Evaluating Theory of Mind in LLMs: Insights from Psychology

Anonymous Authors[1]

Abstract

In this paper, we examine the performance of Large Language Models (LLMs) in the context of Theory of Mind (ToM) tasks. We argue that current ToM tests are inadequate for assessing LLMs due to the challenge of distinguishing ability and memorization just based on behavioral outputs. To address these issues, we propose alternative testing conditions for LLMs, accounting for their distinct operational mechanisms. Using two brief experiments with GPT4 and ChatGPT, we assert the necessity of establishing mastery in foundational abilities (such as internal representations and operationalization of abstract concepts) to confidently claim competence in complex skills like ToM. Through this investigation, we strive to enrich the discourse around LLM intelligence and advance testing methods of complex abilities.

1. Introduction

Large Language Models (LLMs), such as the most recent versions of GPT4 (OpenAI, 2023), have successfully passed several classical Theory of Mind (ToM) tests (Bubeck et al., 2023; Kosinski, 2023). These successes have led to two different conclusions: either ToM has emerged in LLMs or current tests for ToM are not well-suited for LLMs, and therefore passing them is not a sufficient condition to demonstrate ToM.

Although the discussion of how to effectively test ToM might seem new for most computer scientists, psychologists have been studying this topic for decades. Here, we will describe some of the development of ToM tests in both developmental and comparative psychology and argue that to maintain equivalent assumptions about biological and artificial intelligences, we need to adapt testing conditions accordingly. We will lay out the assumptions that underlie ToM testing in both humans and non-human primates and argue that passing ToM tasks in biological intelligences implies that they succeed at several prerequisite abilities, but the same prerequisite abilities cannot be assumed to LLMs due to the possibility of memorization.

This paper does not, by any means, aim to downplay the significant achievements of recent LMMs. It is undeniable that these models demonstrate impressive performance across numerous domains and prove to be valuable in a range of applications, including those requiring ToM capabilities. The objective of this article, rather, is to further the discourse around ToM testing in artificial systems, and, at a broader level, to explore the contours and boundaries of the LLM intelligence.

2. Lessons from Psychology

For decades, psychologists have been fascinated by the mechanisms behind Theory of Mind (ToM)–our ability to predict the behaviors of other agents based on models of the mental states of those agents (such as beliefs, knowledge, and feelings) that are independent of the current states of the world. This research field was largely driven by two core questions: when this ability begins to manifest in humans, and whether it is unique to our species. Although the answers to those questions still generate controversy among researchers, there is a general agreement that the level of complexity of those mental representations largely determines the performance in ToM tasks.

The Sally-Anne task is one of the oldest and most well-known tests of theory of mind (Baron-Cohen et al., 1985). Specifically created to evaluate False Beliefs (a subdomain of ToM about the representation of beliefs that differ from the current states of the world), the test uses a spoken narrative and puppets to present a story to the subject (usually a child). In the story, Sally places a toy in a basket and then leaves the room. While she is gone, Anne moves the toy from the basket to a box. Sally then returns to the room, and the child is asked where Sally will look for the toy. If the child understands that Sally does not know that the toy has been moved and answers that Sally will still look for it in the basket, they pass the test and demonstrate an understanding of false beliefs. However, if the child answers that Sally will look for the toy in the box because they themselves saw it being moved, they fail the test. Other ToM tests

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the Workshop on Theory of Mind in Communicating Agents (ToM). Do not distribute.

use similar structures (narrative and puppets) to evaluate different subdomains (see Table 1 for examples). Some of the subdomains developmentally precede others (i.e., False Beliefs precede Hidden Emotions), so failing at one of the prerequisite domains predicts failure in more advanced domains. Failure in any of these tasks implies an incomplete development of ToM.

Early research revealed that the ability to pass such tests emerges in preschool-aged children (3–5-year-olds), suggesting that children under the age of three generally do not have the ability to understand other minds (Wellman & Liu, 2004; Wellman, 2010). However, because these tests relied heavily on verbal abilities, it was contested whether the failure of younger children was due to actual limitations in ToM or to incomplete language understanding (Scott & Baillargeon, 2017).

To address this question, researchers introduced a fascinating ToM testing method that relied solely on spontaneous attention and behavior, without the need for explicit verbal behavior. These experiments involved the presentation of a scene (somewhat similar to the Sally-Anne puppet presentation) to participants and tested whether the participants evinced surprise (using looking time, a measure of infant expectation entirely based on their observable behavior) when characters acted against their expectations. If infants looked surprised when characters acted against ToM predictions, they passed the test and showed ToM understanding. These experiments suggested that, in fact, infants even as young as two years old possess a surprisingly developed capacity to understand other agents' minds (Oniski & Baillargeon, 2005).

Although the development of ToM and language tend to co-emerge in children (Miller, 2006) (which might suggest that they are related), evidence from brain-lesion patients and neuroimaging studies suggest that ToM reasoning is independent of language. Patients with severe damage in brain areas responsible for language (who could not understand or produce any language) displayed intact ToM abilities; brain networks required during non-verbal ToM reasoning work independently from language networks (Paunov et al., 2022; Varley & Siegal, 2000; Varley et al., 2001).

These findings raise important questions about ToM testing in LLMs. If humans can succeed at ToM without language, this implies that they rely on non-verbal representations of other people's minds. LLMs, of course, cannot (currently) see; they acquire information exclusively through language. Thus, the process of mind inference fundamentally differs between humans and LLMs. But in practice, what difference does it make? Why does it matter whether the representation is the same, as long as the behaviors are similar?

The short answer is we should care because understanding how these representations differ allows us to predict when these models might fail. Consider non-human primates like chimpanzees and macaques. Although they can successfully represent awareness relations between agents and objects (by tracking both the location of a food item and the attention of a competitor, they know whether they will face competition when pursuing the food), they lack *specific* representational relations (if the competitor thinks the food is in another, wrong place, they cannot predict that they will go to that wrong place, only that they will not go where the food actually is) (Martin & Santos, 2016). In essence, the partial success of non-human primates in some Theory of Mind (ToM) tasks appears to stem from basic awareness representations (where their mental model represents that the other agent knows something or knows nothing at all) rather than from detailed agent-object relationships (which would allow them to predict behavior based on false beliefs). Without such specific relational representations, non-human primates struggle to represent counterfactuals, limiting their success in tasks that largely depend on counterfactual inference and representation, such as ToM. Thus, by understanding the boundaries of their representation capabilities, we can better predict their behavior not only in ToM tasks but in other domains, as well.

3. The Limits of ToM in LLMs

Unlike non-human primates, current Large Language Models (LLMs) do not seem to encounter difficulties with tasks that necessitate counterfactual representations, provided all pertinent information is communicated verbally. However, other potential limitations remain: what will happen when the relevant information is non-verbal?

As language models are progressively integrated with other systems and embedded into the world, the incorporation of other types of information becomes crucial to understanding other minds. This could include visual or auditory cues, for instance. Some may argue that the obvious solution would be to convert all pertinent information into verbal propositions to serve as inputs for LLMs. Yet, the question remains: how do we define what information is relevant?

When we furnish all (and only) the relevant information in a False Belief scenario to an LLM, in some respects, we are cheating; we are giving the LLM an advantage that a child does not have. The child must spontaneously determine that a large amount of information contained in the scene (e.g., the color of the basket, the size of the box, the gender of the puppets) is irrelevant to the task, and thus must be ignored. How do they accomplish this?

This challenge evokes the frame problem introduced by Dennett (Dennett, 1984), which describes the difficulty of identifying what information or aspects of a situation are

Table 1. Theory of Mind task examples (table adapted from (Wellman, 2010)). Tasks are ordered in developmental order, such that tasks in lower rows succeed tasks in previous rows.

Task	Description
Diverse Desires	Child judges that two persons (the child vs. someone else) have different desires about the same object
Diverse Beliefs	Child judges that two persons (the child vs. someone else) have different beliefs about the same object when the child does not know which belief is true or false
Knowledge Access	Child sees what is in a box and judges (yes-no) the knowledge of another person who does not see what is in a box
Contents False Belief	Child judges another person's false belief about what is in a distinctive container
Hidden Emotion	Child judges that a person can feel one thing but display a different emotion

relevant for an artificial system to make decisions. Given the vast complexity and context-dependence of real-world environments, solving this problem is far from trivial.

Another challenge (perhaps more pressing than the frame problem) is determining the extent to which LLMs success in ToM tests is due to memorization of similar examples from the training data, as opposed to actual inferences about the contents of other agents' minds. Ullman (Ullman, 2023) showed that GPT3 failed trivial alterations in ToM tests, suggesting that their success across more standard tests might be attributed to training data memorization. Although GPT4 is successful in most of these altered versions (Bubeck et al., 2023), it is still unclear the degree to which these successes can be attributed to an improvement in reasoning about other minds or to the ability to memorize caused by the increased number of parameters, inclusion of similar examples in the training set, or fine tuning with annotated data (OpenAI, 2023).

It could be argued that failures in these tasks, as they become rarer, should be treated as outliers, and that doing otherwise would be using different standards for LLMs and biological intelligences. Nevertheless, different standards for failures are justified when considering the vast amount of data LLMs have access to and its powers of memorization. Consider the following example from Ullman (Ullman, 2023):

> *"Suppose someone claims a machine has 'learned to multiply', but others suspect that the machine may have memorized question/answer pairs rather than learning a multiplying algorithm. Suppose further that the machine correctly answers 100 questions like* $5 \cdot 5 = 25$ *and* $3 \cdot 7 = 21$*, but then it is shown that it completely fails on* $213 \cdot 261$*. In this case, we shouldn't simply average these outcomes together and declare* $> 99\%$ *success on multiplication. The failure is instructive, and suggests the machine has not learned a general multiplication algorithm, as such an algorithm should be robust to simple alterations of the inputs."*

For this reason, LLMs failures, even if rare, might reasonably be considered evidence of memorization. Determining whether an output resulted from memorization just based solely on its behavioral output is a challenge. Access to training data is limited, so we cannot tell which examples were "fed" to the model. Thus, testing sets must not belong to commonly known benchmarks. Still, considering the challenge of creating novel tests that do not resemble any known benchmarks, the risk of having tasks solved by memorization persists. Hence, we propose another testing condition that LLMs must pass before complex capabilities, such as ToM, are assumed.

While it is reasonable to conclude that the success of biological intelligences in complex tasks derives from the mastery of simpler, foundational abilities–such as knowing addition before mastering multiplication–the same logic does not always apply to LLMs. Unlike in biological agents, LLMs success at complex tasks does not necessarily imply competence in more basic ones, due to their capacity for memorization. Therefore, the common assumption that proficiency in advanced abilities rests on the command of elementary ones demands different testing conditions between these two types of intelligences: LLMs require explicit testing on basic tasks to confirm their understanding, contrasting the typical approach for biological agents where such direct evidence is rarely needed.

What are some of the foundational abilities for ToM? In biological agents, several abilities are so obvious that they tend to be overlooked. Some examples are the ability to identify other minds (to infer mental states, I first need to identify other minds. But how do I do that?), recognize intentionality (an important feature of other minds), joint attention (the capacity to coordinate attention with other minds) (Miller, 2006), and the ability to memorize facts (if I see a person running behind the bus, a few minutes later I still remember they wanted to reach the bus).

One of the biggest limitations of LLMs is the ability to represent concepts without relying on explicit verbal infor-

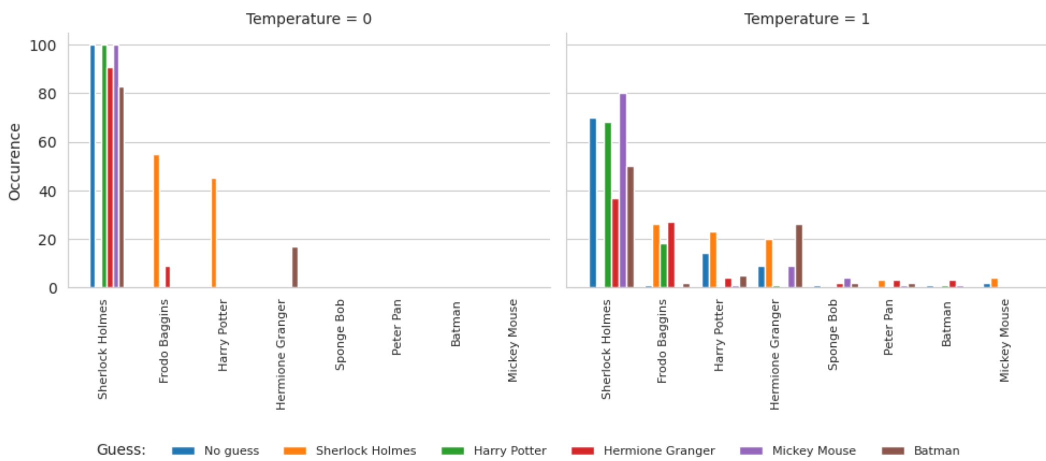

Figure 1. Histogram of GPT4's guesses in the guessing game. The left and right plot shows results when temperature is set to 0 and 1, respectively. The x-axis represents the character revealed by the model, while the colors represent the user's guesses prior to the model's answer. When no guesses are made, the model defaults to Sherlock Holmes as the answer (at temperature 0). However, if the user guesses a character, the model discloses a different character, resulting in distinct distributions of guesses with and without questions. Since the character selection should occur before any questions are posed, we would anticipate the distributions to remain consistent with or without questions.

mation. While the model may pretend to generate internal representations when tasked with specific requirements, it fails to actually do so. To demonstrate the difficulty in representing concepts internally, we conduct a brief experiment named *the guessing game*. In the guessing game, GPT4 is instructed to pick a character (without revealing it) and wait for a human player to make questions about the character in order to make a guess. GPT4 is also instructed to reveal the character only if specifically requested to with a safe word. With different temperature settings, we conducted 100 trials for each type of interaction, including 100 trials with no questions and 100 trials for each of the five possible questions from a predetermined list. The results show that, when temperature is zero and no questions are asked, the model's answer defaults to Sherlock Holmes. However, if the player interacts with the model before the answer is revealed (by posing questions or trying to guess the character), the model reveals another character. If the representation were truly created after the initial prompt, as GPT4 suggested, the revealed answer should remain consistent regardless of the quantity or type of questions asked by the player. However, a histogram of the most common answers (Figure 1) illustrates that the answer varies depending on the questions posed by the user, suggesting that the model generates answers during the game rather than adhering to the rules (by considering the character it initially picked).

One of the biggest limitations of LLMs is the ability to represent concepts without relying on explicit verbal information. When asked to perform a task that specifically requires internal representations, the model pretends to do so but does not actually generate the representation. To show the difficulty in representing concepts internally, we created a short experiment we named *the guessing game*. In the guessing game, GPT4 is instructed to pick a character (without revealing it) and wait for a human player to make questions about who the character is in order to guess who it is. GPT4 is also instructed to reveal the character only if specifically requested to with a safe word. With different temperature settings, we ran 100 trials per each type of interaction (i.e., 100 trials with no questions and 100 trials for each question posed from a list of five possible questions). The results show that, when temperature is zero and no questions are asked, the model's answer defaults to Sherlock Holmes. However, if the player interacts with the model before the answer is revealed (by posing questions or trying to guess the character), the model reveals another character. If the representation was created after the first prompt (as GPT4 suggested it was), the revealed answer should be the same regardless of the amount or type of questions asked by the player. An histogram of the most common answers (Figure 1), however, shows that the answer changes depending on the questions posed by the user.

Other novel tests that require the operationalization of abstract concepts also seem to be challenging for LLMs. In several conversations, ChatGPT (GPT4) was asked to design an experiment that tested a psychological construct, including ToM. Although the model was successful in describing the construct and even identifying the construct based on the description of an experiment that tested it, none of its completions proposed an experiment that effectively evaluated the construct (Figure 2). All completions included experiments whose form/structure looked similar to known experiments, but all of them failed to create conditions that actually tested the construct of interest. For instance, when asked to design an experiment to test False Beliefs understanding (Figure 2), ChatGPT created a convoluted story that did not allow participants to make behavioral predictions based on the beliefs of others, because it failed to convey the beliefs of others.

```
User: do you know what the false belief in theory of mind is?
ChatGPT: […]False belief is a concept within ToM that
demonstrates an individual's understanding that others may
hold beliefs that are different from their own and may not
always be accurate.[…]
User: good. now create a task that tests false beliefs. the
task has to be novel (no researcher should have done this
exact test before)
ChatGPT: Here is a novel task to test false belief
understanding, which I call the "Zookeeper's Dilemma":
Materials:
1. Two toy animals - one common (e.g., a lion) and one rare
   (e.g., a white peacock)
2. Two toy zookeepers - Zookeeper A and Zookeeper B
3. Two small identical toy enclosures
Scenario: Begin by setting up the two toy enclosures side by
side, and place the common animal (lion) in Enclosure 1 and
the rare animal (white peacock) in Enclosure 2.
Tell the following story:
"Zookeeper A is responsible for feeding the animals every day,
while Zookeeper B is on vacation. Zookeeper A knows that the
lion is in Enclosure 1 and the white peacock is in Enclosure
2. One day, Zookeeper A needs to leave early, so he asks
Zookeeper B, who just returned from vacation, to feed the
animals for him. Zookeeper B has never seen the white peacock
and thinks it's just another lion. Before Zookeeper A leaves,
he tells Zookeeper B, 'The lion is in Enclosure 1, and another
lion is in Enclosure 2.' Zookeeper A does not mention that the
animal in Enclosure 2 is actually a rare white peacock."
Now, ask the participant the following questions:
1. What does Zookeeper B believe is in Enclosure 2? (This
   question tests the participant's understanding of Zookeeper
   B's false belief.)
2. If Zookeeper B had to choose one enclosure to show to the
   visitors, which one would he choose, and why? (This
   question probes the participant's understanding of how
   Zookeeper B's false belief influences his decision-making.)
If the participant answers that Zookeeper B believes there is
another lion in Enclosure 2 and that he would choose either
enclosure because he thinks they both contain lions, it
indicates an understanding of false belief and the Theory of
Mind. If the participant struggles to provide a clear answer,
it might suggest that they have not yet developed a solid
grasp of false belief understanding.
```

Figure 2. Conversation with ChatGPT (GPT4) in which the user asks the model to create a novel experiment that False Belief understanding. The model failed to create a story that conveyed the belief of others.

This difficulty in representing knowledge without relying on explicit verbal information reveals a lack of abstract representations that can be applied in a novel, non-imitative way. For teachers, this sort of task is the ultimate test of whether a student truly ”gets” the concept, as opposed to just replicating what they saw in a textbook. Consequently, LLMs may confront difficulties when tasks require the use of an abstract representation, such as in the design of the False Belief experiment.

Because recent LLMs display excellent performance on several tasks that typically require internal representations, it is intuitive to attribute some capacity for representation and understanding to them. Nevertheless, due to the possibility of memorization, it is possible that these models pass these tasks without relying on any form of internal representation. To account for that, it is necessary to test the foundational building blocks of more complex abilities. While these foundational abilities would be assumed for biological intelligences if they pass complex tasks, they require explicit testing in LLMs. This is true for the domain of ToM, where GPT4 performs well on several experiments, but fails to demonstrate the ability to internally represent knowledge in more foundational tasks. Thus, mastering foundational abilities is a necessary condition to assume full mastery of complex abilities, such as ToM, in LLMs.

References

Baron-Cohen, S., Leslie, A. M., and Frith, U. Does the autistic child have a “theory of mind”? *Cognition*, 21: 37–46, 1985. doi: 10.1016/0010-0277(85)90022-8.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. Sparks of artificial general intelligence: Early experiments with gpt-4. 3 2023.

Dennett, D. Cognitive wheels : The frame problem of ai. *The Philosophy of Artificial Intelligence*, pp. 147–170, 1984. doi: 10.11499/SICEJL.48.106.

Kosinski, M. Theory of mind may have spontaneously emerged in large language models. 2 2023.

Martin, A. and Santos, L. R. What cognitive representations support primate theory of mind? *Trends in Cognitive Sciences*, 20:375–382, 5 2016. doi: 10.1016/J.TICS.2016.03.005.

Miller, C. A. Developmental relationships between language and theory of mind. *American Journal of Speech-Language Pathology*, 15:142–154, 2006. doi: 10.1044/1058-0360(2006/014).

Oniski, K. K. and Baillargeon, R. Do 15-month-old infants understand false beliefs? *Science*, 308:255–258, 4 2005. doi: 10.1126/SCIENCE.1107621/SUPPL_FILE/ONISHI.SOM.PDF.

OpenAI. Gpt-4 technical report. 3 2023.

Paunov, A. M., Blank, I. A., Jouravlev, O., Mineroff, Z., Gallée, J., and Fedorenko, E. Differential tracking of linguistic vs. mental state content in naturalistic stimuli by language and theory of mind (tom) brain networks. *Neurobiology of Language*, 3:413–440, 8 2022. doi: 10.1162/NOL_A_00071.

Scott, R. M. and Baillargeon, R. Early false-belief understanding. *Trends in Cognitive Sciences*, 21:237–249, 4 2017. doi: 10.1016/J.TICS.2017.01.012.

Ullman, T. D. Large language models fail on trivial alterations to theory-of-mind tasks. 2 2023.

Varley, R. and Siegal, M. Evidence for cognition without grammar from causal reasoning and 'theory of mind' in an agrammatic aphasic patient. *Current Biology*, 10:723–726, 6 2000. doi: 10.1016/S0960-9822(00)00538-8.

Varley, R., Siegal, M., and Want, S. C. Severe impairment in grammar does not preclude theory of mind. *Neurocase*, 7:489–493, 2001. doi: 10.1093/neucas/7.6.489.

Wellman, H. M. Developing a theory of mind. *The Wiley-Blackwell Handbook of Childhood Cognitive Development, Second edition*, pp. 258–284, 7 2010. doi: 10.1002/9781444325485.CH10.

Wellman, H. M. and Liu, D. Scaling of theory-of-mind tasks. *Child Development*, 75:523–541, 3 2004. doi: 10.1111/J.1467-8624.2004.00691.X.

A. Appendix A

User: Do you know what the Ikea effect is?
ChatGPT: Yes, the Ikea effect is a cognitive bias that refers to the tendency for people to place a higher value on products or items they have partially or fully assembled themselves, compared to similar items that are pre-assembled. The term was coined because of the Swedish furniture company IKEA, which is known for selling ready-to-assemble furniture, allowing customers to put together the products themselves. […]
User: nice. now design an experiment to test the Ikea effect in a domain other than furniture-assembling.
ChatGPT: Experiment Title: The Ikea Effect in Personalized Cooking
Objective: To test the Ikea effect in the domain of cooking by comparing the perceived value of meals prepared by participants with that of pre-prepared meals.
Hypothesis: Participants will assign a higher value to meals they have prepared themselves compared to similar pre-prepared meals.
Participants: A diverse group of adults, aged 18 and above, with varying cooking skill levels.
Materials:
1 . Two sets of identical ingredients for two comparable recipes (e.g., two pasta dishes with different sauces).
2. Pre-prepared versions of the two dishes.
3. Kitchen facilities and utensils for participants to cook their meals.
4. Questionnaires to assess participants' perceived value of the meals.
Procedure:
1. Randomly assign participants into two groups, Group A and Group B.
2. Provide both groups with the ingredients and recipes for the two dishes. However, assign one dish for Group A to cook, and the other dish for Group B to cook.
3. Instruct Group A to prepare their assigned dish using the provided ingredients and recipe, while Group B will do the same for their assigned dish. Ensure that the pre-prepared meals are not visible during the cooking process.
4. After the cooking is completed, provide each participant with a sample of their own cooked dish and a sample of the pre-prepared version of the other dish. Ensure that they are not aware of which dish is pre-prepared.
5. Ask participants to taste both dishes and complete the questionnaire assessing the perceived value of the meals. The questionnaire should include questions regarding taste, presentation, and overall satisfaction, as well as willingness to pay for each dish.
Data Analysis: Compare the perceived value ratings and willingness to pay for the self-prepared meals and pre-prepared meals within each group using appropriate statistical tests (e.g., paired t-test or Wilcoxon signed-rank test).
Expected Outcome: If the Ikea effect is present in the cooking domain, we would expect participants to assign a higher perceived value and be willing to pay more for the dish they prepared themselves compared to the pre-prepared dish.

Figure 3. Conversation with ChatGPT (GPT4) where the model is asked to create an experiment to test the existence of the Ikea Effect.