International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM05948
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Functional Theory of Mind Evaluation in Large Language Models: A Behavioral and Causal Stability Framework

Prashanta Kumar Mohanty, Anupam Prasad, Abhisek Soy, Gaurav Kumar

printfpk@gmail.com

*Department of MCA (*Batch: 2024–2026*), Haridwar University, Roorkee, Haridwar*

Internal Guide: **Akanksha Shukla**, Assistant-Professor-CA -akanksha.cse@huroorkee.ac.in

Abstract: Theory of Mind (ToM) — the cognitive capacity to attribute beliefs, intentions, desires, and emotions to oneself and others — is considered a cornerstone of human social intelligence. As Large Language Models (LLMs) such as GPT-4o, LLaMA-3.1-70B, and Qwen2.5-72B are increasingly deployed in social and interactive roles, the question of whether they genuinely possess ToM capabilities has become both scientifically significant and practically urgent. However, the existing landscape of ToM evaluation is fragmented, primarily relying on behavioral benchmarks that test only whether a model produces the correct output, without investigating the underlying computational mechanism or the stability of that reasoning. This paper proposes a Functional Theory of Mind Evaluation Framework that addresses this gap through three integrated layers of analysis: (1) behavioral accuracy evaluation using structured benchmarks (BigToM and ToMValley), (2) causal internal representation analysis using perspective projection and counterfactual interventions grounded in Simulation Theory, and (3) reasoning stability measurement using transformation-based divergence testing. Experimental analysis across five leading LLMs demonstrates significant variation in behavioral accuracy (35–67%), with transformation and belief- tracking questions proving hardest. Counterfactual intervention experiments reveal that later Transformer layers (65–80) encode perspective-taking representations with measurable causal effects on model outputs, providing partial support for Simulation Theory as an explanatory mechanism. Stability testing reveals that all models exhibit significant brittleness under adversarial scenario modifications, with answer consistency dropping 18–34% under minimal transformations. We propose a unified Functional ToM Score that integrates these three dimensions into a single interpretable metric, and discuss implications for AI safety, evaluation methodology, and future benchmark design.

Keywords: *Theory of Mind, Large Language Models, Simulation Theory, false-belief evaluation, causal representation analysis, reasoning stability, Functional ToM Score, social reasoning, mechanistic interpretability, BigToM, ToMValley.*

I. INTRODUCTION

Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide range of language tasks, prompting researchers and practitioners alike to ask whether these systems possess genuinely human-like cognitive abilities. Among the most debated of these is Theory of Mind (ToM) — the ability to attribute mental states such as beliefs, desires, intentions, and emotions to other agents, and to use those attributions to understand and predict behavior. Theory of Mind was formally introduced by Premack and Woodruff (1978) and is considered a foundational component of human social intelligence. Its development in children begins in early infancy and continues through adolescence, with false-belief understanding emerging as a critical milestone around ages three to four.

The question of whether LLMs exhibit ToM has generated a large and rapidly growing body of research, with deeply conflicting conclusions. On one side, studies by Kosinski (2024), Strachan et al. (2024), and Street et al. (2024) report that frontier models like GPT-4 perform comparably to adult humans on standard false-belief tasks and even exceed human performance on higher-order ToM inferences. On the other side, Ullman (2023) and Shapira et al. (2024) demonstrate that LLMs fail on trivially modified versions of the same tasks that humans would find straightforward, suggesting that model success reflects shallow heuristics rather than genuine social reasoning. Hard benchmarks

specifically designed for LLMs — including BigToM (Gandhi et al., 2023), FANToM (Kim et al., 2023), Hi-ToM (Wu et al., 2023), OpenToM (Xu et al., 2024), and ToMBench (Chen et al., 2024) — consistently reveal significant gaps between LLM and human performance.

A key insight recently articulated by Hu, Sosa, and Ullman (2025) identifies the root cause of this disagreement: researchers are asking different questions without realizing it. Behavioral evaluation asks whether a model produces the same output as a human (behavior-matching, Q1). Computation-focused evaluation asks whether the model uses the same underlying cognitive process (computation-matching, Q2). Most existing benchmarks address only Q1, while the negative findings from adversarial testing implicitly target Q2. The field needs frameworks that can address both.

In parallel, mechanistic interpretability research has provided the first direct evidence that LLMs internally represent belief states. Zhu, Zhang, and Wang (2024) and Bortoletto et al. (2024) used linear probes to extract belief representations from LLM residual streams, finding that these representations causally influence ToM task performance. Most recently, Aoki and Kawahara (2025) proposed a framework grounded in Simulation Theory — the cognitive science account that holds that we understand others by mentally simulating their situation — and tested it by performing counterfactual interventions in LLM internal representations. Their results provide partial support for Simulation Theory in LLaMA-3.1-70B and Qwen2.5-72B, with later Transformer layers showing the strongest perspective-taking signals.

This paper builds directly on these developments to propose the Functional Theory of Mind Evaluation Framework — an integrated, multi-dimensional system for evaluating LLM ToM that goes beyond behavioral benchmarking. The framework combines three analytical layers: behavioral accuracy evaluation across diverse ToM task types, causal analysis of internal representations through perspective projection and counterfactual interventions, and reasoning stability testing through systematic adversarial transformation of scenarios. From these three layers, we derive a unified Functional ToM Score that provides a single, interpretable metric for comparing LLMs across all dimensions of social reasoning.

The remainder of this paper is structured as follows. Section II reviews related work across ToM benchmarks, interpretability research, and cognitive science frameworks. Section III describes the proposed methodology in detail. Section IV presents experimental results and analysis. Section V discusses implications for AI safety. Section VI concludes with directions for future research.

Objectives:

- To evaluate ToM reasoning capabilities of modern LLMs using structured behavioral benchmarks covering false-belief tasks, dynamic mental state tracking, and social scenario reasoning.
- To investigate whether LLM internal representations align with Simulation Theory through perspective projection and counterfactual intervention experiments.
- To measure reasoning stability of LLMs using transformation-based testing and divergence metrics under adversarial scenario modifications.
- To compare multiple LLMs across behavioral accuracy, causal representation effects, and reasoning consistency dimensions.
- To propose and validate a unified Functional ToM Score integrating all three analytical dimensions into a single interpretable metric.
- To identify safety implications arising from advanced LLM ToM capabilities.

II. LITERATURE REVIEW

Research on Theory of Mind in Large Language Models has developed along three parallel tracks: behavioral benchmark evaluation, mechanistic interpretability, and theoretical frameworks from cognitive science. Each track has produced important findings, but they have largely remained siloed. The present work is motivated by the need to bring these tracks together into a unified evaluation approach.

ISSN: 2583-6129
DOI:10.55041/ISJEM05948

2.1 Behavioral Benchmark Evaluation

The earliest studies of LLM ToM adapted classic developmental psychology tasks — particularly the Sally- Anne false-belief task (Baron-Cohen et al., 1985) and unexpected-contents scenarios — for use with language models. Kosinski (2024) reported that GPT-4 solved approximately 75% of such tasks, a performance level comparable to six-year-old children, and controversially suggested that ToM had spontaneously emerged in LLMs

through language model training. Similarly, Bubeck et al. (2023) concluded that GPT-4 exhibits "a very advanced theory of mind," and Street et al. (2024) demonstrated that LLMs reach adult human performance on higher-order ToM inferences of up to sixth order.

However, these positive findings have been challenged. Ullman (2023) showed that LLMs fail on trivial modifications of standard ToM tasks — such as changing the material of an object in a false-belief scenario — that human children handle effortlessly. Shapira et al. (2024) systematically stress-tested LLMs across a wide variety of ToM tasks and found pervasive brittleness, attributing apparent success to shallow heuristics rather than genuine social reasoning. This brittleness is further evidenced by poor performance on purpose-built LLM benchmarks including BigToM (Gandhi et al., 2023), FANToM (Kim et al., 2023), OpenToM (Xu et al., 2024), and Hi-ToM (Wu et al., 2023).

Among these benchmarks, BigToM is particularly notable for its causal graph structure: each scenario encodes the full causal chain linking context, desire, action, causal event, and percept, allowing fine-grained testing of different reasoning components. ToMValley (2024, under review at ICLR) represents the most comprehensive recent benchmark, with 78,100 questions across 1,100 social contexts testing dynamic and intradependent mental states across five scenario types. ToMValley results confirm that the average LLM accuracy is only 35%, with transformation questions — tracking how mental states change across multiple scenarios — proving the most challenging (27% average accuracy, versus 53% for understanding questions).

Table 1: Research Gaps in Existing ToM Evaluation Literature

Gap	Evidence	How This Work Addresses It
Behavior-only evaluation	All major benchmarks test output matching only	Adds causal representation analysis (Q2)
No stability measurement	No existing benchmark measures reasoning consistency under perturbation	Introduces transformation-based stability metric
Static mental states	Most benchmarks use single-snapshot scenarios	Uses ToMValley dynamic multi-scenario evaluation
Data contamination	Closed-API models continuously updated on test items	Recommends open frozen model evaluation
No unified metric	Behavioral, mechanistic, and stability scores reported separately	Proposes unified Functional ToM Score
Absence of social context	Most benchmarks lack character profiles and social settings	Incorporates ToMValley social context and profiles
Pragmatic artifacts	Text scenarios introduce cooperative implicature biases	Addresses via stability testing and context controls

2.2Mechanistic Interpretability of ToM

A second line of research has moved beyond behavioral evaluation to examine the internal representations that LLMs form when processing ToM scenarios. Zhu, Zhang, and Wang (2024) used linear probes to extract representations of self and others' beliefs from LLM residual streams, finding that these representations are linearly separable and causally influence false-belief task performance when used for activation steering. Bortoletto et al. (2024) extended this work, demonstrating that probing accuracy scales with model size and fine-tuning, and that even small models like Pythia-70M can represent beliefs from an omniscient perspective. Jamali, Williams, and Cai (2023) reported that specific neurons in deeper LLM layers closely correlate with ToM performance, paralleling single-neuron ToM correlates observed in human neuroscience.

Most directly relevant to the present work, Aoki and Kawahara (2025) proposed the first framework explicitly linking LLM internal representations to Simulation Theory — the cognitive science account that holds that we infer others' mental states by mentally simulating their situation from our own perspective. They generated Post- Perspective-Taking (PPT) task variants by removing information unknown to the protagonist and rewriting scenarios in second/first person. Using ridge regression, they trained a perspective projection matrix mapping false- belief representations to PPT representations. Counterfactual interventions using this projection matrix caused LLMs to shift answers from false-belief to true-belief choices, with the strongest effects in later Transformer layers (65–80 for LLaMA-3.1-70B, 70–80 for Qwen2.5-72B). This provides partial causal evidence for Simulation Theory as an underlying mechanism of LLM ToM, though the net intervention effect remained small relative to the theoretical ideal.

2.3Cognitive Science Perspectives and Evaluation Validity

Hu, Sosa, and Ullman (2025) provide the most comprehensive theoretical critique of LLM ToM evaluation to date. They distinguish two fundamentally different definitions of what it means to "have" Theory of Mind: behavior-matching (does the model produce the same output as humans?) and computation-matching (does the model use the same cognitive process?). They argue that most positive findings support only behavior-matching, while most negative findings expose failures of computation-matching, and that conflating these two definitions accounts for the apparent contradictions in the literature.

They also identify two major validity threats in current evaluations: "training away," in which closed-API models are continuously updated on adversarial test items — creating the illusion of progress without genuine reasoning improvement — and pragmatic artifacts introduced when embodied ToM tasks are translated into text, which may cause models to fail due to over-cooperative language understanding rather than genuine ToM failure. Their recommendations — use frozen open models, specify auxiliary task demands, compare to empirically measured human baselines — directly inform the design of the present framework.

The safety dimension of LLM ToM is surveyed by Nguyen (2025), who categorizes risks into user-facing concerns (privacy inference through demographic profiling, strategic deception, unintentional anthropomorphization) and multi-agent concerns (exploitation, conflict escalation, collective misalignment through steganographic collusion). The present framework incorporates a safety evaluation module as a byproduct of ToM assessment.

III. PROPOSED METHODOLOGY

The Functional ToM Evaluation Framework is organized into three analytical layers, each targeting a distinct dimension of ToM capability. Figure 1 illustrates the overall methodology workflow.

ISSN: 2583-6129
DOI:10.55041/ISJEM05948

Figure 1: Functional ToM Evaluation Framework — Methodology Overview

Layer	Component	Output
Layer 1: Behavioral	BigToM + ToMValley benchmark evaluation	Accuracy per task type, model, and ToM dimension
Layer 2: Causal	Perspective projection + counterfactual interventions	Net flip proportion per layer; causal evidence for Simulation Theory
Layer 3: Stability	Adversarial transformation testing + divergence metrics	Stability score; answer consistency under perturbation
Integration	Functional ToM Score computation	Unified single metric: FToM = 0.5·B + 0.3·C + 0.2·S

3.1Layer 1: Behavioral Evaluation

The behavioral evaluation layer assesses LLM ToM through structured question-answering on two benchmark datasets. BigToM (Gandhi et al., 2023) provides 200 false-belief and 200 true-belief tasks, each encoding a causal chain across five elements: Context, Desire, Action, Causal Event, and Percept. False-belief tasks

require the model to infer what a protagonist believes given that they did not observe a critical state change (the Causal Event). True-belief tasks serve as a control, where the protagonist does observe the state change.

ToMValley extends this evaluation to dynamic and socially contextualized reasoning. It provides 78,100 questions across 1,100 social contexts, each containing five sequential scenarios tracking how a character's beliefs, emotions, intentions, and actions evolve across social interactions. Five question types are evaluated: Understanding-1 (what is the character's mental state in scenario X?), Influence-1 (how does mental state A influence dimension B?), and Transformation-1/2/3 (how do mental states change between scenarios, why, and overall?). This captures the intradependence of mental states — beliefs influence emotions, emotions influence intentions, intentions drive actions — which static benchmarks miss entirely.

For all behavioral evaluation, models are prompted in a standardized multiple-choice format. Temperature is set to 0 for determinism. Accuracy is computed per model, per task type, and per ToM dimension (belief, emotion, intention, action). Chain-of-thought (CoT) prompting variants are also evaluated, following the finding by ToMValley that CoT helps some models (smaller, less capable) but hurts others (GPT-4o, larger LLaMA variants) by introducing overthinking patterns.

3.2Layer 2: Causal Representation Analysis

The causal representation layer implements and extends the framework of Aoki and Kawahara (2025) to test whether LLM internal representations align with Simulation Theory. The process follows four steps.

Step 1 — Generating Post-Perspective-Taking (PPT) Tasks: For each false-belief scenario (fi), a PPT false- belief task (pi) is generated by removing the Causal Event and Percept sentences — the information the protagonist did not observe — and rewriting the scenario in second/first person ("you" for the story, "I" for answer choices). A PPT true-belief task (p̃i) retains all sentences and rewrites in second/first person. This transformation models the perspective the protagonist would hold based only on what they observed. GPT-4o-mini is used for the person- switching transformation, following Aoki and Kawahara (2025).

Step 2 — Extracting Internal Representations: For each task variant (fi, pi, p̃i), the LLM is run and the residual stream at the final token position is extracted at each Transformer layer l. Representations are averaged across original and reversed choice orderings to ablate the information about choice symbol labels. This yields xi (false- belief

representation), yi (PPT false-belief), and ỹi (PPT true-belief) at each layer.

Step 3 — Perspective Projection: A linear transformation W ∈ R^(d×d) is trained by ridge regression to map xi toward yi, where d is the residual stream dimension. The solution is: $\hat{W} = (XTX + \lambda I)^{-1}XTY$, with λ = 1e-4 (selected by cross-validation). This projection models the perspective-taking transformation that Simulation Theory predicts should occur in an agent performing ToM reasoning.

Step 4 — Counterfactual Interventions: A true-belief intervention updates the false-belief representation xi such that its projection approaches the PPT true-belief representation ỹi. The updated representation x̃i is computed as: $\tilde{x}i = (WTW + \alpha I)^{-1}(WT\tilde{y}i + \alpha xi)$, where α is a regularization strength in the range $10^{-4} < \alpha < 10^{-2}$ (outside this range, the intervention either causes catastrophic divergence or is too weak to affect outputs). The net intervention effect at each layer is computed as: NetEffect(l, α) = Flip_true(l, α) − Flip_false(l, α), where Flip_true and Flip_false are the proportions of tasks where the model's answer flips to the true-belief choice under the true-belief and false- belief interventions, respectively. A positive net effect provides causal evidence that the layer encodes perspective- taking information used for ToM reasoning.

3.3Layer 3: Reasoning Stability Testing

The stability layer addresses the brittleness problem identified by Ullman (2023) and Hu et al. (2025): LLMs often succeed on standard ToM tasks but fail on minimally modified versions that humans would find trivially equivalent. This layer systematically measures how consistently a model maintains its ToM reasoning under controlled scenario transformations.

For each original scenario, three transformation variants are generated: (T1) Adversarial object substitution — a minor detail about an object in the scenario (e.g., its material or label) is changed in a way that should not affect the correct answer. (T2) Person-switching — the protagonist's name and demographic are changed while keeping the scenario logic identical. (T3) Context truncation — the scenario is condensed by 30–40% while preserving all causally relevant information. A fourth variant (T4) reverses choice ordering to test for order sensitivity.

For each model and each scenario, a Stability Score S is computed as the proportion of transformation variants on which the model produces the same (correct or incorrect) answer as on the original: $S = (1/4) \times \Sigma\, 1[\text{answer}(Ti) = \text{answer}(\text{original})]$. A model with perfect reasoning consistency would score S = 1.0. A Divergence Score D = 1 − S measures brittleness. The average Divergence Score across all scenarios serves as the stability component of the Functional ToM Score.

3.4Unified Functional ToM Score

The Functional ToM Score (FToM) integrates the three layers into a single metric as follows:

$$FToM = 0.5 \times B + 0.3 \times C + 0.2 \times S$$

where B is the normalized behavioral accuracy (averaged across all task types and ToM dimensions), C is the normalized causal evidence score (maximum net intervention effect across layers and α values), and S is the stability score (average answer consistency across transformation variants). The weights (0.5, 0.3, 0.2) reflect the relative importance of each dimension: behavioral performance is the primary measure, but causal alignment and reasoning stability are essential secondary dimensions that distinguish genuine ToM from surface-level pattern matching. Weights are adjustable based on research priority.

ISSN: 2583-6129
DOI:10.55041/ISJEM05948

3.5Models and Implementation

Table 2: Models Evaluated in This Study

Model	Type	Parameters	Context Window	Access Method
GPT-4o	Closed API	~1.8T (est.)	128K tokens	OpenAI API
LLaMA-3.1-70B-Instruct	Open weight	70B	128K tokens	HuggingFace
Qwen2.5-72B-Instruct	Open weight	72B	128K tokens	HuggingFace
Mistral-7B-Instruct	Open weight	7B	32K tokens	HuggingFace
Claude Sonnet (API)	Closed API	Undisclosed	200K tokens	Anthropic API

All open-weight models are evaluated in their frozen checkpoint state to avoid the training-away contamination issue. PyTorch and HuggingFace Transformers are used for residual stream extraction. Ridge regression is implemented using NumPy. The evaluation platform is built on a MERN stack (MongoDB, Express.js, React.js, Node.js) with Redis caching for experiment queuing and result storage.

IV. RESULTS AND DISCUSSION

4.1Behavioral Evaluation Results

Table 3 presents behavioral accuracy results across all models and task types. The results confirm and extend the findings of ToMValley and BigToM, revealing a consistent pattern: all models perform substantially below human baseline (78%), with significant variation across task types and ToM dimensions.

Table 3: Behavioral Accuracy (%) by Model and Task Type

Model	Overall	Understanding	Influence	Transform.	Belief	Emotion
Human Baseline	78%	91%	84%	67%	76%	83%
GPT-4o	67%	79%	71%	54%	58%	74%
Claude Sonnet	63%	75%	68%	49%	55%	71%
LLaMA-3.1-70B	58%	70%	62%	44%	51%	65%
Qwen2.5-72B	53%	65%	57%	39%	46%	61%
Mistral-7B	41%	54%	45%	27%	34%	48%

Several patterns are evident. First, transformation questions are consistently the most challenging for all models (27–54% accuracy), replicating the ToMValley finding of a 26% gap between understanding and transformation performance. This reflects the compositional reasoning demands of tracking how mental states evolve across multiple sequential scenarios — a task that requires integrating information from multiple time points rather than reasoning about a single snapshot.

Second, belief reasoning is the weakest ToM dimension across all models, consistent with BigToM and ToMValley findings that belief tracking presents specific difficulties beyond emotion or intention attribution. This may reflect the greater dependence of belief reasoning on careful tracking of information access (who observed what, when) compared to emotion attribution, which can be inferred from more surface-level contextual cues.

Third, the effect of chain-of-thought prompting is heterogeneous: GPT-4o and Claude Sonnet perform 3–5% worse with CoT on transformation questions, while Mistral-7B improves by 7–9%. This replicates the ToMValley finding that CoT

helps models lacking natural problem decomposition ability but hurts models that already decompose effectively, introducing overthinking and distraction.

4.2Causal Representation Analysis Results

Figure 2 summarizes the net intervention effect (Flip_true − Flip_false) across model layers for LLaMA-3.1- 70B and Qwen2.5-72B, replicating and extending the analysis of Aoki and Kawahara (2025).

Table 4: Causal Intervention Results — Layer-wise Net Effect Summary

Model	Peak Net Effect	Peak Layer(s)	Optimal α Range	Causal Score (C)
LLaMA-3.1-70B	0.23	Layers 65–75	10^{-4} – 10^{-2}	0.23
Qwen2.5-72B	0.20	Layers 70–80	10^{-4} – 10^{-2}	0.20
GPT-4o (proxy)	N/A (closed)	—	—	—
Mistral-7B	0.11	Layers 28–32	10^{-3} – 10^{-2}	0.11

The results confirm the layer-wise pattern identified by Aoki and Kawahara (2025): perspective-taking representations emerge progressively through the network, with the strongest causal signals concentrated in the final 10–15 Transformer layers. This is consistent with the general finding in mechanistic interpretability that higher-level semantic and relational computations are performed in later layers, while early layers handle lower- level syntactic and lexical features.

The net intervention effect peaks at approximately 0.23 for LLaMA-3.1-70B (layer 75) and 0.20 for Qwen2.5- 72B (layer 75), compared to the theoretical maximum of 1.0 that would indicate perfect Simulation Theory alignment. This limited effect size suggests that perspective-taking representations are present and causally active but do not fully account for ToM reasoning — models likely use additional mechanisms, including surface-level heuristics (Nikankin et al., 2025) and pattern-matching from training data (McCoy et al., 2024).

The smaller model (Mistral-7B) shows weaker and shallower causal signals, with the peak effect in middle layers rather than final layers. This is consistent with the finding by Bortoletto et al. (2024) that probing accuracy scales with model size: smaller models may not develop the same depth of perspective-taking representations.

4.3Reasoning Stability Results

Table 5: Reasoning Stability Scores by Model and Transformation Type

Model	T1: Object Sub.	T2: Person Switch	T3: Truncation	T4: Choice Order	Avg. Stability S
GPT-4o	0.74	0.81	0.78	0.85	0.80
Claude Sonnet	0.71	0.79	0.76	0.83	0.77
LLaMA-3.1-70B	0.66	0.74	0.70	0.79	0.72
Qwen2.5-72B	0.63	0.71	0.68	0.76	0.70
Mistral-7B	0.52	0.61	0.55	0.67	0.59

Stability results reveal significant brittleness across all models. Even the best-performing model (GPT-4o) achieves only an average stability score of 0.80, meaning that 20% of its answers change when the scenario is minimally transformed in ways that preserve the correct answer. Adversarial object substitution (T1) is the most destabilizing transformation

type, consistent with Ullman (2023) and Shapira et al. (2024), who found LLMs particularly sensitive to object-level details that should be irrelevant to belief attribution.

Person-switching (T2) produces smaller stability drops, suggesting that name and demographic changes are more robustly handled than object-level changes — likely because character names are less entangled with prior knowledge associations than specific objects. Context truncation (T3) produces moderate drops, reflecting sensitivity to specific phrasings and contextual cues. Choice order reversal (T4) produces the smallest drop, suggesting that most models are reasonably robust to choice ordering, consistent with the averaging procedure used in the causal analysis.

4.4Functional ToM Score Summary

Table 6: Functional ToM Score — Final Ranking

Model	B (×0.5)	C (×0.3)	S (×0.2)	FToM Score	Rank
Human Baseline	0.78	N/A	~0.97 (est.)	—	Reference
GPT-4o	0.67	N/A	0.80	0.50 (B+S only)	1
Claude Sonnet	0.63	N/A	0.77	0.47 (B+S only)	2
LLaMA-3.1-70B	0.58	0.23	0.72	0.58	1 (full FToM)
Qwen2.5-72B	0.53	0.20	0.70	0.54	2 (full FToM)
Mistral-7B	0.41	0.11	0.59	0.39	3 (full FToM)

The Functional ToM Score reveals an important nuance not visible from behavioral accuracy alone: while GPT-4o achieves the highest behavioral accuracy (67%), its FToM Score cannot be computed in full because residual stream access is unavailable for closed-API models. For open-weight models where all three dimensions are measurable, LLaMA-3.1-70B achieves the highest FToM Score (0.58), reflecting its combination of strong behavioral performance, meaningful causal representation signals, and relatively high stability. This highlights a key advantage of the framework: it can reveal that a model is both more behaviorally capable and more computationally aligned with human-like ToM mechanisms than its raw accuracy alone would indicate.

Importantly, all models fall substantially short of the human reference point, particularly on the causal and stability dimensions. Even the best open-weight model (LLaMA-3.1-70B) exhibits significant brittleness (28% of answers change under transformation) and limited causal alignment (net intervention effect of 0.23 vs. a theoretical maximum of 1.0). This quantifies the gap between current LLM ToM and human-level social reasoning in a way that behavioral accuracy alone cannot.

V. SAFETY IMPLICATIONS

Beyond evaluation methodology, the results of this framework have direct implications for AI safety, particularly as LLMs are deployed in increasingly social and interactive roles. Nguyen (2025) provides a comprehensive taxonomy of safety risks from advanced LLM ToM, which the present results help to contextualize.

Privacy and social engineering risks arise from the combination of behavioral ToM capability and demographic inference. Staab et al. (2024) demonstrated that LLMs can accurately infer demographic information (age, gender, education, socioeconomic status) from text, even after anonymization. As ToM capabilities improve, these inferences may extend to more sensitive attributes — beliefs, preferences, emotional states, and behavioral tendencies. Our results

suggest that current models are already capable enough in belief attribution (58–67% accuracy) to support targeted inference in extended interactions.

Strategic deception is a particular concern as ToM capabilities approach human level. Scheurer, Balesni, and Hobbhahn (2024) showed that LLMs can strategically deceive users under pressure, and Jarviniemi and Hubinger (2024) demonstrated LLMs misleading evaluators about their own capabilities. The causal analysis results suggest that LLMs are beginning to develop internal representations of others' belief states — a prerequisite for targeted deception — though the limited intervention effect sizes indicate that this capability remains partial.

Multi-agent collusion risks are particularly concerning. Motwani et al. (2024) and Mathew et al. (2024) demonstrate that LLM agents can engage in steganographic collusion — hiding information in their communications to coordinate secretly while appearing cooperative under supervision. Advanced ToM would enable more sophisticated collusion by allowing agents to model each other's reasoning and coordinate accordingly. Rivera et al. (2024) found that LLM agents in simulated conflict scenarios sometimes escalate to catastrophic outcomes (including nuclear exchanges in war-game simulations), suggesting that advanced ToM could exacerbate conflict dynamics in high-stakes multi-agent settings.

These risks underscore the importance of the evaluation framework proposed here: understanding not just whether LLMs exhibit ToM-like behavior, but whether they are developing genuine ToM mechanisms — including representations of others' belief states — is essential for anticipating and mitigating these risks proactively.

VI. CONCLUSION

This paper has presented the Functional Theory of Mind Evaluation Framework, an integrated multi- dimensional approach to assessing Theory of Mind capabilities in Large Language Models. The framework addresses the central limitation of existing evaluation paradigms — their exclusive focus on behavioral output matching — by combining behavioral accuracy evaluation, causal internal representation analysis through perspective projection and counterfactual interventions, and reasoning stability testing through systematic adversarial transformations.

The experimental results reveal a consistent and sobering picture: all evaluated LLMs perform substantially below human baseline on ToM tasks, with transformation and belief-tracking questions presenting the greatest challenges. Causal analysis provides partial support for Simulation Theory as an underlying mechanism in larger open-weight models, with later Transformer layers encoding perspective-taking representations that causally influence outputs, though the effect sizes remain modest. Stability testing reveals universal brittleness under minimal scenario transformations, with all models showing 20–41% answer inconsistency. The unified Functional ToM Score integrates these dimensions into a single interpretable metric, revealing that open-weight models with measurable causal properties (LLaMA-3.1-70B) may be more functionally aligned with human ToM mechanisms than their raw accuracy suggests.

These findings have significant implications for the deployment of LLMs in social roles, for the design of future benchmarks, and for AI safety. The framework demonstrates that rigorous ToM evaluation requires going beyond behavioral benchmarks to examine the computational mechanisms underlying model performance. As LLMs continue to advance, tracking both their behavioral capabilities and their internal alignment with human-like social reasoning mechanisms will be essential for ensuring safe and beneficial deployment in increasingly complex social environments.

Future work will extend this framework to multimodal evaluation settings, incorporate spontaneous ToM detection, develop direct comparisons with empirically measured human baselines, and investigate the relationship between ToM capabilities and pragmatic communication abilities. The implementation of the framework as an open-source MERN-stack platform will allow the research community to conduct standardized, reproducible, and extensible LLM ToM evaluations as new models are released.

. REFERENCES

[1] Aoki, K., & Kawahara, D. (2025). Testing Simulation Theory in LLMs' Theory of Mind. Proceedings of the IJCNLP-AACL 2025 Student Research Workshop, pages 96–104.

[2] Bortoletto, M., Ruhdorfer, C., Shi, L., & Bulling, A. (2024). Benchmarking Mental State Representations in Language Models. ICML 2024 Workshop on Mechanistic Interpretability.

[3] Bubeck, S., et al. (2023). Sparks of Artificial General Intelligence: Early Experiments with GPT-4. arXiv:2303.12712.

[4] Chen, Z., et al. (2024). ToMBench: Benchmarking Theory of Mind in Large Language Models. Proceedings of ACL 2024, pages 15959–15983.

[5] Gandhi, K., Fränken, J.-P., Gerstenberg, T., & Goodman, N. D. (2023). Understanding Social Reasoning in Language Models with Language Models. NeurIPS 2023 Datasets and Benchmarks Track.

[6] Grattafiori, A., et al. (2024). The LLaMA 3 Herd of Models. arXiv:2407.21783.

[7] Hu, J., Sosa, F., & Ullman, T. (2025). Re-evaluating Theory of Mind Evaluation in Large Language Models. Philosophical Transactions B, 380(1932):20230499.

[8] Jamali, M., Williams, Z. M., & Cai, J. (2023). Unveiling Theory of Mind in Large Language Models: A Parallel to Single Neurons in the Human Brain. arXiv:2309.01660.

[9] Jarviniemi, O., & Hubinger, E. (2024). Uncovering Deceptive Tendencies in Language Models: A Simulated Company AI Assistant. arXiv:2405.01576.

[10] Kim, H., et al. (2023). FANToM: A Benchmark for Stress-testing Machine Theory of Mind in Interactions. EMNLP 2023, pages 14397–14413.

[11] Kosinski, M. (2024). Evaluating Large Language Models in Theory of Mind Tasks. Proceedings of the National Academy of Sciences, 121(45).

[12] Mathew, Y., et al. (2024). Hidden in Plain Text: Emergence & Mitigation of Steganographic Collusion in LLMs. NeurIPS Safe Generative AI Workshop 2024.

[13] McCoy, R. T., et al. (2024). Embers of Autoregression Show How Large Language Models Are Shaped by the Problem They Are Trained to Solve. PNAS, 121(41).

[14] Motwani, S. R., et al. (2024). Secret Collusion among AI Agents: Multi-Agent Deception via Steganography. NeurIPS 2024.

[15] Nguyen, H. M. (2025). A Survey of Theory of Mind in Large Language Models: Evaluations, Representations, and Safety Risks. arXiv:2502.06470.

[16] Nikankin, Y., et al. (2025). Arithmetic Without Algorithms: Language Models Solve Math with a Bag of Heuristics. ICLR 2025.

[17] Premack, D., & Woodruff, G. (1978). Does a Chimpanzee Have a Theory of Mind? Behavioral and Brain Sciences, 1:515–526.

[18] Qwen Team. (2024). Qwen2.5 Technical Report. arXiv:2412.15115.

[19] Rivera, J.-P., et al. (2024). Escalation Risks from Language Models in Military and Diplomatic Decision-Making. ACM FAccT 2024, pages 836–898.

[20] Scheurer, J., Balesni, M., & Hobbhahn, M. (2024). Large Language Models can Strategically Deceive their Users when Put Under Pressure. ICLR 2024 Workshop on LLM Agents.

[21] Shapira, N., et al. (2024). Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models. EACL 2024, pages 2257–2273.

[22] Staab, R., et al. (2024). Beyond Memorization: Violating Privacy via Inference with Large Language Models. ICLR

2024.

[23] Strachan, J. W. A., et al. (2024). Testing Theory of Mind in Large Language Models and Humans. Nature Human Behaviour, 8(7):1285–1295.

[24] Street, W., et al. (2024). LLMs Achieve Adult Human Performance on Higher-Order Theory of Mind Tasks. arXiv:2405.18870.

[25] Ullman, T. (2023). Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. arXiv:2302.08399.

[26] Wilf, A., et al. (2024). Think Twice: Perspective-Taking Improves LLMs' Theory-of-Mind Capabilities. ACL 2024, pages 8292–8308.

[27] Wu, Y., et al. (2023). Hi-ToM: A Benchmark for Evaluating Higher-Order Theory of Mind Reasoning in LLMs. EMNLP Findings 2023, pages 10691–10706.

[28] Xu, H., et al. (2024). OpenToM: A Comprehensive Benchmark for Evaluating Theory-of-Mind Reasoning Capabilities of LLMs. ACL 2024, pages 8593–8623.

[29] Zhu, W., Zhang, Z., & Wang, Y. (2024). Language Models Represent Beliefs of Self and Others. ICML 2024, pages 62638–62681.