Rethinking Theory of Mind Benchmarks for LLMs: Towards A User-Centered Perspective

Qiaosi Wang, **Xuhui Zhou**, **Maarten Sap**, **Jodi Forlizzi**, **Hong Shen**
Carnegie Mellon University
{qiaosiw, xuhuiz, msap2, forlizzi, hongs}@andrew.cmu.edu

Abstract

The last couple of years have witnessed emerging research that appropriates Theory-of-Mind (ToM) tasks designed for humans to benchmark LLM's ToM capabilities as an indication of LLM's social intelligence. However, this approach has a number of limitations. Drawing on existing psychology and AI literature, we summarize the theoretical, methodological, and evaluation limitations by pointing out that certain issues are inherently present in the original ToM tasks used to evaluate human's ToM, which continues to persist and exacerbated when appropriated to benchmark LLM's ToM. Taking a human-computer interaction (HCI) perspective, these limitations prompt us to rethink the definition and criteria of ToM in ToM benchmarks in a more dynamic, interactional approach that accounts for user preferences, needs, and experiences with LLMs in such evaluations. We conclude by outlining potential opportunities and challenges towards this direction.

1 Introduction

In recent years, Theory of Mind (ToM) has gained much attention in the evaluation and benchmarks of Large Language Models (LLMs) due to its fundamental role in social cognition. ToM is the human social and cognitive capability of attributing mental states (e.g., knowledge, intentions, desire, emotions) to ourselves and others based on observable behavioral and verbal cues, with the goal of predicting and making sense of others' actions [Baron-Cohen *et al.*, 1985; Baron-Cohen, 1999; Premack and Woodruff, 1978]. Many human social behaviors are enabled by ToM, such as persuasion, teaching, repairing communication breakdowns, building shared plans and goals [Baron-Cohen, 1999], all of which requires us to make conjectures about what's going on in others' minds (e.g., their intentions, knowledge, preference, motivations) to behave accordingly and achieve optimal social interaction outcomes. Given its fundamental role in human social interaction, ToM has been studied extensively across various disciplines especially in developmental and clinical psychology, where researchers have studied the emergence and development of ToM in children as well as the role of ToM in people with autism or schizophrenia who tend to experience difficulty in social interactions with neurotypical people[1] [Milton, 2012; Wellman, 2018; Rakoczy, 2022; Baron-Cohen, 2000]. Throughout these research endeavors, researchers have come up with a number of tasks to assess people's ToM capability. One of the most famous ToM tasks is perhaps the Sally-Anne test (as illustrated in Fig. 1), which presents a scenario to assess the children's understanding of false beliefs, an important indication of the child's ToM ability in recognizing others can have beliefs that the child know to be false.

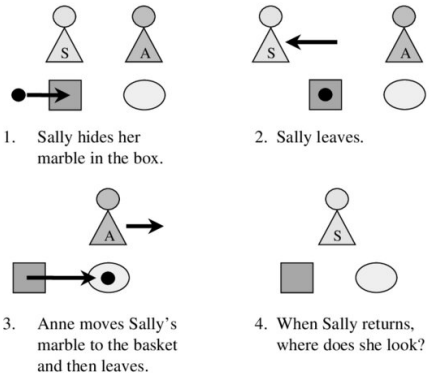

Figure 1: An illustration of the Sally-Anne test commonly used to evaluate children's Theory of Mind. Figure reproduced from Scassellati, 2001.

Recently, such ToM tasks have been appropriated as benchmarks to evaluate LLM's ToM capability, with the goal of assessing their social cognitive abilities. Some studies have directly applied these human-intended tasks to LLMs and drawn bold claims based on model performance on these ToM tasks [Kosinski, 2023; Bubeck *et al.*, 2023]. For example, Kosinski [2023] claimed that "ToM may have spontaneously emerged in LLMs" after models passed over 70% of false belief tasks. Similarly, Bubeck *et al.* [2023] concluded that GPT-4 demonstrates "a very advanced level of ToM" based on its superior performance in false belief, emotion recognition, and intention inference tasks compared to other models. These claims have sparked lively debate within

[1] see the Double Empathy problem

the AI community. In response, there is growing recognition that evaluating ToM in LLMs requires benchmarks grounded in NLP evaluation methods and tailored to the unique affordances and limitations of LLMs, rather than repurposing cognitive assessments designed for humans. Following this notion, many work has created variations based on the existing ToM tasks' structure and content to benchmark LLM's ToM capability (e.g., [Xu *et al.*, 2024; Kim *et al.*, 2023]). Others have pointed out the concerning robustness of drawing claims based on LLM passing human-intended ToM tasks by demonstrating LLM failures when these tasks were modified with trivial alterations (e.g., [Ullman, 2023; Shapira *et al.*, 2023a; Sap *et al.*, 2022]).

Responding to the growing call for examining the limitations of appropriating human-intended ToM tasks to benchmark LLM's ToM [Shapira *et al.*, 2023a; Sap *et al.*, 2022; Ullman, 2023], this position paper aims to surface the number of limitations embedded in the original human-intended ToM tasks— limitations that not only persisted but are amplified when these tasks are repurposed to evaluate LLMs. These inherited limitations cast doubt on the validity of conclusions drawn about LLM's ToM and social capability based on this type of evaluations [Shapira *et al.*, 2023a]. In this paper, we first provide an overview of the various ToM tasks used to evaluate humans and LLMs based on existing work. Drawing from both psychology and AI literature, we summarize the theoretical, methodological, and evaluation limitations of appropriating ToM tasks as benchmarks for LLM's ToM capabilities. Building on this foundation, we take an HCI perspective to rethink why, what, and how we benchmark LLMs' ToM capabilities, highlighting potential opportunities and challenges for designing user-centered ToM evaluations.

2 Evaluating Theory of Mind in Humans and Large Language Models

2.1 Assessing Human ToM through ToM Tasks

While Premack and Woodruff [1978] did not specify what ToM encompasses when they coined the term, decades of psychology literature has established ToM as a multi-faceted construct that includes various cognitive and affective dimensions. Fu *et al.*[2023] distillled four construct dimensions of ToM from a systematic literature review of 127 ToM measures: cognitive-interpersonal, cognitive-intrapersonal, affective-interpersonal, and affective-intrapersonal. Similarly, Beaudoin *et al.* [2020] identified 220 ToM measures used to evaluate children's ToM and pinpointed seven dimensions of ToM: emotions, desires, intentions, percepts, knowledge, beliefs, and mentalistic understanding of non-literal communication. These dimensions are further divided into 39 types of ToM sub-abilities. However, given the broad definition of ToM as the ability to attribute mental states to self and others, researchers have pointed out the nonspecificity of ToM, which can be used to simultaneously include different cognitive constructs such as emotional reactivity and facial expression categorization [Quesque and Rossetti, 2020]. This has led to construct validity issues of certain ToM tasks not actually measuring ToM (e.g., Reading the Mind in the Eyes Test).

Over the years, hundreds of ToM tasks have been proposed and used to evaluate human's ToM capability. In developmental psychology, ToM tasks have been used to identify developmental milestones as well as understanding social deficits in children with autism spectrum disorders [Beaudoin *et al.*, 2020; Baron-Cohen *et al.*, 1985]. While ToM tasks can be administered in varying presentation modes, a typical ToM task often comprises of a social scenario in the form of a story, a comic, or even a video, followed by questions (typically multiple-choice) to the child about the mental state of the characters in the social scenario [Beaudoin *et al.*, 2020; Fu *et al.*, 2023]. Besides the classic Sally-Anne false belief task (shown in Fig. 1), other classic ToM tasks include faux pas, strange stories, second-order false belief, and more [Hayward and Homer, 2017]. Specifically, the faux pas task presents the child with a story, where one character makes a social mistake (e.g., saying that the dish cooked by the dinner host is not good in front of the host), and the child was asked if the character's behavior is appropriate and why [Baron-Cohen *et al.*, 1999]. The strange stories task presents the child with short stories of characters acting strangely by pretending, joking, or lying, and the child must infer the mental state of the characters to explain their behavior [Happé, 1993]. These tasks have far-reaching influences as many have been adapted and extended to measure ToM capabilities beyond children, and more recently, measuring LLMs' ToM.

2.2 Current State of ToM Benchmarks for LLMs

Much like how ToM tasks are administered to children, most ToM benchmarks for LLMs follow a similar approach: presenting static, text-based social scenarios and prompting the models to infer the reality and mental states of the characters [Sap *et al.*, 2022]. Model performance is typically scored based on answer accuracy [Shapira *et al.*, 2023a]. These synthetic scenarios are largely appropriated from ToM tasks originally designed for humans and are adopted by several prominent and highly-cited ToM benchmarks. For example, inspired by the Sally-Anne test, Le *et al.*[2019]'s dataset contains over 1000 distinct stories and questions prompting for the character's memory, reality, and false-beliefs. Similarly, Kosinski [2023]'s ToM benchmark contain 40 tasks focusing exclusively on false-belief scenarios. Shapira *et al.* [2023] used human experts and ChatGPT to generate synthetic social scenario stories based on the faux pas test. Chen *et al.* [2024] constructed eight scenarios covering a range of ToM dimensions to build a more comprehensive benchmark. These scenarios are often paired with multiple-choice questions, where only one option is deemed correct regarding the character's mental state or relevant situational details. Such benchmarks are frequently reused or adapted in subsequent studies to evaluate LLMs' ToM capabilities (e.g., [Shapira *et al.*, 2023a; Sap *et al.*, 2022]), and have had a lasting influence on how claims about LLMs' social reasoning are generated.

Going beyond human ToM tasks, other ToM benchmarks also consist of similar format in presenting synthetic social scenarios to LLMs, followed by question-answering to gauge LLM's understanding of the social scenarios. For

example, Sap *et al.*[2019]'s SocialIQA benchmark contains 38,000 multiple-choice questions about the intents and reactions to daily social interaction scenarios created through crowdsourcing. To better align with real-world scenarios, recent work has either adopted or generated natural real-world human-human conversation dataset to evaluate LLM's ToM capability through question-answering (e.g., [Soubki *et al.*, 2024; Chan *et al.*, 2024; Kim *et al.*, 2023])— some work has gone a step further to examine LLM's applications in using mental state inferences to predict and judge observable behaviors [Gu *et al.*, 2024]. While ToM benchmarks containing text-based static social scenarios provided vast convenience and accessibility for researchers to easily assess LLMs' ToM capability, other work has offered new opportunities to assess more dynamic and interactive social interactions in AI. Zhou *et al.*[2023] presented SOTOPIA, an open-ended environment that can simulate and evaluate social interactions between artificial agents across a wide variety of social scenarios. Chiang *et al.*[2024] created Chatbot Arena, an open platform that can involve humans to crowdsource questions to evaluate LLMs. Shi *et al.*[2024] presented a multi-modal ToM benchmark to simulate the embodied, multimodal, multi-agent social interactions by providing video and text descriptions of people's behaviors in realistic household environment.

The growing body of literature on empathy benchmarks for LLMs [Sorin *et al.*, 2024; Huang *et al.*, 2025; Welivita and Pu, 2024; Zhang *et al.*,] offers valuable insights for ToM benchmark design, given that empathy is inherently a sub-construct of ToM—— requiring the understanding and response to others' mental and emotional states. While most empathy benchmarks follow a similar structure to ToM benchmarks by presenting social scenarios to LLMs, many also involve human annotators to provide baseline comparisons for evaluating the empathy expressed in LLM-generated responses. For example, Welivita and Pu [2024] engaged 1,000 participants to compare the empathetic quality of LLM-generated and human-generated responses. Zhu *et al.* [2024] compared LLM's capability in generating empathic inferences about users' underlying goals and needs from product reviews against a baseline established by human designers. Although these benchmarks involve direct human participation, they typically position humans as a baseline for comparing LLM performance, leaving room to explore opportunities in incorporating user perspectives into the evaluation of LLM ToM capability.

3 The Limitations of Appropriating ToM Tasks as LLM Benchmarks

Although appropriating ToM tasks as LLM benchmarks has been a common practice in the AI benchmark literature, it was not until recently that this practice was put under scrutiny for producing rather controversial claims like "ToM has spontaneously emerged in LLMs"(e.g., [Kosinski, 2023; Bubeck *et al.*, 2023]). Outside of AI literature, however, decades of psychology research on the evaluation of human ToM has revealed certain limitations on ToM tasks, even when used to evaluate human ToM. Some of these limitations naturally persisted when ToM tasks are appropriated to measure LLMs' ToM. In this section, we summarize and consolidate the key limitations of appropriating ToM tasks as LLM benchmarks based on existing psychology (e.g., [Beaudoin *et al.*, 2020; Quesque and Rossetti, 2020; Ahmadi *et al.*, 2015]) and AI literature (e.g., [Ullman, 2023; Ma *et al.*, 2023; Shapira *et al.*, 2023a; Sap *et al.*, 2022]) from three aspects: theoretical limitation, methodological limitation, and evaluation limitation.

3.1 Theoretical Limitation

ToM is a multi-faceted construct, yet it has been mostly measured on one dimension. In the Psychology literature, several systematic literature reviews have pointed out the issue of ToM tasks only measuring one dimension, specifically ToM-beliefs via various false-belief tasks, with few providing comprehensive measures [Beaudoin *et al.*, 2020; Fu *et al.*, 2023; Ahmadi *et al.*, 2015]. In their review of ToM tasks used to evaluate young children, Beaudoin *et al.* [2020] pointed out that a whopping 75.5% among the 220 ToM measures they identified focused solely on Beliefs (i.e., informational states that people believe to be true [Ma *et al.*, 2023]), whereas other ToM dimensions such as Emotions (i.e., emotional or affective states that people experience), Desires (i.e., human desires and wants without committed actions), Intentions (i.e., goals and intentions with committed actions) received far less attention (each accounted for 4.3% to 23.9% of the studies identified). Similarly, Fu *et al.*[2023] were only able to identify four out of the 127 ToM measures for children that cover all construct dimensions of ToM.

Given the wide adoption of human-intended ToM tasks to evaluate LLM ToM, this phenomenon was also observed AI literature [Ma *et al.*, 2023; Ying *et al.*, 2025]. Ma *et al.* [2023] surveyed recent LLM ToM literature and observed "an overwhelming research focus on intention and belief aspects of machine ToM" yet other ToM aspects received little attention [Ma *et al.*, 2023]. For example, prominent ToM benchmarks such as Le *et al.* [2019] were largely inspired by the Sally-Anne false belief tasks; Kosinski [2023]'s test set specifically included only variations of false belief tasks; Shapira *et al.* [2023]'s benchmark, while not focusing on ToM-Beliefs, was also inspired by one ToM task, the faux pas test, that measures only one dimension of ToM—the ability to recognize social gaffe situations. While these benchmarks helped offer valuable insights into certain aspects of LLM's ToM capability, it is important to understand that these tasks only measure one dimension of the multi-faceted construct of ToM, and hence *provide limited insights and evidence in making claims about LLM's overall ToM capability* [Ma *et al.*, 2023; Ying *et al.*, 2025].

3.2 Methodological Limitation

Many ToM tasks lack construct validity and present mixed or lack of reports of test reliability. Premack and Woodruff[1978] first defined ToM as one's ability to attribute/impute mental states to self and others with the goal of predicting actions [Premack and Woodruff, 1978], which has been widely agreed-upon and adopted by researchers across disciplines. However, this definition lacks specificity in the

exact cognitive processes required to generate ToM-enabled behaviors. As a result, *many ToM tasks lack construct validity and can be solved through alternative low-level cognitive strategies* such as pattern recognition or learned association without requiring the participant to engage in mental state attribution when solving ToM tasks [Quesque and Rossetti, 2020]. For example, one of the most used ToM tasks in examining human emotional ascriptions, the Reading the Mind in the Eyes test [Baron-Cohen *et al.*, 2001], measures emotion recognition rather than ToM [Oakley *et al.*, 2016; Quesque and Rossetti, 2020]. While most of such invalid ToM tasks require input and output modalities beyond text (e.g., emotional attribution based on facial expressions) and haven't been used to benchmark LLM's ToM yet, AI researchers already uncovered LLMs leveraging similar tactics to pass ToM tasks— Ullman[2023] found that LLMs fail on trivial alterations to ToM tasks, Shapira *et al.*[2023] pointed out that LLMs rely on shortcuts, heuristics, and spurious correlations to pass ToM tasks, Kim *et al.*[2023] found that LLM ToM reasoning often appears illusory, as models can fail low-level reasoning questions despite correctly recognizing Beliefs. As multi-modal LLMs are being developed to process a variety of inputs beyond texts, researchers should take caution when using existing visual-based ToM tasks to benchmark LLM. In addition, as several AI researchers pointed out that given LLMs are typically trained on data that is readily available on the internet, ToM tasks that have been around for decades might have already been part of the models' training data, leading to *data contamination issues* [Ma *et al.*, 2023; Ullman, 2023] which makes it even more difficult to verify the validity of LLM's claimed ToM capability.

In the past decade, hundreds of ToM tests have been created by psychologists *without reporting important psychometric properties to assess the validity and reliability of the measures (e.g., internal consistency, test-retest reliability)* [Beaudoin *et al.*, 2020; Fu *et al.*, 2023]. Beaudoin *et al.*[2020] found that only 20.2% of their included ToM measures provided any such information; Ahmadi *et al.*[2015] noted that only six of the 11 included ToM measures had been examined for construct validity; Hayward and Homer[2017] identified notable validity and reliability issues of several prominent ToM measures (e.g., second-order false belief test and Strange Stories test has undesirable internal consistency). This has resulted in globally poor replicability of ToM measures in empirical studies [Beaudoin *et al.*, 2020]. Similarly, AI researchers have used human annotators, generative AI, or both to create their own ToM tests. While this enables rapid and large-scale test generation to stress-test LLMs, the process is often opaque and inconsistently reported in terms of validity and reliability. Although inter-rater reliability among the human annotators is commonly noted, the specific processes and measures researchers took to ensure human annotators' correct understanding of the ToM dimensions, internal consistency of the hundreds of tasks generated by each annotator, or the construct validity of the tasks actually measuring the specific ToM dimensions are often buried in appendices or not reported at all. This issue is especially pronounced when datasets are generated through a mix of human annotators and generative AI. This highlights the need to standardize ToM benchmark reporting and documentation to ensure transparency, consistency, and validity in ToM benchmarks.

3.3 Evaluation Limitation

ToM tasks rely on third-person, static, and synthetic scenarios, overlooking the practical use of ToM in dynamic, real-world social interactions. ToM capability enables one to extract social cues embedded in the complex, dynamic, and multi-modal environment to attribute various mental states to self and others when facilitating social interactions. Yet most of the existing ToM tasks take social interactions out of its dynamic context, and consist solely of presenting static social scenarios or stories to examine the respondent's social understanding of the synthatic scenario from a third-person perspective [Byom and Mutlu, 2013; Quesque and Rossetti, 2020]. Quesque and Rossetti[2020] found that 17 out of the 23 classic ToM measures they reviewed only examine children's ToM from a passive observer perspective (i.e., third-person perspective).

As Ma *et al.* [2023] pointed out, this has also been the case in ToM benchmark literature in AI— 12 out of the 21 papers in their survey positioned LLM as a passive observer, with only three papers positioning LLM as an active agent in the benchmark [Ma *et al.*, 2023]. Additionally, Ma *et al.*[2023] highlighted the lack of benchmarks encompassing both the physical and social environments, overlooking other physical and spatial relationships between agents and the object as well as intrinsic motivations on the agent side [Ma *et al.*, 2023]. Ability to understand social scenario is not the exact equivalent or valid predictor of one's ability to engage in actual social interactions, especially when such scenarios are taken out of the social contexts. For instance, even in static social story tasks, Gu *et al.* [2024] found that while most LLMs could accurately infer characters' mental states, they often failed to predict corresponding behaviors and performed worse when judging the reasonableness of those behaviors. Riemer *et al.* [2025] found that top-performing LLMs may possess strong literal ToM capability in predicting others' behaviors, they tend to struggle with functional ToM— the ability to adapt to agents through rational responses based on valid predictions from literal ToM. This has spurred a paradigm shift to encourage the design of ToM tasks based on actual social interactions from a first- or second-person perspective in both psychology and AI literature [Quesque and Rossetti, 2020; Hou *et al.*, 2024; Zhou *et al.*, 2024].

4 Towards User-Centered Theory of Mind Benchmark for LLMs

In the previous section, we summarized key limitations of appropriating human-intended ToM tasks to benchmark LLM's ToM capability based on existing literature: (1) **Theoretical limitation**: ToM is a multi-faceted construct yet it has been mostly measured on one dimension, (2) **Methodological limitation**: Many ToM tasks lack construct validity and present mixed or lack of reports of test validity and reliability, (3) **Evaluation limitation**: ToM tasks rely on third-person, static, and synthetic scenarios, overlooking the practical use

of ToM in dynamic, real-world social interactions. Our goal in surfacing and summarizing these limitations across psychology and AI literature is not to suggest abandoning ToM tasks entirely in LLM benchmarking but to highlight the challenges of repurposing them without further scrutiny in generating broad claims about LLM's general ToM capabilities.

In the process of summarizing and highlighting these limitations, we noticed a recurring pattern across ToM benchmark work: the limited role assigned to humans in ToM benchmarks. Rather than participating as actual end-users of LLM-powered AI applications, humans primarily serve as annotators to generate ToM tasks or provide baseline measurements to benchmark LLM's ToM capabilities. From a user-centered perspective, we must ask: even if LLMs eventually match human ToM capabilities, these models will ultimately power user-facing applications, so shouldn't user perspectives, preferences, and needs inform ToM benchmark design? In the rest of this section, we take an HCI perspective to explore research opportunities and challenges for designing towards user-centered ToM benchmarks for LLMs.

4.1 Defining LLM ToM From a User-Centered Perspective

Recognizing the theoretical limitations of existing ToM benchmarks, recent work has proposed more comprehensive evaluations that go beyond false-belief reasoning to cover multiple ToM dimensions identified in psychology [Beaudoin *et al.*, 2020], such as beliefs, desires, and emotions [**?**, e.g.,]]chen2024tombench. While such efforts represent important progress in understanding LLMs' broader ToM capabilities, they still rest on a foundational assumption: *that frameworks developed for human social cognition are directly applicable to machines.* ToM benchmarks for LLMs— and LLM benchmarks in general— have often borrowed psychology theories, frameworks, constructs, and measurements to evaluate LLM's human-like capabilities. Though this provides a useful foundation, such adaptations also carry over the original goals and assumptions of those theories, which *may not align with the realities of AI design and deployment.* For instance, ToM tasks in developmental psychology were primarily designed to identify deficits in children's social reasoning in interpreting others' mental states [Baron-Cohen *et al.*, 1985]——not to define general-purpose models of social competence applicable to artificial agents interacting with humans in real life.

LLMs are mostly used to power user-facing AI applications, so in practice, it may matter less whether LLMs possess ToM reasoning capabilities and more about the type of downstream behaviors enabled by LLM's ToM capabilities during human-AI interactions [Gu *et al.*, 2024]. In psychology, dimensions of ToM capabilities are good predictors for desirable human social behaviors, but this does not always translate to AI—human-AI interactions differ fundamentally from human-human interactions and thus users may have distinct expectations, preferences, and needs when it comes to AI's ToM-enabled behaviors. Not all ToM-enabled behaviors are desired and needed in every AI applications— Borg and Read [2024] pointed out that mot all empathic behaviors that fall under the umbrella capability of "empathy" will be needed and preferred for different empathic AI applications. Certain ToM-enabled behaviors that are considered socially adept in humans may instead elicit users' discomfort, distrust, or unease when exhibited by LLMs. For instance, an AI that predicts a user's intentions or thoughts too accurately might feel intrusive, raising concerns about user privacy. Similarly, AI systems that recognize and mimic a user's emotions too well might come across as eerie or manipulative rather than empathetic. Some ToM-enabled behaviors may be unnecessary——or even counterproductive——in particular AI application contexts, such as productivity tools or navigation systems, where users prioritize efficiency and reliability over social attunement.

Taking a user-centered perspective, we urge researchers to rethink the definition of ToM in LLM benchmarks by moving beyond "mimicking human behaviors" through adapting psychology theories. Instead, **we advocate for grounding the design of ToM benchmarks in empirical HCI studies that surface the kinds of ToM-enabled AI behaviors that users actually desire and need in real-world interactions.** This would require close collaboration between AI and HCI researchers to envision and implement ToM-enabled AI behaviors across diverse application contexts, conduct user studies or co-design sessions to understand people's interactions and experiences with such ToM-enabled AI systems, and translate those insights into measurable ToM dimensions for designing user-centered ToM benchmarks. Each portion presents its own unique challenges, the most difficult of which is the distillation of a comprehensive and easily-accessible ToM benchmark that meaningfully reflects the diverse user preferences and needs across various human-AI interaction contexts.

4.2 Benchmarking LLM ToM in Dynamic and Interactional Social Contexts

Several recent studies have proposed new approaches for aligning ToM benchmarks more closely with real world social contexts, in response to the limitations of evaluating LLMs using synthetic social scenarios. These approaches include leveraging natural human-human conversation dialogue to generate social scenarios in ToM benchmarks [Soubki *et al.*, 2024; Chan *et al.*, 2024; Kim *et al.*, 2023], expanding the test modality to include both multiple-choice question-answering and free-form responses [Kim *et al.*, 2023; Chan *et al.*, 2024], as well as converting third-person perspective ToM tasks to first-person perspectives in ToM benchmarks [Hou *et al.*, 2024]. Open-ended interaction environment such as Sotopia [Zhou *et al.*, 2024] provides opportunities to assess LLM's ToM in more dynamic, socially complex and active interlocutor perspective. Additionally, Shi *et al.*[2024] created multi-modal ToM benchmark that enables video and text descriptions of people's multi-modal behavior in realistic household environment to probe LLMs in answering about people's goals and beliefs. Work like Yerukola *et al.* [2024] has also highlighted the importance to understand LLMs' capability in interpreting and responding to human intentions beyond the literal meaning of words to achieve "functional ToM" [Riemer *et al.*, 2025].

As this effort towards more socially situated ToM bench-

marks continues, we also want to reflect on the definition and criteria of LLMs "passing" ToM benchmarks when situated in more interactive and dynamic contexts. ToM has traditionally been viewed as a static construct that can be measured through the one-shot "correctness" of one's understanding of social cues through multiple-choice questions in ToM tasks. However, through the lens of Mutual Theory of Mind (MToM) [Wang and Goel, 2022; Wang, 2024; Wang *et al.*, 2021], social interaction is iterative, sometimes requiring multiple back-and-forth between two parties through ToM construction, recognition, and revision for one to achieve the correct understanding of the other's mental states. As described by Wang and Goel [2022] in their MToM framework, each turn of the communication can offer richer social signals through communication feedback, which builds upon the ToM inferences made from the previous turn to eventually arrive at the "correct" social understanding and attribution of mental states.

In this light, when AI systems are embedded in dynamic, interactive environments, should we assess their ToM based on one-shot inference accuracy, or on their ability to iteratively refine their understanding in response to ongoing feedback? If ToM is fundamentally about understanding and predicting others' mental states in dynamic social environments, then **a more meaningful benchmark should account for how well an AI system navigates the iterative nature of real-world social exchanges**. This shift in evaluation criteria would move beyond static correctness toward assessing an AI's ToM based on its adaptability, responsiveness, and ability to integrate evolving social information, all of which are key components to human social intelligence. While readily quantifiable metrics— such as the number of conversational turns required for accurate inference—– offer a starting point, more nuanced measures that track improvements in inference quality based on individual user's feedback may provide deeper insight. Given the increasing deployment of LLM-powered AI applications in global contexts, such nuanced measures will need to include assessments on how well these systems can iteratively infer users' mental states when interacting with users across diverse demographic and cultural backgrounds— something that even humans struggle with during cross-cultural communications. This adds another layer of complexity to ToM benchmark design, requiring methodological innovation that balances the depth of user-centered evaluation with the scalability required for robust ToM assessment.

5 Conclusion

In this position paper, we outlined and summarized limitations of the popular approach in appropriating ToM tasks designed to evaluate children's ToM to benchmark LLM's ToM. Drawing upon existing psychology and AI literature, we argue that these limitations already exist in the original human-intended ToM tasks, and hence persisted and exacerbated when appropriated as LLM benchmarks. Specifically, we summarized three key limitations: (1) Theoretical limitation: ToM is a multi-faceted construct yet it has been mostly measured on one dimension, (2) Methodological limitation: Many ToM tasks lack construct validity and present mixed or lack of reports of test validity and reliability, (3) Evaluation limitation: ToM tasks rely on third-person, static, and synthetic scenarios, overlooking the practical use of ToM in dynamic, real-world social interactions. By identifying these limitations, we caution AI researchers against blind adoption of these ToM tasks and to draw claims about LLM's general ToM capability based on LLM passing such ToM tasks. Based on these limitations, we proposed the future direction towards designing user-centered ToM benchmark for LLMs. We discuss potential opportunities and challenges in this direction and encourage researchers to rethink the definition of LLM ToM based on user needs and preferences, as well as reflecting on the criteria of LLM benchmark in dynamic and interactive social contexts.

References

[Ahmadi *et al.*, 2015] Seyyede Zohreh Ziatabar Ahmadi, Shohreh Jalaie, and Hassan Ashayeri. Validity and reliability of published comprehensive theory of mind tests for normal preschool children: A systematic review. *Iranian journal of psychiatry*, 10(4):214, 2015.

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

[Baron-Cohen *et al.*, 1999] Simon Baron-Cohen, Michelle O'riordan, Valerie Stone, Rosie Jones, and Kate Plaisted. Recognition of faux pas by normally developing children and children with asperger syndrome or high-functioning autism. *Journal of autism and developmental disorders*, 29:407–418, 1999.

[Baron-Cohen *et al.*, 2001] Simon Baron-Cohen, Sally Wheelwright, Jacqueline Hill, Yogini Raste, and Ian Plumb. The "reading the mind in the eyes" test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism. *Journal of child psychology and psychiatry*, 42(2):241–251, 2001.

[Baron-Cohen, 1999] Simon Baron-Cohen. *The evolution of a theory of mind*. na, 1999.

[Baron-Cohen, 2000] Simon Baron-Cohen. Theory of mind and autism: A review. *International review of research in mental retardation*, 23:169–184, 2000.

[Beaudoin *et al.*, 2020] Cindy Beaudoin, Élizabel Leblanc, Charlotte Gagner, and Miriam H Beauchamp. Systematic review and inventory of theory of mind measures for young children. *Frontiers in psychology*, 10:2905, 2020.

[Borg and Read, 2024] Jana Schaich Borg and Hannah Read. What is required for empathic ai? it depends, and why that matters for ai developers and users. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1306–1318, 2024.

[Bubeck *et al.*, 2023] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott

Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[Byom and Mutlu, 2013] Lindsey J Byom and Bilge Mutlu. Theory of mind: Mechanisms, methods, and new directions. *Frontiers in human neuroscience*, 7:413, 2013.

[Chan *et al.*, 2024] Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. Negotiationtom: A benchmark for stress-testing machine theory of mind on negotiation surrounding. *arXiv preprint arXiv:2404.13627*, 2024.

[Chen *et al.*, 2024] Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, et al. Tombench: Benchmarking theory of mind in large language models. *arXiv preprint arXiv:2402.15052*, 2024.

[Chiang *et al.*, 2024] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

[Fu *et al.*, 2023] I-Ning Fu, Kuan-Lin Chen, Meng-Ru Liu, Dai-Rong Jiang, Ching-Lin Hsieh, and Shih-Chieh Lee. A systematic review of measures of theory of mind for children. *Developmental Review*, 67:101061, 2023.

[Gu *et al.*, 2024] Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. Simpletom: Exposing the gap between explicit tom inference and implicit tom application in llms. *arXiv preprint arXiv:2410.13648*, 2024.

[Happé, 1993] Francesca GE Happé. Communicative competence and theory of mind in autism: A test of relevance theory. *Cognition*, 48(2):101–119, 1993.

[Hayward and Homer, 2017] Elizabeth O Hayward and Bruce D Homer. Reliability and validity of advanced theory-of-mind measures in middle childhood and adolescence. *British Journal of Developmental Psychology*, 35(3):454–462, 2017.

[Hou *et al.*, 2024] Guiyang Hou, Wenqi Zhang, Yongliang Shen, Zeqi Tan, Sihao Shen, and Weiming Lu. Entering real social world! benchmarking the theory of mind and socialization capabilities of llms from a first-person perspective. *arXiv preprint arXiv:2410.06195*, 2024.

[Huang *et al.*, 2025] Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. Apathetic or empathetic? evaluating llms' emotional alignments with humans. *Advances in Neural Information Processing Systems*, 37:97053–97087, 2025.

[Kim *et al.*, 2023] Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. Fantom: A benchmark for stress-testing machine theory of mind in interactions. *arXiv preprint arXiv:2310.15421*, 2023.

[Kosinski, 2023] Michal Kosinski. Theory of mind may have spontaneously emerged in large language models. 2023.

[Le *et al.*, 2019] Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 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, 2019.

[Ma *et al.*, 2023] Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. Towards a holistic landscape of situated theory of mind in large language models. *arXiv preprint arXiv:2310.19619*, 2023.

[Milton, 2012] Damian EM Milton. On the ontological status of autism: The 'double empathy problem'. *Disability & society*, 27(6):883–887, 2012.

[Oakley *et al.*, 2016] Beth FM Oakley, Rebecca Brewer, Geoffrey Bird, and Caroline Catmur. Theory of mind is not theory of emotion: A cautionary note on the reading the mind in the eyes test. *Journal of abnormal psychology*, 125(6):818, 2016.

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

[Quesque and Rossetti, 2020] François Quesque and Yves Rossetti. What do theory-of-mind tasks actually measure? theory and practice. *Perspectives on Psychological Science*, 15(2):384–396, 2020.

[Rakoczy, 2022] Hannes Rakoczy. Foundations of theory of mind and its development in early childhood. *Nature Reviews Psychology*, 1(4):223–235, 2022.

[Riemer *et al.*, 2025] Matthew Riemer, Zahra Ashktorab, Djallel Bouneffouf, Payel Das, Miao Liu, Justin D. Weisz, and Murray Campbell. Position: Theory of mind benchmarks are broken for large language models, 2025.

[Sap *et al.*, 2019] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.

[Sap *et al.*, 2022] Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*, 2022.

[Shapira *et al.*, 2023a] Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*, 2023.

[Shapira *et al.*, 2023b] Natalie Shapira, Guy Zwirn, and Yoav Goldberg. How well do large language models perform on faux pas tests? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10438–10451, 2023.

[Shi *et al.*, 2024] Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. Muma-tom: Multi-modal multi-agent theory of mind. *arXiv preprint arXiv:2408.12574*, 2024.

[Sorin *et al.*, 2024] Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. Large language models and empathy: Systematic review. *Journal of Medical Internet Research*, 26:e52597, 2024.

[Soubki *et al.*, 2024] Adil Soubki, John Murzaku, Arash Yousefi Jordehi, Peter Zeng, Magdalena Markowska, Seyed Abolghasem Mirroshandel, and Owen Rambow. Views are my own, but also yours: Benchmarking theory of mind using common ground. *arXiv preprint arXiv:2403.02451*, 2024.

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

[Wang and Goel, 2022] Qiaosi Wang and Ashok K Goel. Mutual theory of mind for human-ai communication. *arXiv preprint arXiv:2210.03842*, 2022.

[Wang *et al.*, 2021] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 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, 2021.

[Wang, 2024] Qiaosi Wang. *MUTUAL THEORY OF MIND FOR HUMAN-AI COMMUNICATION IN AI-MEDIATED SOCIAL INTERACTION*. PhD thesis, Georgia Institute of Technology, 2024.

[Welivita and Pu, 2024] Anuradha Welivita and Pearl Pu. Are large language models more empathetic than humans? *arXiv preprint arXiv:2406.05063*, 2024.

[Wellman, 2018] Henry M Wellman. Theory of mind: The state of the art. *European Journal of Developmental Psychology*, 15(6):728–755, 2018.

[Xu *et al.*, 2024] Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. *arXiv preprint arXiv:2402.06044*, 2024.

[Yerukola *et al.*, 2024] Akhila Yerukola, Saujas Vaduguru, Daniel Fried, and Maarten Sap. Is the pope catholic? yes, the pope is catholic. generative evaluation of non-literal intent resolution in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 265–275, 2024.

[Ying *et al.*, 2025] Lance Ying, Katherine M Collins, Lionel Wong, Ilia Sucholutsky, Ryan Liu, Adrian Weller, Tianmin Shu, Thomas L Griffiths, and Joshua B Tenenbaum. On benchmarking human-like intelligence in machines. *arXiv preprint arXiv:2502.20502*, 2025.

[Zhang *et al.*,] Haoran Zhang, Ling Wang, Zhihao Chen, Yue Liu, and Xinyi Li1 Jing Wu. Developing a comprehensive empathy evaluation benchmark for ai systems.

[Zhou *et al.*, 2024] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. In *ICLR*, 2024.

[Zhu *et al.*, 2024] Qihao Zhu, Leah Chong, Maria Yang, and Jianxi Luo. Reading users' minds from what they say: An investigation into llm-based empathic mental inference. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 88407, page V006T06A018. American Society of Mechanical Engineers, 2024.