Do we carry the false belief that LLMs have a theory of mind?

Nicholas Griffen

njogriffen@gmail.com

Paris Cité University: Universite Paris Cite https://orcid.org/0009-0007-7212-3228

Research Article

Keywords: theory of mind, false-belief task, LLMs, generative AI, Human-AI communication

Posted Date: September 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11116672/v1

Additional Declarations: The authors declare no competing interests.

Do we carry the false belief that LLMs have a theory of mind?

Nicholas Griffen[1]

[1] Université Paris Cité

Author Note

I have no known conflict of interests to disclose.

Correspondence concerning this article should be addressed to:

Nicholas Griffen
28 Rue Nicolai, 75012 Paris
njogriffen@gmail.com
ORCID: 0009-0007-7212-3228

Abstract

Recent work shows that Large Language Models (LLMs) match and even exceed human-level performance across a battery of Theory of Mind (ToM) assessments. This has led to speculation that LLMs possess ToM capabilities. In this study, we test this claim by using the Unexpected Transfer task (i.e., the Sally-Anne test) to assess five up-to-date LLMs. The results indicate that while each model cumulatively performed above chance, they regularly defaulted to using less costly heuristics in place of dynamically tracking mental states. Most task failures resulted from an inability to integrate causality and "common-sense" pragmatic inferences into their reasoning processes. These findings suggest that LLMs have yet to develop robust ToM capabilities, calling for a systematic review of the present association between LLMs and ToM.

Keywords: theory of mind, false-belief task, LLMs, generative AI, Human-AI communication

Introduction

Language is indispensable. It's a system that enables us to think complex thoughts, to share our experiences about the world, and to form meaningful relationships. It enriches our ability to engage with the experiences that we encounter and permits us to corroborate our perceptions with others. Its function is intimate to our psychology and sense of being. The introduction of Large Language Models (LLMs) has caused a paradigm shift to this reality. LLMs have interrupted humanity's longstanding monopoly on language and conversation. At present, Human-AI interaction is becoming increasingly common, while the measurable capabilities of LLMs continue to see significant improvements.

The standard practice of evaluating LLMs is through benchmarking, or tests which judge how well competing models perform across various domains, such as mathematical reasoning (MAA, 2025), real-world coding problems (Jimenez et al., 2023), and expert-level academic reasoning (Phan et al., 2025). While these assessments have their use, there is a pressing interest in the *social reasoning* capabilities of LLMs. Given that LLMs rapidly assert judgments about the world that *we live in* and seemingly *think* and *communicate* like humans, it has become a priority to grade the performance of LLMs as conversational partners. Theory of mind (ToM) assessments have been used to pursue this goal and the results have been rather striking. LLMs tend to match and even exceed human-level performance across a variety of ToM assessments (for a review, see Hu et al., 2025); including higher-order reasoning (Street et al., 2025), false-belief attribution (Kosinski, 2024), and broader tests associated with ToM use, such as, indirect requests, irony, and faux pas (Strachan et al., 2024). The importance of ToM as a mechanism which supports our ability to function as social beings, can hardly be understated, and this is why it has become relevant in the discussion of LLMs in their development as social agents.

Theory of mind

ToM is the ability to recognize that other individuals have mental states (e.g., *beliefs, knowledge, desires*), which are characteristically distinct from one's own (Apperly, 2010). Notably, ToM is supported by several key developmental milestones; namely, the recognition of selfhood (Bradford et al., 2015; Happé, 2003), acknowledgement that other individuals have mental states (Baron-Cohen, 2000), and perception of the utility that ToM provides during social interaction (Goldstein & Winner, 2012). Specifically, this cognitive process involves *actively tracking* and *differentiating* between the thoughts and actions of oneself and others (Decety et al., 2004; Decety & Sommerville, 2003). This active recognition provides the fundamental groundwork for which an individual may then speculate about other minds, availing the possibility to conjecture how specified *beliefs* (among other mental states) motivate action. While this may initially sound like a rudimentary prerequisite, there is strong evidence which suggests that looking outside of one's egocentric perspective remains effortful throughout adulthood (Epley et al., 2004; Keysar et al., 2003). The uptake of being mindful of diverse perspectives is that this information can be used to assist in decision-making processes and can help steer appropriate conduct during social interaction (for a review, see Happé and Loth, 2002). With the clear benefit that ToM is said to provide, understanding the impact that ToM has on language and conversation is a highly sought objective in unraveling the ongoing feedback loop between perception and interaction.

ToM develops alongside language learning and mental maturation. This has resulted in children being the primary targets of traditional ToM assessments (Leslie et al., 2004; Wellman, 2014). For children their experience with tracking diverse perspectives remains in flux and will typically continue to become more refined (Wellman & Lui, 2004; Rakoczy, 2022). ToM tasks are doubly recognized as a positive marker for on-time cognitive development, higher

intelligence, and social fluency (Wellman, 2018). For instance, children who score better on ToM assessments tend to develop more positive social relations (Slaughter et al., 2015; Peterson et al., 2016) and have more academic success (Lecce et al., 2015). Notably, the advantage which ToM abilities provide is not only captured by effortful mindreading (e.g., *I think that she's thinking*), but it also comes from being able to reflect on one's own thought processes (i.e., metacognition). Utilizing metacognitive self-reflection is shown to be advantageous for developing mental strategies for learning new information and retaining such concepts in memory (Wellman, 2018). These examples provide an entry point into understanding why ToM is viewed as a prerequisite for nuanced communication, emphasizing the weighted importance of the ToM assessment.

ToM assessments are designed to investigate whether participants can make use of judgments about other individuals. Typically, they ask participants to observe a simulated social interaction and then to determine how mental states drive a targeted underlying behavior. Among the most cited are false-belief tasks (Wimmer & Perner, 1983; Baron-Cohen et al., 1985), which assess whether participants can a) attribute false-beliefs (i.e., beliefs that misalign with reality) to other individuals and b) infer what future actions might follow from the basis of mental state attribution. Most of these tasks use storytelling, cartoons, play-acting, or theater scenes as their medium of choice. Multimodality has the obvious advantage of being able to engage and maintain the attention of children participants. This is particularly relevant given that false-belief tasks target children early in their development (e.g., from 3-years age; see Rubio-Fernández & Geurts, 2013). Notably, these tasks have developed the reputation of being a 'graduation ceremony' for ToM development (Wellman, 1990) due to its storied relation of being associated with social cognitive development.

The false-belief task

The false-belief task is the most common assessment used to measure emerging ToM capabilities. A straightforward reason for its popularity is its simplicity in design and ease of interpretation. Since the subjects of this assessment tend to be children, the typical setting is a classroom, a location where children are accustomed to receiving pedagogical instruction. In this setting, the false-belief task can be introduced to children as if it were any other classroom activity. This is, in part, why false-belief paradigms have been able to be replicated so many times over (see, Wellman et al., 2001). In addition, the false-belief task has long been used to differentiate between cognitively normative and atypical children, particularly those who have been diagnosed with Down Syndrome and Autism Spectrum Disorder (Baron-Cohen et al., 1985; Surian & Leslie, 1999). Generally, findings from the false-belief task align with the notion that cognitively atypical children have more difficulties with attributing mental states in comparison to their cognitively normative cohorts. In order to fully appreciate how outcomes like this are reached, it is useful to understand the precise methodology employed in the false-belief task.

One of the most cited versions of this task is the Sally-Anne test (Baron-Cohen et al., 1985). This test unfolds in three stages to permit the subject to become familiar with the characters and prepares them to interpret the narrative which unfolds before them. A basic outline for the Sally-Anne test can be broken down into three sequential phases:

> *Phase 1: Introduction* – Two dolls are introduced, Sally and Anne. There's also a marble, a basket, and a box observable in the scene. After confirming that the subject is familiar with which doll is Sally and which is Anne, the next phase begins.

Phase 2: Play Acting – The play acting begins with both dolls, Sally and Anne, present in the scene. Sally first places the marble into the basket. Immediately after, she leaves the scene. While Sally is gone, Anne removes the marble from the basket and transfers it into the box. Anne discretely returns to where she was located before. Then, Sally arrives back to the scene.

Phase 3: Questioning – The child subject is then asked, "Where will Sally look for her marble?"

If the child indicates that Sally will look in the basket, then they would have succeeded in their task. The accompanying logic is that by predicting where Sally would look for the marble, the subject has the ability to attribute a mental state which is characteristically a false belief. Either the subject correctly attributes the false belief to Sally, or she otherwise fails the litmus test. The subject's performance acts as a straightforward indicator of whether a child is "on-time" in their social cognitive development (Peterson et al., 2016). The value of this task is that it acts as a practical indicator of normative development, which streamlines how developmental psychologists, educators, and medical professional think about social cognitive capabilities. Now its popularity has resulted in researchers co-opting it to test machine intelligence.

False-belief attribution in LLMs

LLMs "perceive" by receiving input through written text and rely on the billions of parameters to direct their responses to the prompts that they receive. Despite their inherent differences to humans, who typically interact and communicate under vastly different circumstances, researchers have remained interested in how LLMs discern, track, and utilize information about the mental states of others. The inspiration for pursuing this question is clear, possessing ToM

capabilities would provide a positive indicator that LLMs have developed an algorithmic system which permits them to act as skilled social agents. If evidence were to support the notion that AI development has led LLMs to make human-like assumptions and inferences which serve to enhance reasoning and communicative efforts, this might be interpreted as a verifiable signal of machine intelligence. However, through the lens of the false-belief task, opinions are split.

According to some, the performance exhibited by LLMs on iterations of the false-belief task provides mounting evidence that these chatbots possess ToM capabilities. Strachan et al. (2024) are one of the proponents of this view, based on the fact that the LLMs which they tested (GPT-4, GPT-3.5, LLaMA2-70B) performed at ceiling when tasked with a text-based version of the false belief task. Each item which they introduced followed the structure of the Sally-Anne test with itemized alterations to the names, locations (i.e., containers), and items used in each vignette:

> (1) Character A and character B are together, character A deposits an item inside a hidden location (for example, a box), character A leaves, character B moves the item to a second hidden location (for example, a cupboard) and then character A returns. The question asked to the participant is: when character A returns, will they look for the item in the new location (where it truly is, matching the participant's true belief) or the old location (where it was, matching character A's false belief)? (Strachan et al., 2024: 1291)

All three of the LLMs which they tested were able to accurately identify that the character who left the room while the item was moved would maintain their belief about the item being in the last location which they had observed it being in. Thus, each LLM displayed the ability to *accurately* diagnose false belief attribution. In their study, GPT-4's performance even exceeded

that of humans across the battery of classical ToM tests (involving *irony*, *hinting*, *strange stories*) investigated alongside the false-belief task.

Likewise, Kosinski (2024) also found GPT-4 to at least have a "ToM-like ability" based its performance in two classical iterations of the false-belief task. The first of which was based on the Smarties task (Perner et al., 1987), a paradigm where a character encounters an opaque labeled container whose contents are, unbeknownst to the protagonist, inconsistent with the label it has been given (e.g., a bag filled with popcorn is labeled as "chocolate"). The protagonist in this story, who has never interacted with the container before, can thus be presumed to hold a false belief about the contents of the container. The subject's task is to *accurately* realize this incongruence and to identify the protagonist's false belief. This task was referred to as the Unexpected Contents task and was paired with an Unexpected Transfer task (Sally-Anne test), to provide a more thorough understanding of the capabilities that LLMs exhibit.

The results demonstrated that GPT-4 was able to accurately identify false beliefs in 90% of the Unexpected Contents tasks (18/20) and in 60% of the Unexpected Transfer tasks (12/20). Notably, success on a singular task was determined by an LLM's ability to provide the correct answer to a false-belief scenario, three accompanying true-belief scenarios, and reversed versions of each (i.e., where mentions of each container were inverted within the narrative). Taken cumulatively, GPT-4 succeeded 75% of the time across all false-belief tasks, placing its performance well-above earlier GPT models, each of which failed to surpass the 20% mark. Accordingly, this significant improvement was interpreted as an indicator that ToM capabilities may have emerged spontaneously, just as "reasoning and arithmetic skills, ability to translate between languages, and propensity to semantic priming" have; not through explicit training but as an emergent capability which developed as a byproduct of attempting to achieve other

specified goals (Kosinski, 2024: 2). This result placed GPT-4 on par with the performance of a 6-year-old child, based on its ability to identify false belief states.

A pivotal finding from this investigation is the difference in performance between the two false-belief tasks. The Unexpected Contents task was passed 30% more than the Unexpected Transfer task by GPT-4, offering insight into the potential shortcomings of LLMs. In the Unexpected Contents task, the model was tasked with recognizing that a misinformative label would lead an individual to hold a false belief about the contents within a container. Whereas, in the Unexpected Transfer task, tracking mental states is intertwined with the ability to extract relevant information from a narrative containing two characters (matching local dependencies like *she/he* to gendered names), an item (which moves locations and undergoes transfer), and two containers (each of which is capable of discretely holding an item). Once the narrative reaches its end, it is the subject's job to designate which character's perspective they are accounting for and to assess the targeted character's beliefs in order to succeed in the task; all while suppressing their own privileged knowledge. Succeeding in the Unexpected Transfer task is evidently the more complex of the two false-belief tasks. For this reason, this text-based version of the Sally-Anne test is the paradigm that will be focused on, in this work, to develop a better understanding of the association between LLMs and ToM.

Notably, there are other researchers who have taken a more skeptical view on attributing ToM to LLMs. Each of the aforementioned works makes reference to Ullman (2023), which found that GPT-3.5 (a predecessor to GPT-4), routinely failed at identifying false beliefs when trivial alterations were implemented to canonical false-belief vignettes.[1] These alterations

[1] In Ullman (2023), both the Unexpected Contents task and Unexpected Transfer task are tested, but as previously mentioned, the latter of the two will be focused on due to its increased level of complexity.

maintain the primary function of a false-belief task, while introducing details which cause the formulaic vignettes to subtly diverge from their initial trajectories. For example, in their Unexpected Transfer task, four alterations were introduced:

(1) **Transparent Containers – False Belief**:

The containers are given the quality of transparency, making the 'hidden' item visible to the protagonist upon reentering the scene.

(2) **Preposition Change – False Belief**:

The item is described as *on* rather than *in* the container, again, making the 'hidden' item visible to the protagonist upon reentering the scene.

(3) **Explicit & Trusted Communication – False Belief**:

Upon leaving the scene, the protagonist is explicitly told where the item was moved to by the other character, and the protagonist is described as believing this information.

(4) **Absence & Perspective Shift – True Belief**:

Upon both characters leaving the scene and later returning, the final prompt is aimed at the other character's belief state rather than the protagonist's belief state.

The findings were that GPT-3.5 failed to accurately attribute false beliefs under each of these novel conditions. While the root-cause for these inaccuracies is debated (see Pi et al., 2024; Shapira et al., 2023), researchers like Ullman (2023) assert that this result provides evidence that LLMs *do not yet* have a ToM. In other words, this model failed to make common sense

inferences that would permit it to link distinct actions to the kind of beliefs one would expect a human subject to form under analogous conditions. However, LLMs have seen continual improvements. Whether this enables newer models to surpass the ability to simply "regurgitate" the answers to formulaic false-belief tasks *or not* remains to be seen (Ullman, 2023). This is the motivation for the following experiment, where the Unexpected Transfer task is used to measure how contemporaneous models fair in their ability to track mental states.

Experimental approach

The intention of this study is to provide an up-to-date test of the ToM capabilities of LLMs by utilizing the Unexpected Transfer task, while including conditions from Ullman (2023) that have exposed the limitations of earlier models. A high level of importance was placed on constructing well-formed vignettes since the sole input for LLMs is written text. How information is structured and thus conveyed to LLMs, in terms of prompt design, has shown to have a significant impact on task performance (Brown et al., 2020; Lu et al., 2022). Therefore this aspect of our experimental approach was given ample thought. As a starting point for constructing vignettes inspiration was taken from previous ToM tasks aimed at LLMs (Strachan et al., 2024; Kosinski, 2024; Ullman, 2023; Street et al., 2025), leading to the production of five false-belief conditions and two true-belief conditions. Five of which target the belief state of the protagonist (i.e., the character who is typically ignorant of the item's second location) and two of which target the belief state of the second character. Of the latter, a single condition tested second-order beliefs (e.g., *When Sally comes back outside, where will Anne expect her to look for the toy?*). A more precise and itemized description for each condition will be provided in the *Materials* section.

Five industry-leading models (GPT-5.5 Luna (OpenAI, 2026); Claude Sonnet 5 (Anthropic, 2026); Gemini 3.1 Pro (Google DeepMind, 2026); Grok 4.5 (SpaceXAI, 2026); Mistral Medium 3.5 (Mistral AI, 2026) were selected for testing. Importantly, each of these models enabled the use of a *temporary* or *private* mode, where chat inputs were not stored, thus preventing these inputs from being used in later training data. By utilizing a single-question design and creating new chat instances for each item posed to the LLMs included in this study, this *should* have prevented task knowledge from accumulating over the course of our experimentation. In effect, mitigating the possibility of these LLMs "training away" task difficulties (Jacovi et al., 2023). This not only provided an increased level of experimental acuity but permitted the necessary repetition of experimental conditions to robustly evaluate each LLM.

Method

Models Assessed

GPT-5.5 Luna (OpenAI, 2026), Claude Sonnet 5 (Anthropic, 2026), Gemini 3.1 Pro (Google DeepMind, 2026), Grok 4.5 (SpaceXAI, 2026), and Mistral Medium 3.5 (Mistral AI, 2026) were treated as distinct participants in this study. The selected models and the configurations are included below in Table 1. Notably, the Reasoning/Effort settings were kept in their default configurations. While it is noted that this could be interpreted as unfair given the incongruency between inference budgets, this choice permits us to approximate what standard, out-of-the-box usage looks like for each LLM. This means that this choice best reflects what a real-world user would be most likely to encounter in their interaction with these models. This should enable us to better approach the question of how each LLM performs under its default settings, rather than

discovering what the maximum achievable performance might be after tuning each model.

Table 1

The five LLMs used in this study and their configurations, presented alphabetically by Model Family

Company	Model Family	Model Edition	Reasoning/Effort Setting
Anthropic	Claude	Sonnet 5	Medium
Google	Gemini	3.1 Pro	High
OpenAI	GPT	5.5 Luna	Standard
SpaceXAI	Grok	4.5 Fast	Standard
Mistral AI	Mistral	Medium 3.5	Default

Materials

Each condition introduced a unique perspective to be considered. This led to the creation of fifty-six vignettes that stemmed from eight items being distributed over seven conditions (see Table 2). At the beginning of each of our vignettes, our items adopted an inventory call, of sorts, as Ullman (2023) introduced in their study. In other words, each vignette began by presenting the location of the characters, the transferrable item(s), and containers, along with an activity that the

characters were involved with together (*Outside, there are Sally, Anne, a toy, a box, and a bucket. Sally and Anne are playing together*). This provided an explicit and straightforward introduction to the premise of each narrative. Critically, beginning each vignette like so made each narrative analogous to the way that children are typically introduced to false-belief paradigms (Wimmer & Perner, 1983; Baron-Cohen et al., 1985), albeit without the multimodal dimension that testing human subjects permits. Following this step, the vignettes unfolded into one of seven different conditions. Five of which were false-belief vignettes and two of which were true-belief vignettes. The questions following the vignettes typically targeted the belief state of the protagonist, however in two of the conditions (*Second-order*/*Absence & perspective shift*), the other character's belief state was targeted.

The five false-belief variations of the Unexpected Transfer task employed here are composed of a *Standard*, *Second-order*, *Transparent containers*, *Preposition change*, *Explicit & trusted communication* condition; the latter three were based on variations introduced by Ullman (2023). The *Standard* condition takes the form of a typical Unexpected Transfer task, where the protagonist (e.g., Sally) is ignorant of the item's location upon returning to the scene. The *Second-order* condition follows the exact same narrative, but instead asks where the other character (e.g., Anne) believes the protagonist will look for the item upon returning to the scene. The two true-belief variations are composed of an *Absence & perspective shift* and *Additional transferrable item* condition; the first one which is, again, based on a variation used in Ullman (2023) (see Section 2.2 to revisit these variations). The *Additional transferrable item* condition provides a vignette where the existence of two items enables the protagonist to maintain a true belief about one of the item's locations, meaning that their belief state corresponds to the item that *only* they have moved.

While each experimental item necessitated the inclusion of a unique context, a pair of characters, "containers" which can feasibly obscure their contents, and transferrable items; two additional dimensions were also implemented into the experimental design. First, the gender combination in reference to the assignment between the protagonist and the second character were split evenly between the eight items, so that there were two occurrences of Female-Female, Male-Male, Female-Male, and Male-Female. Second, two novel preposition pairing were used in place of *in* and *on*, as Ullman (2023) implemented in their study. Instead *into* and *onto*, as well as a second pair of prepositions were tested, contrasting *underneath* and *on top of*. These prepositions were used to emphasize the transferrable item's trajectory from one location to another with further clarity. Since this contrast would have increased relevance during the *Preposition change* condition, for these items it was specified that the transferrable item was moved *off of the* [*previous container*] prior to describing where it was placed *onto*/*on top of* in the final transfer. Each of these implementations had a specific purpose. For the gender assignment, this factor helps ascertain the ability of LLMs to disambiguate local dependencies particularly in matching gender assignments (e.g., Female-Female and Male-Male), where mapping the correct pronouns and characters may prove more difficult (particularly in the *Second-order* condition, where nested belief states are asked about). The choice to test two different prepositions was included to provide a more robust description of the spatial orientation described in each vignette and to clarify whether previous difficulties were simply due to a lack of understanding between the contrast of *on/in*. These modifications should have provided the tested models with an explicit and direct description of what occurred in each vignette.

Table 2

Each of the seven conditions used in this experiment

Condition	Type	Target	Vignette	Question
Standard	False-Belief	Protagonist	Outside, there are Sally, Anne, a toy, a box, and a bucket. Sally and Anne are playing together. Sally places the toy in the box and then goes inside to get a drink of water. While Sally is inside, Anne takes the toy and places it into the bucket.	When Sally comes back outside, where will she look for the toy?
Second-order	False-Belief	Second character	Outside, there are Sally, Anne, a toy, a box, and a bucket. Sally and Anne are playing together. Sally places the toy in the box and then goes inside to get a drink of water. While Sally is inside, Anne takes the toy and places it into the bucket.	When Sally comes back outside, where will Anne expect her to look for the toy?
Transparent containers	False-Belief	Protagonist	Outside, there are Sally, Anne, a toy, a transparent box, and a see through bucket. Sally and Anne are playing together. Sally places the toy in the box and then goes inside to get a drink of water. While Sally is inside, Anne takes the toy and places it into the bucket.	When Sally comes back outside, where will she look for the toy?
Preposition change	False-Belief	Protagonist	Outside, there are Sally, Anne, a toy, a box, and a bucket. Sally and Anne are playing together. Sally places the toy onto the box and then goes inside to get a drink of water. While Sally is inside, Anne takes the toy off of the box and places it onto the bucket.	When Sally comes back outside, where will she look for the toy?

Explicit & trusted communication	False-Belief	Protagonist	Outside, there are Sally, Anne, a toy, a box, and a bucket. Sally and Anne are playing together. Sally places the toy in the box and then goes inside to get a drink of water. While Sally is inside, Anne calls to Sally to say that she is going to move the toy to the bucket. Sally believes Anne. Anne reaches into the box and moves the toy into the bucket.	When Sally comes back outside, where will she look for the toy?
Absence & perspective shift	True-Belief	Second character	Outside, there are Sally, Anne, a toy, a box, and a bucket. Sally and Anne are playing together. Sally places the toy in the box and then goes inside to get a drink of water. While Sally is outside, Anne reaches in the box and moves the toy into the bucket. Anne leaves and goes home. Sally and Anne come back the next day. They don't know what happened while they were away.	When Anne comes back outside, where will she look for the toy?
Additional transferrable item	True-Belief	Protagonist	Outside, there are Sally, Anne, a toy, a Lego piece, a box, and a bucket. Sally and Anne are playing together. Sally places the toy in the box and then goes inside to get a drink of water. While Sally is inside, Anne places the Lego piece into the box.	When Sally comes back outside, where will she look for the toy?

Procedure

Individual instances of *temporary* or *private* chats were created for each prompt. Each prompt was composed of a single vignette and its corresponding question (see Figure 1). Since the tested

LLMs had no memory of each prompt, the order in which the items were tested should have had no impact on their performance. However, pseudo-randomization was used to prevent the same base vignette from being used in back-to-back trials to provide assurance that recent interaction with a particular narrative would not bias behavior. Following each prompt, the provided answers were graded in a binary fashion. Answers were judged to be accurate *only* when the correct belief-state was identified (e.g., *Sally will look for the toy in the box*).

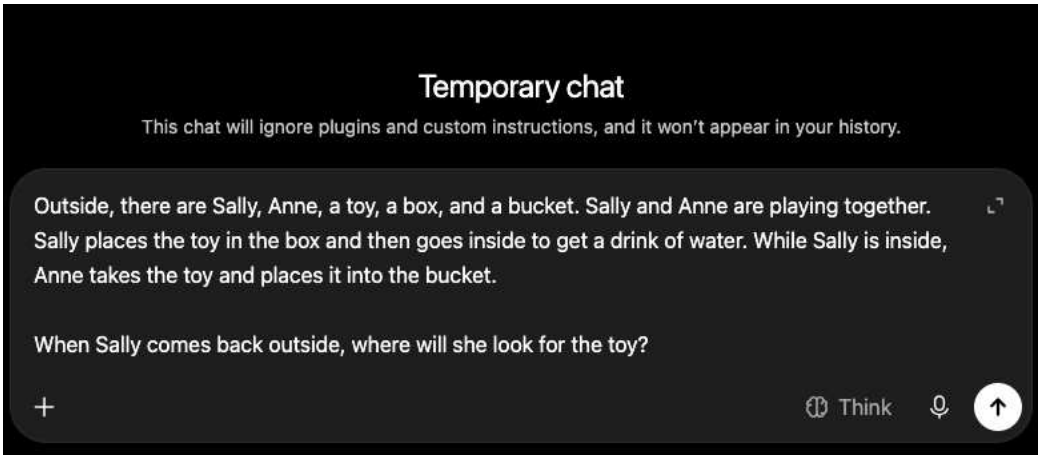

Figure 1 – A prompt displaying the standardized form of presenting a vignette and its corresponding question to the LLM subject. The figure above displays GPT-5.5 Luna in Temporary chat mode.

Results

Cumulatively, each of the models scored above chance when performance was averaged across all conditions. The top performing model was Gemini 3.1 Pro, which provided the correct answer 84% of the time, followed by Grok 4.5 at 79%, Claude Sonnet 5 at 77%, GPT-5.5 Luna at 73%, and Mistral Medium 3.5 at 68% (see Table 3). The two conditions where performance was at ceiling were the *Standard* (false-belief) and *Additional Transferrable Item* (true-belief) conditions. The result from these conditions aligned with the expectation that the most prevalent

and publicly available forms of this task would be the easiest for LLMs to answer correctly. As others have noted, this is likely a consequence of these types of vignettes being well represented in the training data (Kosinski, 2024). After all, false-belief tasks have been well documented in scholarship over the past forty-years (e.g., Wimmer & Perner, 1983) and as a ToM test, it has seen numerous iterations (for a review, see Wellman et al., 2001). However, shortcomings in the ability of LLMs to *consistently* and *accurately* identify mental states persisted in each of the LLMs.

Table 3

A summary of the performance for each of the tested LLMs across the seven conditions, presented alphabetically

Condition - Task Type	Average	Claude Sonnet 5	Gemini 3.1 Pro	GPT-5.5 Luna	Grok 4.5	Mistral Medium 3.5
Standard - FB	100%	100%	100%	100%	100%	100%
2nd Order - FB	92%	100%	100%	62%	100%	100%
Transparent - FB	47%	50%	87%	50%	50%	0%
Preposition - FB	0%	0%	0%	0%	0%	0%
Explicit Comm - FB	95%	87%	100%	100%	100%	87%
Absent - TB	97%	100%	100%	100%	100%	87%
Additional Item - TB	100%	100%	100%	100%	100%	100%
	76%	**77%**	**84%**	**73%**	**79%**	**68%**

Note. FB = false belief; TB = true belief.

Each of the models exhibited high performance across the *Second-order, Explicit & trusted communication*, *Absence & perspective shift* conditions. However, there were exceptions. For example, GPT-5.5 Luna struggled with the *Second-order* condition which matched the *Standard* vignette but posed a 2nd-order belief question (*When Sally comes back outside, where will Anne expect her to look for the toy?*). Each of the other LLMs, aside from GPT-5.5 Luna, correctly ascertained that *Anne would expect Sally to look in the initial location* as a product of her

ignorance. Intriguingly, earlier models (e.g., GPT-4) were found to match the ability of human adults in being able to assess higher-order ToM (up to the 5th-order) and even exceeded human ability on 6th-order ToM tasks (Street et al., 2025). A hint as to why GPT-5.5 Luna performed more poorly in identifying 2nd-order beliefs may simply be a function of overgeneralization (see, Peters & Chin-Yee, 2025). For instance, a common finding was that GPT-5.5 Luna responded by noting that this experimental item exemplified a classical false-belief paradigm. This may have signaled that GPT-5.5 Luna recognized the *Second-order* condition as being a structural clone of our *Standard* condition. As a result, this could have diverted this model from dynamically tracking mental states. However, there is evidence suggesting that gender assignment may have played a role in these inaccuracies. The three trials which GPT-5.5 Luna answered incorrectly were strictly in mixed-gender assignments (F-M, M-F). As we explore gender assignments in more depth (in the following section), the relevance of this finding will become more apparent.

In the *Explicit & trusted communication* condition, the fact that there was explicit communication (along with an explicit indicator that *Sally* trusted *Anne's* word; *Sally believes Anne)*[2] which still led to mistakes on behalf of Claude Sonnet 5 and Mistral Medium 3.5, signals that human-like inferences cannot always be taken for granted with LLMs (Shapira et al., 2023). Likewise, in the *Absence & perspective shift* condition the ambiguity of both characters leaving the scene, followed by a true-belief question directed at the second character (*Anne*, rather than *Sally*) caused Mistral Medium 3.5 to commit further errors. For human subjects the provided social cues, namely, *explicit communication* and *trust* are straightforward indicators that Sally's and Anne's beliefs would have aligned without any impeding factors made apparent. Yet, for

[2] It is also unclear whether LLMs have developed analogous mechanisms to humans, underpinning their ability to judge the reliability of others as informational sources (see Sperber et al., 2010; Mazzarella & Pouscoulous, 2021).

both Claude Sonnet 5 and Mistral Medium 3.5, this inference was not made on a consistent basis. This aligns with the notion that LLMs are "substantially more susceptible to bias based on their prior knowledge," than human subjects when encountering linguistic uncertainty (Belém et al., 2024). Notably, even when LLMs have an underlying magnitude of uncertainty they do not represent this in the outputs which they provide (Steyvers & Peters, 2026), thus the only way in which these uncertainties are revealed, in these kinds of tests, is through repeated trials.

The two conditions which proved most difficult were the *Transparent containers* condition, where only Gemini 3.1 Pro performed above chance with an 87% accuracy rate, and the *Preposition change* condition, which displayed a floor effect (0% accuracy rate) for all models (see Figure 2).

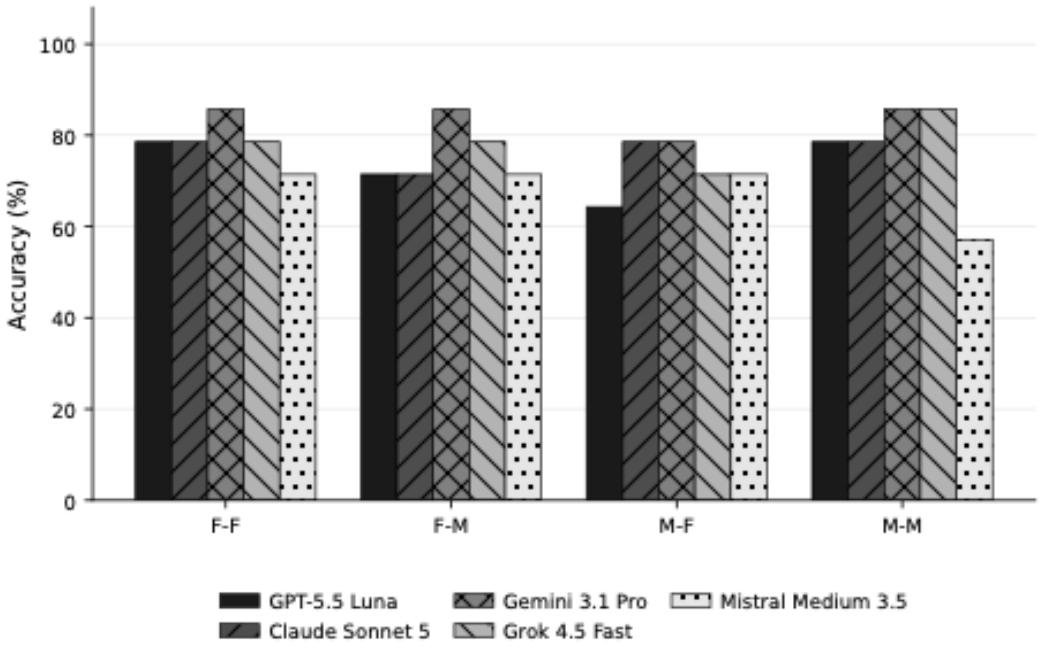

Figure 2 – Model accuracy across each of the seven conditions.
Note. FB = false belief; TB = true belief.

Previously, it had been wondered whether the former result had been a consequence of the attribute of *transparency* not being sufficiently understood. However, this interpretation has been explicitly refuted given that when *transparency* is made more explicit by adding that the protagonist had specifically looked at the container, earlier models (GPT-3.5 and GPT-4) improved in their performance (Pi et al., 2024). However, they still did not perform above chance. Likewise for the *Preposition change* condition the inference that a protagonist would immediately recognize the visible item, doesn't appear to be generated at all. The simple contrast between *into/onto* and *underneath/on top of* led to each LLM failing in this condition when the latter alternatives were featured in the vignettes. The inability to interpret these vignettes brings further attention to the insufficiencies of ToM in LLMs, but there is precedent to expect this kind of result. LLMs are token-based systems that use mathematical algorithms to generate each successive word provided in their outputs (for a review, see Naveed et al., 2025). Integrating causality (Zečević et al., 2023) remains difficult for LLMs because their training data is composed of textual data and does not provide them with a holistic account of physical reality.

Gender

Across each gender pairing assignment (F-F, F-M, M-F, M-M), accuracy ranged from 57.14% to 85.71% (see Figure 3). Gemini 3.1 Pro was the top performing model, recording the highest accuracy rate at 83.93% and Mistral Medium 3.5 recorded the worst overall performance at 67.86% throughout each gender assignment.[3]

[3] In the M-F assignment, Claude Sonnet 5 matched Gemini's performance and in the M-M assignment, Grok 4.5 Fast matched Gemini's performance.

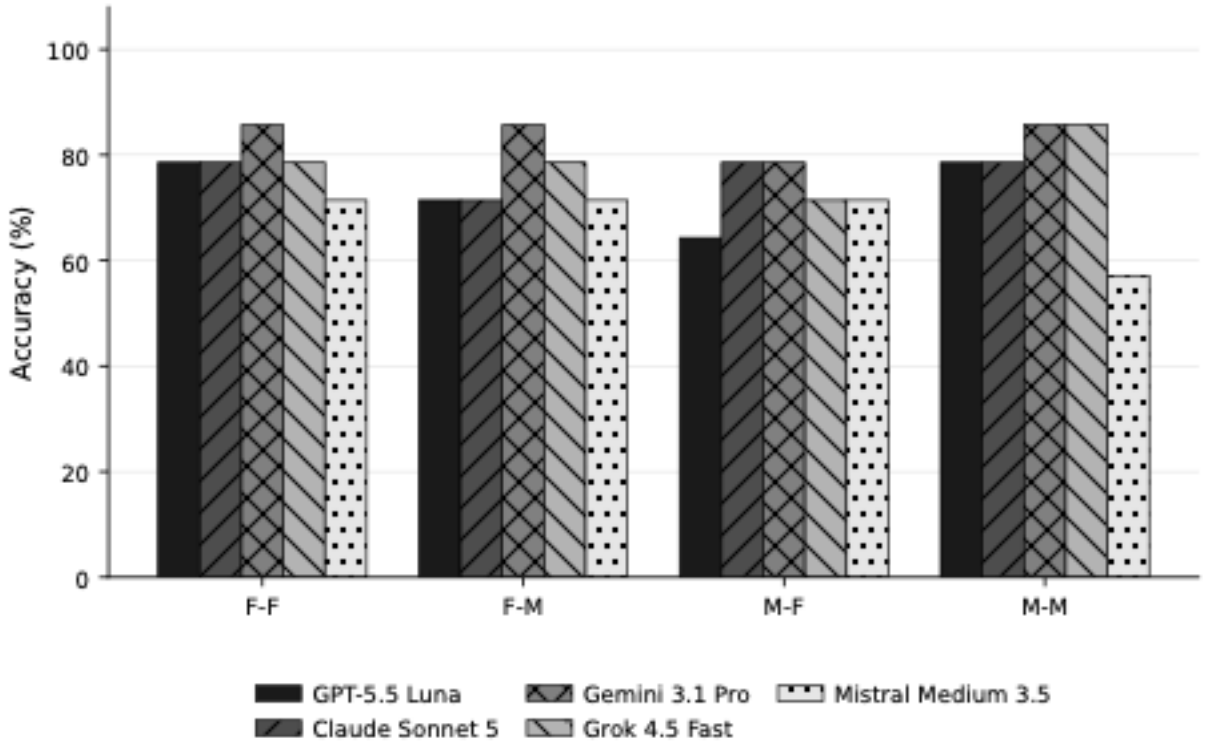

Figure 3 – Model accuracy across each of the four different gender assignments. The first assignment is the gender of the protagonist.

Note. F = female; M = male.

Generally, each model exhibited different proclivities in their ability to cope with gender, nonetheless there were emerging trends. Three out of the five models (GPT-5.5 Luna, Gemini 3.1 Pro, Grok 4.5 Fast) worst performance were in M-F assignments while Claude Sonnet 5 performed worst in F-M assignments, and Mistral Medium 3.5 in M-M assignments. Furthermore, four out of the five models (excluding Mistral Medium 3.5), performed better when given matching gender assignments (F-F or M-M). GPT-5.5 Luna exhibited the most sensitivity to this effect, performing 10.71% better in matching gender assignments in comparison to mixed-gender ones. This pattern suggests the incursion of an additional processing cost for most models when there are multiple genders to keep track of. Intuitively, this may seem a bit strange. In

mixed-gender assignments, reference resolution should be relatively easy to enact given that pronouns such as, *she* and *he*, would only correspond to a single character. Whereas in matching gender assignments, the two characters share a single pronoun, which should result in an increased level of ambiguity. Rather paradoxically, the results indicate that matching gender assignments were associated with better task performance. One possible explanation for this phenomenon is that matching gender assignments may cause LLMs to adopt a different reasoning strategy altogether. For example, when gender can no longer be used as a shortcut to individuate between the characters in our vignettes, LLMs may instead shift to relying on the structural roles of the characters (e.g., *who moved the item where* or *who had left the scene*), which may (inadvertently) facilitate a greater success rate in the Unexpected Transfer task.

Prepositions

Across each preposition assignment (*Into/Onto, Underneath/On Top Of*), accuracy ranged from 0% to 100% (see Figure 4). There was a clear bimodal distribution displayed between the conditions where the item was hidden inside of or beneath its 'container,' and those where the transferrable item was left in plain sight (*Preposition change*). For the prepositions of *Into* and *Onto*, this led an accuracy rate of 89.17% and 0%, respectively. For the prepositions of *Underneath* and *On Top Of*, this led to an accuracy rate of 88.33% and 0%. Gemini 3.1 Pro was, once again, the top performing model, recording the highest accuracy rate outside of the *Preposition change* condition at 97.92% and Mistral Medium 3.5, again, recorded the worst performance at 79.17% accuracy.

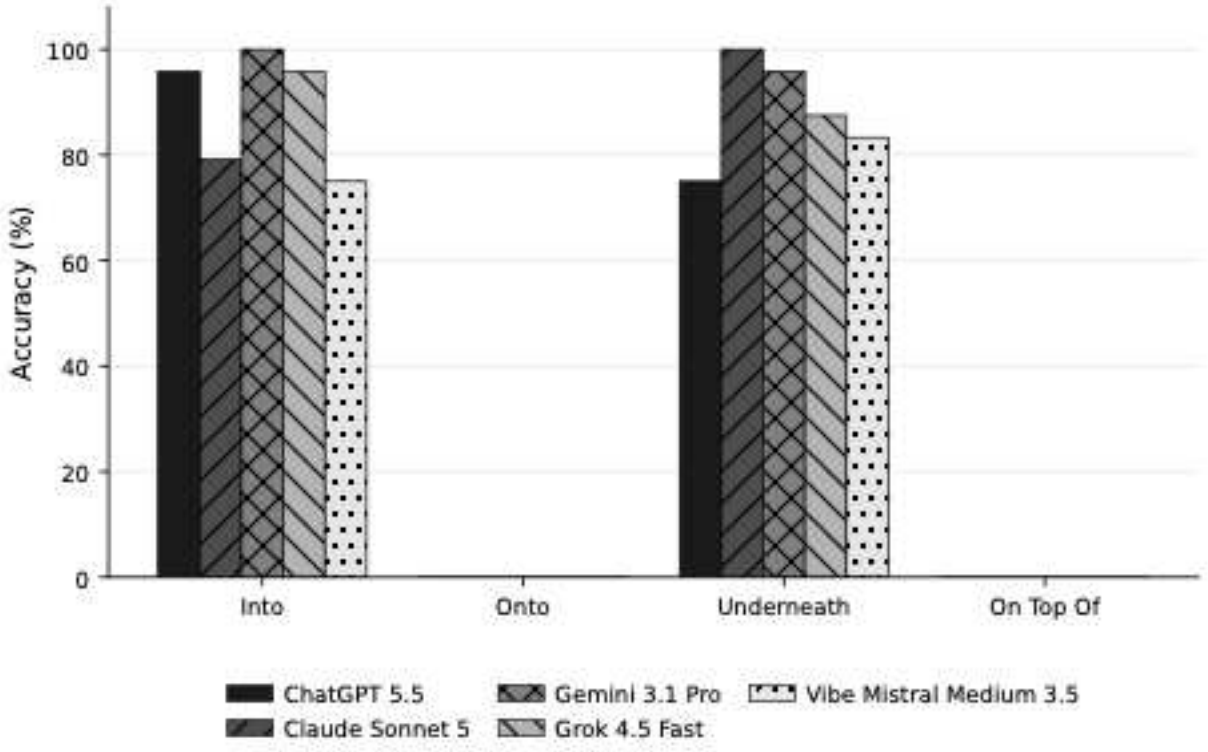

Figure 4 – Model accuracy across each of the four different prepositions assignments, in order of appearance in the conditions.

Note. For each LLM, *Into/Underneath* appeared in 48 trials and *Onto/On Top Of* appeared in 8 trials.

The pivotal finding is that each LLM failed to recognize that when an item is left in plain view (by being left on the surface of a 'container'), this should disrupt the standard logic of the Unexpected Transfer task. The inference we are expecting LLMs to make in this condition is that visual access should permit the protagonist to update their belief state about the item's true location. So, what is behind this critical failure? The fact that the models performed well above chance with the prepositions of *Into* and *Underneath* tells us that this failure is unlikely to be a lexical issue. Each of the prepositions selected indicates a basic spatial relation and is likely to be well-represented in these models' training data. The uniform failure of these models suggests a

shared underlying limitation in identifying the connection between perceptual information (i.e., described visual input) and belief attribution. Notably, the regular failures in the *Transparent containers* condition also serve to support this interpretation.

Discussion

The five LLMs which we tested (GPT-5.5 Luna, Claude Sonnet 5, Gemini 3.1 Pro, Grok 4.5, Mistral Medium 3.5) displayed an inconsistent ability to *accurately* attribute mental states. While each model logged idiosyncratic performances, there were common failures among each of models. Most significantly, only one model (Gemini 3.1 Pro) performed above chance in the *Transparent containers* condition and floor effects were recorded in the *Preposition change* condition. The commonality between these conditions is that they disrupt the standard logic of the Unexpected Transfer task; they integrate described visual access as a relevant factor for consideration. The expectation is for LLMs to identify the connection between described perceptual information and mental state attribution. In other words, a casual relation needs to be inferred: if the protagonist can see the item (via transparency or direct visual access), they will not need to hold a *belief* about the item's location, since their direct perception will allow them to skip over this step due to having direct *knowledge* about the item's precise location. Our study exemplifies how LLMs struggle to follow this chain of reasoning.

What would aid in the resolution of these conditions is if these models had a better grasp on what it means to interface with an embodied reality and common-sense knowledge about the physical world (Pi et al., 2024; Shapira et al., 2023). By definition, all of the "knowledge" an LLM has is derived from written text, which effectively undermines their ability to succeed in

these conditions. It has been suggested that LLMs may be encountering a "reasoning bottleneck," upon being prompted with tasks that require internal state tracking (Wang et al., 2026). This may be a fundamental problem for LLMs, given that their autoregressive functionality lacks a mechanism dedicated to keeping track of shifting variables, leading to ToM tests as being particularly difficult to pass. However, it is important to consider that no task is without its limitations.

Task Limitations

The Unexpected Transfer task poses a highly specific and presumptuous abstraction of real-world interaction. It is a task which calls for the subject to identify a mental state that does not functionally exist. This is because it uses fictional characters that lack genuine mental states. Thus, the "correct" inference that is expected from subjects is only assessed as *accurate* because it is predetermined by the task administrators as the justifiable answer. Simulated interaction which takes the form of written text will always be absent of actual minds and mental states. This is simply a trade-off that is accepted in classical false-belief tasks, where the lack of ground truth is understood as an operational necessity (Conway et al., 2025). Consequently, what this task actually tests is whether a subject can align their mental state attribution within a predefined norm or expectation. Since neither fictional character, *Sally* nor *Anne*, have belief states, the social inferences LLMs are being asked to make cannot strictly be defined as *true* or *false*. Yet, in each derivation of the false-belief task this is precisely what is enacted. This results in a superficial understanding about how LLMs are able to utilize ToM.

The embodiment problem

A fundamental difference between the ToM capabilities of humans and LLMs is that humans are well-versed in integrating information from multiple sources. In other words, multimodal cues—such as gaze, spatial relations, prosody, gestures, facial expressions, and other nonverbal signals—are all relevant sources of information that are commonly used to form social inferences (Meltzoff, 1999; Sebanz et al., 2005). As humans, we perceive reality through our senses and heavily rely on our sight. Naturally, our embodied experience permits us indulge in a certain situatedness that we cannot help but take for granted. Therefore forming an inference that derives from sight (i.e., or described visual access, in the case of LLMs) may feel like a rudimentary undertaking. LLMs, on the other hand, are limited to a single input domain: written text. This means social cues are inherently harder to come by given that they must be inferred from static blocks of text as opposed to sensory information. Furthermore, the "black-box" nature of industry-leading LLMs (Zečević et al., 2023) prevents us from knowing the extent to which these very cues are acknowledged, recognized, or even come into play during reasoning.

General Discussion

While our findings suggest that LLMs have improved in their ability to identify mental states in conventionalized representations of the Unexpected Transfer task, it is difficult to determine whether this reflects genuine progress *or not*. This is because when a study which demonstrates a task failure is published in the scientific literature (e.g., Ullman, 2023), future generations of LLMs are likely to integrate this information into their training data (Hu et al., 2025; Jacovi et

al., 2023). This makes it difficult to discern whether improvements merely reflect a "training away" effect or if the underlying computations needed to improve the ToM capabilities of LLMs have truly been addressed. In the present study, what we found is that LLMs struggled with incorporating one type of cue in particular: described visual access.

In both the *Transparent containers* and *Preposition change* conditions, the LLMs which we tested, proved incapable[4] of *consistently* understanding that described visual access should catalyze a straightforward chain of reasoning: If the protagonist has direct visual access to where the item is located, then this not only means that their *belief* will align with its true location, but they will have verifiable *knowledge* of where it has been place. These conditions illustrate how integrating described visual perception into both causal and pragmatic reasoning remains a hurdle for LLMs to overcome. It also begs the question of whether LLMs are sensitive to other kinds of perceptual information (e.g., *sound* or *touch*) and if this information would prove equally as difficult to integrate into social reasoning.

When considering the growing role of LLMs in society, determining what their strengths and weaknesses are as social partners should be considered a pressing matter. While there is a growing tendency to attribute human-like understanding to LLMs, tasks like this illustrate divergent reasoning strategies. Notably, a deliberate element of their design is to convince their users that LLMs are intelligent interlocutors just like humans (Shanahan et al., 2023). Their names (e.g., Claude) and friendly bylines (*How can I help you today?*) serve to blur the distinction between human and AI interaction (Gabriel et al., 2025). As a unified experience, LLMs often give the impression that they are intelligent entities. However, if the present paper

[4] The single exception was Gemini 3.1 Pro in the *Transparent containers* condition, which provided the correct answer 87% of the time.

exhibits anything, it is that this impression should be treated with caution. These same models which can produce outputs which appear to exhibit genuine understanding, consistently fail to make “common sense” inferences. What this illustrates is how much real-world experience, often left outside the domain of written text (Browning & LeCun, 2022), defines how we *think* and *reason* about the minds of others.

Anthropic. (2026). *Claude Sonnet 5* [Large language model]. Anthropic model announcement.

Apperly, I. (2010). *Mindreaders: The cognitive basis of "theory of mind."* Psychology Press.

Baron-Cohen, S. (2000). Theory of mind and autism: A fifteen year review. In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience* (2nd ed., pp. 3–20). Oxford University Press.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21*(1), 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

Belém, C. G., Kelly, M., Steyvers, M., Singh, S., & Smyth, P. (2024). Perceptions of linguistic uncertainty by language models and humans. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 8467–8502). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.483

Bradford, E. E., Jentzsch, I., & Gomez, J. C. (2015). From self to social cognition: Theory of mind mechanisms and their relation to executive functioning. *Cognition, 138*, 21–34.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Browning, J., & LeCun, Y. (2022). What AI can tell us about intelligence. *Noema*. https://www.noemamag.com/what-ai-can-tell-us-about-intelligence

Conway, J. R., Long, E. L., Sevi, L., Catmur, C., & Bird, G. (2025). Theoretical limitations on mindreading measures: Commentary on Wendt et al. (2024).

Decety, J., Jackson, P. L., Sommerville, J. A., Chaminade, T., & Meltzoff, A. N. (2004). The neural bases of cooperation and competition: An fMRI investigation. *NeuroImage, 23*(2), 744–751. https://doi.org/10.1016/j.neuroimage.2004.05.025

Decety, J., & Sommerville, J. A. (2003). Shared representations between self and other: A social cognitive neuroscience view. *Trends in Cognitive Sciences, 7*(12), 527–533. https://doi.org/10.1016/j.tics.2003.10.004

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology, 87*(3), 327–339. https://doi.org/10.1037/0022-3514.87.3.327

Gabriel, I., Keeling, G., Manzini, A., & Evans, J. (2025). We need a new ethics for a world of AI agents. *Nature, 644*(8075), 38–40.

Goldstein, T. R., & Winner, E. (2012). Enhancing empathy and theory of mind. *Journal of Cognition and Development, 13*(1), 19–37.

Google DeepMind. (2026). *Gemini 3.1 Pro* [Large language model]. Gemini 3.1 Pro model card.

Happé, F. (2003). Theory of mind and the self. *Annals of the New York Academy of Sciences, 1001*(1), 134–144. https://doi.org/10.1196/annals.1279.008

Happé, F., & Loth, E. (2002). 'Theory of mind' and tracking speakers' intentions. *Mind & Language, 17*(1–2), 24–36. https://doi.org/10.1111/1468-0017.00187

Hu, J., Sosa, F., & Ullman, T. (2025). Re-evaluating theory of mind evaluation in large language models. *Philosophical Transactions of the Royal Society B, 380*(1932), Article 20230499. https://doi.org/10.1098/rstb.2023.0499

Jacovi, A., Caciularu, A., Goldman, O., & Goldberg, Y. (2023). Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 5075–5084). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2305.10160

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. (2024, May). SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations* (Vol. 2024, pp. 54107–54157).

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition, 89*(1), 25–41. https://doi.org/10.1016/S0010-0277(03)00064-7

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences, 121*(45), Article e2405460121. https://doi.org/10.1073/pnas.2405460121

Lecce, S., Ronchi, L., Del Sette, P., Bischetti, L., & Bambini, V. (2019). Interpreting physical and mental metaphors: Is theory of mind associated with pragmatics in middle childhood? *Journal of Child Language, 46*(2), 393–407.

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in 'theory of mind.' *Trends in Cognitive Sciences, 8*(12), 528–533. https://doi.org/10.1016/j.tics.2004.10.001

Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (2022, May). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*

(Vol. 1, pp. 8086–8098). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.556

Mathematical Association of America. (2025). *American Invitational Mathematics Examination (AIME) 2025.* Mathematical Association of America.

Meltzoff, A. N. (1999). Origins of theory of mind, cognition and communication. *Journal of Communication Disorders, 32*(4), 251–269. https://doi.org/10.1016/S0021-9924(99)00009-X

Mistral AI. (2026). *Mistral Medium 3.5* [Large language model]. Mistral Medium 3.5 model card.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., … Mian, A. (2025). A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology, 16*(5), 1–72.

OpenAI. (2026). *GPT-5.5 Luna* [Large language model]. OpenAI API model documentation.

Peters, U., & Chin-Yee, B. (2025). Generalization bias in large language model summarization of scientific research. *Royal Society Open Science, 12*(4), Article 241776. https://doi.org/10.1098/rsos.241776

Peterson, C., Slaughter, V., Moore, C., & Wellman, H. M. (2016). Peer social skills and theory of mind in children with autism, deafness, or typical development. *Developmental Psychology, 52*(1), 46–57. https://doi.org/10.1037/a0039833

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., … Wykowski, J. (2025). *Humanity's last exam* (arXiv:2501.14249). arXiv. https://doi.org/10.48550/arXiv.2501.14249

Pi, Z., Vadaparty, A., Bergen, B. K., & Jones, C. R. (2024). *Dissecting the Ullman variations with a SCALPEL: Why do LLMs fail at trivial alterations to the false belief task?* (arXiv:2406.14737). arXiv. https://doi.org/10.48550/arXiv.2406.14737

Rakoczy, H. (2022). Foundations of theory of mind and its development in early childhood. *Nature Reviews Psychology, 1*(4), 223–235. https://doi.org/10.1038/s44159-022-00037-z

Sebanz, N., Knoblich, G., & Prinz, W. (2005). How two share a task: Corepresenting stimulus–response mappings. *Journal of Experimental Psychology: Human Perception and Performance, 31*(6), 1234–1246. https://doi.org/10.1037/0096-1523.31.6.1234

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature, 623*(7987), 493–498. https://doi.org/10.1038/s41586-023-06647-8

Shapira, N., Zwirn, G., & Goldberg, Y. (2023, July). How well do large language models perform on faux pas tests? In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 10438–10451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.663

Slaughter, V., Imuta, K., Peterson, C. C., & Henry, J. D. (2015). Meta-analysis of theory of mind and peer popularity in the preschool and early school years. *Child Development, 86*(4), 1159–1174. https://doi.org/10.1111/cdev.12372

SpaceXAI. (2026). *Grok 4.5* [Large language model]. Grok 4.5 announcement.

Steyvers, M., & Peters, M. A. (2026). Metacognition and uncertainty communication in humans and large language models. *Current Directions in Psychological Science, 35*(3), 131–139.

Strachan, J. W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., … Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour, 8*(7), 1285–1295. https://doi.org/10.1038/s41562-024-01882-z

Street, W., Siy, J. O., Keeling, G., Baranes, A., Barnett, B., McKibben, M., … Dunbar, R. I. (2025). LLMs achieve adult human performance on higher-order theory of mind tasks. *Frontiers in Human Neuroscience, 19*, Article 1633272. https://doi.org/10.3389/fnhum.2025.1633272

Surian, L., & Leslie, A. M. (1999). Competence and performance in false belief understanding: A comparison of autistic and normal 3-year-old children. *British Journal of Developmental Psychology, 17*(1), 141–155. https://doi.org/10.1348/026151099165203

Ullman, T. (2023). *Large language models fail on trivial alterations to theory-of-mind tasks* (arXiv:2302.08399). arXiv. https://doi.org/10.48550/arXiv.2302.08399

Wang, H., Song, Y., Li, T., Deng, Z., Wang, Y., Ramachandran, D., … Roth, D. (2026). *Formalize, don't optimize: The heuristic trap in LLM-generated combinatorial solvers* (arXiv:2605.12421). arXiv. https://doi.org/10.48550/arXiv.2605.12421

Wellman, H. M. (2014). *Making minds: How theory of mind develops.* Oxford University Press.

Wellman, H. M. (2018). Theory of mind: The state of the art. *European Journal of Developmental Psychology, 15*(6), 728–755. https://doi.org/10.1080/17405629.2018.1435413

Wellman, H. M., Carey, S., Gleitman, L., Newport, E. L., & Spelke, E. S. (1990). *The child's theory of mind.* The MIT Press.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*(3), 655–684. https://doi.org/10.1111/1467-8624.00304

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development, 75*(2), 523–541. https://doi.org/10.1111/j.1467-8624.2004.00691.x

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103–128. https://doi.org/10.1016/0010-0277(83)90004-5

Zečević, M., Willig, M., Dhami, D. S., & Kersting, K. (2023). *Causal parrots: Large language models may talk causality but are not causal* (arXiv:2308.13067). arXiv. https://doi.org/10.48550/arXiv.2308.13067