protagonist is dealing with *unexpected content*, i.e., unaware of the actual contents of a container because of false labeling.

### 2.2 Faux Pas Test

Faux Pas occurs when “*a speaker says something without considering if it is something that the listener might not want to hear or know, and which typically has negative consequences that the speaker never intended*” (Baron-Cohen et al., 1999). An example of a faux pas situation is when a guest tells their hosts that they “like cakes except for apple pie”, without realizing that the hosts have made an apple pie for them. The complexity of the situation depends not only on the content of the statement (“except for apple pie”) but also on the context in which it was made (e.g., the host had made an apple pie and the guest was unaware). Faux pas is the “uh-oh!” emotion most people would feel when they reveal the reality of the context. In this context, the statement wouldn’t be problematic if the hosts made a cheesecake instead.

In the original test, the subject is told 10 stories that contain faux pas. At the end of each story, the subject is asked 4 questions: **detection** - *In the story did someone say something that they should not have said?*; **identification** - *What did they say that they should not have said?*; And two questions that differ by story: **comprehensive** - e.g., *Where does the event take place?*, and **false belief** - *did they know or remember that?*

### 2.3 From Human Tests to Machine Tests

Studies have explored the use of NLP techniques to model basic ToM skills. For example, in detecting mental states and emotions (Tausczik and Pennebaker, 2010; Guntuku et al., 2017; Gordon and Hobbs, 2017; Rashkin et al., 2018a,b; Shapira et al., 2021) or by generating a humorous response when the interlocutor is in a playful mood (Shani et al., 2022; Shapira et al., 2023a). Recent work is focused around creating datasets testing whether and to what extent models have ToM (see §3). **It is important to note that the consequences of the success of these tests do not straightforwardly transfer from humans to models** (see §6). One may say that it looks like a discrepancy - what

Dataset	Inspired by Theory/Test	Test Size	Construction	Example
<b>Triangle COPA</b> Gordon (2016)	Interpreting the social behaviour of fictional characters	100	Experts	A circle is in the house moving around. A triangle bursts in through the door. The circle turns around and freezes. How does the circle feel? (a) <b>The circle is surprised by the triangle’s sudden entrance into the room.</b> (b) The circle is excited to see the triangle.
<b>SocialIQA</b> Sap et al. (2019)	Reasoning about motivations, what happens next and emotional reaction	400 random sample out of 37,588	Crowdsourcing	In the school play, Robin played a hero in the struggle to the death with the angry villain. How would others feel afterwards? (a) sorry for the villain (b) <b>hopeful that Robin will succeed</b> (c) like Robin should lose
<b>ToMi</b> Le et al. (2019)	Unexpected transfer task, first and second order false belief; (Baron-Cohen et al., 1985)	400 random sample out of above 1000	Synthetic	Jackson entered the hall. Chloe entered the hall. The boots is in the bathtub. Jackson exited the hall. Jackson entered the dining_room. Chloe moved the boots to the pantry. (Memory) Where was the boots at the beginning? ( <i>bathtub</i> ) (Reality) Where is the boots really? ( <i>pantry</i> ) (First order) Where will Chloe look for the boots? ( <i>pantry</i> ) (Second order) 7 Where does Chloe think that Jackson searches for the boots? ( <i>bathtub</i> )
<b>ToMi’</b> This paper, based on ToMi adjustments		180 questions 30 stories	Experts	<Same story as in ToMi> (Memory) At the beginning, the boots were in the ( <i>bathtub</i> ) (Reality) The boots are really in the ( <i>pantry</i> ) (First order) Chloe will look for the boots in the ( <i>pantry</i> ) (Second order) Chloe thinks that Jackson searches for the boots in the ( <i>bathtub</i> )
<b>epistemic reasoning</b> Cohen (2021)	Verbs, factive and non-factive, that describe epistemic mental states; intra-personal, inter-personal and inference reasoning; (Wimmer and Perner, 1983; Hintikka, 1962)	2000	Experts with 10 templates	Premise: John knows that Ann thinks that there is milk in the kitchen. Hypothesis: Ann thinks that there is milk in the kitchen. ( <i>Entailment = 1</i> ) Hypothesis: John thinks that there is milk in the kitchen. ( <i>Entailment = 0</i> )  Premise: John thinks that Ann knows that there is milk in the kitchen. Hypothesis: Ann thinks that there is milk in the kitchen. ( <i>Entailment = 0</i> ) Hypothesis: John thinks that there is milk in the kitchen. ( <i>Entailment = 1</i> )
<b>Adv-CSFB</b> This paper, based on Kosinski & Ullman (2023)	Unexpected content or transfer task, integrate commonsense reasoning, first-order false belief; (Baron-Cohen et al., 1985; Perner et al., 1987)	183 questions 40 stories	Experts	On the shelf, there is a bottle. It is full of beer and the label on this bottle says “beer”. Mark walks into the room looking for beer and notices the bottle. He has never seen it before. He reads the label. (a) He opens the bottle and looks inside, He can clearly see that it is full of ( <i>beer</i> ) (b) He believes that it is full of ( <i>beer</i> ) (c) He calls his friend to tell them that he has just found a bottle full of ( <i>beer</i> )
<b>FauxPas-EAI</b> Shapira et al. (2023b)	Recognition of faux pas (Baron-Cohen et al., 1999)	176 questions 44 stories	Experts and AI+Experts	Jeff was in an interview. When he finished the interview he sank into a couch in the lobby. Sarah and Tim, the executives who interviewed him, went out of the room and into the lobby while Sarah said: “He asked for 179K, could have asked for much more. We need to make him an offer quickly at the salary he asked for”. Tim saw Jeff and said “Oh, goodbye”. (Faux Pas) In the story did someone say something that they should not have said? ( <i>Yes</i> ) (Identification) What did they say that they should not have said? (“ <i>He asked for.</i> ”) (Comprehensive) Who was interviewed? ( <i>Jeff</i> ) (False Belief) Did Sarah know Jeff is sitting on one of the couches in the lobby? ( <i>No</i> )

Table 1: Theory of Mind (ToM) datasets used in this work.

would be the point of testing ToM tasks if there is no direct transfer between human and LLMs. Half of our tests aren’t meant for humans. In addition, ultimately, we argue that when LLMs achieve positive results on benchmarks designed for humans, we should be careful about interpreting them as “LLMs have this human capability.” To that end, we used these tests designed for humans to show that the performance of LLMs is sensitive to the introduction of adversarial examples, which indicates reliance on shallow heuristics and/or familiarity with specific examples from their pretraining (as opposed to N-ToM). As a general recommendation, we think future dataset creators should take this in mind and not blindly apply tests designed for humans to models.

### 3 Data

We used all datasets listed in Table 1 in our experiments. A brief description of each dataset is in

Appendix §7.3. The creation of **ToMi’** (based on ToMi) is described immediately after ToMi (§7.3). Below is the creation of **Adv-CSFB** (§3.1).<sup>3</sup>

#### 3.1 Creation of Adv-CSFB

Inspired by the disagreeing conclusions reached by prior work, we introduce the *ADversarial Commonsense with False-Belief* (Adv-CSFB) dataset. Adv-CSFB contains 110 examples of the unexpected contents task and 73 examples of the unexpected transfer task (§2.1). Each manually-created example in the dataset consists of a short paragraph describing two objects  $O_1$  and  $O_2$ , and is followed by questions pertaining to **reality**, i.e. whether a certain container contains  $O_1$  or  $O_2$ , and the protagonist’s **belief** regarding the content.

The examples in Adv-CSFB are categorized to *false belief*, i.e. the original examples from ToM-

<sup>3</sup>The code and data is available at: [https://github.com/salavi/Clever\\_Hans\\_or\\_N-ToM](https://github.com/salavi/Clever_Hans_or_N-ToM)

k (Kosinski, 2023), *true belief*, and *adversarial* examples inspired by Ullman (2023).

**False Belief.** In the false-belief examples from Kosinski (2023), the protagonist’s belief about the container’s contents is different from its actual contents. The examples are variants of the corresponding original tests, e.g., the unexpected contents examples are variants of the Sally-Anne test. Notably, Kosinski only created false-belief scenarios.

**True Belief.** For a more fair evaluation setup, we enhance the unexpected contents task with *true belief* examples, i.e. in which the protagonist’s belief about the content of the container is the same as its actual contents. We do so by modifying each of the *false belief* examples such that the label now indicates the true content of the container,  $O_1$ . We mention the alternative content  $O_2$  in a way that doesn’t change the answer, e.g. Mark walks into the room looking for  $O_2$  but finds a bag with  $O_1$  labelled as “ $O_1$ ”. One author of this paper created a variation for each applicable example, which was then verified by another author.

**Adversarial Examples.** Ullman (2023) showed that LLMs that achieve near-perfect performance on the *false belief* examples fail to solve a number of adversarial examples where new information is introduced. In particular, LLMs still predict false belief even when new information suggests that the protagonist should know the truth. For example, the LLM predicts that a protagonist looking at a bag full of popcorn that is labelled as “chocolate” believes the bag is full of chocolate, even if the bag is transparent or if the protagonist cannot read. Ullman’s counter examples are sufficient in showing that LLMs did not robustly acquire ToM abilities. To further quantify the LLMs’ abilities, we created up to 4 additional examples for each of the *false belief* examples, following each of the alterations suggested by Ullman (2023): transparent access, uninformative label, trustworthy testimony, and late labels for the unexpected contents task, and transparent access, in→on, trustworthy testimony, and other person for the unexpected transfer task (see Appendix 7.2 for an example for each variation). Again, the examples were created by one author and verified by another.

## 4 Experiments & Results

To investigate the ToM abilities of LLMs, we explore various aspects. The first experiment presents

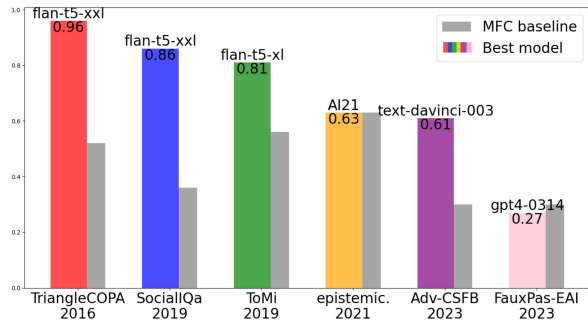


Figure 1: Accuracies of top-performing models on each of the ToM tasks, compared to a most frequent class (MFC) baseline. For several datasets, the best model achieves performance comparable to the MFC baseline, suggesting very limited ToM ability.

a meta-evaluation of 15 LLMs evaluated on multiple ToM-related datasets in a zero-shot manner (§4.1).<sup>4</sup> We then investigate to what extent LLMs are sensitive to the probing method (§4.2). We examine the performance of 15 different LLMs. LLMs’ list and technical details regarding prompting and decoding parameters in Appendix 7.4.

### 4.1 How well do LLMs perform on ToM tasks? Meta-Evaluation

We conducted an evaluation of the performance of 15 LLMs in a zero-shot manner (Liu et al., 2021) on all ToM-related datasets considered (§3), and compare to a most-frequent-class (MFC) baseline that always predicts the most frequent answer in each dataset. The summary of the results is presented in Figure 1, and the complete results, the exhaustive accuracy for all LLMs, in Table 2.

Our findings demonstrate that while some LLMs achieve near perfect accuracies on some datasets (e.g., TriangleCOPA with 96% accuracy by `flan-t5-xxl`), others datasets remain challenging for LLMs with considerably lower performance. For instance, the best performing LLM on the FauxPasEAI dataset achieved 27% accuracy which is inferior to a simple most-frequent-class baseline.

Notably, the best LLMs performance seems correlated to the dataset’s age (i.e., the older the dataset, the better the performance). This trend could be attributed to the fact that the increasing sophistication of LLMs is driving the creation of more challenging datasets, prompting researchers

<sup>4</sup>In the initial stage of the research, we tested for varied temperature settings  $\{0, 0.01, 0.1, 0.2, 0.3, 0.6, 0.9, 1\}$ , each setting 100 times with different random seeds. This phase helped us determine that the most preferable results came from fixing the temperature to zero (the results and the explanation are detailed in Appendix 7.5). We later continued for simplicity with the zero fixed temperature results only.

Model	Triangle COPA	SocialIQa	ToMi	Epistemic Reasoning	Adv-CSFB	FauxPas EAI
MFC	52	36	56	<b>63</b>	32, 25	55, <b>30</b>
Flan-ul2	95	84	81	60	65, 23	60, 07
Flan-T5-xxl	<b>96</b>	<b>86</b>	75	57	60, 17	68, 18
Flan-T5-xl	92	81	<b>81</b>	61	65, 21	68, 14
Flan-T5-large	92	79	78	44	61, 24	53, 07
Flan-T5-base	84	67	81	52	58, 36	52, 07
Flan-T5-small	58	46	52	54	42, 26	58, 07
gpt4-0314	94	79	70	43	75, 57	<b>74, 27</b>
gpt-3.5-turbo-0301	84	67	70	45	70, 42	73, 25
text-davinci-003	95	60	67	59	<b>79, 61</b>	67, 07
text-davinci-002	92	19	39	58	76, 53	63, 14
j2-grande-instruct	06	08	00	37	64, 47	58, 00
j2-jumbo-instruct	48	04	13	47	63, 45	45, 00
j2-grande	75	45	50	<b>63</b>	57, 40	45, 00
j2-jumbo	68	40	54	<b>63</b>	57, 36	38, 00
j2-large	58	38	53	<b>63</b>	57, 41	31, 00

Table 2: Accuracy (%) of LLMs on different datasets compared to a most frequent class baseline. For Adv-CSFB and FauxPas-EAI we report two metrics: question level and story level.

		LM	MC	CoT
Siqa	davinci-003	55	60	68
	GPT-3.5	-	67	69
	GPT-4	-	79	72
ToMi	davinci-003	67	67	71
	GPT-3.5	-	70	73
	GPT-4	-	70	73

Table 3: Accuracy of the recent GPT models on a random sample of 400 instances from SocialIQa (Siqa) and ToMi. The probing method affects the performance. For example, in Siqa there is a 7% difference in the accuracy of GPT-4 between MC-probing and CoT-probing.

to set a higher bar. Another possibility is that LLMs have had more opportunities to train on the older datasets, resulting in better performance (see §7.6).

Based on this meta-evaluation, our results suggest are that while some models exhibit strong ToM abilities on some datasets, **no model robustly exhibits ToM on all datasets**. These findings are consistent with Sap et al. (2022) and Ullman (2023).

## 4.2 How sensitive are LLMs to the probing technique?

We examine the effect of the different probing methods detailed below on LLM performance. Certain techniques have shown to be superior to others (e.g., Wei et al., 2023). However, we argue that to claim that a model has N-ToM abilities, it is essential that it performs well across probing techniques. On one hand, the most efficient method can potentially reveal latent capabilities, while on the other hand, there is a reasonable expectation for LLMs

to succeed in the tasks regardless of the probing approach used to extract information.

**LM-probing** predicts the option with the highest probability (Brown et al., 2020; Sap et al., 2022).

**MC-probing** prompts the LLM with the context, question, and answer choices, and asks it to generate the answer in the form of “a, b, c”. This method is applicable for LLMs such as GPT-3.5 and GPT-4 that don’t produce probabilities (Hu et al., 2022).

**CoT-probing** asks the model to first “reason” about the question step-by-step and then give a final answer, which generally contributes to better performance (Wei et al., 2023).<sup>5</sup>

Table 3 shows that the probing techniques influence the LLM performance on both datasets. CoT generally demonstrates enhanced performance (Camburu et al., 2018; Schwartz et al., 2020; Wei et al., 2023). Nonetheless, there are cases where this trend does not hold, since the reasoning may result in erroneous conclusions (Jung et al., 2022).

## 5 Clever Hans vs. Generalized Reasoning

We conducted a series of experiments to enhance our understanding of the factors influencing performance in the context of N-ToM tasks. The research question that guided us was: Do the models that solve the tasks possess a general ability or do they rely on memorization and shallow heuristics (“Clever Hans”; Kavumba et al., 2019)?

<sup>5</sup>We used zero-shot without providing reasoning examples.

Dataset	ToMi'		ToM-k
Subset	All question	No second order	All questions
text-davinci-003	10	21	87
GPT-3.5	27	48	65
GPT-4	20	52	87

Table 4: Comparison of LLMs’ accuracy on ToM-k vs. ToMi’. Both datasets examine the unexpected transfer task. ToM-k contains only first-order questions. ToMi’ also contains simple alternations such as omission or duplication of information, and second-order questions. The subset “No second order” was created manually to better compare to ToM-k dataset. Lower accuracy suggests the dataset is more robust to spurious correlations.

We showed the existence of shallow heuristics and that they are not negligible. Our investigations, especially with ToMi’ (§5.1) and Adv-CSFB (§5.2) are meant to isolate the use of heuristics by design by introducing adversarial examples that could reveal the LLMs’ use of heuristics. We follow extensive prior work on using adversarial examples to show reliance on spurious correlations (e.g., Hendrycks et al., 2021; Wu et al., 2021; Kaushik et al., 2019; Geirhos et al., 2020; Glockner et al., 2018).

### 5.1 Do LLMs Rely on Spurious Correlations?

ToMi and ToM-k datasets examine the unexpected transfer task. While ToM-k contains variants of the original Sally-Annie test, ToMi also contains simple alternations such as omission or duplication of information that create negative examples (e.g., Appendix 7.1) and second-order questions.

To ensure a fair comparison between the question answering format of ToMi and the sentence completion format of ToM-k (see the effect of probing methods on performance in §4.2), we adjusted ToMi to match the sentence completion format (details about the adjustments can be found at §7.3). Additionally, we analyzed the results separately for second-order questions in order to facilitate a more accurate comparison with the ToM-k dataset.

Table 4, shows significantly lower scores in ToMi’. The notable discrepancy between the performance of the two datasets suggests that the model’s abilities are not based on generalization. Instead of true understanding of the problem at hand, such as accurately determining one’s exact thoughts, the model might be recognizing patterns from the Sally-Anne story in other ToM-k examples and generating responses based on those patterns. Conversely, the performance on ToMi’ is worse because it is more robust to spurious correlations.

### 5.2 Is N-ToM Robust to Adversarial Changes?

To test the robustness of the LLMs’ N-ToM, we test the performance of GPT models on each of the categories in Adv-CSFB (§3.1), using MC-probing. To ensure correct formatting and prevent unintended outputs (e.g., explanation of why the answer is correct), we prepend to the prompt *one* out-of-domain example from ToMi, which has a similar format. We report the average accuracy of questions 2 and 3, both focusing on an agent’s belief rather than objective truth. Finally, to ensure maximum reproducibility of the results, we set the temperature to 0. **Our main finding is that LLMs don’t exhibit robust performance across different categories. In particular, later LLMs excel in some categories while completely failing on others.** Details below.

Figure 2 illustrates the performance of GPT models on different categories within the unexpected transfer segment of Adv-CSFB. It is evident that both *false belief* (i.e. the original examples from ToM-k) and *trusted testimony* (i.e., someone tells the protagonist that the object has been moved) have improved in newer models. GPT-4 achieves 97.5% and 83.3% on the two categories respectively. Nevertheless, there has been a gradual decline in the performance of subsequent models on other categories, such as *other person* (from 93.8% by davinci-002 to 68.8% by GPT-4), *in→on* (from 71.4% by davinci-002 to 0% by GPT-4), and *transparent access* (from 66.7% by davinci-002 to 0% by GPT-4).

Figure 3 showcases the performance of the GPT family on various categories within the unexpected contents segment. It becomes apparent that, akin to the unexpected transfer segment, newer models such as GPT-3.5-Turbo and GPT-4 demonstrate improved performance in handling samples that involve *false belief* and *transparent access* (i.e., the container is transparent). Furthermore, nearly all models since text-davinci-002 exhibit strong performance on *true belief* samples. However, both GPT-3.5-Turbo and GPT-4 experience a substantial decline in performance compared to their earlier counterparts when it comes to *transparent access*, *late label* (e.g., the protagonist is the one who wrote the label), and *uninformative label* (i.e., the protagonist can’t read the label).

We regenerated the responses multiple times, consistently obtaining similar results, so we can conclude that the models exhibit confidence in their predictions, even if they are incorrect. It is impor-

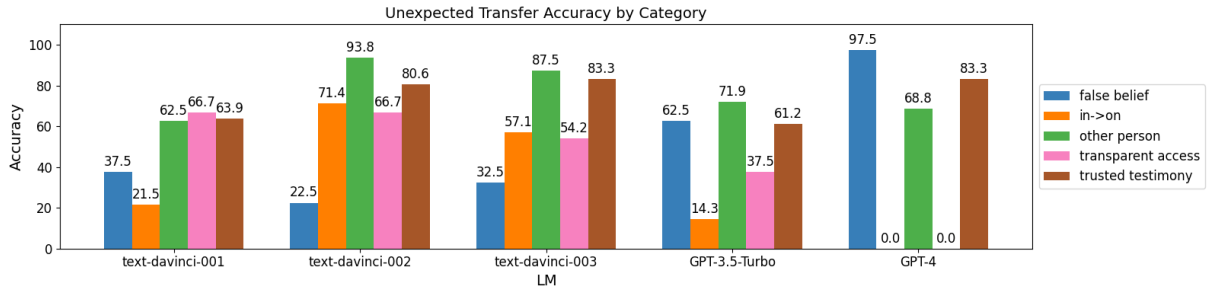


Figure 2: Performance of GPT models on various categories within the unexpected transfer segment of Adv-CSFB. Results are the average accuracy of question 2 (e.g., *Maria thinks that the bananas are in the \_*) and question 3 (e.g., *When Maria comes back, she will first look for the bananas in the \_*), which focus on agent’s beliefs rather than objective truth. Notably, GPT-4 achieves an accuracy of 97% on the subset of false belief samples (original examples from ToM-k), while failing on adversarial samples as transparent access or relationship change (in→on).

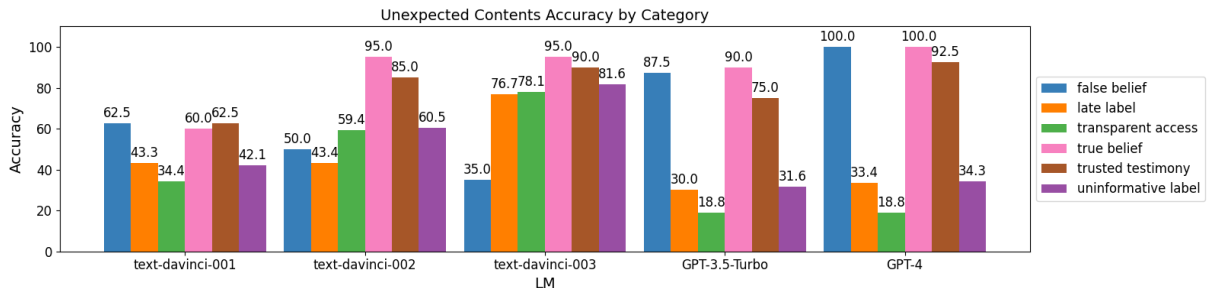


Figure 3: Performance of GPT models on various categories within the unexpected content segment of Adv-CSFB. The results are average accuracy of question 2 (e.g. *He believes that it is full of \_*) and question 3 (e.g. *He calls his friend to tell them that he has just found a bottle full of \_*), that focus on agent’s beliefs rather than objective truth.

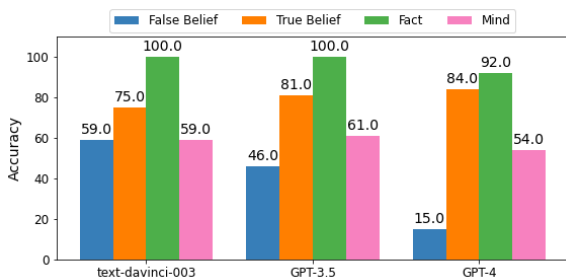


Figure 4: ToMi’s accuracies with different splits of the dataset. While GPT-3.5 (the best-performing model) achieves a total of 0.7 accuracy score (see Figure 1), it achieves only 0.46 on the subset questions “false belief”.

tant to note, however, that the results obtained from LM-probing may slightly differ from MC-probing. In MC-probing, even with our 1-shot setup, the model may produce responses that are not applicable, such as “none of the above” or “both”. This is particularly noticeable in verbose models like GPT-3.5-Turbo and GPT-4. These models tend to be careful to avoid providing incorrect answers and, as a result, generate longer phrases. With that said, as we argue in §4.2, a LLM exhibiting robust N-ToM ability should be able to answer questions correctly regardless of the probing method.

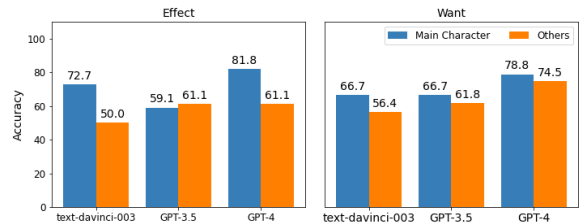


Figure 5: SocialIqa’s accuracies for the questions that focus on the main character vs. others. While GPT-4 (the best-performing model) achieves a total of 79% accuracy score, it achieves only 61% on the subset questions of “others effect”.

### 5.3 Are Spurious Correlations a Trend?

In previous experiment §5.2, we saw that Adv-CSFB contain both difficult and easy questions. Here we show this recurring phenomenon across two more datasets, inspired by Sap et al. (2022).

Figure 4 describes ToMi accuracies on different question types: (1) facts vs. beliefs (mind), and (2) true or false beliefs. While GPT-3.5 (the best-performing model) achieves 81% accuracy, on the subset questions “false belief”, it achieves close to random performance (46%).

Figure 5 shows the SocialIqa accuracies for questions focusing on the main character vs. others. While GPT-4 (best-performing model) achieves a total of 79% accuracy score, on the subset questions of “others”, it achieves only 74.5%.

## 6 Summary of Findings and Insights

We investigated whether LLMs robustly display N-ToM abilities. By quantifying their performance on 6 N-ToM benchmarks, we found that while some datasets have been nearly “solved” (e.g., Triangle-COPA with 96% accuracy by `flan-t5-xxl`), others remain challenging for LLMs with considerably lower performance (e.g., FauxPas-EAI with 27% accuracy by GPT-4, which is even below the majority baseline). We also created Adv-CSFB, a new ToM benchmark designed to uncover whether LLMs solve ToM questions for the right reasons, or merely rely on surface cues and shallow heuristics.

**So... Do LLMs have ToM?** Our results show that while some datasets have been successfully solved, others remain challenging for LLMs. Thus, **models do not have robust N-ToM abilities**. These findings are inconsistent with Kosinski (2023), who claimed that ToM has emerged in LLMs as a byproduct of their development, a claim further echoed by Bubeck et al. (2023). We argue that these conclusions were over-generalized based on a specific aspect of ToM and a small number of examples (40 for Kosinski (2023) and 10 for Bubeck et al. (2023)). Following Ullman (2023), we empirically showed that even the best models fail on small variations of the original tasks, proving that even GPT-4 does not display robust N-ToM abilities.

**Clever Hans, Heuristics & Shortcuts** The performance gaps between different question types suggests that **LLMs rely on shortcuts, heuristics, and spurious correlations**, which often lead them astray. In Adv-CSFB (§5.2), the bad performance on some of the adversarial categories might be partly attributed to *reporting bias* (Gordon and Van Durme, 2013; Shwartz and Choi, 2020). People don’t share obvious facts (Grice, 1975), so it is likely that LLMs are biased towards generating surprising rather than unsurprising continuations. In most of these categories, the protagonist belief is the same as the truth, making a boring story.

Furthermore, the newer models such as GPT-3.5 and GPT-4 are trained in addition to the LM objective to follow natural language instructions and generate helpful answers. This might make them cooperative and lead to LLMs assuming that all details are important, rather than that the input is adversarial. For example, they might pay too much attention to the mention of the false label in the unexpected contents task, failing to see that

the label doesn’t matter if the person can’t read it or if the container is transparent. The fact that LLMs perform reasonably well on true belief examples (Figure 3) might be attributed to recency bias (O’Connor and Andreas, 2021), since the correct content is typically the last one to be mentioned.

Finally, we reassess the finding of Sap et al. (2022) that LLMs perform better on predicting the mental states of the main character vs. others (§5.1); Sap et al. suggested that this might be due to centering theory (Grosz et al., 1995) i.e., texts tend to focus on describing a single protagonist.

### **ELIZA Effect & Anecdotal Generative vs. Automatic Large-Scale Multiple-Choice Testing**

The impressive anecdotal examples produced by LLMs (e.g., ChatGPT/GPT4 web-demo; Bubeck et al., 2023), tends to captivate non-expert. However, it is important to recognize that these models are designed to generate text that appears high-quality to human observers (Ouyang et al., 2022). This inherent bias in their design can lead to the “ELIZA effect” (Weizenbaum, 1976; Shapira et al., 2023b), i.e. the human assumption that computer behaviors are analogous to human behaviors. Thus, **the illusion that a LLM has acquired human-like N-ToM often says more about the human reader than the model itself** (Whang, 2023).

Moreover, later models are by design trained to practice “epistemic humility” (i.e., hedge and provide multiple possible answers; Ouyang et al., 2022, p.17). This often leads them to provide rationales for each given answer without committing to actually answering the question. But humans might fall prey to confirmation bias and simply see the right answer and its rational and conclude that the model has gotten it correctly. We thus argue that in order to conclude whether a certain model possesses a certain ability, it is crucial to quantify the performance across multiple large-scale datasets, preferably using an automatic evaluation method.

### **Using psychological tests designed for humans on LLMs**

Clinical psychology tests designed for humans are carefully constructed and vetted to ensure that they have external and internal validity, i.e., they measure what they aim to measure (Frank et al., 2023). While there is evidence that a person’s success in one ToM task can indicate their ToM abilities (e.g., Milligan et al., 2007), this does not transfer to models. Therefore, it is important to be cautious when drawing conclusions about ToM in models based on their performance on a few



tasks (Marcus and Davis, 2023). In general, when a system succeeds on an instrument designed for humans, we can't draw the same conclusions as we would for humans (e.g., that they have ToM). Instead, we need to consider other explanations (e.g., that they are relying on heuristics). The same holds in the other direction, when analyzing how models work in order to learn about the human brain.

### **Machine intelligence and Anthropomorphism**

There is also a need for caution when discussing the abilities of machines in relation to concepts referring to human cognition, such as ToM. While it is common in computer science to use human-related concepts and metaphors for AI systems, we caution readers to interpret “neural ToM” carefully and without aiming to make claims about “AI cognition,” especially since given our propensity for anthropomorphizing non-human animals and computers (Epley et al., 2007; Kim and Sundar, 2012); our performance measures on these benchmarks is not meant as an endorsement of the pursuit of a human-like social intelligence for AI systems.<sup>6</sup> Instead, in light of the hype around AI and its “intelligence,” we sought out to provide a more sober look at the empirical performance of LLMs on tasks related to social intelligence and ToM.

**“Solving” a ToM benchmark is necessary but not sufficient** Methodologically, if a model fails at least one ToM task, it does not have ToM in general. Success on one example or task is not a sound proof that a model has ToM. **Future work will need to continue to develop benchmarks testing various ToM aspects**, and these benchmarks will need to be designed to assess LLMs directly rather than using clinical tests designed for humans.

Additionally, reporting the aggregated performance of LLMs on benchmarks obscures the performance differences across questions of different types and complexities. To overcome this, one approach is to pair a difficult question with an easy question, requiring model to answer both correctly. This methodology resembles the “joint score” employed in FauxPas-EAI, Adv-CSFB, and ToMi. In situations where pairing is challenging, a recommendation for future works is that dataset difficulty could be evaluated by calculating the final score across different splits of the dataset. The difficulty level of the dataset can then be determined based on the lowest score obtained among these splits.

<sup>6</sup>We leave the question of whether LLMs could develop human-like cognition and ToM up to philosophers.

**Emergence vs. Supervised Learning vs Training on the Test set** Prior work claimed that ToM abilities emerged as a byproduct of the LLM training (Kosinski, 2023). We argue that **claims about emergence are (i) unfounded, and (ii) unfalsifiable without access to the LLMs’ training data**.

To make a statement regarding emergent ToM, a careful experiment is needed to ensure that ToM did indeed appear spontaneously and not as a result of other factors such as training on related datasets, exposure to descriptions of clinical tests online, interactions with users, and more.<sup>7</sup> However, since the data used to train the GPT models is not publicly available, it is impossible to quantify the degree of the potential data leakage.<sup>8</sup> We echo calls by Dodge et al. (2021) for increased transparency and open-access to the training data of LLMs, which is crucial for scientifically valid and reproducible experiments (Rodgers, 2023).

**Conclusion** Based on our research and replication studies, we conclude that contemporary LLMs demonstrate an enhanced yet limited degree of N-ToM abilities. We find that their abilities are not robust, and in some instances, we identify evidence of their over-reliance on simple heuristics rather than generalized reasoning. In our research, we concentrated on analyzing the current models and highlighted pitfalls that the current literature encounters. We caution against concluding anecdotal examples, limited benchmark testing, and using human-designed psychological tests to evaluate models.

### **Limitations**

**Benchmark scope and more versions of ToM tests** The datasets used in this study were limited in scope and size; ToM is required in most human interaction, and thus unbounded in scope. In our study, we used only 6 different test sets. There are many other versions of ToM clinical tests (For a detailed review see Osterhaus and Bosacki, 2022) or new NLP tasks that check different aspects of N-ToM e.g., through intersectional ToM (Kim et al., 2023; Zhou et al., 2023).

**Ambiguous Test Questions** Parts of the datasets could be ambiguous, either due to lack of context or inherent ambiguity (Plank, 2022). We used existing

<sup>7</sup>OpenAI acknowledged that GPT-4 was trained on test data from BIGBench (OpenAI, 2023, footnote 5).

<sup>8</sup>See Appendix 7.6 for an attempt to quantify data leakage.

clinical tests formulated by previous researchers and tested on humans. There is an inherent ambiguity in every question that has some judgment. Perhaps if we had asked clearly, LLMs might have been more successful, but we want to ask in ways that are conducted in front of humans because we aim for LLMs to understand language at the same level as humans.

Sharma et al. (2023) in their recent study suggests that human feedback may also encourage model responses that match user beliefs over truthful ones (a behavior known as sycophancy) and that analyzing existing human preference data, they found that when a response matches a user’s views, it is more likely to be preferred, and humans prefer convincingly written sycophantic responses over correct ones a non-negligible fraction of the time.

Due to this potential ambiguity, some LLMs were safeguarded and refused to answer certain questions; while we attempted to instruct them to respond in the correct format, some LLMs still did not output the right format. This was only an issue for MC-probing, but probability distributions were not available for all LLMs. Future work should investigate how to mitigate this issue via better instructions or methods that map generated answers to multiple choice better (e.g., Niu et al., 2021; Bulian et al., 2022).

### Improving N-ToM (by CoT or other methods)

We engaged with the concept that to determine that LLMs display robust N-ToM **abilities in its outputs**, it should perform accurately across many different tasks and many different probing styles. As long as we find an adversary method that makes the models fail in N-ToM tasks (e.g., by different probing methods or by different test sets) this indicates that the models don’t hold a robust N-ToM (intrinsic abilities, if exist, don’t contradict this claim). One of the goals of our work is to critically examine claims made by Kosinski (2023) and Bubeck et al. (2023), which all rely on zero-shot setups, so we kept zero-shot to be experimentally rigorous. Some works use CoT prompting, and others e.g., algorithms that involve world tracking (Sclar et al., 2023). However, these works are all focused on improving neural ToM as measured by those benchmarks, which is not our goal. Our goal is to examine zero-shot ToM abilities of LLMs without explicit handholding, to rebut previous claims that ToM abilities have somehow “emerged” in zero-shot format. It is feasible that techniques such

as CoT would enhance the performance of GPT-4 where it currently performs poorly. Nevertheless, we need to exercise caution to **ensure that the utilization of methods like CoT or others does not excessively guide the models** by essentially revealing the task structure to them—just like Clever Hans who appeared proficient in math merely due to subtle hints given by the owner.

**Limited text-only LLMs** Our experiments were conducted with a limited number of LLMs that were accessible at the time of writing, and we did not explore the full spectrum of LLMs that are currently available. Future work could explore the N-ToM abilities displayed by other LLMs, and additionally, explore multimodal models.

### ToM tasks that require low-level processes

There is a range of ToM abilities that require low-level processes that models/humans have to master to answer true ToM questions, some of the failures in tests point to failures in lower-level skills (e.g., Ullman’s transparent container access requires knowledge of physics, which presumably a blind baby would not have either). ToM should be constrained in given requirements and reasoning level, like in advanced clinical psychological ToM tasks that require world-model too, e.g., the faux pas test for 7-11 years-old children (Baron-Cohen et al., 1999). Note that we also looked at “pure” ToM tasks and showed that current models failed on them too (Table 4).

### Ethical Statement

**Data.** All the existing and new datasets used in this study are publicly available. The narratives were evaluated by the authors to ensure that they do not contain offensive content.

**Models.** LLMs may generate offensive content if prompted with certain inputs. However, we used them for evaluation only, with non-offensive inputs, and we did not record their responses.

### References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- 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(5):407–418.
- Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- 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*.
- Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. [Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 291–305, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.
- Michael Cohen. 2021. Exploring roberta’s theory of mind through textual entailment.
- Hal Daumé. 2017. *A course in machine learning*. Hal Daumé III.
- Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. [Documenting large webtext corpora: A case study on the colossal clean crawled corpus](#). In *EMNLP*.
- Nicholas Epley, Adam Waytz, and John T Cacioppo. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, 114(4):864.
- Michael C Frank, Mika Braginsky, Julie Cachia, Nicholas Coles, Tom Hardwicke, Robert Hawkins, Maya B Mathur, and Rondeline Williams. 2023. Experimentology: An open science approach to experimental psychology methods.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.
- Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. *arXiv preprint arXiv:1805.02266*.
- Andrew Gordon. 2016. Commonsense interpretation of triangle behavior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.
- Andrew S Gordon and Jerry R Hobbs. 2017. *A formal theory of commonsense psychology: How people think people think*. Cambridge University Press.
- Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.
- Erin Grant, Aida Nematzadeh, and Thomas L Griffiths. 2017. How can memory-augmented neural networks pass a false-belief task? In *CogSci*.
- Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.
- 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.
- Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.
- Fritz Heider and Marianne Simmel. 1944. An experimental study of apparent behavior. *The American journal of psychology*, 57(2):243–259.
- Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271.
- Jaakko Hintikka. 1962. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Ithaca: Cornell University Press.
- Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2022. A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv preprint arXiv:2212.06801*.

- Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. [Maieutic prompting: Logically consistent reasoning with recursive explanations](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1266–1279, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.
- Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. 2019. [When choosing plausible alternatives, clever hans can be clever](#). In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 33–42, Hong Kong, China. Association for Computational Linguistics.
- Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. Fantom: A benchmark for stress-testing machine theory of mind in interactions. *arXiv preprint arXiv:2310.15421*.
- Youjeong Kim and S Shyam Sundar. 2012. Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior*, 28(1):241–250.
- Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.
- 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.
- David K Lewis. 1966. An argument for the identity theory. *The Journal of Philosophy*, 63(1):17–25.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.
- Gary Marcus and Ernest Davis. 2023. How not to test GPT-3. <https://garymarcus.substack.com/p/how-not-to-test-gpt-3>. Accessed: 2023-2-19.
- Karen Milligan, Janet Wilde Astington, and Lisa Ain Dack. 2007. Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child development*, 78(2):622–646.
- Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Thomas L Griffiths. 2018. Evaluating theory of mind in question answering. *arXiv preprint arXiv:1808.09352*.
- Yilin Niu, Fei Huang, Jiaming Liang, Wenkai Chen, Xiaoyan Zhu, and Minlie Huang. 2021. [A semantic-based method for unsupervised commonsense question answering](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3037–3049, Online. Association for Computational Linguistics.
- Joe O’Connor and Jacob Andreas. 2021. [What context features can transformer language models use?](#) In *ACL*, pages 851–864, Online. Association for Computational Linguistics.
- OpenAI. 2023. [GPT-4 technical report](#).
- Christopher Osterhaus and Sandra L Bosacki. 2022. Looking for the lighthouse: A systematic review of advanced theory-of-mind tests beyond preschool. *Developmental Review*, 64:101021.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *NeurIPS*.
- Josef Perner, Susan R Leekam, and Heinz Wimmer. 1987. Three-year-olds’ difficulty with false belief: The case for a conceptual deficit. *British journal of developmental psychology*, 5(2):125–137.
- Barbara Plank. 2022. [The “problem” of human label variation: On ground truth in data, modeling and evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.
- Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018a. [Modeling naive psychology of characters in simple commonsense stories](#). In *ACL*.
- Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018b. [Event2mind: Commonsense inference on events, intents, and reactions](#). In *ACL*.
- Anna Rodgers. 2023. [Closed ai models make bad baselines](#). *Towards Data Science*. Accessed 2023-05-23.
- Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. [Neural theory-of-mind? on the limits of social intelligence in large lms](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialliqa: Commonsense reasoning about social interactions. In *AAAI*.
- 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. *arXiv preprint arXiv:2306.00924*.
- Chen Shani, Alexander Libov, Sofia Tolmach, Liane Lewin-Eytan, Yoelle Maarek, and Dafna Shahaf. 2022. “alexa, do you want to build a snowman?” characterizing playful requests to conversational agents. In *CHI 2022*.
- Natalie Shapira, Dana Atzil-Slonim, Daniel Juravski, Moran Baruch, Dana Stolorowicz-Melman, Adar Paz, Tal Alfi-Yogev, Roy Azoulay, Adi Singer, Maayan Revivo, Chen Dahbash, Limor Dayan, Tamar Naim, Lidar Gez, Boaz Yanai, Adva Maman, Adam Nadaf, Elinor Sarfati, Amna Baloum, Tal Naor, Ephraim Mosenkis, Badreya Sarsour, Jany Gelfand Morgenshteyn, Yarden Elias, Liat Braun, Moria Rubin, Matan Kenigsbuch, Noa Bergwerk, Noam Yosef, Sivan Peled, Coral Avigdor, Rahav Obercyger, Rachel Mann, Tomer Alper, Inbal Beka, Ori Shapira, and Yoav Goldberg. 2021. Hebrew psychological lexicons. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 55–69.
- Natalie Shapira, Oren Kalinsky, Alex Libov, Chen Shani, and Sofia Tolmach. 2023a. Evaluating humorous response generation to playful shopping requests. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part II*, pages 617–626. Springer.
- 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.
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askill, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.
- Vered Shwartz and Yejin Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629, Online. Association for Computational Linguistics.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.
- Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.
- Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby, and Donald Metzler. 2022. Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*.
- Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.
- Joseph Weizenbaum. 1976. *Computer power and human reason: From judgment to calculation*. WH Freeman & Co.
- Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.
- Oliver Whang. 2023. Can a machine know that we know what it knows? *The New York Times*.
- 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.
- Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. *arXiv preprint arXiv:2101.00288*.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

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

## 7 Appendices

### 7.1 Example of ChatGPT Failure

Table 5 shows an example inspired by the ToMi dataset. The unexpected transfer test discusses an unexpected (false belief) rather than trivial (true belief) case. ChatGPT solves the more complex task (false belief) while failing on the trivial task, likely due to its exposure to the Sally-Anne task.

### 7.2 Ullman’s Variations



Figure 6: An illustration of Ullman’s Variations for the unexpected contents task. Image is taken from Ullman (2023).

Figures 6 and 7 illustrate the variations proposed by Ullman for the examples in ToM-k.

### 7.3 Existing Benchmarks & Variants

**Triangle COPA (Gordon, 2016).** A set of 100 problems, each describes a short sequence of events involving the characters of the Heider and Simmel (1944) film: two triangles and a circle moving around a box with a hinged opening. The questions

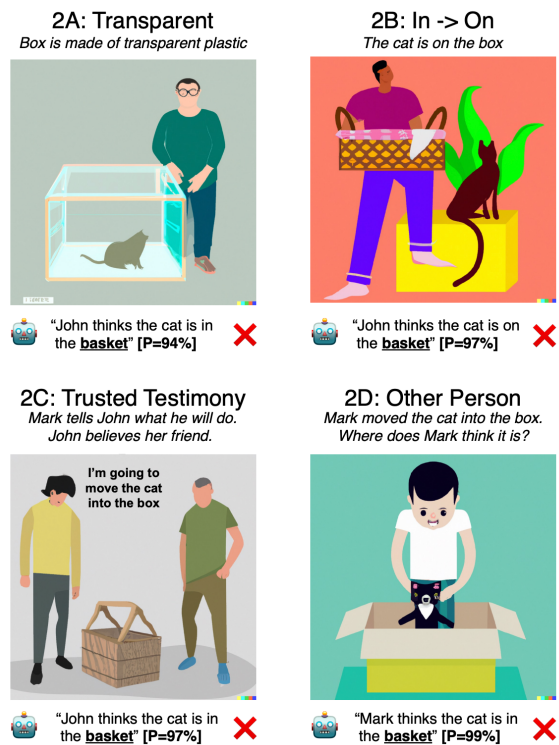


Figure 7: An illustration of Ullman’s Variations for the unexpected transfer task. Image is taken from Ullman (2023).

require understanding the action sequence and social reasoning, and two answer choices are given.

**SocialIQA (Sap et al., 2019).** A large-scale (38k) dataset for commonsense reasoning about social situations. Questions in SocialIQA require reasoning about people’s motivations and mental states, causes and effects. The questions in SocialIQA were crowdsourced along with correct and incorrect answers. Additional distractors were added by using the correct answer for a different question on the same context, using a framework that mitigates stylistic artifacts.

**ToMi (Le et al., 2019).** Inspired by the Sally-Anne test, ToMi is an improved iteration of prior datasets (Weston et al., 2015; Grant et al., 2017; Nematzadeh et al., 2018), comprising over 1,000 distinct stories and questions regarding memory, reality, and first and second-order false belief. This synthetic dataset was automatically generated for a range of essential objects and actions and was further processed for artifact prevention.<sup>9</sup>

**ToMi Adjustments (ToMi’)** ToMi stories are in question-answering format. We randomly sampled

<sup>9</sup>See Appendix 7.1 for an example.

False Belief (Positive Example)	True Belief (Negative Example)
Natalie and Michal are in the same room. Natalie puts a ring in <b>box A</b> . Natalie leaves the room. Michal opens box A, takes the ring from inside and transfers it to box B. Natalie enters the room. Where will Natalie look for the ring?	Natalie and Michal are in the same room. Natalie puts a ring in box A. Michal opens box A, takes the ring from inside and transfers it to <b>box B</b> . Where will Natalie look for the ring?
ChatGPT: Natalie will look for the ring in <b>box A</b> , where she initially put it. She is not aware of Michal’s actions and assumes that the ring is still in the original box. Therefore, her natural inclination would be to look in the place where she left it.	ChatGPT: Natalie will most likely look in <b>box A</b> , where she initially put the ring. Since Natalie did not witness Michal moving the ring to box B, she would assume that the ring is still in box A. It is possible that Natalie may eventually check box B, but her initial instinct would be to look in box A.

Table 5: An example of ChatGPT failure on an instance inspired by ToMi.

30 stories (each story has 6 questions, 180 questions in total) from the ToMi dataset and modified them to match a sentence completion format with the same meaning.<sup>10</sup> For example the question: “Where does Oliver think that Emma searches for the grapes?”. Was adjusted to the following sentence completion task: “Oliver thinks that Emma searches for the grapes in the”.

**Epistemic Reasoning (Cohen, 2021).** This dataset is part of BIG bench (Srivastava et al., 2022). It combines ToM with natural language inference. The tests pertain to epistemic mental states (Wimmer and Perner, 1983) and epistemic logic (Hintikka, 1962). This is done by using specific verbs related to knowledge and belief: factive (i.e., know, understand, recognize, see, remember, learn), and non-factive (i.e., believe, think, suspect, assume). The dataset contains 3 types of tests: (1) **intra-personal tests**: reasoning about the mental states of a single agent; (2) **inter-personal tests**: reasoning about the mental states of multiple agents; and (3) **inference reasoning**: recognizing that other agents are making inferences (i.e., if X entails Y, and Bob believes that X, then, it is reasonable to conclude that Bob believes Y).

**FauxPas-EAI (Shapira et al., 2023b).** Based on the clinical faux pas test (Baron-Cohen et al., 1999), the set contains 44 stories (22 faux pas and 22 equivalent control) with 4 corresponding questions. The stories require both social reasoning skills and detecting false belief. The stories were created by experts and a small part of the stories was created by ChatGPT with rephrasing and fixes by experts.

<sup>10</sup>This was done manually by one of the authors.

## 7.4 Generative LLMs

We provide the technical details regarding the prompts (§7.4.1) and decoding parameters (§7.4.2).

### 7.4.1 Prompts

As input to the LLMs, we used (unless written otherwise) an MC-probing setup (§4.2), i.e., concatenation of the original test with all possible answers and an instruction to choose an option. Table 7 exemplifies the prompt for each task.

### 7.4.2 Decoding Parameters

A single sample (the first) was selected from each model for the analysis of the stories. We used the hyperparameters detailed below. We chose hyperparameters that minimize randomness, predict the most probable answer (i.e., low temperature, sampling method), and allow for a sufficient number of tokens.

**FlanT5 (Chung et al., 2022).** Python package *transformers* implementation (AutoModelForSeq2SeqLM, AutoTokenizer); torch; Generation by *generate* function; do\_sample=True; max\_length=50, from\_pretrained:google/flan-t5-small, google/flan-t5-base, google/flan-t5-large, google/flan-t5-xl, google/flan-t5-xxl; temperature=0.0001

**FlanUI2 (Tay et al., 2022).** Python package *transformers* implementation (T5ForConditionalGeneration, AutoTokenizer); torch; Generation by *generate* function; do\_sample=True; max\_length=50; temperature=0.0001

<p>On the shelf, there is a <b>bottle</b>.  It is full of <b>beer</b>; there is no <b>wine</b> in it.  Yet, the label on this bottle says ‘<b>wine</b>’ and not ‘<b>beer</b>’.  Mark walks into the room and notices the bottle.  He has never seen it before. He reads the label.</p>	<p>On the shelf <b>in the company’s headquarters</b>, there is a <b>hard drive</b>  that contains only <b>audio files</b> and no <b>video files</b>.  Yet, <b>confusingly</b>, its label <b>clearly states</b> ‘<b>video files</b>’ and not ‘<b>audio files</b>’.  <b>The newly hired computer engineer finds the hard drive on the shelf.</b>  She has never seen <b>this hard drive</b> before. She reads the label.</p>
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Table 6: Variations that demonstrate the pattern similarity. Besides the lexical match (black) there are also semantic, grammatical, and pragmatic matches e.g., “beer” and “audio files” both share the same POS-tag and place in the parsing tree; “full of” and “contains” share the same semantic meaning for the purpose of the question.

Dataset	Example Prompt
<b>Triangle COPA</b>	<p>A circle knocks on the door. A triangle goes to the door, but hesitates to open it.  Why does the triangle hesitate to open the door?</p> <p>a. The triangle hesitates to open the door because it is unsure if it wants to let the circle in.  b. The triangle hesitates to open the door because it is excited to see the circle.  <b>Answer with “a” or “b”.</b>  <b>Answer:</b></p>
<b>epistemic reasoning</b>	<p>Premise: Olivia suspects that Evelyn understands that a man plays a piano painted with an image of a woman on it.  Hypothesis: Evelyn understands that a man plays a piano painted with an image of a woman on it.  <b>Is entailment? answer with “0” or “1”.</b>  <b>Answer:</b></p>
<b>FauxPas EAI</b>	<p>Abby’s father publishes short stories in the newspaper every week.  She told him “Dad, I want to learn how to write just like you!”  and he replied: “Well then my dear why don’t you go to a writing class?”  Abby goes to a first lesson in a class she found and David the teacher says:  “Today we’ll look at some bad and good examples of story-writing”.  He shows the class a story Abby’s father wrote last week and says:  “And now I’ll give you a good example of the writer Pichnik and you will say what the differences are.”  In the story did someone say something that they should not have said?  <b>Answer with “Yes” or “No” only, without explanations.</b>  <b>In case of doubt, answer according to the most probable answer.</b>  <b>Answer:</b></p>

Table 7: An example prompt used for each task.

**GPT (Brown et al., 2020).** Python package *openai* model=text-davinci-002, text-davinci-003; Generation by *Completion.create* function; temperature=0, max\_tokens=50

**ChatGPT (Achiam et al., 2023).**<sup>11</sup> Python package *openai* model=gpt-3.5-turbo-0301, gpt-4-0314; Generation by *ChatCompletion.create* function; temperature=0

**AI21.**<sup>12</sup> Python package *ai21* model=j2-jumbo-instruct, j2-grande-instruct, j2-jumbo, j2-grande, j2-large; Generation by *Completion.execute* function; temperature=0, max\_tokens=50, topKReturn=0, topP=1, without any panalty

## 7.5 Complete Results

Table 8 contains the exhaustive accuracy results for varied temperature settings {0<sup>13</sup>, 0.01, 0.1, 0.2, 0.3,

0.6, 0.9, 1}, each setting 100 times with different random seeds. As can be seen, there is a linear decrease trend in the accuracy as the temperature increases. For reproducible reasons, zero temperature is preferred because the zero temperature is constant. Worth noting, that since the zero temperature is always the lowest (the most accurate), this allows a fair reference point between all models. Additional models and other data sets showed similar behavior. This initial phase helped us determine that the most preferable results came from fixing the temperature to zero. We later continued for simplicity with the zero fixed results only.

Running the well-organized code provided by Kosinski (2023) we found that task 2 (Unexpected Transfer Task) scored lower than reported for GPT 3.5. Specifically, two samples resulted in clear mispredictions and one sample had borderline predictions that provided the correct answer but in a format that differed from the expected answer (i.e., the first word was not the expected answer). As a result, the score for task 2 was either 85% or

<sup>11</sup><https://chat.openai.com/chat>

<sup>12</sup><https://www.ai21.com/blog/introducing-j2>

<sup>13</sup>Note that there are models that can’t be set to exactly 0, in those cases we set the temperature to 0.00001.



Temperature	0	0.01	0.1	0.2	0.3	0.6	0.9	1
Flan-t5-small	58.00	57.77	56.94	55.96	55.3	53.58	52.67	51.03
Flan-t5-base	84.00	83.99	83.74	83.24	82.37	79.31	76.35	75.15
Flan-t5-large	92.00	92.00	91.70	91.07	90.55	89.39	87.63	86.94
Flan-t5-xl	92.00	92.00	92.24	92.30	92.43	92.32	91.80	91.49
Flan-t5-xxl	96.00	96.00	95.94	95.89	95.97	95.52	94.47	94.05

Table 8: Accuracy (%) of different temperatures, each setting 100 times with different random seeds, for the Flan-t5 for the “TriangleCOFA” test set.

	Average score				Joint score	
	Reality	Memory	First order	Second order	w.o Second order	All
Devinci003	100	96.6	61.6	25.0	20.6	10.3
Turbo-0301	100	90.0	73.3	40.0	41.3	17.2

Table 9: ToMi’ zero-shot subsets comparison. All numbers are percentages.

90%, and the average score across the two tasks was either 85% or 87.5%, which is lower than the reported average of 93%.

## 7.6 “Emergence” or test data contamination?

We would like to determine whether LLMs generalize or memorize when they solve the ToM tasks (Daumé, 2017). We explored the possibility that the increase in performance is a result of training on the test data itself. For that purpose we used a second, secret, test set for SocialIQa that was purposefully kept hidden to avoid data contamination and is only available to the original SocialIQa authors as well as through the AI2 leaderboard.<sup>14</sup> For each test set (i.e., the standard and secret test sets) we randomly sample 11 subsets of 100 questions on which we evaluate gpt3.5-turbo-0301 and gpt-4-0314. Comparing the performance of both models on both test sets samples with a T-test, we found no significant differences, making it inconclusive whether the models were trained on the normal test set or not. As we discuss in Sec 6, this doesn’t mean that ToM has “emerged” in LLMs, since they may have been exposed to training data or similar examples.

## 7.7 ToMi’ subsets analysis

Table 9 provides the complete results from the evaluation of GPT-3.5 on the ToMi’ dataset. The same

overall conclusion can be drawn from this table as well: although the model can correctly answer simple reading comprehension questions, it doesn’t answer questions that require ToM skill (first and second order) with similar accuracy.

We divided the results into the average score and joint score. The average score is calculated as a simple average on the different types of questions, while the joint score is considers the prediction as correct only if the model answered correctly all the questions from the same story (with a total of 30 stories). The average results emphasize the major gaps between the model’s accuracy on reading comprehension questions to first order questions (“Chloe will look for the boots in the”) and between the first order questions to the second order questions (“Chloe think that Jackson searches for the boots in the”). The joint score reveals that even when the model correctly answers questions about the story, it might still fail to answer more complex questions.

<sup>14</sup><https://leaderboard.allenai.org/socialiqa/submissions/public>