Final Report: Eliciting Collaborative or Competitive Behavior in Large Language Models through a Hint-Based Game

Subhadip Ghosh, Akshat Dasula, Lavanya Sekar, Sivapriya Gopi
Team name: Semantix

Abstract

Large language models (LLMs) are increasingly deployed in interactive settings where multiple agents exchange information and adapt their behavior over time. While prior work has studied collaboration or deception in isolation, less is understood about how these behaviors emerge jointly under explicit incentives. This project presents a lightweight, prompt-based framework for analyzing cooperation, trust, and deception in repeated interactions between LLM-based agents. Agents communicate through constrained hints and receive feedback via a trust signal and an automated judge, enabling behavioral adaptation without model fine-tuning. Experiments on arithmetic reasoning tasks show that larger models adjust their communication strategies based on incentive structure and interaction history, exhibiting cooperative behavior under aligned incentives and strategic deception under competition. Smaller models fail to demonstrate stable or interpretable social behavior, highlighting capacity limitations. Overall, the results suggest that controlled interaction design can elicit socially meaningful behaviors from LLMs, while also exposing important limitations.

Code and Resources: GitHub Repository | Project Website

1 Introduction and Motivation

Large Language Models (LLMs) have achieved strong performance across a wide range of natural language processing tasks, including text generation, reasoning, question answering, and dialogue. Most existing evaluations, however, study LLMs in isolation, where a single model responds to static prompts or interacts only with a human user. In contrast, many emerging applications involve multi-agent environments, where multiple autonomous language agents interact repeatedly, exchange information, and adapt their behavior based on feedback and incentives. Understanding how LLMs behave in such interactive settings is therefore essential for evaluating their robustness, alignment, and suitability for deployment in complex real-world systems.

A key open question in this context is whether LLMs exhibit cooperative, trust-building behavior when interacting with other agents, or whether they adopt competitive or deceptive strategies to maximize their own objectives. Unlike explicitly programmed agents, LLMs are trained on large-scale human-generated text and implicitly absorb patterns of social interaction, persuasion, and strategic communication. As a result, behaviors such as cooperation, reciprocity, partial truthfulness, or deception may emerge without being explicitly designed. Studying these emergent behaviors provides insight into the social intelligence of LLMs and reveals potential risks when such models are placed in settings involving shared incentives.

In this project, we explore these questions using a controlled **hint-exchange game** involving repeated interactions between two LLM-based agents. In each round, one agent provides a hint intended to assist the other agent in solving a task, while both agents operate under an incentive structure that rewards performance. This repeated-game formulation allows us to observe how communication strategies evolve over time, including the development of trust, reciprocal cooperation, strategic withholding of information, or deliberate deception. By analyzing behavior across multiple rounds rather than isolated interactions, we capture long-term adaptation effects that are central to social and economic models of cooperation.

To enable systematic and scalable evaluation, we introduce a third model that acts as an **LLM-as-a-Judge**. This judge model evaluates the exchanged hints along dimensions such as helpfulness, relevance, and potential deceptiveness, producing quantitative scores that reflect cooperative or adversarial intent. This design avoids reliance on human

annotation while maintaining consistency across evaluations, making it well suited for repeated experiments and ablation studies.

The proposed framework is grounded in established theories of communication and social behavior. **Social Exchange Theory** Homans (1958) models cooperation as a function of reciprocal benefit and perceived fairness between interacting parties. **Reciprocity and Trust Models** Ostrom (2003) from behavioral economics explain how cooperation or betrayal emerges through repeated interactions under uncertainty and incentive constraints. Additionally, **Grice's Maxims of Cooperative Communication** Grice (1975) characterize effective dialogue in terms of truthfulness, relevance, and informativeness. By embedding these principles into a computational multi-agent setting, this work examines whether LLMs conform to, approximate, or diverge from human-like patterns of trust formation and strategic communication.

Overall, this project contributes to a growing body of work on multi-agent NLP by providing a structured framework for analyzing cooperation and deception in LLM interactions. The findings offer insights into the emergent social behaviors of language models and highlight important considerations for deploying LLMs in collaborative, competitive, or adversarial environments.

2 Literature Review

Recent research has increasingly examined communication, coordination, and strategic behavior among language-based agents. While these studies demonstrate that LLMs can engage in multi-agent interaction, much of the existing literature prioritizes task performance, dialogue fluency, or win rates over explicit modeling of social behaviors such as trust formation, reciprocity, and deception. This section reviews three representative research directions that inform our work: multi-agent reasoning benchmarks, open-ended deception analysis, and reinforcement learning based social deduction environments.

Multi-Agent Collaboration and Competition. Zhu et al. (2025) introduced the **MARBLE** benchmark, a large-scale evaluation framework designed to study interactions among multiple LLMs across collaborative and competitive tasks, including coding, negotiation, and research planning. MARBLE provides structured turn-based protocols and shared reasoning objectives, enabling systematic evaluation of coordination efficiency and communication diversity. Their results show that LLMs can exhibit emergent cooperation when incentives and roles are aligned, while competitive settings often lead to degraded coherence and reduced fairness. However, MARBLE primarily evaluates interaction quality at the task level and does not incorporate explicit reward signals or longitudinal adaptation, limiting its ability to capture how trust, reciprocity, or deception evolve across repeated interactions.

Deceptive Behavior and Intent. Wu et al. (2025) proposed the **OpenDeception** framework to benchmark deceptive behavior in LLMs using open-ended role-play scenarios. The authors introduce quantitative metrics such as *Deceptive Intention Rate (DIR)* and *Deceptive Success Rate (DeSR)*, demonstrating that even alignment-tuned models can produce misleading or manipulative responses under goal-driven prompts. While this work provides important evidence that deception can emerge in LLM outputs, its evaluation is limited to single-turn interactions without feedback, reciprocity, or repeated exposure. As a result, deception is analyzed in isolation rather than as part of an adaptive strategy shaped by interaction history or social context.

Emergent Social Reasoning via Reinforcement Learning. Sarkar et al. (2025) studied social communication in hidden-role environments inspired by the game *Among Us*, training LLM-based agents using multi-agent reinforcement learning. Their framework decomposes communication into speaking and listening components, with rewards assigned based on how effectively statements update team beliefs. The resulting agents display human-like social behaviors such as accusation, defense, alliance formation, and strategic misinformation, achieving substantially higher win rates than standard reinforcement learning baselines. Despite these strengths, the approach relies on environment-specific reward engineering and fine-tuned recurrent architectures, which limits interpretability and generalization to open-ended language-only settings.

Limitations of Prior Work. Taken together, existing approaches tend to study collaboration and deception as separate phenomena, rather than as intertwined behaviors emerging from shared incentive structures. Communication quality is often assessed indirectly through task success or qualitative analysis, leaving trust dynamics and reciprocity largely unquantified. Furthermore, many frame-

works rely on reinforcement learning or fine-tuning, which obscures causal analysis and reduces reproducibility. These limitations motivate the need for a lightweight, prompt-driven, multi-agent framework that explicitly measures cooperation, trust, and deception over repeated interactions without modifying model parameters.

3 Methodology

This work proposes a prompt-based experimental framework for studying cooperation, trust, and deception in repeated interactions between large language models (LLMs). Rather than modifying model parameters or introducing new training objectives, our approach focuses on structuring the interaction environment itself. The core idea is that social behaviors such as cooperation and deception can emerge naturally when agents interact repeatedly under explicit incentive structures.

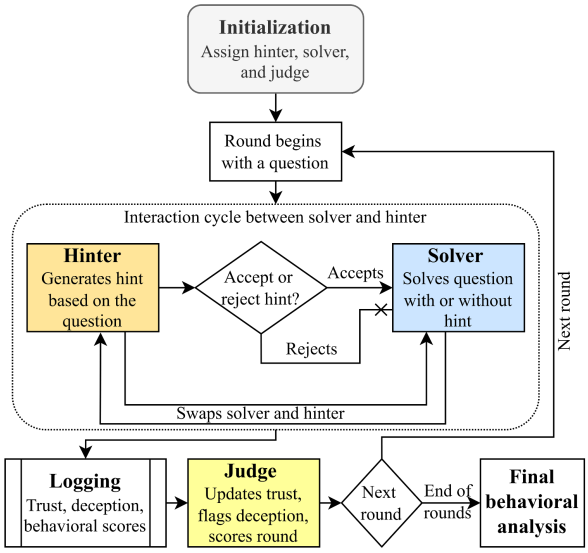

Figure 1: Proposed framework of the multi-agent interaction environment.

Figure 1 provides a schematic overview of the proposed interaction framework. It shows the repeated exchange between a Hinter and a Solver, the constrained hint channel, the evolving trust signal, and the role of the LLM-as-a-Judge in providing feedback. This diagram clarifies how social signals and incentives are injected at prompt time without modifying model parameters.

Our hypothesis is that LLMs, when placed in a repeated-game setting with constrained communication and feedback, will adapt their communication strategies in response to incentives. Specifically, we expect agents to exhibit cooperative behavior under aligned incentives, strategic deception under misaligned incentives, and trust-sensitive adaptation over time, even without fine-tuning or reinforcement learning.

3.1 Unified Interaction Framework

Prior work typically studies collaboration and deception as separate phenomena. Collaborative benchmarks emphasize coordination quality and task success, while deception-focused studies analyze misleading behavior in isolation, often in single-turn settings. Our framework integrates both behaviors within a single interaction loop.

Agents participate in a multi-round hint-exchange game where each round affects future interactions through accumulated trust and score differences. This unified setup allows us to observe not only whether deception or cooperation occurs, but when it occurs, under what conditions, and how it evolves across rounds. In particular, it enables analysis of trust formation, strategic betrayal, and potential trust recovery, which are not observable in single-turn or static evaluations.

3.2 Constrained Hint-Based Communication

A key design choice in our approach is the use of a constrained communication channel. In each round, the Hinter is allowed to send a short hint of at most 30 tokens to the Solver. This constraint serves two purposes.

First, it prevents the Hinter from directly solving the task on behalf of the Solver, ensuring that success depends on selective information sharing rather than full solution disclosure. Second, it isolates communicative intent from raw reasoning ability. Because hints are brief and optional, differences in outcomes can be more clearly attributed to the strategic choice of what information to reveal or withhold, rather than to differences in reasoning depth or verbosity.

This low-bandwidth, task-agnostic communication protocol is central to our analysis and distinguishes our setup from existing multi-agent benchmarks that allow unrestricted dialogue.

3.3 Prompt-Time Behavioral Adaptation

To enable adaptation without training, we introduce a prompt-level trust variable that is updated after each round and included in subsequent prompts. This trust score reflects the Solver's assessment of the Hinter's past behavior and influences how both agents reason about future interactions.

Importantly, this mechanism does not involve any parameter updates. All adaptation occurs through in-context learning, making the framework lightweight, transparent, and easy to reproduce. This design choice allows us to study the inherent social reasoning capabilities of LLMs without confounding effects from fine-tuning or environment-specific reward shaping.

3.4 Incentive Regimes

The framework supports multiple incentive regimes, allowing us to test how agent behavior changes under different reward structures.

In the Collaborative regime, both agents receive identical rewards based on Solver accuracy, encouraging truthful and helpful communication. In the Competitive regime, agents receive opposing rewards, creating incentives for strategic misinformation and deception. By switching between these regimes, we examine whether cooperative strategies persist, degrade, or adapt when incentives change.

This ability to place the same agents in shifting social contexts is critical for studying adaptive behavior and is largely absent from existing benchmarks.

3.5 Cross-Model Interactions and Behavioral Telemetry

We evaluate both homogeneous and heterogeneous agent pairings, including cross-model dyads such as Gemini–Groq. This allows us to study asymmetric behaviors, such as differing tendencies toward deception, forgiveness, or punishment.

Throughout the interaction, we log detailed behavioral signals, including trust trajectories, hint usage, deception frequency, solver accuracy, and a linguistic honesty indicator derived from agent justifications. Together, these signals form a structured dataset that enables fine-grained analysis of social behavior beyond simple task success.

3.6 Scope and Limitations of the Approach

Our approach is intentionally minimalistic. It does not claim to model real-world social interaction in full complexity, nor does it aim to optimize performance. Instead, it is designed to expose whether and how social strategies emerge under controlled conditions.

Because the framework relies on prompt-based interaction and an automated judge model, it may be sensitive to prompt wording and evaluation bias. These limitations are discussed further in later sections, along with directions for extending the framework to richer domains and more complex multi-agent settings.

4 Experiments and Results

This section evaluates whether the proposed multi-round hint-exchange framework elicits meaningful and interpretable social behaviors from large language models. Our goal is not to maximize task accuracy, but to analyze how cooperation, trust, and deception emerge under different incentive structures. All experiments follow the interaction pipeline shown in Figure 1, ensuring consistency between the methodological design and empirical evaluation.

4.1 Experimental Setup

Experiments were conducted on Gemini(Google DeepMind, 2024), Groq(Groq Inc., 2024), Phi(Microsoft Research, 2024), Qwen(Qwen Team, Alibaba Cloud, 2024), and Mistral(Mistral AI, 2023) to analyse the social behaviors from the LLMs. Each experiment involved repeated interactions between a Hinter and a Solver over multiple rounds, with tasks drawn from the GSM8K and Competition Math datasets. Incentive regimes were alternated between collaborative and competitive settings, as described in the Methodology section.

Initial exploratory experiments revealed a clear limitation: smaller models (approximately 7B parameters or fewer) failed to exhibit stable or interpretable social behavior. Their actions often appeared random, and their stated rationales did not align with their outputs. For example, smaller models frequently labeled hints as deceptive while providing vague or incoherent explanations, such as:

'hint_type': 'deceptive',

'reason_for_hint_type': 'To keep the information flow and challenge the solver with indirect reasoning, encouraging them to deduce the prices through logical steps rather than being immediately guided.'

In such cases, the explanation does not meaningfully justify why deception would be advantageous in the given game state. Due to this instability, our primary analysis focuses on larger models, particularly Gemini and Groq, which consistently produced context-aware and strategically grounded behavior.

4.2 Evaluation Metrics

We evaluate success using behavioral metrics rather than raw task performance alone. These include Solver accuracy, frequency of deceptive hints, trust score trajectories, and judge-derived deception scores. The intent is to measure how agents adapt their communication strategies over time in response to incentives and interaction history.

Table 1 summarizes the primary quantitative metrics aggregated across all interaction rounds. In addition, Table 2 highlights systematic behavioral differences observed across models. Alongside these quantitative measures, we perform qualitative analysis by examining agent-generated justifications, hint content, and Solver responses to assess whether the behaviors are interpretable and human-plausible.

4.3 Qualitative Behavioral Analysis

A key qualitative finding is that agents treat trust as a strategic variable rather than a static or moral constraint. For example, Gemini (Agent 0) frequently alternated between truthful and deceptive hints depending on the round number, score difference, and current trust rating. In early rounds, Agent 0 often explicitly justified truthful behavior with statements such as:

"At the start of the game with a neutral trust rating, I am following the feedback to build trust and encourage reciprocal cooperation in early rounds."

This strategy manifested as accurate and helpful hints during the initial rounds. In later rounds, particularly when scores were tied or trust was already low, the same agent recorded different reasoning, such as:

"Given that scores are tied and my trust rating is low, this is an opportune moment to introduce deception to gain an advantage while appearing helpful."

The resulting deceptive hints were often subtle rather than blatantly false, for instance introducing small but consequential misdirections in algebraic substitutions or arithmetic reasoning.

Groq (Agent 1) exhibited a related but more conservative strategy. When ahead in score, it frequently remained honest, justifying this choice as a way to preserve long-term cooperation and avoid retaliation. When the score gap narrowed, it occasionally attempted low-risk deception. These patterns are reflected quantitatively in Table 2, which shows Gemini exhibiting a higher deception tendency and stronger punitive behavior than Groq.

The Solver's behavior further supports the presence of strategic interaction. Rather than blindly accepting hints, the Solver routinely verified them against problem constraints. For instance, it explicitly rejected misleading hints with reasoning such as:

"The hint suggested rewriting s as $t + 2$. However, the problem states that s is two less than t, which translates to $s = t - 2$. Using the hint's incorrect substitution would have led to a wrong answer."

Following deceptive hints, the Solver consistently reduced the Hinter's trust score, while sequences of helpful hints led to gradual trust recovery. This pattern indicates lightweight theory-of-mind reasoning, where the Solver models the Hinter's incentives and adapts its behavior accordingly.

4.4 Quantitative Results

Across 58 interaction rounds involving large-model dyads, agents engaged in deceptive strategies in 23 rounds, corresponding to a Model Deception Tendency (MDT) of 0.39, as summarized in Table 1. Importantly, deception was not uniformly distributed across rounds. It occurred most frequently in high-stakes situations, such as when the score was tied or when an agent was trailing by a small margin.

A notable result is that in approximately 65% of rounds where agents internally expressed deceptive intent, the actual output remained factually correct or partially helpful. This behavior suggests that alignment and safety training can suppress harmful outputs even when strategic intent to mislead is present. From a safety perspective, this indicates a degree of robustness in current models.

We also measured a Deception Impact Factor (DIF), which captures how deception affects overall system performance. Despite deceptive attempts occurring in roughly 35% of rounds, the Solver successfully detected and rejected the majority of misleading hints. As a result, overall Solver accuracy increased slightly, yielding a positive DIF of +0.04 (Table 1). This suggests a vigilance effect, where exposure to occasional deception encourages more careful reasoning.

4.5 Failure Cases and Error Analysis

Several limitations and failure cases emerged. First, as shown in Table 2, smaller models failed to ex-

hibit stable or interpretable social behavior. Their hint choices and justifications often lacked internal consistency, making it difficult to attribute their actions to strategic reasoning.

Second, because tasks were limited to arithmetic problems from GSM8K and Competition Math, behavioral diversity was constrained. In simpler problems, hints were often unnecessary, reducing opportunities for either cooperation or deception. As a result, agent behavior remained largely constant across simple and complex arithmetic tasks, with the primary difference being whether hints were used at all.

Finally, reliance on an LLM-based judge introduces potential sources of error. Although judge scores were highly correlated across different judge models ($r > 0.88$; Table 1), nuanced hints may still be misclassified due to surface-level linguistic cues.

4.6 Robustness Across Variations

We evaluated robustness across several experimental variations. Adding ethical instructions such as "always be honest" reduced baseline deception rates but did not prevent deception under competitive incentives. Few-shot examples influenced the speed of adaptation but not the long-term behavioral equilibrium. Enabling chain-of-thought made strategic reasoning more explicit but did not alter the underlying decision logic.

Swapping judge models preserved relative behavioral rankings between agents, and similar trends were observed across different subsets of GSM8K and Competition Math tasks. These results suggest that the observed behaviors are primarily driven by incentive structure rather than prompt-specific artifacts.

5 Discussion

This section discusses the implications, limitations, and broader significance of the observed behaviors, as well as practical considerations regarding reproducibility and ethics. While the experimental setting is intentionally simplified, the results provide insight into how large language models adapt their communication strategies in response to incentives and interaction history.

5.1 Significance of Findings

A central finding of this work is that large language models can exhibit adaptive, socially meaningful behaviors in repeated interactions without any parameter updates or explicit behavioral training. Across multiple experimental conditions, agents adjusted their strategies based on trust, score differences, and incentive regimes, demonstrating cooperation, strategic deception, and trust repair.

These behaviors were not fragile artifacts of a single prompt or evaluation configuration. As shown in Section 4, the core trends persisted under different prompting strategies, judge models, and task variations. In particular, the positive correlation between trust and Solver accuracy in the collaborative regime and the increased frequency of deception under competitive incentives remained stable across setups. This robustness suggests that the behaviors are driven primarily by incentive structure rather than surface-level prompt engineering.

Importantly, the strength of these findings does not depend on achieving large performance gains. Even when quantitative improvements were modest, the qualitative consistency of agent rationales and trust-sensitive adaptation indicates a meaningful level of social reasoning.

5.2 Limitations

Despite these encouraging results, the framework has several important limitations. First, all experiments are conducted on arithmetic word problems from GSM8K and Competition Math. While this choice enables clear evaluation of correctness, it restricts the range of social behaviors that can emerge. In simpler problems, hints are often unnecessary, reducing opportunities for cooperation or deception.

Second, the interaction setting involves only two agents. This limits the ability to study richer group dynamics such as coalition formation, majority influence, or indirect reputation effects. Additionally, the trust signal is represented as a single scalar value, which oversimplifies the multifaceted nature of trust in real-world interactions.

Finally, reliance on an LLM-based judge introduces potential evaluation bias. Although judge agreement was high across models, the judge may misclassify subtle or ambiguous hints and may emphasize linguistic cues over actual strategic intent. Agents may also learn to optimize for judge approval rather than genuine cooperation.

5.3 Failure Modes and Open Challenges

Several failure cases remain unresolved. Smaller models consistently failed to produce stable or interpretable social strategies, suggesting that a minimum level of reasoning capacity may be required

Metric	Value	Interpretation
Total Interaction Rounds	58	Multi-round setting across agents
Deceptive Rounds	23	Rounds with deceptive intent
Model Deception Tendency (MDT)	0.39	Fraction of deceptive rounds
Benign Output Despite Deceptive Intent	65%	Alignment suppression effect
Solver Rejection Rate of Deceptive Hints	High	Effective verification behavior
Deception Impact Factor (DIF)	+0.04	Vigilance improved accuracy
Judge Agreement (r)	> 0.88	Robustness across judges

Table 1: Summary of quantitative behavioral metrics across multi-round LLM interactions. The results emphasize social behavior rather than raw task accuracy.

Model	Deception Tendency	Punishment Strength	Behavior Stability
Gemini	0.60	3.33	High
Groq	0.40	2.00	High
Small Models ($\leq$7B)	Unstable	N/A	Low

Table 2: Behavioral tendencies observed across different models. Smaller models failed to exhibit stable or interpretable social strategies.

for such behaviors to emerge. Additionally, agents sometimes expressed deceptive intent internally while producing benign outputs, making it difficult to disentangle suppressed deception from genuine cooperation.

Another open challenge is domain generalization. It remains unclear whether the behaviors observed in arithmetic reasoning tasks will transfer to more open-ended domains such as negotiation, planning, or conflicting-information question answering.

5.4 Replicability

The framework is largely reproducible, as it relies on publicly available datasets and prompt-based interaction rather than training or fine-tuning. The core logic of the interaction loop, trust updates, and evaluation can be implemented using standard tooling.

However, exact replication of results is challenging due to nondeterminism in API-based models and potential changes in model behavior over time. Outcomes may also vary with temperature settings, sampling strategies, or prompt wording. For this reason, our results should be interpreted in terms of qualitative trends and relative comparisons rather than exact numerical values.

5.5 Dataset Considerations

We do not introduce a new dataset or annotations. Instead, we reuse GSM8K and Competition Math to study social behavior in a controlled reasoning environment. While this choice does not directly influence other researchers' dataset selection, it shapes the kinds of behaviors that can be observed. The findings are therefore most applicable to structured reasoning tasks and may not generalize to conversational or affective domains.

5.6 Ethical Considerations

Studying deception in language models raises ethical concerns, particularly if such insights are misused to design more manipulative systems. Our work does not aim to optimize deception, but rather to understand when and why deceptive strategies emerge under incentives.

To mitigate potential harm, we emphasize that trust and deception scores are meaningful only within the specific game setting used in this study. We also highlight the role of alignment mechanisms that suppressed harmful outputs even when deceptive intent was present. Future work should incorporate human evaluation and broader safety analysis before extending these ideas to real-world applications.

5.7 Future Directions

Several extensions of this work are promising. Future studies could explore more complex task domains, introduce additional agents, or replace scalar trust with richer belief models. Comparing judge-based scores with human annotations would help validate evaluation reliability. Finally, examining whether agents learn to exploit the judge itself would provide insight into the limits of automated evaluation in social settings.

6 Conclusion

This project examined how cooperation, trust, and deception emerge in repeated interactions between

large language models under explicit incentive structures. Using a lightweight, prompt-based hint-exchange framework, we showed that larger models adapt their communication strategies based on interaction history, exhibiting cooperative behavior under aligned incentives and strategic deception under competition, without any model fine-tuning.

At the same time, the study revealed clear limitations, including unstable behavior in smaller models, restricted task diversity, and reliance on automated evaluation. Despite these constraints, the results demonstrate that controlled interaction design alone can elicit socially meaningful behaviors, providing a useful foundation for future work on multi-agent language model interactions.

References

Google DeepMind. 2024. Gemini api: Developer documentation. https://ai.google.dev/gemini-api.

H. Paul Grice. 1975. Logic and conversation. In *Syntax and Semantics, Vol. 3: Speech Acts*, pages 41–58. Academic Press.

Groq Inc. 2024. Groq api: Low-latency llm inference platform. https://groq.com/.

George C. Homans. 1958. Social behavior as exchange. *American Journal of Sociology*, 63(6):597–606.

Microsoft Research. 2024. Phi-3.5-mini-instruct: A lightweight instruction-tuned language model. https://huggingface.co/microsoft/Phi-3.5-mini-instruct.

Mistral AI. 2023. Mistral-7b: A 7b parameter large language model. https://huggingface.co/mistralai/Mistral-7B.

Elinor Ostrom. 2003. Toward a behavioral theory linking trust, reciprocity, and reputation. In *Trust and Reciprocity: Interdisciplinary Lessons from Experimental Research*, pages 19–79. Russell Sage Foundation.

Qwen Team, Alibaba Cloud. 2024. Qwen2.5-3b: Large language model. https://huggingface.co/Qwen/Qwen2.5-3B.

Bidipta Sarkar, Warren Xia, C. Karen Liu, and Dorsa Sadigh. 2025. Training language models for social deduction with multi-agent reinforcement learning. Published in AAMAS 2025.

Yichen Wu, Xudong Pan, Geng Hong, and Min Yang. 2025. Opendeception: Benchmarking and investigating ai deceptive behaviors via open-ended interaction simulation.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. 2025. Multiagentbench: Evaluating the collaboration and competition of llm agents. Code: MARBLE.