MovieGraph-ToM: Evaluating Long-Range Theory of Mind in Large Language Models via Implicit Social-Causal Graphs

Tingjiang Wei[1*], Qin Ni[2*], Rong Gao[1], Yingying Wang[1], Liang He[3]
[1]Lab of Artificial Intelligence for Education, East China Normal University
[2]Institute of Language Sciences, Shanghai International Studies University
[3]School of Computer Science and Technology, East China Normal University
{52275901031, 52295901036, 52215901020}@stu.ecnc.edu.cn, niqin@shisu.edu.cn, lhe@cs.ecnu.edu.cn

Abstract

The capacity for social reasoning, particularly Theory of Mind (ToM), is a foundational prerequisite for aligning Large Language Models (LLMs) with human values. However, current evaluations are predominantly confined to simplistic, short-text scenarios, obscuring their true capabilities and potential failure modes in complex, long-range social dynamics. To address this deficit, we introduce MovieGraph-ToM, a large-scale benchmark for evaluating long-range ToM and social cognition within extended, multimodal narratives. We employ a "scaffold-and-probe" methodology, and we construct a ground-truth Social-Causal Graph offline, which maps the narrative's latent mental states and causal chains. During evaluation, the model is denied access to this graph and must reason directly from raw multimodal inputs. This decoupling forces genuine inference over superficial pattern matching. Reasoning is probed via a hierarchical questioning framework designed to differentiate spontaneous understanding from logical robustness. Our empirical results reveal systematic vulnerabilities in even state-of-the-art models. We identify a critical multiple-choice pitfall, where accuracy plummets against well-crafted distractors, and a stark "generative-discriminative divide," where models fail to construct coherent explanations for answers they correctly identify. These findings highlight a latent risk, as models that feign comprehension could lead to unpredictable and misaligned behaviors. MovieGraph-ToM thus offers a rigorous platform for assessing and advancing the robust social intelligence required for safely aligned AI systems.

Code — https://github.com/mxdlzg/MovieGraph-ToM

Introduction

Large Language Models (LLMs) have achieved unprecedented success across a vast array of natural language tasks (Brown et al. 2020; Grattafiori et al. 2024; OpenAI et al. 2023; Comanici et al. 2025), igniting a vibrant scientific debate on whether advanced cognitive abilities, such as a Theory of Mind (ToM), can "emerge" in these systems (Bubeck et al. 2023; Sejnowski 2023). ToM, the ability to attribute and reason about the unobservable mental states of others—including their beliefs, desires, and intentions (Premack and Woodruff 1978)—is a cornerstone of human social intelligence and a prerequisite for aligning AI with human values (Wang et al. 2024a). While recent studies suggest that state-of-the-art LLMs can pass classic false-belief tasks (Strachan et al. 2024; Kosinski 2024), the validity and depth of these capabilities remain highly contested.

We contend that the current evaluation paradigm is fundamentally inadequate for measuring the true robustness of these nascent social reasoning abilities. Existing benchmarks are overwhelmingly static and myopic, relying on artificially constructed, isolated short-text stories that fail to capture the complexity of real-world social dynamics (Gandhi et al. 2023a; Chen et al. 2024). This "laboratory-like" environment primarily tests "literal ToM" (predicting behavior in simple scenarios) rather than the adaptive, "functional ToM" required for genuine understanding (Riemer et al. 2025). Compounding this issue, these benchmarks are highly susceptible to data contamination (Deng et al. 2024) and largely ignore the rich, non-verbal cues from vision that are indispensable for human social inference (Villa-Cueva et al. 2025).

To address these limitations, we introduce MovieGraph-ToM, a large-scale benchmark for evaluating long-range, multimodal social reasoning in ecologically valid, full-length narratives. It employs a "scaffold-and-probe" methodology: an offline, ground-truth Social-Causal Graph is constructed to map the narrative's latent mental states and causal chains. During evaluation, the model is denied access to this graph and must reason from raw multimodal inputs (keyframes and script dialogue) alone. This decoupling of the reasoning scaffold from the input forces the model to perform genuine inference, rather than rely on superficial pattern matching.

Our evaluation reveals systematic vulnerabilities in even state-of-the-art models. We identify a critical "multiple-choice pitfall": model accuracy plummets on multiple-choice questions, exposing a fragility to well-crafted distractors not visible in simpler formats. Furthermore, a stark "generative-discriminative divide" emerges, where models can select correct answers but fail to generate coherent, well-supported explanations. Conversely, we observe a promising and uniform robustness in counterfactual reasoning, suggesting that models possess a decoupled logical faculty.

*Corresponding authors.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Together, these results demonstrate that benchmarks must evolve beyond simple accuracy to probe the structural integrity and failure modes of AI reasoning.

Our primary contributions are:

1. **A large-scale, multimodal benchmark, MovieGraph-ToM,** to evaluate the long-range Theory of Mind and social cognition of LLMs. By leveraging full-length films, it provides an unprecedented level of ecological validity and narrative complexity.
2. **A scaffold-and-probe evaluation methodology** that decouples the ground-truth reasoning structure (a Social-Causal Graph) from the model's raw input, compelling genuine inference while mitigating shortcut learning from memorized patterns.
3. **A multi-faceted evaluation framework** combining hierarchical questioning with a two-phase protocol. This framework distinguishes between a model's spontaneous reasoning capabilities and its logical robustness when challenged with targeted counterfactuals.

Related Work

Benchmarks for Theory of Mind and Social Reasoning

The evaluation of ToM in LLMs has progressed from early systematic benchmarks like ToMBench (Chen et al. 2024) and BigToM (Gandhi et al. 2023a), which first highlighted performance gaps and methodological issues. More recent work has sought greater naturalism through LLM role-playing (ToMATO (Shinoda et al. 2025)) or moved towards multimodality with videos (MoMentS (Villa-Cueva et al. 2025)) and images (VCR (Zellers et al. 2019), Social-IQ (Sap et al. 2019)). Concurrently, a critical view has emerged, questioning if benchmarks assess genuine "functional ToM" or merely pattern matching of "psychological dramas"(Riemer et al. 2025; Sejnowski 2023). Concerns over data contamination(Deng et al. 2024) and the oversimplification of mental states like perception (Jung et al. 2024) are also prevalent. MovieGraph-ToM directly addresses these critiques by leveraging feature-length films to evaluate ToM at an unprecedented scale of temporal and causal complexity.

Evaluation of Long-Range Narrative Understanding

From early benchmarks on books and scripts (NarrativeQA (Kočiský et al. 2018), QuALITY (Pang et al. 2022)) to recent evaluations for massive context windows (LongBench v2 (Bai et al. 2025)). However, popular tests like "Needle in a Haystack" are often criticized for testing simple retrieval over reasoning(Wang et al. 2024b). Models still struggle to integrate multiple, disparate pieces of information for causal inference. MovieGraph-ToM is specifically designed to evaluate this deeper, integrative reasoning over long contexts. Our graph construction methodology also draws from techniques for building causal and temporal knowledge graphs from text (Heddaya et al. 2024; Ban et al. 2025; Luo et al. 2024).

Interactive and Agent-based Evaluation

A parallel line of research argues that true social intelligence requires interactive, agent-based evaluation, not just static observation. This approach tests "functional ToM" in complex social games (Zhang et al. 2025; FAIR) and is systematized in frameworks like WebArena (Zhou et al. 2024), AgentGym (Xi et al. 2025), and SocialMaze (Xu et al. 2025). While MovieGraph-ToM is a static benchmark, it serves as a crucial complement to this research. It deeply assesses an agent's ability to comprehend an existing, complex social narrative, a prerequisite for its ability to generate effective social actions. Bridging these two evaluation paradigms—understanding and action—is a key direction for future work.

Methodology

Our methodology unfolds in three stages: (1) constructing a structured, graph-based benchmark from raw movie data; (2) generating hierarchical question trees from the graph to probe complex reasoning; and (3) employing a novel two-phase protocol for model evaluation.

Benchmark Construction Pipeline

The construction of MovieGraph-ToM is a multi-stage pipeline, depicted in Figure 1, that transforms raw cinematic data into a structured, queryable evaluation dataset. Our process begins with the MovieNet (Huang et al. 2020), from which we select a set of movies $\mathcal{M}$. For each movie $m \in \mathcal{M}$, we gather its raw assets: a set of keyframes V_m, the screenplay S_m, and associated metadata A_m (e.g., character lists, scene boundaries). The initial preprocessing stage focuses on structuring and aligning these heterogeneous sources. We parse each screenplay S_m into a temporally ordered sequence of elements $\{e_1, e_2, \ldots, e_k\}$, where each element e_i is tagged with its type (e.g., scene heading, action description, character dialogue). The critical alignment step aims to create synchronized multimodal blocks. By leveraging subtitle timestamps as initial anchors, we employ a hybrid alignment model combining embedding-based sentence similarity with fuzzy string matching to precisely map script elements e_i to their corresponding shot sequences from V_m. This results in a set of synchronized blocks $\{B_j\}$, where each block $B_j = (s_j, v_j, t_j)$ contains a script segment s_j, a sequence of keyframes v_j, and the corresponding timestamp interval t_j.

The core of our methodology is the semi-automated construction of a large-scale Social-Causal Graph $G_m = (N, E)$ for each movie. The nodes N represent key narrative entities, categorized into characters (N_{char}), objects (N_{obj}), locations (N_{loc}), and pivotal events (N_{evt}). The edges E represent the complex relationships between them and are partitioned into two primary types: ToM-based edges E_{ToM} and causal edges E_{causal}. ToM edges capture mental states, such as $e_{tom} = (n_{char}, \text{believes}, n_{evt}, \rho)$, where $n_{char} \in N_{char}$, $n_{evt} \in N_{evt}$, and ρ is the textual or visual evidence. Causal edges represent influence, such as $e_{causal} = (n_{evt,1}, \text{causes}, n_{evt,2}, \rho)$. This graph construction is initiated by a powerful multimodal LLM, Gemini 2.5

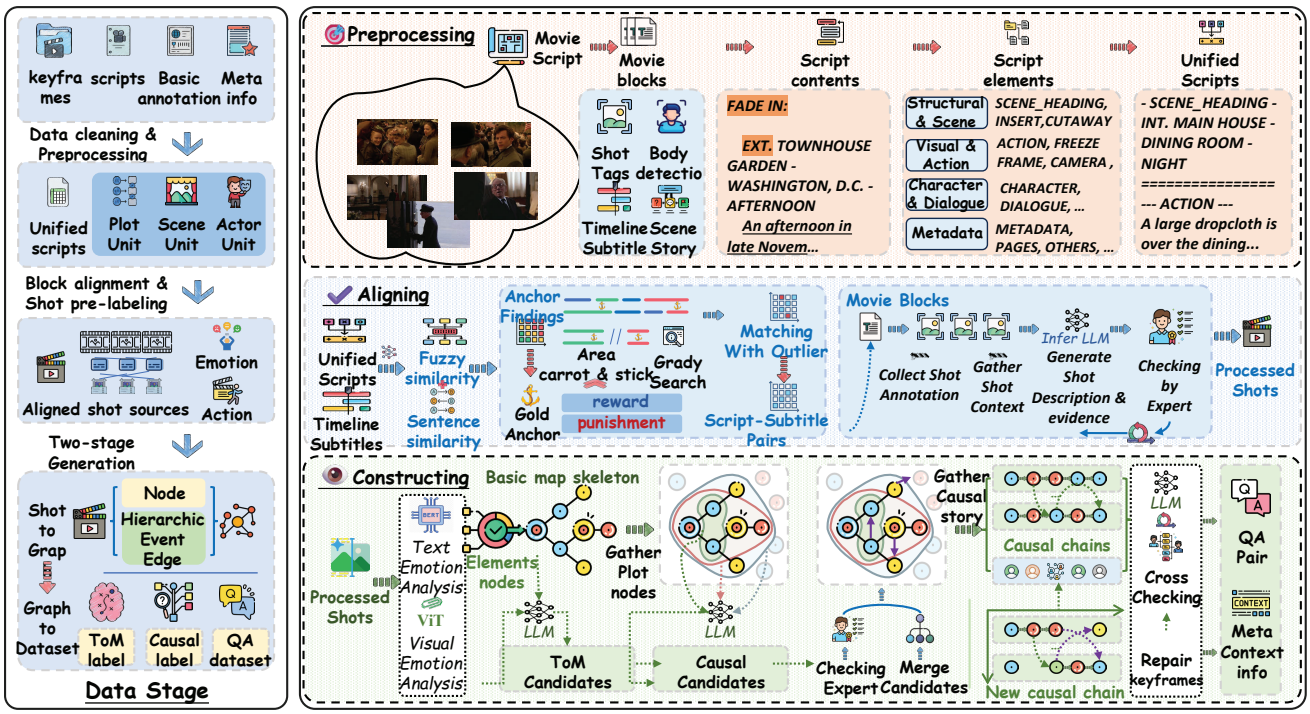

Figure 1: The data construction pipeline for MovieGraph-ToM, from raw movie data to the final graph-based QA dataset.

Pro, which we denote as the generator $\mathcal{L}_{gen}$. For each block B_j, $\mathcal{L}_{gen}$ proposes a set of candidate nodes and edges. To ensure high fidelity, these candidates undergo a rigorous, multi-agent verification process. Each candidate edge is independently evaluated by at least two distinct auditor LLMs, $\mathcal{L}_{aud_1}$ and $\mathcal{L}_{aud_2}$. A candidate is accepted if a consensus is reached. In cases of conflicting assessments, the candidate is escalated to a final adjudication step involving either a human domain expert or a specialized judge model $\mathcal{J}_{exp}$ fine-tuned on expert annotations.

With the verified Social-Causal Graph G_m serving as a knowledge backbone, we generate the final evaluation dataset. Instead of isolated question-answer pairs, we construct hierarchical Question Trees $\mathcal{T}$ to facilitate a deep, multi-step evaluation of a model's reasoning. Each tree is anchored to a significant subgraph, typically a long-range causal chain like $n_1 \xrightarrow{r_1} n_2 \rightarrow \ldots \rightarrow n_k$ or a pivotal ToM-related node. The root of the tree, Q_1, poses a high-level question about the anchor. Each potential answer $A_{1,i}$ to Q_1 (including plausible distractors) leads to a new branch with a follow-up question $Q_{2,i}$ that probes for justification or explores consequences. A path through the tree thus represents a coherent line of reasoning. The generation of these questions is guided by the graph's structure, with questions specifically designed to test understanding of ToM (e.g., "Why did Character A believe X?") and causality ("What was the main reason for Event Y?"). Subsequent questions on branching paths are conditioned on the answers to the previous level, creating a multi-turn, logical dependency. This tree-like structure allows for a deep and contextual evaluation of a model's multi-step reasoning abilities, moving beyond simple fact retrieval to assess its understanding of intricate social and causal dynamics.

Dataset Statistical Analysis

The resulting MovieGraph-ToM benchmark is a large-scale, deeply annotated dataset. Our current version is built upon a corpus of **30** feature-length films, from which we have constructed a substantial underlying knowledge base and a vast set of evaluation questions.

Social-Causal Graph Statistics The foundation of our benchmark is the Social-Causal Graph, which provides the structural backbone for question generation. As detailed in Table 2, the construction process yielded a dense graph structure. Across all films, we generated over 100,000 ToM state candidates and nearly 9,000 potential causal links between events. After a rigorous human-in-the-loop verification process, we retained approximately 48,000 high-quality ToM nodes and 2,300 causal edges. On average, each film is represented by a complex graph containing over 2,600 ToM candidates and 229 causal links, providing a rich substrate for generating nuanced questions.

Question Dataset Distribution From the social-causal graph, we generated a dataset of approximately **65,600** questions designed to probe deep, hierarchical social reasoning.

Cognitive Framework Coverage. As shown in Figure 2, the question distribution is concentrated on high-level cognition, aligning with our research goals. The dominant Level 1 categories are *Behavior Understanding* (28.0%) and *Theory*

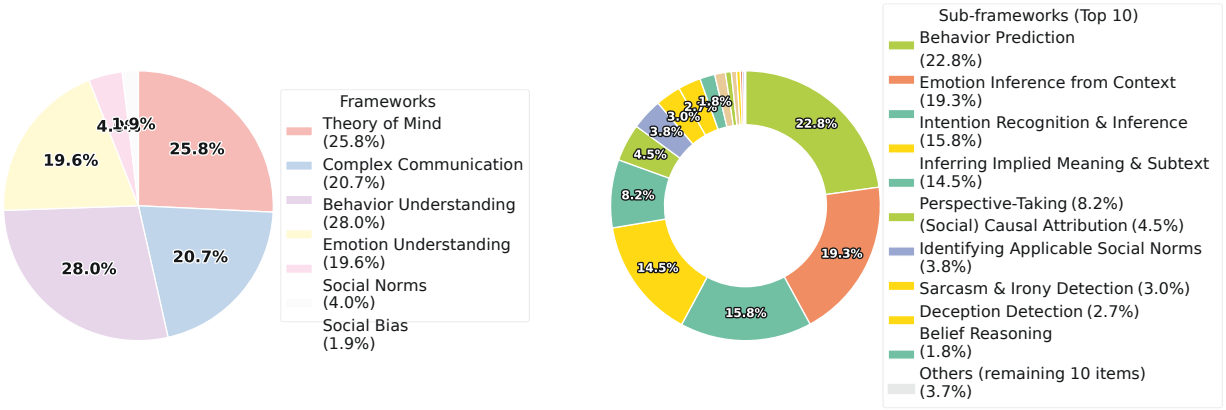

Figure 2: The data distribution of QA types based on multi-level scenes.

Dataset	#QA Pairs	Multi-level QA	Multimodal	Primary QA Types	Source
MovieGraph-ToM (Ours)	**˜65k**	**Yes (Hierarchical)**	**Yes (Image+Text)**	**ToM, Behavior, Social Inference**	**Full-length Movies**
Text-based Theory of Mind & Social Reasoning Benchmarks					
BigToM (Gandhi et al. 2023b)	5k eval.	Primarily 1st-order	No	Belief/Behavior Inference	Synthetic (Controlled)
Hi-ToM (Wu et al. 2023)	1.2k-1.8k stories	Yes (up to 4th-order)	No	Higher-order Beliefs	Stories (Human-AI)
SimpleToM (Gu et al. 2024)	˜3.4k Qs	No (Application-focused)	No	State Inference, Prediction	Stories
PersuasiveToM (Yu et al. 2025)	˜788 Qs	Yes (1st/2nd-order)	No	Belief/Desire/Intent, Strategy	Dialogues
Multimodal Theory of Mind & Social Reasoning Benchmarks					
MoMentS (Villa-Cueva et al. 2025)	˜2.3k	No (7 ToM facets)	Yes (Video)	Emotion, Intent, Belief, etc.	Long Video Clips
SoMi-ToM (Fan et al. 2025)	N/A	Yes (Multi-view)	Yes (Video+State)	Goal/Belief Inference	Embodied Gameplay

Table 1: Comparison of MovieGraph-ToM with recent representative reasoning benchmarks (2022-2025). Our work significantly advances the state-of-the-art in context length, reasoning depth, and multimodal cognitive complexity.

Statistic	Total Count	Average per Movie
ToM Candidates	103,232	2647.0
Plot Point Nodes	2,823	72.4
Causal Links	8,936	229.1
Verified ToM Nodes	48,406 (46.9%)	-
Verified Plot Points	1,614 (57.2%)	-
Verified Causal Links	2,282 (25.5%)	-

Table 2: Overall and per-movie statistics of the Social-Causal Graph construction.

of Mind (25.8%). At the more granular Level 2, key sub-frameworks include *Behavior Prediction* (22.8%), *Emotion Inference* (19.3%), and *Intention Recognition & Inference* (15.8%), ensuring the benchmark rigorously tests core social intelligence.

Question Depth and Format. The dataset's structure is designed to challenge models beyond simple, single-turn QA (Table 3). A significant majority of questions (**59.3%**) are deep follow-ups at Depth 3, requiring understanding of prior context. Furthermore, the prevalence of open-ended questions (**51.1%**) poses a more difficult and realistic challenge by demanding generative reasoning over simple classification.

Depth Distribution		Format Distribution	
Depth 1	˜16.3k (26.7%)	Open-Ended	˜31.5k (51.1%)
Depth 2	˜12.7k (20.5%)	Single-Choice	˜23.6k (38.3%)
Depth 3	˜36.5k (59.3%)	Multi-Choice	˜10.4k (17.0%)
Total	**˜65k**	**Total**	**˜65k**

Table 3: Distribution of questions by their depth in the reasoning tree and by their format.

Task Formulation

The core task is a multimodal, hierarchical question-answering challenge. For each question in a reasoning tree, models are provided not with the entire movie, but with a targeted context snippet containing a montage of crucial keyframes and the relevant script segment. The evaluation proceeds dynamically: starting from the tree's root, the path is determined by the model's answers to a sequence of multiple-choice (MCQ) or open-ended questions. This entire process is governed by our Two-Phase Evaluation Protocol, detailed below.

Two-Phase Evaluation Protocol To comprehensively evaluate a model's reasoning abilities and distinguish between its natural performance and its logical robustness, we designed a novel Two-Phase Evaluation Protocol. This protocol is executed for each question tree in our dataset. The entire interaction occurs within a single, continuous session to maintain conversational context.

Notation for Algorithm 1:

- c_{node}: The current node being processed.
- p_{node}: The parent of a given node.
- N_{exp}: The set of nodes already explored in Phase 1.
- q, A, ctx: Shorthand for a node's question, answer, and context.
- opt_{trig}: The trigger option for a counterfactual question.
- P_{cf}, A_{cf}: The counterfactual prompt and its corresponding answer.
- FormatCFPrompt$(\cdot)$: A function that formats the long counterfactual prompt string.

Phase 1: Dynamic Path Evaluation This phase measures the model's authentic, unguided reasoning by traversing a **single, dynamic path** through the question tree, guided solely by the model's answers (Algorithm 1, lines 3-11). The resulting trajectory is used exclusively to calculate the *Reasoning Depth Score (RDS)* and *Causal Chain Fidelity (CCF)*, capturing the model's natural performance.

Phase 2: Forced Exploration & Counterfactual Challenge This phase assesses the model's logical robustness by systematically exploring all branches **left unvisited** after Phase 1 (Algorithm 1, lines 14-22). We present each unexplored branch as a "thought experiment," providing a counterfactual premise (e.g., "Assuming the previous answer was X...") that forces the model to reason from a new, hypothetical state. The responses are used solely to compute the *Counterfactual Recovery Score (CRS)*.

This two-phase protocol allows us to cleanly separate the data reflecting the model's integrated performance (Phase 1) from the data reflecting its pure, abstract logical reasoning ability (Phase 2), providing a more comprehensive and insightful evaluation.

Algorithm 1: Two-Phase Evaluation Protocol

1: **procedure** EVALUATEQUESTIONTREE($Tree, Model$)
Phase 1: Dynamic Path Evaluation
2: $c_{node} \leftarrow Tree.root$
3: $Path \leftarrow []$
4: **while** $c_{node} \neq$ NULL **do**
5: $A \leftarrow Model.ask(c_{node}.q, c_{node}.ctx)$
6: $Path.append((c_{node}, A))$
7: **if** A is Correct **then**
8: $c_{node} \leftarrow c_{node}.getCorrectFollowUp()$
9: **else**
10: $c_{node} \leftarrow c_{node}.getFollowUpFor(A)$
11: **end if**
12: **end while**
13:
Phase 2: Forced Exploration & Counterfactual Challenge
14: $N_{exp} \leftarrow \{node \mid (node, ans) \in Path\}$
15: **for** $node \in Tree.getNodes()$ **do**
16: **if** $node \notin N_{exp}$ **then**
17: $p_{node} \leftarrow node.getParent()$
18: $opt_{trig} \leftarrow node.getTriggerOption()$
19: $P_{cf} \leftarrow$ FormatCFPrompt$(opt_{trig}, node.q)$
20: $A_{cf} \leftarrow Model.ask(P_{cf}, node.ctx)$
21: ▷ Store A_{cf} for CRS calculation
22: **end if**
23: **end for**
24: **end procedure**

Experiments

Configuration

For each question, the model received sanitized textual context and, for multimodal queries, up to five relevant visual keyframes. All ground-truth ToM labels and causal annotations were stripped from the context to prevent data leakage and ensure a genuine test of reasoning.

The evaluation of open-ended questions employed a hybrid scoring system. A final score was computed as a weighted average from a Qwen3-30B-A3B judge model (70% weight) and a lexical ROUGE-L F1-score (30% weight). We established a correctness threshold of 0.7 for this combined score to balance semantic understanding and factual accuracy. While our primary reported metric is accuracy, we employ a suite of novel metrics to conduct a fine-grained analysis. Detailed definitions are provided in Appendix.

We select a range of state-of-the-art large language and multimodal models for evaluation, including representatives from the deepseek (DeepSeek-AI et al. 2025), gpt (OpenAI et al. 2023), gemini (Comanici et al. 2025), llava (Li et al. 2025), mistral (Mistral-AI et al. 2025), and InternVL (Zhu et al. 2025) families. The specific models tested, such as gemini-2.5-pro and gpt-4.1, are detailed in Appendix.

Results and Comparisons

Our comprehensive evaluation reveals significant nuances in the reasoning capabilities of leading AI models. The results, detailed in Table 4, 5, and 6, For some human evaluation results, see Appendix.

Overall Performance Hierarchy. As shown in Ta-

Model	Overall		Behavior		ToM		Emotion		Communication		Social Norms	
	Acc. (%)	Score	Acc. (%)	Score	Acc. (%)	Score	Acc. (%)	Score	Acc. (%)	Score	Acc. (%)	Score
deepseek-chat-v3-0324	87.8	0.765	86.8	0.737	88.6	0.803	91.2	0.728	88.0	0.762	90.0	0.835
gpt-4.1	89.8	0.784	89.3	0.760	90.4	0.821	92.2	0.738	88.4	0.761	90.7	0.841
gemini-2.5-flash	89.4	0.784	89.2	0.761	90.2	0.822	91.2	0.736	87.2	0.769	90.0	0.831
gemini-2.5-pro	**92.0**	**0.797**	**91.8**	**0.775**	**92.9**	**0.832**	**93.3**	**0.743**	**89.6**	**0.775**	**94.6**	**0.851**
gemini-2.0-flash-001	85.6	0.759	85.2	0.737	86.7	0.798	86.5	0.705	81.6	0.730	87.7	0.815
llava-onevision-qwen2-7b	75.1	0.684	76.8	0.675	72.5	0.693	79.4	0.661	79.7	0.708	74.2	0.728
gpt-4o-mini	85.1	0.762	85.0	0.739	84.4	0.784	88.2	0.719	85.2	0.760	85.4	0.816
InternVL3-8B	83.4	0.746	83.7	0.725	81.5	0.771	86.7	0.704	85.2	0.748	83.1	0.796
qwen2.5-vl-72b	85.9	0.771	86.2	0.749	87.7	0.808	88.0	0.723	83.2	0.756	88.5	0.829
mistral-small-3.2-24b	88.2	0.776	88.5	0.755	88.1	0.806	90.7	0.736	87.2	0.762	91.5	0.835

Table 4: Performance comparison of various models on the top 5 categories.

Model	Format			Type				Depth					
	Single	Multi.	Open	Basic ToM		Complex		Depth 1		Depth 2		Depth 3	
	Acc. %	Acc. %	Score	Acc. %	Score	Acc. %	Score	Acc. %	Score	Acc. %	Score	Acc. %	Score
deepseek-chat-v3-0324	91.6	61.5	0.662	81.5	0.845	86.9	0.838	83.0	0.843	84.6	0.746	90.9	0.740
gpt-4.1	93.1	66.9	0.675	88.2	0.892	88.5	0.879	88.3	0.890	84.1	0.745	92.3	0.754
gemini-2.5-flash	94.5	66.2	0.669	84.0	0.855	88.5	0.865	85.2	0.857	89.0	0.789	91.3	0.752
gemini-2.5-pro	**96.4**	**70.8**	**0.674**	**88.9**	**0.891**	**93.4**	**0.894**	**90.1**	**0.892**	**90.1**	**0.792**	**93.3**	**0.760**
gemini-2.0-flash-001	94.2	55.4	0.638	80.9	0.840	77.1	0.812	79.8	0.833	82.4	0.741	89.1	0.735
llava-onevision-qwen2-7b	84.6	26.9	0.622	52.0	0.629	45.6	0.586	50.2	0.617	77.7	0.711	84.5	0.702
gpt-4o-mini	93.8	35.4	0.664	74.1	0.807	75.4	0.793	74.4	0.803	84.1	0.761	89.8	0.745
InternVL3-8B	91.9	33.1	0.656	69.1	0.760	68.9	0.789	69.1	0.768	84.6	0.750	88.9	0.736
qwen2.5-vl-72b	95.5	50.8	0.654	75.9	0.830	85.3	0.843	78.5	0.833	83.5	0.762	89.8	0.749
mistral-small-3.2-24b	95.1	53.0	0.667	78.3	0.816	86.8	0.860	80.7	0.828	85.7	0.773	92.2	0.756

Table 5: Detailed performance breakdown by question format, type, and reasoning depth. "Complex" refers to Complex Reasoning.

Model Name	Reasoning Depth		Causal Chain Fidelity						Counterfactual Recovery Score
	Score	Weighted Score	Depth 1		Depth 2		Depth 3		
			Acc.	Avg. Score	Acc.	Avg. Score	Acc.	Avg. Score	
deepseek-chat-v3-0324	**2.918**	1.754	0.830	0.843	0.846	0.746	0.909	0.740	0.880
gpt-4.1	2.852	1.789	0.883	0.889	0.840	0.745	0.923	0.753	0.893
gemini-2.5-flash	2.820	1.796	0.852	0.857	0.890	0.789	0.913	0.752	0.902
gemini-2.5-pro	2.852	**1.818**	**0.901**	**0.892**	**0.905**	**0.792**	**0.933**	**0.760**	**0.918**
gemini-2.0-flash-001	2.623	1.742	0.798	0.832	0.824	0.741	0.891	0.735	0.863
llava-onevision-qwen2-7b	2.175	1.621	0.502	0.617	0.776	0.711	0.845	0.702	0.828
gpt-4o-mini	2.787	1.759	0.744	0.803	0.841	0.761	0.898	0.745	0.875
InternVL3-8B	2.787	1.731	0.691	0.768	0.846	0.750	0.889	0.736	0.866
qwen2.5-vl-72b	2.852	1.773	0.785	0.833	0.835	0.762	0.898	0.749	0.870
mistral-small-3.2-24b	2.836	1.790	0.807	0.828	0.857	0.774	0.920	0.757	0.895

Table 6: Advanced and evaluative metrics. Counterfactual Recovery Score measures the model's ability to adapt to counter-intuitive scenarios. Reasoning Depth Scores quantify the complexity of the model's reasoning chain.

ble 4, `gemini-2.5-pro` establishes itself as the definitive state-of-the-art model, securing the highest overall accuracy (92.0%) and score (0.797). It consistently leads across all five major categories, demonstrating a well-rounded and superior capacity for social and causal reasoning. A competitive second tier, including `gpt-4.1` (89.8%), `gemini-2.5-flash` (89.4%), and `mistral-small-3.2-24b` (88.2%), follows closely. Smaller models, while capable, show a discernible performance gap, underscoring the role of scale in achieving advanced reasoning.

The Multiple-Choice Pitfall. One of the most striking findings, presented in Table 5, is the dramatic performance collapse when transitioning from single-choice (`Single`)

to multiple-choice (Multi.) questions. While most models perform admirably on single-choice tasks (often ¿90% accuracy), their ability to correctly answer multiple-choice questions plummets. For example, gpt-4o-mini's accuracy catastrophically drops from 93.8% to 35.4%, and llava-onevision-qwen2-7b falls from 84.6% to a near-chance level of 26.9%. Even the top-performing gemini-2.5-pro experiences a significant decline from 96.4% to 70.8%. This "multiple-choice pitfall" suggests that models are highly susceptible to carefully crafted distractors. They may excel at identifying a correct statement in isolation but struggle with the more complex task of discernment and elimination required by multiple-choice formats.

Generative vs. Discriminative Reasoning. Performance on open-ended (Open) questions, evaluated by a hybrid score, reveals another dimension of model capability. Notably, no model's average score on open-ended questions surpasses the 0.7 correctness threshold, with the top models like gpt-4.1 (0.675) and gemini-2.5-pro (0.674) hovering just below it. This indicates a significant gap between discriminative ability (selecting an answer) and generative ability (constructing a high-quality, factually accurate, and semantically coherent explanation from scratch).

Reasoning Depth and Complexity. Our analysis of reasoning depth (Table 5 and 6) confirms that performance does not degrade linearly with causal chain length. While most models exhibit a slight dip in accuracy at Depth 2, they often recover or even improve at Depth 3. However, the explanation quality, measured by the 'Avg. Score', shows a more consistent downward trend as depth increases (gemini-2.5-pro's score drops from 0.892 at Depth 1 to 0.760 at Depth 3). This divergence suggests that while models can identify correct outcomes in longer reasoning chains, their ability to articulate the underlying causal steps deteriorates, potentially indicating a reliance on heuristics rather than step-by-step reasoning.

Discussion

The Frontier of Reasoning and the Role of Scale. The superior performance of gemini-2.5-pro reinforces the hypothesis that advanced reasoning is an emergent property of scale, sophisticated architectures, and high-quality training data. Its consistent lead suggests a more generalized and robust reasoning faculty. However, our results also challenge a simplistic view of "capability," demonstrating that even frontier models have significant and systematic weaknesses.

Question Format as a Revealing Probe for Reasoning Fragility. The "multiple-choice pitfall" is arguably our most critical finding. It suggests that high accuracy on single-choice questions may be a misleading indicator of true understanding. The models' vulnerability to distractors implies a potential reliance on shallow semantic matching rather than deep, structural reasoning. This has profound implications for real-world deployment: in high-stakes environments where decisions involve weighing multiple plausible but flawed options, these models could be dangerously unreliable. Their failure highlights a critical need to move beyond simple accuracy metrics and develop benchmarks that specifically test for **reasoning robustness against adversarial or distracting information**.

The Generative-Discriminative Divide. The contrast between high accuracy in discriminative tasks and lower scores in open-ended generation points to a fundamental divide between two cognitive functions: recognition/discrimination and recall/generation. Models appear far more adept at identifying correctness within a constrained set of options than they are at articulating a correct and coherent line of reasoning from an unconstrained space. Bridging this gap is essential for building AI systems that can not only find answers but also explain, teach, and collaborate with humans in a meaningful way.

Deconstructing the Reasoning Process: Depth, Complexity, and Heuristics. The model's reasoning process appears to degrade with complexity, even when final accuracy remains high. An accuracy dip at moderate (Depth 2) causal chains suggests a "cognitive trap," where simple heuristics fail but deeper reasoning is not yet engaged. Furthermore, the divergence between stable accuracy and declining explanation scores at greater depths indicates that models may be "jumping to conclusions"—predicting outcomes without faithfully modeling the intermediate causal steps. This gap between a correct outcome and a correct process is a critical barrier to developing explainable and trustworthy AI.

The Promise of Logical Robustness. On a positive note, the consistently high Counterfactual Recovery Scores (CRS) across models are highly encouraging. The ability to reason coherently from a hypothetical, non-factual premise demonstrates a crucial form of abstract logical machinery. This confirms that models possess a foundational capability for flexible reasoning that is decoupled from memorized facts, a cornerstone for tasks like planning, creativity, and safe adaptation to novel scenarios.

Conclusion

Our evaluation reveals that while leading AI models achieve high overall performance on complex reasoning tasks, their competence is fragile. This fragility manifests in two key failure modes: a "multiple-choice pitfall," where accuracy collapses against well-crafted distractors, and a "generative-discriminative divide," where models fail to construct sound arguments for answers they correctly select. In contrast, these models exhibit a promising and uniform robustness in counterfactual reasoning, suggesting a decoupled logical faculty.

A key limitation of this work is the reliance on an LLM-based judge, which may introduce evaluation biases. Our findings point to three critical directions for future research: 1) designing adversarial benchmark tasks, particularly for multiple-choice formats, to systematically probe and improve model robustness; 2) conducting mechanistic interpretability studies to understand the cause of the paradoxical reasoning behaviors observed in our experiments. 3) designing more complex and adversarial benchmark tasks rather than single-choice, particularly for multiple-choice and open-ended formats;

Acknowledgments

This work was supported by the Special Project on Artificial Intelligence-Driven Research Paradigm Reform and Discipline Leapfrog Development Empowerment by Shanghai Municipal Education Commission.

References

Bai, Y.; Tu, S.; Zhang, J.; Peng, H.; Wang, X.; Lv, X.; Cao, S.; Xu, J.; Hou, L.; Dong, Y.; Tang, J.; and Li, J. 2025. LongBench v2: Towards Deeper Understanding and Reasoning on Realistic Long-context Multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3639–3664. Association for Computational Linguistics.

Ban, T.; Chen, L.; Lyu, D.; Wang, X.; Zhu, Q.; Tu, Q.; and Chen, H. 2025. Integrating Large Language Model for Improved Causal Discovery. *IEEE Transactions on Artificial Intelligence*, 6(11): 3030–3042.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712.

Chen, Z.; Wu, J.; Zhou, J.; Wen, B.; Bi, G.; Jiang, G.; Cao, Y.; Hu, M.; Lai, Y.; Xiong, Z.; and Huang, M. 2024. ToMBench: Benchmarking Theory of Mind in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15959–15983.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv:2507.06261.

DeepSeek-AI; Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; et al. 2025. DeepSeek-V3 Technical Report. arXiv:2412.19437.

Deng, C.; Zhao, Y.; Tang, X.; Gerstein, M.; and Cohan, A. 2024. Investigating Data Contamination in Modern Benchmarks for Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8706–8719. Association for Computational Linguistics.

(FAIR)†, M. F. A. R. D. T.; Bakhtin, A.; Brown, N.; Dinan, E.; Farina, G.; Flaherty, C.; Fried, D.; Goff, A.; Gray, J.; Hu, H.; Jacob, A. P.; Komeili, M.; Konath, K.; Kwon, M.; Lerer, A.; Lewis, M.; Miller, A. H.; Mitts, S.; Renduchintala, A.; Roller, S.; Rowe, D.; Shi, W.; Spisak, J.; Wei, A.; Wu, D.; Zhang, H.; and Zijlstra, M. 2022. Human-level play in the game of ¡i¿Diplomacy¡/i¿ by combining language models with strategic reasoning. *Science*, 378(6624): 1067–1074.

Fan, X.; Zhou, X.; Jin, C.; Nottingham, K.; Zhu, H.; and Sap, M. 2025. SoMi-ToM: Evaluating Multi-Perspective Theory of Mind in Embodied Social Interactions. arXiv:2506.23046.

Gandhi, K.; Fraenken, J.-P.; Gerstenberg, T.; and Goodman, N. 2023a. Understanding Social Reasoning in Language Models with Language Models. In *Advances in Neural Information Processing Systems*, volume 36, 13518–13529. Curran Associates, Inc.

Gandhi, K.; Fraenken, J.-P.; Gerstenberg, T.; and Goodman, N. 2023b. Understanding Social Reasoning in Language Models with Language Models. In *Advances in Neural Information Processing Systems*, volume 36, 13518–13529. Curran Associates, Inc.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Gu, Y.; Tafjord, O.; Kim, H.; Moore, J.; Bras, R. L.; Clark, P.; and Choi, Y. 2024. SimpleToM: Exposing the Gap between Explicit ToM Inference and Implicit ToM Application in LLMs. arXiv:2410.13648.

Heddaya, M.; Zeng, Q.; Zentefis, A.; Voigt, R.; and Tan, C. 2024. Causal Micro-Narratives. In *Proceedings of the 6th Workshop on Narrative Understanding*, 67–84. Association for Computational Linguistics.

Huang, Q.; Xiong, Y.; Rao, A.; Wang, J.; and Lin, D. 2020. MovieNet: A Holistic Dataset for Movie Understanding. In *Computer Vision – ECCV 2020*, 709–727. Springer International Publishing.

Jung, C.; Kim, D.; Jin, J.; Kim, J.; Seonwoo, Y.; Choi, Y.; Oh, A.; and Kim, H. 2024. Perceptions to Beliefs: Exploring Precursory Inferences for Theory of Mind in Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 19794–19809. Association for Computational Linguistics.

Kočiský, T.; Schwarz, J.; Blunsom, P.; Dyer, C.; Hermann, K. M.; Melis, G.; and Grefenstette, E. 2018. The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association for Computational Linguistics*, 6: 317–328.

Kosinski, M. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45): e2405460121.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; and Li, C. 2025. LLaVA-OneVision: Easy Visual Task Transfer. *Transactions on Machine Learning Research*.

Luo, K.; Zhou, T.; Chen, Y.; Zhao, J.; and Liu, K. 2024. Open Event Causality Extraction by the Assistance of LLM in Task Annotation, Dataset, and Method. In *Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning*

(NeusymBridge) @ LREC-COLING-2024, 33–44. ELRA and ICCL.

Mistral-AI; :; Rastogi, A.; Jiang, A. Q.; Lo, A.; Berrada, G.; Lample, G.; Rute, J.; et al. 2025. Magistral. arXiv:2506.10910.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.; et al. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Pang, R. Y.; Parrish, A.; Joshi, N.; Nangia, N.; Phang, J.; Chen, A.; Padmakumar, V.; Ma, J.; Thompson, J.; He, H.; and Bowman, S. 2022. QuALITY: Question Answering with Long Input Texts, Yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5336–5358. Association for Computational Linguistics.

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

Riemer, M.; Ashktorab, Z.; Bouneffouf, D.; Das, P.; Liu, M.; Weisz, J. D.; and Campbell, M. 2025. Position: Theory of Mind Benchmarks are Broken for Large Language Models. In *Forty-second International Conference on Machine Learning Position Paper Track*.

Sap, M.; Rashkin, H.; Chen, D.; Le Bras, R.; and Choi, Y. 2019. Social IQa: Commonsense Reasoning about Social Interactions. 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)*, 4463–4473. Association for Computational Linguistics.

Sejnowski, T. J. 2023. Large Language Models and the Reverse Turing Test. *Neural Computation*, 35(3): 309–342.

Shinoda, K.; Hojo, N.; Nishida, K.; Mizuno, S.; Suzuki, K.; Masumura, R.; Sugiyama, H.; and Saito, K. 2025. ToMATO: Verbalizing the Mental States of Role-Playing LLMs for Benchmarking Theory of Mind. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2): 1520–1528.

Strachan, J. W.; Albergo, D.; Borghini, G.; Pansardi, O.; Scaliti, E.; Gupta, S.; Saxena, K.; Rufo, A.; Panzeri, S.; Manzi, G.; et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7): 1285–1295.

Villa-Cueva, E.; Ahmed, S. M. M.; Chevi, R.; Cruz, J. C. B.; Elzeky, K.; Cristobal, F.; Aji, A. F.; Wang, S.; Mihalcea, R.; and Solorio, T. 2025. MoMentS: A Comprehensive Multimodal Benchmark for Theory of Mind. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 22591–22611. Suzhou, China: Association for Computational Linguistics.

Wang, Q.; Walsh, S.; Si, M.; Kephart, J.; Weisz, J. D.; and Goel, A. K. 2024a. Theory of Mind in Human-AI Interaction. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24. Association for Computing Machinery.

Wang, W.; Zhang, S.; Ren, Y.; Duan, Y.; Li, T.; Liu, S.; Hu, M.; Chen, Z.; Zhang, K.; Lu, L.; Zhu, X.; Luo, P.; Qiao, Y.; Dai, J.; Shao, W.; and Wang, W. 2024b. Needle In A Multimodal Haystack. In *Advances in Neural Information Processing Systems*, volume 37, 20540–20565. Curran Associates, Inc.

Wu, Y.; He, Y.; Jia, Y.; Mihalcea, R.; Chen, Y.; and Deng, N. 2023. Hi-ToM: A Benchmark for Evaluating Higher-Order Theory of Mind Reasoning in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10691–10706. Singapore: Association for Computational Linguistics.

Xi, Z.; Ding, Y.; Chen, W.; Hong, B.; Guo, H.; Wang, J.; Guo, X.; Yang, D.; Liao, C.; He, W.; Gao, S.; Chen, L.; Zheng, R.; Zou, Y.; Gui, T.; Zhang, Q.; Qiu, X.; Huang, X.; Wu, Z.; and Jiang, Y.-G. 2025. AgentGym: Evaluating and Training Large Language Model-based Agents across Diverse Environments. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 27914–27961. Association for Computational Linguistics.

Xu, Z.; Wang, Y.; Huang, Y.; Ye, J.; Zhuang, H.; Song, Z.; Gao, L.; Wang, C.; Chen, Z.; Zhou, Y.; Li, S.; Pan, W.; Zhao, Y.; Zhao, J.; Zhang, X.; and Chen, X. 2025. SocialMaze: A Benchmark for Evaluating Social Reasoning in Large Language Models. In *First Workshop on Social Simulation with LLMs*.

Yu, J.; Li, Z.; Chen, Z.; Zhan, H.; Yang, Z.; Hu, Z.; Tsujii, J.; Xiao, C.; and Xing, E. P. 2025. PersuasiveToM: A Benchmark for Evaluating Machine Theory of Mind in Persuasive Dialogues. arXiv:2502.21017.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6713–6724.

Zhang, Z.; Lan, Y.; Chen, Y.; Wang, L.; Wang, X.; and Wang, H. 2025. DVM: Towards Controllable LLM Agents in Social Deduction Games. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Zhou, S.; Xu, F. F.; Zhu, H.; Zhou, X.; Lo, R.; Sridhar, A.; Cheng, X.; Ou, T.; Bisk, Y.; Fried, D.; Alon, U.; and Neubig, G. 2024. WebArena: A Realistic Web Environment for Building Autonomous Agents. In *The Twelfth International Conference on Learning Representations*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; et al. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. arXiv:2504.10479.