Minding Language Models' (Lack of) Theory of Mind: A Plug-and-Play Multi-Character Belief Tracker

Melanie Sclar[1] **Sachin Kumar**[2] **Peter West**[1] **Alane Suhr**[3]
Yejin Choi[1,3] **Yulia Tsvetkov**[1]
[1]Paul G. Allen School of Computer Science & Engineering, University of Washington
[2]Language Technologies Institute, Carnegie Mellon University
[3]Allen Institute for Artificial Intelligence
msclar@cs.washington.edu

Abstract

Theory of Mind (ToM)—the ability to reason about the mental states of other people—is a key element of our social intelligence. Yet, despite their ever more impressive performance, large-scale neural language models still lack basic theory of mind capabilities out-of-the-box. We posit that simply scaling up models will not imbue them with theory of mind due to the inherently *symbolic* and *implicit* nature of the phenomenon, and instead investigate an alternative: can we design a decoding-time algorithm that enhances theory of mind of off-the-shelf neural language models without explicit supervision? We present SYMBOLICTOM, a plug-and-play approach to reason about the belief states of multiple characters in reading comprehension tasks via explicit symbolic representation. More concretely, our approach tracks each entity's beliefs, their estimation of other entities' beliefs, and higher-order levels of reasoning, all through graphical representations, allowing for more precise and interpretable reasoning than previous approaches. Empirical results on the well-known ToMi benchmark (Le et al., 2019) demonstrate that SYMBOLICTOM dramatically enhances off-the-shelf neural networks' theory of mind in a zero-shot setting while showing robust out-of-distribution performance compared to supervised baselines. Our work also reveals spurious patterns in existing theory of mind benchmarks, emphasizing the importance of out-of-distribution evaluation and methods that do not overfit a particular dataset.

1 Introduction

Reasoning about other people's intentions, desires, thoughts, and beliefs is a cornerstone of human social intelligence. Children naturally develop an understanding of every individual's unique mental state and how it might impact their actions (Frith et al., 2003). Known as *Theory of Mind (ToM)* (Premack and Woodruff, 1978), this ability is crucial for efficient and effective communication.

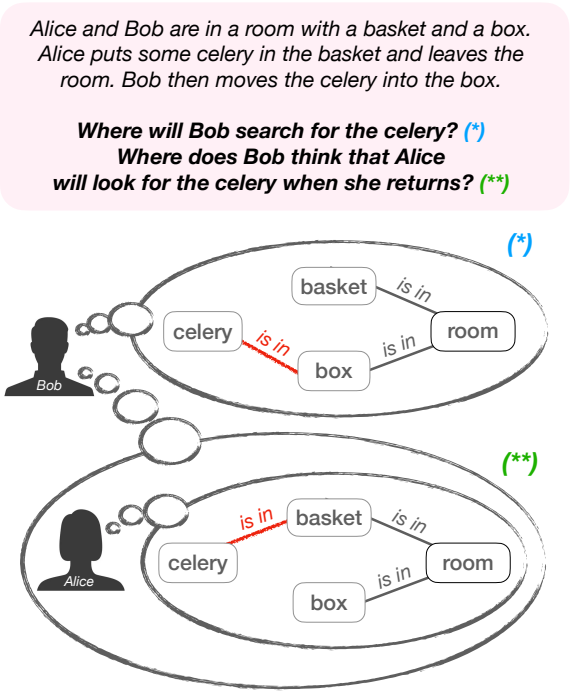

Figure 1: A simple story requiring theory of mind. Note that Alice's belief of the celery's location differs from reality (i.e. Alice holds a *false belief*). Readers must reason that Alice will look for the celery where she left it, and that Bob will make that same assumption. Questions shown require different depths of mental state modeling.

Cognitive and literary studies have extensively argued theory of mind's key role in understanding stories, in order to explain and predict each character's actions (Zunshine, 2006; Carney et al., 2014; Leverage et al., 2010; van Duijn et al., 2015, inter alia). As exemplified in Figure 1, readers need to model Bob's mental state (called *first-order ToM*), as well as Bob's estimation of Alice's mental state (*second-order ToM*) to answer questions.

Despite recent progress in language understanding abilities, large language models have been shown to lack theory of mind skills (Sap et al., 2022). Existing efforts to enable them have primarily relied on supervised methods (e.g., Grant et al.,

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 13960–13980
July 9-14, 2023

2017; Nematzadeh et al., 2018; Arodi and Cheung, 2021). However, current reading comprehension datasets for theory of mind reasoning are simplistic and lack diversity, leading to brittle downstream models which, as we show, fail in the presence of even slight out-of-distribution perturbations.

We introduce SYMBOLICTOM, an inference-time method that improves large language models' theory of mind capabilities by augmenting them with an explicit symbolic graphical representation of each character's beliefs. Unlike prior efforts, our approach does not require training and instead divides the problem into simpler subtasks, leveraging off-the-shelf models to solve them, and carefully consolidating their results. This makes SYMBOLICTOM significantly more robust than existing models trained specifically for theory of mind behavior.

While beliefs about the world state differ among people, most existing work on encoding belief states do not model this behavior relying on singular graphs (Jansen, 2022; Jacqmin et al., 2022). SYMBOLICTOM, instead, utilizes a *set of graphs*, each representing what the character p_1 *thinks that* p_2 *believes that [...]* p_m *assumes to be the current state of the world*, where m is the maximum reasoning depth as determined by the user. This explicit, recursive mental state representation enables the model to answer questions from the perspective of each character. SYMBOLICTOM's process of selecting and querying a particular character's graph grounds it in cognitive science research arguing theory of mind as an essential mechanism of selective attention (Leslie et al., 2004). Our approach also instills desirable inductive biases, such as object permanence—for example, object locations (represented by edges in the graphs) are assumed to be constant until the method can infer a change. Although existing NLP datasets only test up to second-order reasoning (i.e., $m \leq 2$), SYMBOLICTOM is designed to work at any depth.

SYMBOLICTOM dramatically improves the performance of large language models in theory of mind reading comprehension tasks. For example, GPT-3-Davinci's (Brown et al., 2020) accuracy on the ToMi benchmark (Le et al., 2019) increases by 38 absolute points using SYMBOLICTOM (yielding 92% accuracy averaging across question types). Furthermore, we extend the ToMi test sets with diverse story structures and sentence paraphrases and demonstrate that our approach is significantly more robust than supervised approaches.

2 Motivation and Background

Although large-scale language models have recently shown improvements in some classic theory of mind examples, they are still far from reliably showing theory of mind capabilities (Sap et al., 2022; Yu et al., 2022; Ullman, 2023; Shapira et al., 2023). While the training data for these models includes human-written stories which require theory of mind reasoning, this information is largely implicit and hence difficult for models to learn. ChatGPT and GPT3-Davinci's incorrect answers to Figure 1's question #2 are shown below.[1]

ChatGPT (`gpt-3.5-turbo`): Based on the information provided, Bob would likely think that Alice will look for the celery in the box when she returns. Since Bob moved the celery from the basket to the box, he would assume that Alice would expect to find it in its new location.
GPT3 (`text-davinci-003`): Bob will likely think that Alice will look for the celery in the box, since that is where he moved it.

Natural stories which make theory of mind explicit are scarce, necessitating automatically generated, template-based datasets like ToM-bAbI (Nematzadeh et al., 2018) and ToMi (Le et al., 2019). However, templated narratives cover limited types of interactions, and include only simplistic discourse and sentence structures. On the other hand, relying on human-generated data, e.g., in situated dialogue (Bara et al., 2021), leads to barriers in dataset size due to high annotation costs. Moreover, another source of data—text-based games with multiple characters—also faces limitations; in particular, modeling mental states is required mainly to infer intents (Zhou et al., 2022) and to maintain a consistent style of each character (Qiu et al., 2022). Rather, in this work, we aim to study and evaluate differences in knowledge and beliefs among multiple characters, traditional *cognitive* aspects of theory of mind.

To the best of our knowledge, the only available datasets for measuring theory of mind in reading comprehension tasks are ToM-bAbI and ToMi. Because of their templated nature, supervised training on them is prone to overfitting to spurious artifacts in the data. While ToMi was developed to counter this behavior in ToM-bAbI by introducing noise in the form of flexible sentence ordering and distractor sentences and characters, we show it still faces the same pitfalls.

[1] Queried on May 22, 2023 with `top_p=1` and `temperature=0`. Given the non-deterministic and continuously changing nature of these models, exact examples may not produce the same response we report.

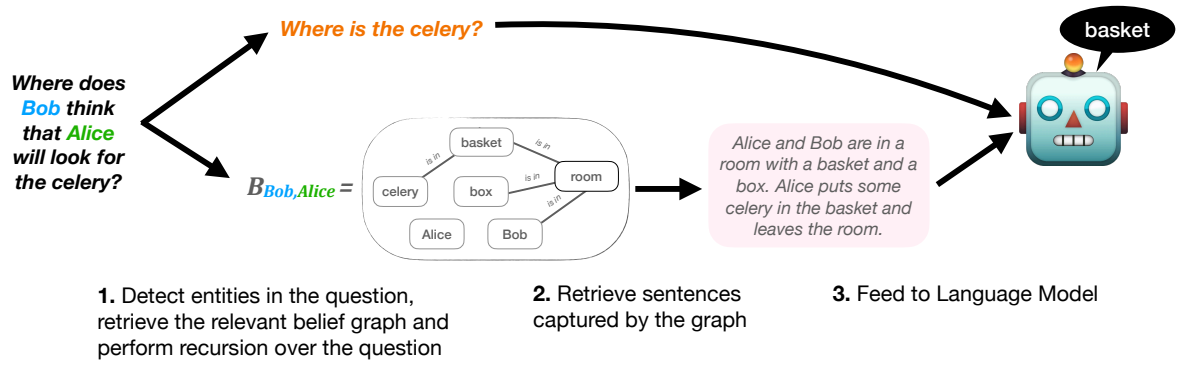

Figure 2: Pipeline overview of SYMBOLICTOM, a decoding-time algorithm that enhances large language models' theory of mind capabilities. SYMBOLICTOM does not require training: it divides the problem into smaller subtasks and uses off-the-shelf models to solve them. Given a passage, SYMBOLICTOM constructs explicit symbolic graphical representations of each character's belief states (step 1). To answer ToM questions, it retrieves relevant sentences from the graph (step 2) and then queries the LLM in a zero-shot manner (step 3).

Due to theory of mind's inherently implicit nature and limited naturally available data, in this work, we argue against supervision as a way forward and instead call for unsupervised, or inference-time approaches that combine modern neural models and traditional symbolic algorithms.

3 Methods

3.1 SYMBOLICTOM: Algorithm Overview

Our goal is to automatically answer reading comprehension questions given a story involving multiple characters, without requiring any supervised training or fine-tuning on this task. We first introduce key notation, then provide a high-level overview of SYMBOLICTOM (Algorithm 1).

Notation We use the term *k-th order theory of mind* to refer to an estimate of what a character p_1 *thinks that* p_2 *thinks that [...]* p_k *thinks about the world state*. We denote this belief by $B_{p_1,\ldots,p_k}$. We let $k \leq m$, where m is a maximum reasoning depth. This is a user-specified limit, denoting the maximum recursion that the reader is assumed to be capable of performing. For instance, in Figure 1, questions #1 and #2 measure 1st- and 2nd-order theory of mind respectively; B_{Bob} refers to Bob's beliefs about the current world state, and $B_{\text{Bob,Alice}}$ represents Bob's estimation of Alice's beliefs about the world state. In this work, $B_{p_1,\ldots,p_k}$ only represents beliefs about the current world state, without additional modeling of other characters' mental states, such as their opinions.

A benefit of this notation is that any belief state can be represented as an m-th order one. We assume that *what* p_k *thinks that* p_k *thinks* is equivalent to *what* p_k *thinks*, and by induction, $B_{p_1\ldots p_k} \equiv B_{p_1,\ldots,p_k,p_k,\ldots,p_k}$, where the last p_k is repeated $m-k$ times. We adopt this notation going forward, denoting all states as m-th order. As a conceptual note, the set of belief states $\{B_{p_1\ldots p_k,q_{k+1}\ldots q_m} \mid \forall q_{k+1},\ldots,q_m\}$ represents the mental state from the perspective of $p_1,\ldots,p_k$, using $m-k$ order of theory of mind.

Local and Global Context We represent each $B_{p_1\ldots p_k}$ as a graph (a simplified version is depicted in Figure 1) where each node represents an entity (e.g. a character, object, room, container) and each edge connects two nodes with a stated relationship in the story. We construct the graphs by iterating through a story one sentence at a time, and adding both nodes and edges to the graph (BELIEFTRACKINGSTRUCTURE; described in §3.2 and Algorithm 2). Each edge is also paired with the sentence from the story from which it was constructed. We refer to the set of all belief state graphs as the *local contexts*. We also maintain a *global context* graph, denoted by G, which contains the true world state. G has an identical structure to $B_{p_1\ldots p_k}$. See A.1 for a detailed definition of G.

Question Answering After parsing a story and constructing the complete set of belief-tracking structures, we can use these structures to answer questions by querying the appropriate graph and considering it as the real-world state. For example, if the question is "Where will Bob think that Alice will look for the celery?", we retrieve $B_{\text{Bob, Alice}}$, but if instead the question were "Where will Bob

look for the celery?", we would retrieve B_{Bob}. In both cases, we would ask "Where is the celery?" on the retrieved graph. Figure 2 shows an example of the full pipeline.

Given a question, we identify the relevant characters $p_1, \ldots, p_k$ mentioned in order heuristically, and rephrase the question to ask directly about the world state (PROCESSQUESTION; owing to the questions' templatic nature in our evaluation data, this approach rephrases all questions correctly).[2] We then retrieve the corresponding graph; i.e., $B_{p_1,\ldots,p_k}$, of which we can simply ask the question "Where is the celery?". To obtain the answer, we first reconstruct a subset S' of sentences in the original story, consisting of those represented by the retrieved graph (SENTENCESREPRESENTEDBYGRAPH). We then use a large language model $\mathcal{L}$ to answer the simplified question zero-shot given S', using as input the sentences in S' in the same order as they appeared in the original text, and preserving phrasing. We optionally further filter S' based on the entities mentioned in the question (FILTERBASEDONQUESTION). An ablation study showed this last step can often be skipped (see Appendix C.1).

Algorithm 1 SYMBOLICTOM

$B \leftarrow$ BELIEFTRACKINGSTRUCTURE($sentences$)
$p_1,\ldots,p_k, question' \leftarrow$ PROCESSQUESTION($question$)
$S' \leftarrow$ SENTENCESREPRESENTEDBYGRAPH($B_{p_1,\ldots,p_k}$)
$S'' \leftarrow$ FILTERBASEDONQUESTION($S', question$)
return $S'', question'$

3.2 Computing the Belief Graphs $B_{p_1 \ldots p_k}$

Assuming each story is told chronologically, SYMBOLICTOM processes each sentence s sequentially in two stages (Algorithm 2). First, it extracts all actions in s and updates the global context G from an omniscient point of view while identifying the characters ($\mathcal{W}$) who witnessed actions and world state changes described in the sentence. Second, for each witness $w \in \mathcal{W}$, it propagates this new information to update w's local contexts; i.e., we only update $B_{p_1,\ldots,p_m}$ with, for $1 \leq i \leq m$, each $p_i \in \mathcal{W}$, and leave the rest unchanged.

As an example, when processing the last sentence in Figure 3, we update Bob and Charles's state (B_{Bob} and B_{Charles}) and the perception of others' respective state ($B_{\text{Bob,Charles}}$, $B_{\text{Charles, Bob}}$), but we need not update Alice's state, or Bob and Charles's perception of Alice's mental state, because she did not witness the actions described.

[2] Our explorations show that GPT3 is also capable of rephrasing the questions zero-shot (see §A.3), but we refrained from this solution due to budget concerns.

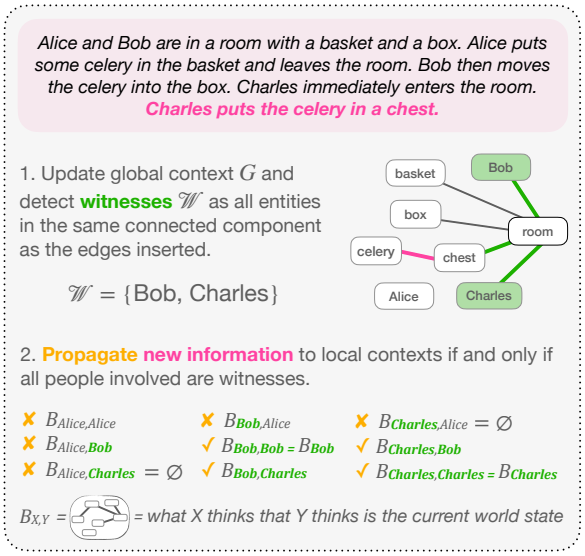

Figure 3: High-level depiction of the belief update procedure for $m = 2$. $B_{p_1,\ldots,p_k}$ denotes a graph, and the graph updating procedure is detailed in the main text.

Algorithm 2 Belief Tracking

function BELIEFTRACKINGSTRUCTURE($sentences$)
  for $s \in sentences$ **do**
    $G, \mathcal{W} \leftarrow$ GLOBALCONTEXTUPDATE(G, s)
    for all $[p_1, \ldots, p_m] \in \mathcal{W}^m$ **do**
      $B_{p_1 \ldots p_m} \leftarrow$ LOCALCONTEXTUPDATE($B_{p_1 \ldots p_m}, G, s$)
    end for
  end for
end function

3.2.1 Detecting Witnesses, Updating Graphs, and Propagating Knowledge

Starting with an empty graph, for each new sentence s, we update the global context G by combining off-the-shelf models in four steps (Algorithm 3; GLOBALCONTEXTUPDATE). **First,** we detect the existing edges E in G that contradict s. This is implemented as detecting Natural Language Inference (NLI) contradictions, considering s as the premise, and every edge in G as a hypothesis. **Second,** we augment G with new edges and nodes, by first deriving a natural language representation r of the state resulting from the actions described in s, and then extract new nodes and edges from r as OpenIE triples (Stanovsky et al., 2018). For example, for "Bob then moves the celery to the box", the resulting state r would be the sentence

"The celery is in the box". To obtain r from s, we prompt a language model such as GPT3 (see Appendix A.2 for details). After obtaining r, we use the corresponding triple (e.g., `(celery, box, is in)`) to add new nodes and edges to G if not already present (e.g., the nodes "celery" and "box", and a directed edge connecting them labeled by "is in"). Importantly, we only add edges that represent positive relations between nodes; i.e., there will not be an edge representing "The celery is not in the box". **Third,** we detect the witnesses $\mathcal{W}$ of the actions described in s. Since each character will be a node in G, we identify $\mathcal{W}$ as all the characters that are in the same connected component as the newly added edges. **Finally,** we remove all edges E that are no longer valid in G as identified by the NLI contradictions. This step is done last to ensure all witnesses are found before their edges are deleted.

Algorithm 3 World State Beliefs Graphs Update

function GLOBALCONTEXTUPDATE(G, s)
  $E \leftarrow$ DETECTCONTRADICTINGEDGES(G, s)
  $G \leftarrow G \cup$ TRIPLES(RESULTINGSTATE(s))
  $\mathcal{W} \leftarrow$ FINDWITNESSES(G)
  $G \leftarrow G \setminus E$
  return $G, \mathcal{W}$
end function

function LOCALCONTEXTUPDATE(C, G, s)
  $E \leftarrow$ DETECTCONTRADICTINGEDGES(G, s)
  $C \leftarrow C \cup$ TRIPLES(RESULTINGSTATE(s))
  $C \leftarrow$ PROPAGATEKNOWLEDGE(G, C, s)
  $C \leftarrow C \setminus E$
  return C
end function

The local contexts ($B_{p_1,\ldots,p_k}$) are updated similarly (LOCALCONTEXTUPDATE in Algorithm 3), except for an additional step of knowledge propagation. While performing an action, a character may implicitly gain information not described in the text. For example, when entering a room, a character may gain knowledge of the people and visible objects in the room. This knowledge (already present in G, which tracks the omniscient world state) needs to be propagated to each $B_{p_1,\ldots,p_k}$ with each $p_i \in \mathcal{W}$. As G represents the true world state, we simplify the problem: if a character p_i is in a specific connected component D of G, then it possesses all knowledge encoded in D. To model implicit knowledge gain, we add all edges in D to $B_{p_1,\ldots,p_k}$. As D represents the latest global context information, we remove from the local context edges that are in $B_{p_1,\ldots,p_k}$ but not in D (representing outdated beliefs about the world state).

3.3 Notes on Memory Efficiency

Memory requirements grow exponentially with m, the maximum order of theory of mind considered. However, m in practice is small, as humans find tasks increasingly challenging as m increases. For example, psychological tests for $m = 3$ are aimed at teenagers and adults (Valle et al., 2015). All experiments in this work are done with $m = 2$, the maximum order of theory of mind reasoning that current datasets evaluate. If memory were a concern, one could process the questions first for memory efficiency, and compute only the graphs $B_{p_1,\ldots,p_k}$ required for target queries.

4 Fundamental Issues in Existing ToM Datasets

Construction of ToMi As introduced in §2, the sole large-scale theory of mind dataset for reading comprehension tasks is ToMi (Le et al., 2019). Barring its added distractor characters and sentences, ToMi strictly mimics the Sally-Anne test, a widely adopted evaluation for assessing children's social cognitive ability to reason about others' mental states (Wimmer and Perner, 1983; Baron-Cohen et al., 1985). Stories are structured are as follows: characters A and B are in a room, and A moves an object from an opaque container to another; B may or may not leave the room before A moves the object. B will know the object's new location if and only if they were in the room at the time it was moved. Four types of ToM questions are posed: first-order or second-order, probing a character about either a true or a false belief (i.e, belief that matches reality or not). ToMi also includes questions probing about reality (or *zeroth-order* ToM, Sclar et al., 2022) and memory.

ToMi has six types of sentences (i.e. six *primitives*) with set phrasing. These include someone (a) entering or (b) exiting a room; the location of (c) an object or (d) a person; (e) someone moving an object; and (f) someone's opinion about an object (distractors). Primitives are combined into stories with a finite list of possible orderings. Despite the limited types of primitives, correctly answering questions requires high-order levels of reasoning.

Templated stories are filled with randomly sampled objects, locations, containers, and rooms from a set list. ToMi implicitly assumes that questions about the story do not depend on these decisions, only on the underlying story template. Yet, in a small-scale human study, we find physical com-

1. Oliver entered the front yard.
2. Ethan entered the front yard.
3. Liam entered the kitchen.
4. **objectA** is in the basket.
5. Ethan exited the front yard.
6. Ethan entered the kitchen.
7. Oliver moved **objectA** to the **containerX**.
8. Where does Ethan think **objectA** is?
ToMi Gold Label: basket

Table 1: Interpretation of ambiguities in ToMi can be affected by commonsense. In the above template, the correct label is that Ethan thinks **objectA** is in the *basket*, as this is where he last saw it. Setting **objectA** to *hat* and **containerX** to *box* results in 80% human accuracy. However, setting these to *apple* and *pantry*, accuracy drops to 20%. Physical commonsense suggests the pantry is likely in the kitchen, changing the answer to *pantry*, but regardless of the identity of **objectA** or **containerX**, the correct label in ToMi is *basket*.

monsense leads human answers to change, and disagree with ToMi's labels depending on the noun. Table 1 presents an example where the object and container have a large effect on human responses.[3]

Resolving Unintentional Ambiguities ToMi's story construction process often leaves object locations ambiguous, which forces humans to (incorrectly) rely on their physical commonsense. For example, the location of the *basket* in line 4 of Table 1 is ambiguous. This ambiguity is at times resolved at a later step in the story (Arodi and Cheung, 2021), but it is not true for all cases, and these resolutions were not expressly intended by ToMi's original design. This complicates the task beyond theory of mind. For example, in Table 1, the reader must conclude from *"Oliver is in front yard"*, *"Oliver moved the objectA (...)"*, and *"The objectA is in basket"* that the basket is in the front yard, and hence that Ethan saw it there. This requires 3-hop reasoning, and knowing ahead of time that, in ToMi, characters do not change rooms unless explicitly stated.

To solve these unintentional ambiguities and additional 3-hop reasoning requirements, and instead solely measure theory of mind reasoning skills, we automatically add a sentence that disambiguates the location of each container immediately after each primitive (c) or (e) (e.g., adding *"The basket is in the front yard"* as line 5 in Table 1). Finally, as reported in Arodi and Cheung (2021); Sap et al. (2022), ToMi contains some mislabeled second-order questions, which we also correct.

[3]Using Amazon Mechanical Turk, we present 20 humans with the template in Table 1, using either *(hat,box)* or *(apple, pantry)*. Workers are paid $1 per HIT.

5 Experiments

We experiment with several base LMs, and evaluate each of them both out-of-the-box via zero-shot prompting, and by applying SYMBOLICTOM to ToMi stories to produce answers. We evaluate Macaw-3B (Tafjord and Clark, 2021), GPT3-{Curie,Davinci} (Brown et al., 2020), Flan-T5-{XL,XXL} (Chung et al., 2022), LLaMA-{7B, 13B} (Touvron et al., 2023), GPT3.5 (OpenAI, 2022), and GPT4 (OpenAI, 2023). We use WANLI (Liu et al., 2022) for identifying NLI contradictions, and the AllenNLP library (Gardner et al., 2018) for OpenIE. We additionally refine each subject and object in extracted triples to remove any stopwords that may be accidentally included by OpenIE.

We first evaluate SYMBOLICTOM's performance as a plug-and-play method for different base LMs on ToMi (§5.1). We then test whether performance gains are robust to ToMi story structure modifications (§5.2). Finally, we explore SYMBOLICTOM's robustness to linguistic diversity (§5.3).

Supervised Models For comparison, we train two supervised models: Textual Time Travel (TTT) (Arodi and Cheung, 2021), and a fine-tuned GPT3-Curie. TTT is a modification of EntNet (Henaff et al., 2017) designed for theory of mind tasks; GPT3-Curie is finetuned on 6000 ToMi examples for one epoch. GPT3-Curie achieves near-perfect performance when finetuned on ToMi (98.5% accuracy when averaging all questions; Table 5). Interestingly, GPT3-Curie achieves a higher accuracy than the theory of mind-motivated TTT (accuracy 92.3%). We explore model robustness in §5.2.

5.1 In-Domain Evaluation

We evaluate all base LMs comparing their performance out-of-the-box, versus when adding SYMBOLICTOM. Figure 4 shows results by question type, showing dramatic improvements for all theory of mind questions: +62 points in accuracy for first-order false-belief questions for Flan-T5-XL, +78 points in accuracy for second-order false-belief questions for GPT3.5, among other improvements. In addition, we observe all models maintain near-perfect performance with and without SYMBOLICTOM in memory questions. Supervised models

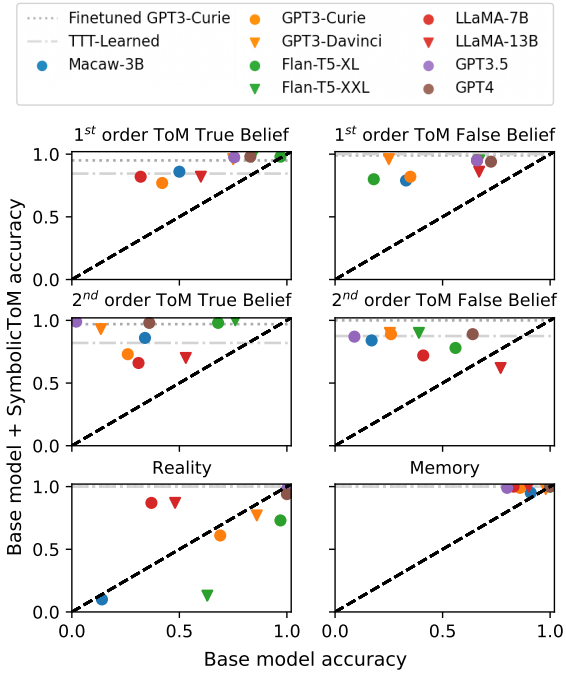

Figure 4: Accuracy for each ToMi question type and base model (higher is better). Dots in the upper triangle have higher performance with SYMBOLICTOM than the base model out-of-the-box. Horizontal lines give supervised models' performance. Full results in Table 5.

show high accuracy for all question types.

We only see significant decreases in performance for reality questions in Flan-T5 models. This can be partially attributed to the questions' phrasing: questions are posed as "Where is the celery *really*?". Removing *really* results in 96% accuracy for Flan-T5-XL. Flan-T5-XXL empirically shows a bias towards providing a room rather than container as an answer when only one container is mentioned, which is often the case for SYMBOLICTOM-filtered stories. Rooms are invalid answers in ToMi. An ablation on the final filter function of Algorithm 1 suggests that keeping more containers in the final story reduces this bias and still yields significant improvements for false-belief questions across all models (see §C.1). Besides *reality* questions, Flan-T5-XXL with SYMBOLICTOM achieves results comparable to the supervised TTT.

5.2 Story Structure Robustness Test Sets

We create three test sets by modifying ToMi's stories structures without adding new types of actions or linguistic diversity. These tests were only evaluated once, after finishing development of SYMBOLICTOM. Test sets are defined below. See Appendix B.2 for concrete examples.

	D_1	D_2	D_3
Off-the-shelf models			
Macaw-3B	8	12	30
Flan-T5-XL	86	51	68
Flan-T5-XXL	69	59	52
GPT3-Curie	37	39	57
GPT3-Davinci	20	25	39
GPT3.5[4]	1	0	48
GPT4	58	62	97
LLaMA-7B	17	17	17
LLaMA-13B	26	36	37
SYMBOLICTOM + *Off-the-shelf models*			
Macaw-3B	89 (+81)	71 (+60)	70 (+41)
Flan-T5-XL	76 (-10)	96 (+46)	**100** (+33)
Flan-T5-XXL	93 (+24)	**100** (+41)	**100** (+49)
GPT3-Curie	84 (+48)	81 (+42)	73 (+16)
GPT3-Davinci	92 (+73)	91 (+66)	90 (+50)
GPT3.5	**100** **(+99)**	**100** **(+99)**	99 (+51)
GPT4	**100** (+42)	**100** (+38)	**100** (+4)
LLaMA-7B	99 (+82)	92 (+75)	88 **(+71)**
LLaMA-13B	78 (+52)	84 (+48)	84 (+47)
Supervised models			
TTT	49	65	78
Finetuned GPT3	51	68	32

Table 2: Precision using SYMBOLICTOM on all questions from 100 stories for each of the modified test sets D_i. Supervised models were trained on ToMi; all others do not require training. Parenthesis reflect differences between using and not using SYMBOLICTOM: **bold** reflects higher overall performance, and **green** reflects the highest net improvements when using SYMBOLICTOM.

Double Room False Belief Story (D_1) Two false belief substories involving the same two characters p_1, p_2 are concatenated to yield a longer, more complex story. Each substory has different objects being moved, across different containers. The system is probed using all four combinations of second-order theory of mind questions involving the two characters and locations. Questions are evenly split between the first and second substory.

Three Active Characters Story (D_2) Three characters p_1, p_2, p_3 are in the same room, where an object o_1 and three containers c_1, c_2, c_3 are available. The story is as follows: p_2 leaves before p_1 moves o_1 from c_1 to c_2, but p_3 witnesses the move.

[4]Low scores are due to the model refusing to answer, e.g. answering *"There is no information in the given text to determine where Bob thinks Alice searches for the celery."*

Then, p_1 leaves the room. Later, p_3 moves the object to container c_3 without any witnesses. The system is probed using all combinations of second-order theory of mind questions.

Multiple Object Movements Across Four Containers (D_3) Two characters p_1, p_2 are in a room, with a single object, and four containers $c_1, \ldots, c_4$. p_1 moves the object from c_1 to c_2 and right before leaving the room, p_2 enters. p_2 then moves the object to c_3, and then c_4. We probe with all first and second-order theory of mind questions.

Results Supervised models significantly overfit to ToMi's original story structures (Table 2). In contrast, all models had high accuracy when equipped with SYMBOLICTOM, especially larger models, such as GPT3.5, LLaMA-{7B,13B}, among others.

D_2 may also be used to test third-order ToM reasoning, asking questions such as "Where does p_1 think that p_2 thinks that p_1 will search for the o_1?". Third-order ToM is a reasoning depth currently untested by available NLP benchmarks. SYMBOLICTOM consistently enhances the performance of off-the-shelf LLMs and outperforms supervised methods in the third-order ToM setting. See details in Appendix C.2. This experiment showcases how extensions of ToMi may be used to test higher-order reasoning. This is the first approach towards testing third-order ToM in LLMs; a benchmark to comprehensively test such order of reasoning exceeds the scope of this paper.

5.3 Paraphrasing Robustness Evaluation

We assess the robustness of all models when utilizing various wordings for each sentence. We reword all templates using GPT3-Davinci, utilizing different choices of objects, rooms, and names, and manually excluded incorrect paraphrases. The resulting dataset—ParaphrasedToMi—exhibits much greater complexity, as these rewordings can express actions in a less straightforward way. All paraphrases are shown in Appendix B.1.

Figure 5 demonstrates significant performance decreases for supervised models transferring to ParaphrasedToMi. TTT's average accuracy drops 54 points from ToMi, with losses across all question types. Finetuned GPT3 exhibits significant losses in false-belief questions (-40 average accuracy) but is robust for other question types.

Methods without supervision also suffer significant losses, but SYMBOLICTOM still results in large improvements for theory of mind questions. Models equipped with SYMBOLICTOM perform significantly better than the supervised TTT model across all theory of mind questions. ParaphrasedToMi is significantly more difficult for SYMBOLICTOM since it triggers more errors in edge removal (due to errors in NLI classification), as well as errors in edge insertion (due to errors in the resulting state's triple extraction). Although computing RESULTINGSTATE by prompting the base LMs was successful with original phrasings (as defined in §3.2.1), we observed differences in robustness when prompting with paraphrases. We found implementing RESULTINGSTATE with GPT3 reliable, and thus we use it for all models. Results using other models are included in §C.3: false-belief performance is even better for models like LLaMA, GPT3.5, or GPT4.

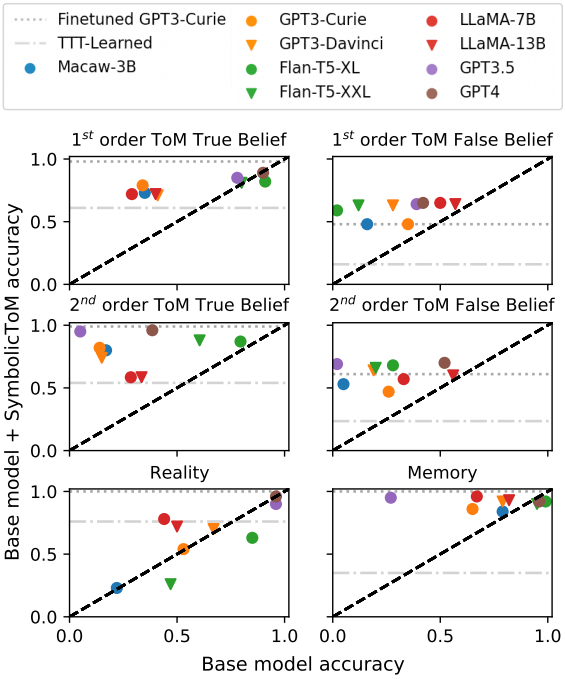

Figure 5: Results for ParaphrasedToMi when prompting GPT3 as implementation of RESULTINGSTATE (Davinci for all except for Curie). Dots in the upper triangle imply performance with SYMBOLICTOM is higher than using the base model out-of-the-box. Horizontal lines reflect supervised models' performance (higher is better).

6 Related Work

Existing Approaches Classical reasoning tasks require achieving some goal, e.g., proving a statement, given a set of facts and universally valid rules (e.g., Tafjord et al., 2021). A common approach is to decompose the target reasoning task into subtasks, for example by using off-the-shelf

LMs (Creswell et al., 2023; Kazemi et al., 2022; Nye et al., 2021). We use a similar technique in SYMBOLICTOM, breaking the higher-level reasoning task into graph reasoning subtasks. Nonetheless, these approaches cannot be simply ported to our domain: stories' facts (i.e. the world state) change over time and are not universally accessible to all characters, and commonsense rules and assumptions like object permanence must made explicit. SYMBOLICTOM's design addresses these challenges by maintaining and updating graphs about facts and beliefs as a story progresses.

In scenarios where world state changes over time, such as in text-based games, existing approaches maintain and update structured world representations as the world state changes (Ammanabrolu and Riedl, 2021; Adhikari et al., 2020). However, while these approaches could potentially be applied in our scenario to update G, they would not address the problems of multiple-belief representation or knowledge propagation to witnesses' graphs, with some approaches even being explicitly impossible for modeling second-order ToM (Qiu et al., 2022).

ToM beyond NLP Theory of mind is also crucial in multi-agent reinforcement learning (Rabinowitz et al., 2018), including in bidirectional symbolic-communication (Wang et al., 2022; Sclar et al., 2022), unidirectional natural-language settings (Zhu et al., 2021); and recently, by combining reinforcement learning, planning, and language, to create a human-level Diplomacy player (, FAIR). It has also received increased attention in human-computer interaction (Wang et al., 2021) and explainable AI (Akula et al., 2022).

Psychologists divide theory of mind into two types of reasoning: affective (emotions, desires) and cognitive (beliefs, knowledge) (Shamay-Tsoory et al., 2010), with the former developing earlier in children (Wellman, 2014). Our work focuses on the latter, but the principle of multiple belief representation could also be applied to affective theory of mind reasoning. Existing work has shown that humans are proficient at second-order or higher false-belief reasoning, also referred to as *advanced ToM* (Białecka-Pikul et al., 2017), with evidence that we can perform even third- and fourth-order reasoning (Valle et al., 2015; Osterhaus et al., 2016). While, to best of our knowledge, no dataset requires beyond second-order ToM, SYMBOLICTOM explicitly models the recursive reasoning that supports queries of any reasoning order.

7 Conclusions

Theory of mind is an essential social intelligence ability. Developing agents with theory of mind is requisite for a wide range of applications, including reading comprehension, tutoring, dialogue, personalization, and negotiation. For example, in reading comprehension settings (and broadly for natural language understanding), having a multi-level understanding of texts is crucial for providing meaningful and contextualized answers: stories often rely on theory of mind reasoning to create conflict (e.g., in murder mysteries, drama, and romances, as in the final acts of *Romeo and Juliet*).

We present SYMBOLICTOM, a plug-and-play method to enable theory of mind reasoning in language models via explicit symbolic representations in the form of nested belief states. SYMBOLICTOM requires no training or fine-tuning, a key aspect for a domain with scarce supervised data and limited success in learning from massive unlabeled text alone. With experiments on reading comprehension tasks, our approach demonstrates dramatic improvement in the accuracy of base language models, especially for false-belief scenarios.

We also show that, in contrast to supervised methods, SYMBOLICTOM is highly robust to story perturbations and out-of-domain inputs where supervised methods suffer significant degradations (as in, e.g., Yu et al., 2022).[5] Our results show the promise of augmenting neural language models with symbolic knowledge for improving their social reasoning skills. We leave to future work to investigate similar approaches for other types of social intelligence; as well as develop new datasets that cover a more diverse set of interactions.

Limitations

SYMBOLICTOM assumes stories are written chronologically, which may not hold for some human-written stories. This may be alleviated using time-stamping models like Faghihi and Kordjamshidi (2021). Furthermore, since we use off-the-shelf models (WANLI (Liu et al., 2022) and OpenIE (Stanovsky et al., 2018)) to create and update the graphs, the presented approach may propagate errors as revealed in the linguistic diversity experiments. However, these issues can be largely alle-

[5] As a part of out-of-domain testing, we also create a more challenging version of the available ToM datasets, available at https://github.com/msclar/symbolictom along with a corrected version of ToMi.

viated by using more sophisticated models, even the LLMs like GPT3 themselves. We do not experiment with them due to budgetary restrictions.

Currently, all NLP datasets available for theory of mind reasoning describe Sally-Anne tests. In these datasets, the concept of large distances is absent, meaning that anyone specified to be in a location is assumed to be a witness of the actions that occur there. This assumption can be violated in realistic settings. For example, *"Anne is in the USA"* does not imply she is a witness to every action happening in the USA. In future work, this approach can be improved by refining the witnesses detection algorithm to incorporate physical commonsense reasoning. We could also refine the witness detection algorithm by sampling paths between the inserted edge and each node referring to a person, to query an LM directly on that substory by asking if the person witnessed the action. To be able to test both of these ideas, we would need to obtain new theory of mind datasets with significantly more types of interactions and physical commonsense in the stories.

Ethics Statement

Theory of mind research at its core deals with reasoning about the mental states of others. In this work, we focus on reading comprehension, a task which can similarly be exposed to ethical concerns: for example, when a model makes erroneous predictions about the mental states of characters in the description, when it is misused to reason about private situations, and when it makes predictions which reinforce social biases. This issue can be exacerbated if the characters are actual people. In this work, however, we experiment with simple, prototypical character references from a public dataset, and not with actual people. This decision is intentional. Furthermore, we focus on reasoning about physical objects and observers' knowledge about their location in space, which is less prone to ethical concerns. This data can nonetheless lead to biased decisions, such as imbalanced decisions correlated with social attributes like gender (often correlated with names). Future work in this area may include scenarios with more realistic human-agent interaction, such as dialogue tasks, where parties involved may not have the same incentive structure. These scenarios will need to be handled with special care as they could lead to agents learning to deceive humans by exploiting a predicted (lack of) knowledge. The state-of-the-art in machine theory of mind is still far from these capabilities, but we believe it is important to consider these risks when designing experiments.

Acknowledgements

We thank Lucille Njoo and Tianxing He for the valuable discussions, and Akshatha Arodi for the support in running the Textual Time Travel code base. S.K. gratefully acknowledges support from Google Ph.D. Fellowship. We also thank OpenAI for providing academic access to their language model API. This material is based upon work partly funded by the DARPA CMO under Contract No. HR001120C0124, by DARPA MCS program through NIWC Pacific (N66001-19-2-4031), by NSF DMS-2134012, by NSF CAREER Grant No. IIS2142739, and an Alfred P. Sloan Foundation Fellowship. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily state or reflect those of the United States Government or any agency thereof.

References

Ashutosh Adhikari, Xingdi Yuan, Marc-Alexandre Côté, Mikuláš Zelinka, Marc-Antoine Rondeau, Romain Laroche, Pascal Poupart, Jian Tang, Adam Trischler, and Will Hamilton. 2020. Learning dynamic belief graphs to generalize on text-based games. *Advances in Neural Information Processing Systems*, 33:3045–3057.

Arjun R. Akula, Keze Wang, Changsong Liu, Sari Saba-Sadiya, Hongjing Lu, Sinisa Todorovic, Joyce Chai, and Song-Chun Zhu. 2022. Cx-tom: Counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models. *iScience*, 25(1):103581.

Prithviraj Ammanabrolu and Mark Riedl. 2021. Learning knowledge graph-based world models of textual environments. In *Advances in Neural Information Processing Systems*.

Akshatha Arodi and Jackie Chi Kit Cheung. 2021. Textual time travel: A temporally informed approach to theory of mind. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4162–4172, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Cristian-Paul Bara, CH-Wang Sky, and Joyce Chai. 2021. Mindcraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125.

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

Marta Białecka-Pikul, Anna Kołodziejczyk, and Sandra Bosacki. 2017. Advanced theory of mind in adolescence: Do age, gender and friendship style play a role? *Journal of Adolescence*, 56:145–156.

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.

James Carney, Rafael Wlodarski, and Robin Dunbar. 2014. Inference or enaction? the impact of genre on the narrative processing of other minds. *PloS one*, 9(12):e114172.

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*.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2023. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations*.

Hossein Rajaby Faghihi and Parisa Kordjamshidi. 2021. Time-stamped language model: Teaching language models to understand the flow of events. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4560–4570.

Meta Fundamental AI Research Diplomacy Team (FAIR), Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

C.D. Frith, D.M. Wolpert, Uta Frith, and Christopher D. Frith. 2003. Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1431):459–473.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

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

Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. 2017. Tracking the world state with recurrent entity networks. In *International Conference on Learning Representations*.

Léo Jacqmin, Lina M Rojas Barahona, and Benoit Favre. 2022. "do you follow me?": A survey of recent approaches in dialogue state tracking. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 336–350.

Peter Jansen. 2022. A systematic survey of text worlds as embodied natural language environments. In *The Third Wordplay: When Language Meets Games Workshop*.

Seyed Mehran Kazemi, Najoung Kim, Deepti Bhatia, Xin Xu, and Deepak Ramachandran. 2022. Lambada: Backward chaining for automated reasoning in natural language. *arXiv preprint arXiv:2212.13894*.

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.

Alan M Leslie, Ori Friedman, and Tim P German. 2004. Core mechanisms in 'theory of mind'. *Trends in cognitive sciences*, 8(12):528–533.

Paula Leverage, Howard Mancing, and Richard Schweickert. 2010. *Theory of mind and literature*. Purdue University Press.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and ai collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Maxwell Nye, Michael Tessler, Josh Tenenbaum, and Brenden M Lake. 2021. Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning. *Advances in Neural Information Processing Systems*, 34:25192–25204.

OpenAI. 2022. ChatGPT: Optimizing language models for dialogue.

OpenAI. 2023. GPT-4 technical report.

Christopher Osterhaus, Susanne Koerber, and Beate Sodian. 2016. Scaling of advanced theory-of-mind tasks. *Child development*, 87(6):1971–1991.

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

Liang Qiu, Yizhou Zhao, Yuan Liang, Pan Lu, Weiyan Shi, Zhou Yu, and Song-Chun Zhu. 2022. Towards socially intelligent agents with mental state transition and human value. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 146–158, Edinburgh, UK. Association for Computational Linguistics.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. In *International conference on machine learning*, pages 4218–4227. PMLR.

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 Association for Computational Linguistics: EMNLP 2022*, page 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Melanie Sclar, Graham Neubig, and Yonatan Bisk. 2022. Symmetric machine theory of mind. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19450–19466. PMLR.

Simone G Shamay-Tsoory, Hagai Harari, Judith Aharon-Peretz, and Yechiel Levkovitz. 2010. The role of the orbitofrontal cortex in affective theory of mind deficits in criminal offenders with psychopathic tendencies. *Cortex*, 46(5):668–677.

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

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895.

Oyvind Tafjord and Peter Clark. 2021. General-purpose question-answering with macaw. *arXiv preprint arXiv:2109.02593*.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Annalisa Valle, Davide Massaro, Ilaria Castelli, and Antonella Marchetti. 2015. Theory of mind development in adolescence and early adulthood: The growing complexity of recursive thinking ability. *Europe's journal of psychology*, 11(1):112.

Max J van Duijn, Ineke Sluiter, and Arie Verhagen. 2015. When narrative takes over: The representation of embedded mindstates in shakespeare's othello. *Language and Literature*, 24(2):148–166.

Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Yuanfei Wang, fangwei zhong, Jing Xu, and Yizhou Wang. 2022. Tom2c: Target-oriented multi-agent communication and cooperation with theory of mind. In *International Conference on Learning Representations*.

Henry M Wellman. 2014. *Making minds: How theory of mind develops*. Oxford University Press.

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.

Ping Yu, Tianlu Wang, Olga Golovneva, Badr Alkhamissy, Gargi Ghosh, Mona Diab, and Asli Celikyilmaz. 2022. Alert: Adapting language models to reasoning tasks. *arXiv preprint arXiv:2212.08286*.

Pei Zhou, Andrew Zhu, Jennifer Hu, Jay Pujara, Xiang Ren, Chris Callison-Burch, Yejin Choi, and Prithviraj Ammanabrolu. 2022. An ai dungeon master's guide: Learning to converse and guide with intents and theory-of-mind in dungeons and dragons. *arXiv preprint arXiv:2212.10060*.

Hao Zhu, Graham Neubig, and Yonatan Bisk. 2021. Few-shot language coordination by modeling theory of mind. In *International Conference on Machine Learning*, pages 12901–12911. PMLR.

Lisa Zunshine. 2006. *Why we read fiction: Theory of mind and the novel*. Ohio State University Press.

A Additional Details on SYMBOLICTOM

A.1 Detailed Description of Information Contained in Global Context G

In the main paper, we define G as a graph containing the true world state (as opposed to beliefs about the current world state). This means that G will represent where people and objects are truly located, regardless of beliefs. G will in general contain only the *observable* true world state. Thus, information passed verbally would not be stored in the global context (e.g. someone speaking in a room is not observable after they finished talking), and would instead be stored in the local contexts of the people that heard the speech. Since verbal interactions are not tested by available datasets, this distinction is not relevant in ToMi.

A.2 Prompts for Resulting State Extraction

For GPT3-Curie we 2-shot prompt with the following prompt (both for original and linguistic diversity experiments):

`John quit his job. The resulting state after this action is that John no longer has a job.\n\nJohn signed a contract. The resulting state after this action is that the contract is signed.\n\n<sentence>. The resulting state after this action is that`

We find that GPT3-Davinci, Flan-T5-XL, GPT3.5, and GPT4 are able to zero-shot answer to this subtask just by describing the instruction, but smaller models benefit from few-shot. We were unable to query Macaw for this task, so we instead rely on GPT3-Curie, a model of comparable size. Zero-shot instruction is as follows:

`<sentence>. What is the resulting state after this action? Do not add new information. The resulting state after this action is that`

We observe that GPT3 is significantly more robust to paraphrases than Flan-T5: Flan-T5 models are poor at detecting the resulting state for florid paraphrases, although the original phrasings are a straightforward task for Flan-T5.

Larger models like GPT3.5 and GPT4 are able to perform the task well zero-shot, similarly to GPT3; LLaMA models require fewer demonstrations than Flan-T5. We ran all main experiments implementing Resulting State Extraction with GPT3.

A.3 Solving PROCESSQUESTION using GPT3

Our explorations suggest that GPT3 (Curie and GPT3-Davinci `text-davinci-002`—the version used in all our experiments) can successfully extract entities and rephrase the question. See Figure 6 for an example prompt.

```
Original: Where will Anne look for the apple?
Rephrased without people: Where is the apple?
People mentioned in order: Anne

Original: Where will John think that Anne will search for the eggplant?
Rephrased without people: Where is the eggplant?
People mentioned in order: John, Anne
```

Figure 6: GPT3 shows one-shot generalization abilities from first-order to second-order questions.

B Details on Out-Of-Domain Evaluation

B.1 Linguistic Diversity Per ToMi Template

Sentence type	*Count*
Object's Position	38
Distractor Negative Sentiment	36
Distractor Positive Sentiment	31
Person Entered Room	21
Person Exited Room	19
Person Moved Object	18
Person's Position	9

Table 3: Number of paraphrases per original sentence template. Paraphrases were obtained from prompting GPT3-Davinci (`text-davinci-002`).

B.1.1 All Paraphrases of `PersonX entered the RoomY.`

```
PersonX entered the RoomY.
PersonX approached the RoomY.
PersonX arrived at the RoomY.
PersonX arrived in the RoomY.
PersonX bounded into the RoomY.
PersonX came by the RoomY.
PersonX came into the RoomY.
PersonX came to the RoomY.
PersonX crept into the RoomY.
PersonX entered the RoomY.
PersonX leapt into the RoomY.
PersonX showed up at the RoomY.
PersonX shuffled into the RoomY.
PersonX sidled into the RoomY.
PersonX slithered into the RoomY.
```

PersonX stepped into the RoomY.
PersonX tiptoed into the RoomY.
PersonX visited the RoomY.
PersonX walked into the RoomY.
PersonX went into the RoomY.
PersonX went to the RoomY.

B.1.2 All Paraphrases of `PersonX exited the RoomY.`

Prompted with the prompt: *Find 30 alternative ways of expressing the following sentence: Abigail exited the bedroom.* and manually filtering results (with this and other name/location selection.

PersonX exited the RoomY.
PersonX left the RoomY.
PersonX walked out of the RoomY.
PersonX stepped out of the RoomY.
PersonX departed the RoomY.
PersonX went out of the RoomY.
PersonX came out of the RoomY.
PersonX emerged from the RoomY.
PersonX quit the RoomY.
PersonX took off from the RoomY.
PersonX bolted from the RoomY.
PersonX flew from the RoomY.
PersonX ran from the RoomY.
PersonX sprinted from the RoomY.
PersonX jogged from the RoomY.
PersonX hurried from the RoomY.
PersonX crawled from the RoomY.
PersonX crept from the RoomY.
PersonX tiptoed from the RoomY.

B.1.3 All Paraphrases of `The Object1 is in the Container1.`

Prompted with `Object1=apple, Container1={fridge, envelope, bathtub}`. Then filtered to remove object-specific wording.

The Object1 is in the Container1.
The Object1 is stored in the Container1.
The Object1 is kept in the Container1.
The Object1 is located in the Container1.
The Object1 is situated in the Container1.
The Object1 is set in the Container1.
The Object1 is placed in the Container1.
The Object1 is found in the Container1.
The Object1 is positioned in the Container1.
The Object1 is set upon in the Container1.
The Object1 is put in the Container1.
The Object1 is laid in the Container1.
The Object1 is deposited in the Container1.
The Object1 is stationed in the Container1.
The Object1 is put to rest in the Container1.
The Object1 is set to rest in the Container1.
The Object1 is rested in the Container1.
The Object1 is set aside in the Container1.
The Object1 is stowed in the Container1.
The Container1 contains the Object1.
The Object1 is inside the Container1.
The Object1 is within the Container1.
The Container1 is where the Object1 is.
The Container1 has the Object1.
The Container1 is holding the Object1.
The Container1 is keeping the Object1.
The Container1 is safeguarding the Object1.
The Container1 is storing the Object1.
The Container1 has the Object1 within it.
The Container1 has the Object1 inside of it.
The Container1 is holding the Object1 within it.
The Container1 is keeping the Object1 inside of it.
The Container1 is safeguarding the Object1 inside of it.
The Container1 is storing the Object1 inside of it.
There is a Object1 in the Container1.
A Object1 is in the Container1.
The Container1 has a Object1 in it.
Inside the Container1 is a Object1.

B.1.4 All Paraphrases of `PersonX moved the Object1 to the Container1.`

PersonX moved the Object1 to the Container1.
PersonX relocated the Object1 to the Container1.
PersonX transferred the Object1 to the Container1.
PersonX shifted the Object1 to the Container1.
PersonX placed the Object1 in the Container1.
PersonX set the Object1 in the Container1.
PersonX put the Object1 in the Container1.

PersonX stowed the Object1 in the Container1.
PersonX stored the Object1 in the Container1.
PersonX hid the Object1 in the Container1.
PersonX shoved the Object1 into the Container1.
PersonX pushed the Object1 to the Container1.
PersonX carried the Object1 to the Container1.
PersonX conveyed the Object1 to the Container1.
PersonX led the Object1 to the Container1.
PersonX transported the Object1 to the Container1.
PersonX brought the Object1 to the Container1.
PersonX took the Object1 to the Container1.

B.1.5 All Paraphrases of `PersonX is in the RoomY.`

PersonX is in the RoomY.
PersonX is inside the RoomY.
PersonX is located in the RoomY.
PersonX is situated in the RoomY.
PersonX is present in the RoomY.
PersonX is to be found in the RoomY.
PersonX is contained in the RoomY.
The RoomY holds PersonX.
The RoomY shelters PersonX.

B.1.6 All Paraphrases of positive distractor sentences

PersonX has a bad case of Object1 fever.
PersonX is Object1 crazy.
PersonX is Object1-crazed.
PersonX is Object1-obsessed.
PersonX is a Object1 fiend.
PersonX is a Object1 maniac.
PersonX is a Object1-aholic.
PersonX is always thirsty for a Object1.
PersonX is besotted with the Object1.
PersonX is captivated by the Object1.
PersonX is charmed by the Object1.
PersonX is crazy about the Object1.
PersonX is crazy for the Object1.
PersonX is eager for the Object1.
PersonX is enamored with the Object1.
PersonX is enthusiastic about the Object1.
PersonX is entranced by the Object1.
PersonX is fascinated by the Object1.
PersonX is fond of the Object1.
PersonX is in love with the Object1.
PersonX is infatuated with the Object1.
PersonX is keen on the Object1.
PersonX is mad about the Object1.
PersonX is never seen without a Object1.
PersonX is nuts about the Object1.
PersonX is smitten with the Object1.
PersonX is spellbound by the Object1.
PersonX is taken with the Object1.
PersonX is wild about the Object1.
PersonX loves to drink from a Object1.
PersonX would do anything for a Object1.

B.1.7 All Paraphrases of positive negative sentences (`PersonX hates ObjectY`)

PersonX hates Object1.
PersonX can't stand the Object1.
PersonX despises the Object1.
PersonX detests the Object1.
PersonX is annoyed by the Object1.
PersonX is bothered by the Object1.
PersonX is concerned by the Object1.
PersonX is disconcerted by the Object1.
PersonX is discouraged by the Object1.
PersonX is disgusted by the Object1.
PersonX is disheartened by the Object1.
PersonX is disquieted by the Object1.
PersonX is grieved by the Object1.
PersonX is horrified by the Object1.
PersonX is irritated by the Object1.
PersonX is offended by the Object1.
PersonX is pained by the Object1.
PersonX is repelled by the Object1.
PersonX is revolted by the Object1.
PersonX is scandalized by the Object1.
PersonX is shocked by the Object1.
PersonX is sorrowful by the Object1.
PersonX is terrified by the Object1.
PersonX is troubled by the Object1.
PersonX is vexed by the Object1.
PersonX loathes the Object1.
The Object1 horrifies PersonX.
The Object1 is abhorrent to PersonX.
The Object1 nauseates PersonX.
The Object1 offends PersonX.
The Object1 repulses PersonX.
The Object1 revolts PersonX.
The Object1 scandalizes PersonX.
The Object1 shocks PersonX.

The Object1 sickens PersonX.
The Object1 terrifies PersonX.
The Object1 turns PersonX's stomach.

B.2 Structure of Story Structure Robustness Test Sets

B.2.1 Double Room False-Belief Episode

person1 entered the room1.
person2 entered the room1.
The object1 is in the container1.
The container1 is in the room1.
person2 exited the room1.
person1 moved the object1 to the container2.
The container2 is in the room1.
person1 exited the room1.
person2 entered the room2.
person1 entered the room2.
The object2 is in the container3.
The container3 is in the room2.
person1 exited the room2.
person2 moved the object2 to the container4.
The container4 is in the room2.
person2 exited the room2.

B.2.2 Three Active Characters Story

person1 entered the room1.
person2 entered the room1.
person3 entered the room1.
The object1 is in the container1.
The container1 is in the room1.
person2 exited the room1.
person1 moved the object1 to the container2.
The container2 is in the room1.
person1 exited the room1.
person3 moved the object1 to the container3.
The container3 is in the room1.
person3 exited the room1.

B.2.3 True-Belief Interaction, Falsified by Unwitnessed Third-Person Story

person1 entered the room1.
person2 entered the room1.
The object1 is in the container1.
The container1 is in the room1.
person1 moved the object1 to the container2.
The container2 is in the room1.
person2 exited the room1.
person1 exited the room1.
person3 entered the room1.
person3 moved the object1 to the container1.

B.2.4 Four Containers with Multiple Movements

person1 is in the room1.
The object1 is in the container1.
The container1 is in the room1.
person1 moved the object1 to the container2.
The container2 is in the room1.
person2 entered the room1.
person1 exited the room1.
person2 moved the object1 to the container3.
The container3 is in the room1.
person2 moved the object1 to the container4.
The container4 is in the room1.

C Expanded Results

Experimental Note: All zero-shot GPT3 (`text-curie-001` and `text-davinci-002`) experiments were performed between November 2022 and January 2023. GPT3.5 (`gpt-3.5-turbo`) and GPT4 (`gpt-4`) were added in May 2023.

C.1 Ablating FILTERBASEDONQUESTION from SYMBOLICTOM

FILTERBASEDONQUESTION definition This function filters the story S' to obtain an even shorter subset of the original story S'' by only keeping edges where at least one of the endpoints represents an entity mentioned in the question.

Last step of Algorithm 1 is applying FILTERBASEDONQUESTION, which yields an even shorter story to feed language models. We evaluate the effect this final filter has on the final performances reported by SYMBOLICTOM.

FILTERBASEDONQUESTION has a positive effect on Macaw-3B, GPT3, Flan-T5-XXL, and LLaMA-7B (+7, +3.5, +12.8, and +15 points in average accuracy gain across all question types), and a mild negative one on Flan-T5-XL, and GPT4 (-5.3, and -4 points of accuracy on average). See Table 7 for all differences between executing SYMBOLICTOM using this final filtering or not. Figure 7 visually represents the accuracy of all models by

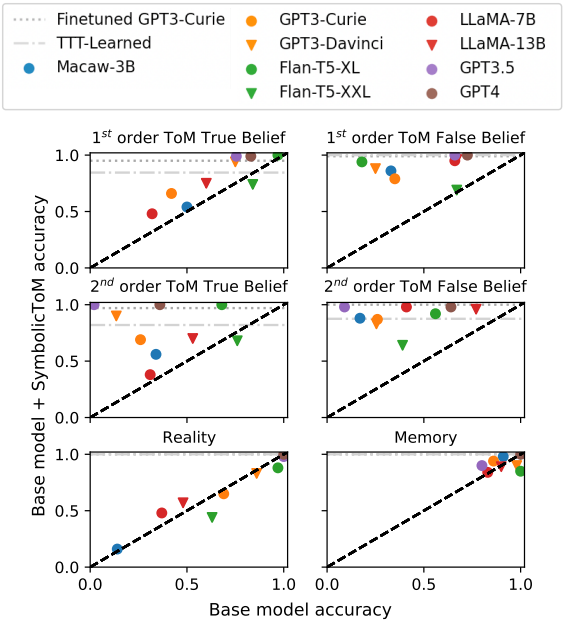

Figure 7: Precision using SYMBOLICTOM on ToMi, for several language models without the final filter function. Performance is shown for each question type, dots in upper triangle imply performance improvements. Full results table may be found in Table 6.

question type. Regardless of the final filter application, GPT4+SYMBOLICTOM significantly outperforms out-of-the-box GPT4 in all four ToM question types and maintains performance on Reality and Memory questions. For Flan-T5-XL, Flan-T5-XL+SYMBOLICTOM outperforms Flan-T5-XL significantly in all four ToM question types (e.g. +76 and +36 points in accuracy for first and second-order false belief questions), and shows slight declines for Reality and Memory questions—in line with findings on the full algorithm, but with less stark declines, suggesting that having more entities may help reduce bias towards answering rooms instead of containers. See Table 6 for the full table of accuracy differences.

Regardless of the final filtering application, SYMBOLICTOM shows improvements in theory of mind questions for all models. We only find the filter application to be relevant to beat the base model in theory of mind questions for Flan-T5-XXL.

C.2 Third-Order Theory of Mind Evaluation

We ask all third-order theory of mind questions for each D_2 story, such as "Where does p_1 think that p_2 thinks that p_1 will search for the o_1?". Questions involving p_2 will have a final answer c_1, since everyone saw p_2 leaving. We ask all six possible

	D_2's THIRD-ORDER TOM QUESTIONS
Off-the-shelf models	
Macaw-3B	13
Flan-T5-XL	32
Flan-T5-XXL	62
GPT3-Curie	28
GPT3-Davinci	19
GPT3.5	8
GPT4	26
LLaMA-7B	22
LLaMA-13B	39
SYMBOLICTOM + *Off-the-shelf models*	
Macaw-3B	85 (+72)
Flan-T5-XL	97 (+65)
Flan-T5-XXL	**100** (+38)
GPT3-Curie	89 (+61)
GPT3-Davinci	90 (+71)
GPT3.5	**100** **(+91)**
GPT4	**100** (+73)
LLaMA-7B	90 (+68)
LLaMA-13B	95 (+57)
Supervised models	
TTT	52
Finetuned GPT3	76

Table 4: Precision using SYMBOLICTOM on all questions from 100 stories for each of the modified test sets D_i. Supervised models were trained on ToMi; all others do not require training. Parenthesis reflect differences between using and not using SYMBOLICTOM: **bold** reflects higher overall performance, and **green** reflects the highest net improvements when using SYMBOLICTOM.

questions involving p_2. We also ask the two third-order theory of mind questions that do not involve p_2 nor repeats the same person twice consecutively ("Where does p_1 think that p_3 thinks that p_1 will search for the o_1?" and "Where does p_3 think that p_1 thinks that p_3 will search for the o_1?"), totaling eight questions per D_2 story.

Table 4 shows results for all models using $k = 2$ representations (same depth as in the main paper). Using SYMBOLICTOM significantly outperforms the supervised baselines and yields dramatic improvements with respect to using the LLMs off-the-shelf. We hypothesize that although the task theoretically requires $k = 3$, the second-order theory of mind representation already helps models avoid attending to parts of the story that are inaccessible to relevant characters.

C.3 Alternative RESULTINGSTATE Implementations

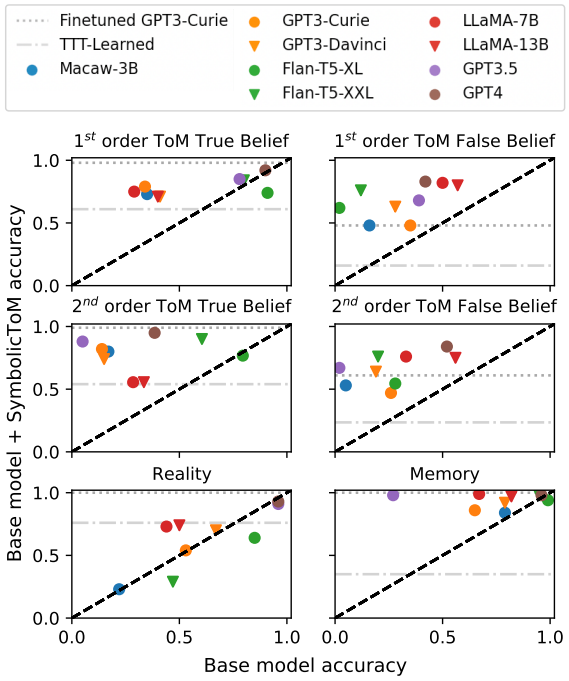

Figure 8: Results for ParaphrasedToMi when using the same model for implementing the RESULTINGSTATE function as in the final question-answering task (except using Davinci for Macaw, who did not show reliable enough few shot-prompting). Dots in upper triangle imply performance with SYMBOLICTOM is higher than using the base model out-of-the-box. Horizontal lines reflect supervised models' performance (higher is better).

RESULTINGSTATE(s) refers to the state of the world after s has been performed. For example, if "Oliver moved the apple to the box", then the resulting state is that "The apple is in the box". If "Oliver exited the bedroom", the resulting state would be that "Oliver is no longer in the bedroom". These are the relationships that we may insert in a context graph—actions are instantaneous and do not reflect an observable state.

In this section, we explore using the same LLM for implementing RESULTINGSTATE as well as the final inference. In the main text, we use Davinci for all non-GPT3-based models.

We find GPT3 to be among the most reliable to answer the resulting state of a given action in a zero-shot (Davinci) or two-shot (Curie) manner. Similarly, GPT3.5 and GPT4 perform well zero-shot: for experiments, we use GPT3.5 zero-shot and GPT4 two-shot to improve the resulting phrasing stability.

Additional exploration shows that although Flan-T5 models perform worse zero-shot than GPT models, they are capable of performing this task with more careful prompting. Figure 8 shows the results after nine-shot prompting Flan-T5-XL and eleven-shot prompting Flan-T5-XXL. Our explorations show that LLaMA models require fewer demonstrations than the Flan-T5 models to compute the resulting state: we observe highly reliable results when using six-shot prompting for LLaMA-7B, and seven-shot prompting for LLaMA-13B. Accuracy using LLaMA was even higher than when using GPT3.

C.4 Detailed Result Tables

All results in the appendix show accuracy as a ratio (between 0 and 1). For simplicity of reading, in the main text, they are referred to in percentages (values 0 to 100, higher is better). Figures 5, 6, and 7 show performances when applying the final filtering function, when not applying it, and the difference in performance between the two, respectively.

	1st TB	1st FB	2nd TB	2nd FB	Reality	Memory
Macaw-3B	0.86 [0.50]	0.79 [0.33]	0.86 [0.34]	0.84 [0.17]	0.10 [0.14]	0.95 [0.91]
GPT3-Curie	0.77 [0.42]	0.82 [0.35]	0.73 [0.26]	0.89 [0.26]	0.61 [0.69]	0.99 [0.86]
GPT3-Davinci	0.96 [0.75]	0.96 [0.25]	0.93 [0.14]	0.90 [0.26]	0.77 [0.86]	0.98 [0.98]
Flan-T5-XL	0.98 [0.97]	0.80 [0.18]	0.98 [0.68]	0.78 [0.56]	0.73 [0.97]	1.00 [1.00]
Flan-T5-XXL	0.98 [0.84]	0.95 [0.67]	1.00 [0.76]	0.90 [0.39]	0.13 [0.63]	1.00 [1.00]
LLaMA-7B	0.82 [0.32]	0.95 [0.66]	0.66 [0.31]	0.72 [0.41]	0.87 [0.37]	1.00 [0.83]
LLaMA-13B	0.82 [0.60]	0.86 [0.67]	0.70 [0.53]	0.62 [0.77]	0.87 [0.48]	1.00 [0.90]
GPT3.5	0.97 [0.76]	0.95 [0.66]	0.99 [0.02]	0.87 [0.09]	0.98 [1.00]	0.99 [0.80]
GPT4	0.98 [0.83]	0.94 [0.73]	0.98 [0.36]	0.89 [0.64]	0.94 [1.00]	1.00 [1.00]
Finetuned GPT3	0.95	0.99	0.97	1.00	1.00	1.00
TTT-learned	0.84	1.00	0.82	0.88	1.00	1.00

Table 5: Performance per model and question using SYMBOLICTOM, with out-of-the-box performance shown in brackets (100 samples per question type). Bottom rows represent supervised baselines.

	1st TB	1st FB	2nd TB	2nd FB	Reality	Memory
Macaw-3B	0.54 [0.50]	0.86 [0.33]	0.56 [0.34]	0.88 [0.17]	0.16 [0.14]	0.98 [0.91]
GPT3-Curie	0.66 [0.42]	0.79 [0.35]	0.69 [0.26]	0.87 [0.26]	0.65 [0.69]	0.94 [0.86]
GPT3-Davinci	0.94 [0.75]	0.88 [0.25]	0.90 [0.14]	0.83 [0.26]	0.83 [0.86]	0.90 [0.98]
Flan-T5-XL	1.00 [0.97]	0.94 [0.18]	1.00 [0.68]	0.92 [0.56]	0.88 [0.97]	0.85 [1.00]
Flan-T5-XXL	0.74 [0.84]	0.69 [0.67]	0.68 [0.76]	0.64 [0.39]	0.44 [0.63]	1.00 [1.00]
LLaMA-7B	0.48 [0.32]	0.95 [0.66]	0.38 [0.31]	0.98 [0.41]	0.48 [0.37]	0.84 [0.83]
LLaMA-13B	0.75 [0.60]	0.96 [0.67]	0.70 [0.53]	0.96 [0.77]	0.57 [0.48]	0.89 [0.90]
GPT3.5	0.99 [0.76]	1.00 [0.66]	1.00 [0.02]	0.98 [0.09]	0.98 [1.00]	0.90 [0.80]
GPT4	0.99 [0.83]	1.00 [0.73]	1.00 [0.36]	0.98 [0.64]	1.00 [1.00]	1.00 [1.00]
Finetuned GPT3	0.95	0.99	0.97	1.00	1.00	1.00
TTT-learned	0.84	1.00	0.82	0.88	1.00	1.00

Table 6: Performance per model and question using SYMBOLICTOM without FILTERBASEDONQUESTION, with out-of-the-box performance shown in brackets (100 samples per question type). Bottom rows represent supervised baselines.

	1st TB	1st FB	2nd TB	2nd FB	Reality	Memory
Macaw-3B	0.32	-0.07	0.30	-0.04	-0.06	-0.03
GPT3-Curie	0.11	0.03	0.04	0.02	-0.04	0.05
GPT3-Davinci	0.02	0.08	0.03	0.07	-0.06	0.08
Flan-T5-XL	-0.02	-0.14	-0.02	-0.14	-0.15	0.15
Flan-T5-XXL	0.24	0.26	0.32	0.26	-0.31	0.00
LLaMA-7B	0.34	0.00	0.28	-0.26	0.39	0.16
LLaMA-13B	0.07	-0.10	0.00	-0.34	0.30	0.11
GPT3.5	-0.02	-0.05	-0.01	-0.11	0.00	0.09
GPT4	-0.01	-0.06	-0.02	-0.09	-0.06	0.00

Table 7: Differences between accuracy of base models using SYMBOLICTOM with the final FILTERBASEDONQUESTION filter, and without using the final filter. As shown in Table 5 and 6, both versions are still far superior to not using SYMBOLICTOM.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section "Limitations" after Conclusions but before the references, as required by ACL 2023 guidelines.

☑ A2. Did you discuss any potential risks of your work?
Section "Ethics Statement" after Conclusions but before the references, as required by ACL 2023 guidelines.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract + Section 1

☑ A4. Have you used AI writing assistants when working on this paper?
GPT3-Davinci, for brainstorming paraphrases of sentences in Section 1 and Section 2. We later edited these paraphrases, but GPT3-Davinci gave interesting suggestions.

B ☒ Did you use or create scientific artifacts?

Left blank.

☑ B1. Did you cite the creators of artifacts you used?
Section 1, Section 4, Section 5, Abstract.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Artifact is an NLP research dataset.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 4

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The dataset is artificially generated.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 5

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Section 5

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Model does not require training, it is inference-time only.

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Model does not require training, it is inference-time only.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 5

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 4

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Annotators are only used to contribute to a small comment and are not used in evaluating our method.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Left blank.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Left blank.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not necessary for the small-scale experiment ran.