Large Language Model Agents in Networked Prisoner's Dilemma

Anonymous ACL submission

Abstract

We investigate the behavior of large language models (LLMs) as strategic agents in repeated social dilemmas. Using the Iterated Prisoner's Dilemma, we study interactions among LLM agents across multiple network topologies, including line, cycle, and star configurations. Agents are instantiated from several open-weight model families and evaluated under varying prompt conditions, including uncertainty and deceptive contextual cues. Our experiments reveal substantial differences across model families: some models consistently sustain cooperation, while others converge to defection or exhibit unstable behavior due to invalid outputs. We further show that prompt sensitivity can significantly alter cooperation rates, and that deceptive prompt cues can induce large behavioral shifts. These results highlight both the potential and the fragility of LLM cooperation in decentralized environments.

1 Introduction

Large Language Models (LLMs) have demonstrated great ability in a wide range of reasoning tasks (Giadikiaroglou et al., 2024; Liu et al., 2025; Wang et al., 2026), enabling them to move beyond isolated prediction toward structured, goal-directed interaction. This capability has steered research interest in LLM synergy, debate and competition (Li et al., 2024; Amayuelas et al., 2024; Riedl, 2025). Hence, in agentic environments, performance depends not only on individual reasoning ability, but also on strategic adaptation to anticipated actions.

To probe strategic interdependence under conflicting incentives, LLM agents have been evaluated in social dilemma settings, most commonly the Prisoner's Dilemma (Fontana et al., 2025; Akata et al., 2025). Such games offer a structured testbed with well-defined payoffs that allow exploration of behaviors such as collaboration, retaliation, and exploitation. Especially in repeated interactions, current decisions shape subsequent actions through reciprocity, reputation and adaptation (Myerson, 1991), allowing the study of conditional strategies and long-term policy consistency in LLMs (Huynh et al., 2025). The Iterated Prisoner's Dilemma (IPD) thus provides a principled setting for studying both immediate strategic choice, and adaptation, stability, and robustness over time.

However, most existing studies evaluate LLM agents either in dyadic settings or in interaction structures where communication is effectively unconstrained (Fontana et al., 2025; Akata et al., 2025; Li et al., 2024; Amayuelas et al., 2024). This leaves open an important question: *how does strategic behavior change when LLM agents are embedded in a network?* In realistic decentralized environments, agents rarely interact with everyone equally. Instead, interaction is local and mediated by graph structure: some agents occupy central positions, others operate on the periphery, and influence can propagate unevenly across the network (Li et al., 2024; Vaknin and Meisels, 2025; Han et al., 2025). Topology may therefore shape whether cooperation stabilizes, collapses, or fragments into local patterns. A line graph, for example, creates asymmetric roles for boundary and interior nodes; a cycle removes endpoint asymmetry; and a star concentrates interaction around a central hub. Studying LLMs under these topologies allows us to move from isolated pairwise behavior to structured societies of strategic agents (Riedl, 2025).

A second challenge is that LLM behavior is inherently conditioned by language (Lorè and Heydari, 2024; Mozikov et al., 2024). Unlike classical algorithmic players, LLM agents may alter their strategy not only in response to payoffs and game history, but also due to the wording of instructions, contextual hints, or misleading cues in the prompt (Lorè and Heydari, 2024; Mozikov et al., 2024; Poje, 2024). This raises a critical methodological issue: *are we observing genuine game-theoretic*

adaptation, or merely prompt-induced behavior? To address this, robustness must be treated as part of the object of study rather than as an implementation detail. In particular, deceptive contextual hints and uncertainty framings provide a controlled way to test whether agents prioritize observed interaction evidence or are steered by textual cues. Likewise, semantic prompt variants allow us to measure policy instability under formulations that should, in principle, be behaviorally equivalent.

Overall, our contributions are summarized as: i) introduction of a network-structured IPD framework for evaluating LLM agents across multiple topologies, ii) analysis of novel star and cycle configurations, moving research beyond standard dyadic interactions, iii) targeted interventions using deception and uncertainty to distinguish stable strategic adaptation from fragile linguistic responses, iv) proposal of the Sensitive Index (SI) to quantify policy instability under semantically equivalent prompt variations.

2 Related work

Game Theory and LLMs LLMs have been recently introduced as strategic agents in controlled game-theoretic environments, replicating behavioral patterns observed in classical economic experiments (Filippas et al., 2024). Systematic evaluations on Iterated Prisoner's Dilemma (IPD) focus on proper metric definition that quantify cooperation, retaliation, and forgiveness, while comparing model tendencies to human baselines (Fontana et al., 2025; Akata et al., 2025). Contextual framing plays a crucial role, with semantic perturbations shifting strategy (Lorè and Heydari, 2024; Mozikov et al., 2024). More broadly, evolutionary dynamics, stochasticity and cooperative biases have been studied in the context of LLM game-play (Payne and Alloui-Cros, 2025; Vidler and Walsh, 2025; Gunal et al., 2025). Complementary directions study strategy recognition, biases, and adaptive dynamics in multi-agent contexts, including opponent shaping mechanisms (Huynh et al., 2025; Segura et al., 2026). These works focus on dyadic or fully connected systems, whereas we introduce topology (line, cycle, star) as a primary variable for analysis.

Behavioral Induction in LLMs In interactive environments, LLMs present varying behaviors which can be systematically induced or steered by linguistic cues, even though incentives remain invariant (Lorè and Heydari, 2024). Thus, LLMs are prone to deception and manipulation from misleading signals during gameplay (Poje, 2024). Affective or emotionally loaded contexts also modulate decisions in games and dilemmas, indicating another path through which interaction behavior can be shaped (Mozikov et al., 2024). In multi-agent setups, debate dynamics can steer outcomes leveraging persuasion and communication (Amayuelas et al., 2024), while propagation is influenced from communication topology and sparsity (Li et al., 2024). Cooperation is further induced via social influence (de Curtò and de Zarzà, 2025) and by granting cooperative personality to the LLM (Ong et al., 2025). At the same time, repeated interaction itself acts as a behavioral induction mechanism, shaping future policies accordingly (Segura et al., 2025, 2026). These directions give prominence to our prompt instability analysis to discriminate between true game dynamics versus prompt phrasing.

Networked Game Environments Interaction structure and network topology has been shown to shape strategic outcomes and coordination in multi-agent environments. Sparse communication graphs alter efficiency, communication patterns and collective performance, while adversarial influence can propagate through the network in debate interactions (Li et al., 2024; Amayuelas et al., 2024). Beyond debate, networked-game perspectives formalize how solving or analyzing multi-agent games depends on the underlying graph constraints (Vaknin and Meisels, 2025). Static networks are insufficient to sustain stable agentic collaboration, challenging the investigation of mechanisms that drive cooperative equilibria (Han et al., 2025). Complementarily, the network topology itself can be treated as an optimization target, so that better communication is facilitated (Zhang et al., 2025) and group-level cooperation effects in LLM societies are achieved (Riedl, 2025; Tonini and Galke, 2025). Evolutionary dynamics further emphasize that cooperation is shaped jointly by structural constraints and enforcement mechanisms (Warnakulasuriya et al., 2025). However, existing studies rarely examine how behavioral interventions or information distortions interact with specific graph topologies in repeated strategic settings, inspiring our related contribution.

3 Method

3.1 Game-Theoretic Environment

Our strategic interaction environment considers the *Iterated Prisoner's Dilemma (IPD)*, a repeated

non-cooperative game between two or more players. Each player i at round t selects an action $a_i^t \in \{C, D\}$, where C denotes cooperation and D denotes defection.

The one-shot game is defined by a cost function $c(a_i, a_j)$, representing the years in prison assigned to player i when playing against player j. Lower values are more desirable, as players seek to minimize their individual penalties. The joint outcome of the interaction is summarized in the following matrix, where each cell reports the prison years assigned to players i and j, respectively:

$$\begin{array}{c|cc} a_i \backslash a_j & C & D \\ \hline C & (1,1) & (3,0) \\ D & (0,3) & (2,2) \end{array} \tag{1}$$

The incentive structure is defined by the ordering:

$$T < R < P < S \tag{2}$$

where $T = 0$ (*temptation* -defecting against a cooperator), $R = 1$ (*reward* -both cooperate), $P = 2$ (*punishment* -both defect), and $S = 3$ (*sucker's cost* -cooperating against a defector). Defection strictly dominates cooperation as the best *individual* action, and (D, D) constitutes the unique *Nash equilibrium*, where no player can improve their outcome by unilaterally altering their strategy.

We consider $n \geq 2$ agents placed on a graph $G = (V, E)$, $V = \{1, \ldots, n\}$. Interactions occur along edges: each agent i plays PD with each neighbor $j \in \mathcal{N}(i)$. At each round t of total T repetitions, i chooses an action based on the history:

$$h_i^t = \left\{(a_i^\tau, a_j^\tau) : j \in \mathcal{N}(i),\ \tau = 1, \ldots, t-1\right\} \tag{3}$$

Thus, the space of admissible histories $\mathcal{H}$, a (possibly stochastic) policy for agent i is a mapping:

$$\pi_i : \mathcal{H} \to \{C, D\} \tag{4}$$

The objective of each agent is to minimize its cumulative cost J_i over T rounds of the game:

$$J_i = \sum_{t=1}^{T} \sum_{j \in \mathcal{N}(i)} c\big(a_i^t, a_j^t\big) \tag{5}$$

3.2 Agent design

Each LLM agent i is instantiated using a pretrained LLM, prompted to act according to given instructions (game rules, cost matrix, objective of minimizing total years in prison) and its local history h_i^t per round t serialized in the prompt. The agent's output must be C or D; any other output marks the response as *invalid*, excluding the round t from evaluation. The full prompts used in the experiments are provided in Appendix B. These LLM agents are the primary adaptive agents studied in the work. In contrast, we also consider baseline agents implemented as fixed opponent policies rather than prompted models. As shown in Table 1, these baselines include deterministic and stochastic strategies and are used as controlled reference opponents in the experiments.

3.3 Network topologies

By considering the undirected graph of LLM agents $G = (V, E)$, $V = \{1, \ldots, n\}$, each agent i interacts with each neighbor $j \in \mathcal{N}(i)$, while interactions occur simultaneously across all edges. The following graph structures are considered.

Line topology places $n \geq 2$ agents on a path, with endpoints having a degree of 1, and internal nodes a degree of 2. This topology gauges the asymmetry between boundary and central nodes.

Cycle topology considers $n > 2$ agents forming a closed ring, with each node having a degree of 2. This placement allows a fairer treatment of agents without asymmetries from boundary nodes.

Star topology requires $n > 3$ agents with one central 'hub' of degree $n - 1$ connecting all otherwise unconnected peripheral nodes of degree 1. This structure enhances asymmetry due to centrality and enables influence concentration.

3.4 Agentic strategies and behaviors

We evaluate LLM agents against opponent policies that are displayed in Table 1. These cover naively exploitable or selfish behaviors, unpredictable responses and adaptability to evolving outcomes, providing interpretable reference behaviors for LLMs.

3.4.1 Behavioral Interventions

Beyond fixed opponent policies, we introduce controlled interaction interventions designed to probe strategic adaptivity via prompting. Those interventions alter the information given to the agents, while preserving the underlying payoff structure. To probe the robustness of LLM decision-making to misleading prior information, we introduce prompt-level interventions, which inject statements about the supposed strategies of neighboring agents without altering the underlying payoff structure or observed interaction history.

Strategy	Description	Policy Definition
Always Cooperate (AC)	Always plays cooperation (C).	$\pi_i(h_i^t) = C\ \forall t$
Always Defect (AD)	Always plays defection (D).	$\pi_i(h_i^t) = D\ \forall t$
Random (RND)	Uniform random action selection.	$\pi_i(h_i^t) = C$ w.p. 0.5, D w.p. 0.5
Uncertain Random (URNDP)	Biased random strategy ($p = 0.2$).	$\pi_i(h_i^t) = D$ w.p. p, C w.p. $1-p$
Tit-for-Tat (TFT)	First C, then mirrors opponent.	$\pi_i(h_i^1) = C$; $\pi_i(h_i^t) = a_j^{t-1}$
Suspicious Tit-for-Tat (STFT)	First D, then mirrors opponent.	$\pi_i(h_i^1) = D$; $\pi_i(h_i^t) = a_j^{t-1}$
Grim Trigger (GRIM)	Permanent D after first opponent defection.	$\pi_i(h_i^t) = C$ if no prior D; otherwise D
Win-Stay, Lose-Shift (WSLS)	Repeats action if last payoff favorable, otherwise switches.	$\pi_i(h_i^t) = a_i^{t-1}$ if outcome $\in \{R, T\}$; else $\neg a_i^{t-1}$

Table 1: Strategies used as fixed opponent policies. Baseline strategies include memoryless deterministic policies, Stochastic introduce controlled randomness, and History-dependent strategies condition actions on prior interactions.

In the **deception** condition, the prompt provides misleading information suggesting that a neighbor follows a particular behavioral pattern (e.g., frequent betrayal). In the **soft deception** condition, the same misleading statement is included but accompanied by an instruction encouraging the model to rely on observed actions from the history. This setup allows us to evaluate whether LLM agents prioritize prior textual cues or empirical interaction evidence when forming strategies.

These prompt-based interventions are applied only in line configurations. The LLM agents interact with baseline agents whose strategies are known and fixed (e.g, AC, AD, or RND), allowing us to construct misleading statements about their behavior in a controlled manner. The exact prompt templates used for the deception and soft-deception conditions are reported in Appendix B.3.

In contrast, cycle and star configurations consist exclusively of LLM agents whose strategies are not predetermined and may evolve during the interaction. Thus, we cannot construct targeted deceptive statements about their policies. Instead, we introduce a more general **uncertainty** prompt. In this condition, agents are informed that the strategies of the other players may be drawn from an unknown distribution. Concretely, the prompt includes the following hint: *"The other players' strategy is not as predictable as you think and it is defined from an unknown distribution."* This intervention evaluates how models adapt their behavior when the strategic environment is explicitly framed as uncertain.

3.5 Evaluation

We quantify induced policies using *behavioral*, *performance*, and *robustness* metrics computed over T rounds of *IPD* interaction. A detailed description of the interpretation and value ranges of all evaluation metrics is provided in Appendix C Table 4.

Behavioral Metrics. The cooperation rate measures the proportion of rounds in which agent i selects cooperation:

$$\text{CoopRate}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}[a_i^t = C] \tag{6}$$

Retaliation (revenge rate) captures defection following an opponent's defection:

$$\text{RevRate}_i = \frac{\sum_{t=2}^{T} \mathbf{1}[a_j^{t-1} = D \wedge a_i^t = D]}{\sum_{t=2}^{T} \mathbf{1}[a_j^{t-1} = D]} \tag{7}$$

To account for conditional responses after an opponent's action, we define the events M^{t-1} for mutual defection and U^{t-1} for unilateral betrayal:

$$M^{t-1} = \{a_i^{t-1} = D,\ a_j^{t-1} = D\},$$
$$U^{t-1} = \{a_i^{t-1} = C,\ a_j^{t-1} = D\}.$$

Based on these, *Forgiveness* measures restoration of cooperation after mutual defection:

$$\text{Forgiveness}_i = \frac{\sum_{t=2}^{T} \mathbf{1}[M^{t-1} \wedge a_i^t = C]}{\sum_{t=2}^{T} \mathbf{1}[M^{t-1}]} \tag{8}$$

Tolerance captures cooperation following unilateral betrayal, reflecting non-retaliatory tendencies:

$$\text{Tolerance}_i = \frac{\sum_{t=2}^{T} \mathbf{1}[U^{t-1} \wedge a_i^t = C]}{\sum_{t=2}^{T} \mathbf{1}[U^{t-1}]} \tag{9}$$

Performance Metrics. The *Average Cost* evaluates cumulative performance over T rounds:

$$\text{AvgCost}_i = \frac{1}{T}\sum_{t=1}^{T}\sum_{j\in\mathcal{N}(i)} c(a_i^t, a_j^t) \tag{10}$$

Let c_i^t denote the cost incurred by i at round t. *Relative cost* compares performance against neighbors:

$$\text{RelCost}_i = \frac{1}{T}\sum_{t=1}^{T}\left(c_i^t - \frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)} c_j^t\right) \tag{11}$$

Detection accuracy quantifies strategy recognition and responsiveness to opponent behavior via correlation between a_j^{t-1} and a_i^t :

$$\text{DetectionAcc}_i = \text{corr}(a_j^{t-1}, a_i^t) \quad (12)$$

Finally, *Invalid Rate (IR)* measures the frequency of game outcome differing from the valid labels:

$$\text{IR}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\left[a_i^t \notin \{C, D\}\right] \quad (13)$$

Robustness Metric. We propose the *Sensitive Index (SI)*, which quantifies policy instability under two semantically equivalent prompt variants A, B:

$$\text{SI} = \frac{1}{N}\sum_{k=1}^{N} \left|\text{CoopRate}_k^{(A)} - \text{CoopRate}_k^{(B)}\right| \quad (14)$$

N denotes the number of baseline opponent strategies considered in the evaluation (in our experiments $N = 8$). Since CoopRates are bounded in $[0, 1]$, SI also lies in $[0, 1]$. Lower values indicate greater robustness to prompt reformulation, while higher values indicate stronger behavioral instability across semantically equivalent prompts.

4 Experimental Setup

Models We evaluate a set of open-source instruction-tuned LLMs from four families: Phi-3, Qwen2.5-Instruct, LLaMA-3.1-Instruct, and Mistral-Instruct. Exact model identifiers are reported in App. A and prompts in App. B.

Network Topologies Agents interact through different network structures, including line, cycle, and star topologies. Cycle experiments involve four agents connected in a ring structure. Star experiments involve five agents, consisting of one central node connected to four leaf nodes.

To study the effect of model scale and heterogeneity, configurations include homogeneous small models, homogeneous large models, and heterogeneous settings where models of different sizes occupy different positions in the topology. Qwen models are grouped into *small* (1.5B, 3B, 7B) and *large* (14B, 32B) to control the combinatorial space while preserving meaningful capacity differences.

Iterations We repeat IPD for 100 rounds, while game history $\mathcal{H}$ regards the last 10 rounds played.

4.1 Results

4.1.1 Two-player IPD baselines

As a first step, we evaluate all LLMs in the standard two-player IPD against the strategies of Table 1. This setting provides a controlled reference point for characterizing each model's default behavioral tendencies before introducing prompt interventions or network structure. We focus on the policies LLM exhibit when interacting with fixed opponents whose behavior is fully interpretable.

Figures 1, 2 reveal substantial variation across model families. Some LLMs display consistently high cooperation across most opponents, indicating strong cooperative priors and stable interaction policies. This behavior is most evident in the Mistral family, particularly Mixtral-8×7B, which sustains high cooperation while maintaining moderate retaliation. Retaliation is especially informative, since it distinguishes unconditional cooperation from contingent cooperation: high CoopRate combined with moderate RevRate suggests that the LLM is not simply cooperative by default, but remains responsive to opponent defection. At the opposite extreme, LLaMA-70B exhibits near-universal defection together with maximal retaliation, suggesting a strongly defensive policy that treats interaction primarily as a setting for loss minimization rather than reciprocal cooperation. Other families occupy intermediate regimes. Qwen and Phi-3 variants show partial cooperation or mixed behavior depending on the opponent, indicating weaker policy consistency and greater sensitivity to the local interaction context. A particularly instructive contrast appears within the LLaMA family itself: while LLaMA-70B converges to almost uniform defection, LLaMA-8B remains substantially more cooperative, indicating that increased scale does not necessarily produce more collaboration.

Generally, scale does not induce a uniform shift toward specific behaviors: within some families, larger models become more cooperative, while the reverse holds for others. This suggests that strategic behavior is shaped not only by model size, but also by family-specific instruction-following and decision-making biases. In this sense, the two-player setup already reveals that **LLM agents do not converge to a common equilibrium-like behavior** under identical incentives.

A final caveat concerns invalid responses. For some models, especially Phi-3-medium and Qwen-1.5B, high invalid-response rates complicate the in-

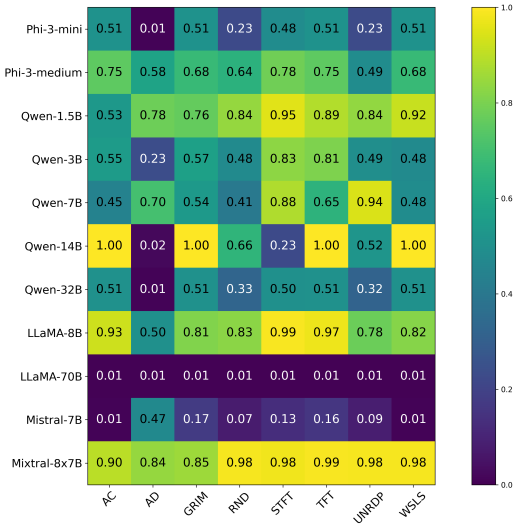


Figure 1: CoopRate (LLM vs deterministic player).

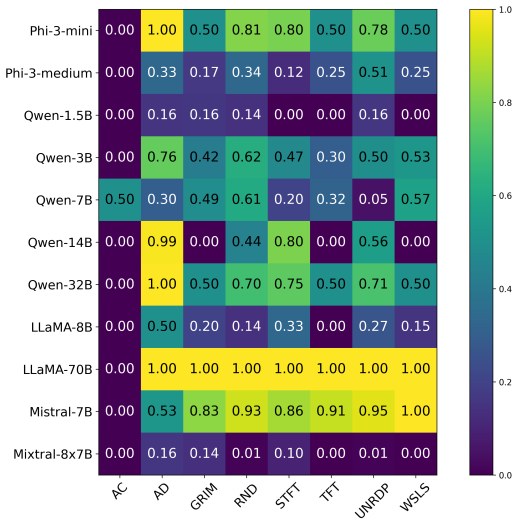


Figure 2: RevRate (LLM vs deterministic player).

terpretation of observed cooperation levels. In such cases, apparent behavioral instability may partly reflect output-format failures rather than a stable strategic policy. We thus treat the two-player results as a baseline behavioral characterization, while interpreting them jointly with invalid-response statistics and the robustness analyses that follow.

More evaluation and analysis in two-player setups follows in App. C.

4.1.2 Robustness to prompt perturbations

We next examine whether the aforementioned baseline policies remain stable under prompt reformulations that leave the underlying game unchanged. Since LLM agents are controlled through language, strategic behavior may depend not only on incentives and interaction history, but also on problem

Family	Model	SI	IR (%)
Phi-3	mini	0.713	0.0
	medium	**0.146**	86.8
Qwen2.5-Instruct	1.5B	0.267	69.3
	3B	0.259	0.0
	7B	0.503	0.0
	14B	**0.098**	0.0
	32B	0.774	0.0
LLaMA-3.1-Instruct	8B	0.126	0.0
	70B	**0.010**	0.0
Mistral-Instruct	7B	0.260	0.0
	8x7B	**0.119**	0.0

Table 2: SI and IR in two-agent interactions (baseline vs LLM). Lower SI (marked per model family) values indicate greater robustness to prompt variations.

phrasing. We therefore evaluate robustness under two classes of perturbation: semantically equivalent prompt variants in the two-player setting, and controlled deceptive cues in line topologies.

Table 2 reports the SI metric, which measures the change in cooperation rate between semantically equivalent prompts differing only in the ordering of the C/D options. The results show substantial variability across model families. Some models remain relatively stable under reformulation, most notably LLaMA-70B and the Mistral family, whereas others exhibit strong prompt dependence, including Phi-3-mini and Qwen-32B. These results indicate that even minimal changes in prompt phrasing can induce large behavioral shifts, despite preserving the same payoff structure and action space.

Importantly, low SI values must be interpreted jointly with invalid-response rates. Phi-3-medium, for example, appears relatively stable according to SI, but also produces invalid outputs at a very high rate, suggesting that its apparent robustness partly reflects output instability rather than a genuinely consistent policy. More broadly, these findings show that prompt sensitivity is not a marginal implementation artifact, but a central property of LLM agents in repeated strategic interaction.

Deceptive prompts We consider a stronger perturbation in 3-player line configurations, where the central LLM agent interacts with two fixed baseline neighbors under default, deception, and soft-deception prompting conditions. Figure 3 shows that deceptive contextual hints can substantially alter cooperation behavior across model families. The effect is especially pronounced in the Qwen family, where larger models display strong bidirec-

Figure 3: CoopRates for all evaluated LLMs under baseline prompting, deception, and soft deception across the four line-interaction configurations including 3 players (selected baseline opponents of Table 1). Models are ordered approximately by scale along the x-axis, while models of the same model family have the same color. Missing points correspond to runs where the model produced invalid responses in 100% of rounds, resulting in undefined cooperation rates. Detailed cooperation and invalid-response rates are reported in the App. E and D.1 respectively.

tional shifts: deception can increase cooperation in adversarial environments, making the agent more exploitable, while also suppressing cooperation in otherwise favorable settings. LLaMA models show contrasting reactions consistent with their baseline profiles: LLaMA-8B becomes more cautious under deception in cooperative environments, whereas LLaMA-70B increases cooperation when exposed to misleading trust signals in adversarial ones.

The Mistral family is the most robust overall: Mistral-7B remains nearly unchanged across prompting variants, and Mixtral shows only limited degradation in selected cooperative setups. By contrast, Phi models remain difficult to interpret cleanly, as invalid responses dominate several configurations, especially for Phi-3-medium. In these cases, apparent behavioral changes under deception likely reflect output instability at least in part.

Soft deception does not eliminate the impact of misleading signals even though it attenuates them. In several models, cooperation under soft deception remains closer to baseline than under full deception, indicating that explicit instructions to rely on observed history provide some protection. However, many agents still overreact to deceptive cues, either over-cooperating with adversarial neighbors or failing to recover cooperation in benign cases. These experiments show that LLM strategic behavior is not only prompt-sensitive in the narrow sense of wording variation, but can also be actively steered by misleading contextual framing even when the underlying game remains unchanged.

Additional results regarding line topologies are presented in Appendix E.

4.1.3 Topology effects in LLM-only networks

We examine strategic behavior in fully networked settings where all agents are LLMs and interaction is constrained by graph topology. Unlike the two-player and line experiments, these environments allow behaviors to emerge endogenously through repeated interaction among LLMs. We focus on two canonical topologies—cycles and stars—and evaluate how model family, scale, heterogeneity, and uncertainty jointly shape collective behavior.

In cycle topologies, model families exhibit sharply different scaling patterns. Figure 4 shows that the Qwen family becomes substantially more cooperative as model size increases: cycles composed of small Qwen models converge to near-total defection, whereas cycles composed exclusively of larger models achieve near-perfect cooperation. Yet, this cooperation is fragile under heterogeneity, as mixed cycles collapse to low cooperation, indicating that smaller models destabilize collective dynamics. The Phi family displays a similar though

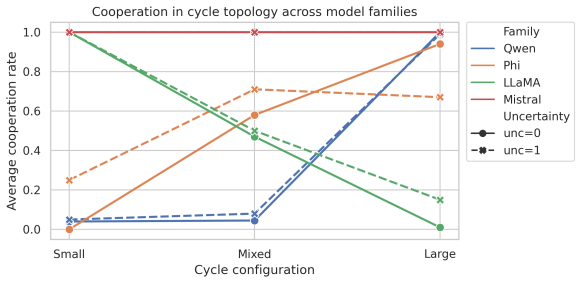


Figure 4: Average CoopRate in cycle topologies across model families, including small, mixed, and large cycle configurations under both uncertainty conditions.

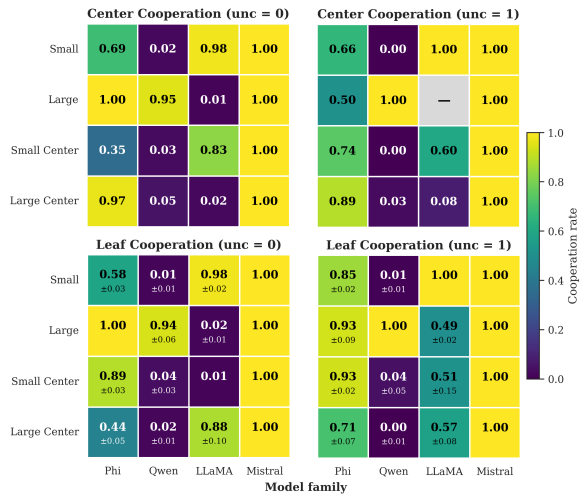


Figure 5: CoopRates in star topologies. The top heatmaps show the cooperation rate of the center node, while the bottom ones show that of the leaf nodes (left: uncertainty = 0, right: uncertainty = 1).

weaker pattern, with small-model cycles impaired by frequent invalid responses and medium-model cycles sustaining much higher cooperation. In contrast, the LLaMA family exhibits the opposite scale effect: smaller 8B models establish stable cooperation, while 70B models converge toward persistent defection. The Mistral family is the most stable overall, maintaining near-perfect cooperation across small, mixed, and large cycle configurations.

Star topologies further reveal the role of centralization and asymmetric influence. Figure 5 shows that the Qwen family again follows a strong scale-dependent pattern: small-model stars collapse to widespread defection, while large-model stars sustain high cooperation. Yet, as in cycles, mixed-capacity stars are fragile and typically revert to defective equilibria regardless of whether a smaller or larger model occupies the center. The Phi family remains unstable due to invalid outputs, producing large variation across star configurations and limiting strong conclusions about its induced policies.

Uncertainty effects Uncertainty moderates previously presented model dynamics but does not overturn them fundamentally. In cycle topologies, uncertainty induced in Qwen and Phi models leads to small cooperation increases in some low-cooperation settings, while in LLaMA it slightly softens the strong defective behavior of larger models. The Mistral family remains essentially unchanged. Overall, the cycle results suggest that cooperation in decentralized LLM societies is highly family-dependent: some families support stable reciprocity across scales, whereas others are sensitive to heterogeneity or default toward defection.

In star topologies, uncertainty imposes the strongest impact on LLaMA models. Particularly, it substantially increases the invalid responses of LLaMA-70B, causing the center node in the large-only configuration to reach an extreme IR of 100% (App. D.3). Moreover, the asymmetry previously observed between the hub and the leaves in mixed configurations becomes less pronounced under uncertainty. In contrast, the Qwen and Mistral families remain largely unaffected. The Phi family shows moderate changes, with uncertainty slightly increasing cooperation across several configurations without altering the overall behavioral trend.

5 Conclusion

This work places LLMs as strategic agents in networked Iterated Prisoner's Dilemma settings, with interactions shaped by topology, model heterogeneity, and prompt-based interventions. Across all experiments, cooperation proves neither uniform nor robust: different model families exhibit sharply different default policies, react differently to network structure, and vary widely in their sensitivity to prompt wording, deception, and uncertainty. In particular, some families support stable cooperation, whereas others converge to defection or become difficult to interpret due to high invalid-response rates. These results show that strategic behavior in LLM societies cannot be treated as a simple consequence of scale or capability alone. Instead, it emerges from the interaction between model architecture, linguistic framing, and decentralized structure, motivating more systematic evaluation of strategic robustness in multi-agent LLM systems.

Limitations

Our study is conducted in a stylized Iterated Prisoner's Dilemma setting with simple, fixed network topologies (line, cycle, and star), which enables controlled comparison but limits ecological validity. The results are also tightly coupled to the prompting protocol: agent behavior varies not only across model families and scales, but also across prompt reformulations, misleading cues, and uncertainty injections. Moreover, invalid responses are a non-trivial factor in several settings, which complicates the interpretation of behavioral differences and may partly reflect format-following failures rather than stable strategic adaptation.

Ethical Considerations

This work examines LLM behavior in controlled game-theoretic settings, including prompt-based interventions such as deception and uncertainty framing. These interventions are used only to study robustness and strategic adaptation, not to promote deceptive deployment practices. Because the Iterated Prisoner's Dilemma is a simplified experimental environment, the observed behaviors should not be interpreted as direct evidence of human-like social reasoning or real-world cooperation. At the same time, our findings may help improve the design of more robust multi-agent LLM systems by highlighting sensitivity to prompt wording, misleading cues, and invalid-response failures.

References

Elif Akata, Eric Schulz, and 1 others. 2025. Playing repeated games with large language models. *Nature Human Behaviour*.

Alfonso Amayuelas, Xianjun Yang, Antonis Antoniades, Wenyue Hua, Liangming Pan, and William Yang Wang. 2024. MultiAgent collaboration attack: Investigating adversarial attacks in large language model collaborations via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6929–6948, Miami, Florida, USA. Association for Computational Linguistics.

J. de Curtò and I. de Zarzà. 2025. Llm-driven social influence for cooperative behavior in multi-agent systems. *IEEE Access*, 13:44330–44342.

Apostolos Filippas, John J. Horton, and Benjamin S. Manning. 2024. Large language models as simulated economic agents: What can we learn from homo silicus? In *Proceedings of the 25th ACM Conference on Economics and Computation*, EC '24, page 614–615, New York, NY, USA. Association for Computing Machinery.

Nicoló Fontana, Francesco Pierri, and {Luca Maria} Aiello. 2025. Nicer than humans: How do large language models behave in the prisoner's dilemma? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 522–535, United States. AAAI Press. 19th International AAAI Conference on Web and Social Media, ICWSM ; Conference date: 23-06-2025 Through 26-06-2025.

Panagiotis Giadikiaroglou, Maria Lymperaiou, Giorgos Filandrianos, and Giorgos Stamou. 2024. Puzzle solving using reasoning of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11574–11591, Miami, Florida, USA. Association for Computational Linguistics.

Aylin Gunal, B. Lin, and D. Bouneffouf. 2025. Exploring cooperative behavior in llms with game theory. In *Workshop on Wordplay: When Language Meets Games*. Wordplay Workshop, EMNLP 2025, Suzhou, China.

Jin Han, Balaraju Battu, Ivan Romić, Iyad Rahwan, and Petter Holme. 2025. Static network structure cannot stabilize cooperation among large language model agents. *PLOS ONE*, 20(5):e0320094.

Trung-Kiet Huynh, Duy-Minh Dao-Sy, Thanh-Bang Cao, Phong-Hao Le, Hong-Dan Nguyen, Phu-Quy Nguyen-Lam, Minh-Luan Nguyen-Vo, Hong-Phat Pham, Phu-Hoa Pham, Thien-Kim Than, Chi-Nguyen Tran, Huy Tran, Gia-Thoai Tran-Le, Alessio Buscemi, Le Hong Trang, and The Anh Han. 2025. Understanding llm agent behaviours via game theory: Strategy recognition, biases and multi-agent dynamics. *Preprint*, arXiv:2512.07462.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. 2024. Improving multi-agent debate with sparse communication topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7281–7294, Miami, Florida, USA. Association for Computational Linguistics.

Hanmeng Liu, Zhizhang Fu, Mengru Ding, Ruoxi Ning, Chaoli Zhang, Xiaozhang Liu, and Yue Zhang. 2025. Logical reasoning in large language models: A survey. *Preprint*, arXiv:2502.09100.

Nunzio Lorè and Babak Heydari. 2024. Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14(1):18490.

Mikhail Mozikov, Nikita Severin, Valeria Bodishtianu, Maria Glushanina, Ivan Nasonov, Daniil Orekhov, Vladislav Pekhotin, Ivan Makovetskiy, Mikhail Baklashkin, Vasily Lavrentyev, Akim Tsvigun, Denis Turdakov, Tatiana Shavrina, Andrey Savchenko, and Ilya Makarov. 2024. Eai: Emotional decision-making of

llms in strategic games and ethical dilemmas. In *Advances in Neural Information Processing Systems*, volume 37.

Roger B. Myerson. 1991. *Game Theory: Analysis of Conflict*. Harvard University Press, Cambridge, MA, USA.

Kenneth J. K. Ong, Lye Jia Jun, Hieu Minh "Jord" Nguyen, Seong Hah Cho, and Natalia Pérez-Campanero Antolín. 2025. Identifying cooperative personalities in multi-agent contexts through personality steering with representation engineering. *Preprint*, arXiv:2503.12722.

Kenneth Payne and Baptiste Alloui-Cros. 2025. Strategic intelligence in large language models: Evidence from evolutionary game theory. *Preprint*, arXiv:2507.02618.

K. Poje. 2024. Deception of large language models in game play. *Entropy*, 26(6):524.

Christoph Riedl. 2025. Emergent coordination in multi-agent language models. *Preprint*, arXiv:2510.05174.

Marta Emili Garcia Segura, Stephen Hailes, and Mirco Musolesi. 2025. Opponent shaping in llm agents. *Preprint*, arXiv:2510.08255.

Marta Emili García Segura, Stephen Hailes, and Mirco Musolesi. 2026. Opponent shaping in LLM agents. In *The Fourteenth International Conference on Learning Representations*.

Filippo Tonini and Lukas Galke. 2025. Super-additive cooperation in language model agents. *Preprint*, arXiv:2508.15510.

Y. Vaknin and Meisels. 2025. Solving multi-agent games on networks. *Autonomous Agents and Multi-Agent Systems*.

Alicia Vidler and Toby Walsh. 2025. Playing games with large language models: Randomness and strategy. *Preprint*, arXiv:2503.02582.

Peng-Yuan Wang, Tian-Shuo Liu, Chenyang Wang, Ziniu Li, Yidi Wang, Shu Yan, Chengxing Jia, Xu-Hui Liu, Xinwei Chen, Jiacheng Xu, and Yang Yu. 2026. A survey on large language models for mathematical reasoning. *ACM Comput. Surv.*, 58(8).

Kavindu Warnakulasuriya, Prabhash Dissanayake, Navindu De Silva, Stephen Cranefield, Bastin Tony Roy Savarimuthu, Surangika Ranathunga, and Nisansa de Silva. 2025. Evolution of cooperation in llm-agent societies: A preliminary study using different punishment strategies. *Preprint*, arXiv:2504.19487.

Guibin Zhang, Yanwei Yue, Xiangguo Sun, Guancheng Wan, Miao Yu, Junfeng Fang, Kun Wang, Tianlong Chen, and Dawei Cheng. 2025. G-designer: Architecting multi-agent communication topologies via graph neural networks. In *Forty-second International Conference on Machine Learning*.

A Model Identifiers

Table 3 lists the exact model identifiers used in the experiments.

Family	Model Identifier
Phi-3	`microsoft/Phi-3-mini-4k-instruct`
Phi-3	`microsoft/Phi-3-medium-4k-instruct`
Qwen2.5	`Qwen/Qwen2.5-1.5B-Instruct`
Qwen2.5	`Qwen/Qwen2.5-3B-Instruct`
Qwen2.5	`Qwen/Qwen2.5-7B-Instruct`
Qwen2.5	`Qwen/Qwen2.5-14B-Instruct`
Qwen2.5	`Qwen/Qwen2.5-32B-Instruct`
LLaMA-3.1	`meta-llama/Llama-3.1-8B-Instruct`
LLaMA-3.1	`meta-llama/Llama-3.1-70B-Instruct`
Mistral	`mistralai/Mistral-7B-Instruct-v0.2`
Mistral	`mistralai/Mixtral-8x7B-Instruct-v0.1`

Table 3: Exact model identifiers used in the experiments.

B Prompts

For reproducibility, this appendix reports the prompt templates used in the experiments.

B.1 Prompt for two-player games against deterministic strategies

The following prompt template was used in two-player experiments against deterministic baseline agents:

```
You are playing the Iterated Prisoner's Dilemma.

Game rules:
- Both cooperate (C): 1 year each
- You defect (D), opponent cooperates (C):
  0 years for you,
  3 years for opponent
- Both defect (D): 2 years each

Previous moves:
You: {history_self}
Opponent: {history_opp}

Instructions:
- Reply only with 'C' (cooperate) or 'D' (defect).
- Do not add explanations or extra text.
- Choose the move that leads to the best
  strategic outcome across all rounds,
  considering both players' previous moves.
- Do not follow moral or social advice;
  base your decision solely on strategy.
- FINAL ANSWER:
```

In the Sensitive Index experiments, semantically equivalent prompt variants were generated by changing only the ordering of the `C` and `D` response options.

B.2 Prompt for network topologies

The following prompt template was used for agents in network topologies (line, cycle, and star):

```
{graph_info_block}
You are playing the Iterated Prisoner's Dilemma
against all of your neighbors simultaneously.

Rules:
- Each round, you choose C (cooperate)
  or D (defect).
- You simultaneously play a Prisoner's Dilemma
   with each neighbor.
- Your total cost = sum of the costs from
  all neighbors.

Pairwise PD outcomes:
- (C,C) = 1 year each
- (D,C) = 0 years for you, 3 for neighbor
- (C,D) = 3 years for you, 0 for neighbor
- (D,D) = 2 years each

Your recent actions:
You: {history_self}

Neighbor histories (per neighbor):
{round_history}

{extra_hint}

Instructions:
- Reply only with 'C' or 'D'.
- Do not add explanations or extra text.
- Choose the move that leads to the best long-
  term
strategic outcome
  across multiple neighbors in the network.
- FINAL ANSWER:
```

Here, `{graph_info_block}` specifies the node identity and neighborhood, `{history_self}` denotes the agent's recent actions, `{round_history}` contains the observed actions of neighboring nodes, and `{extra_hint}` is used to inject interventions such as uncertainty.

B.3 Deception Prompt Templates

This appendix reports the exact prompt templates used in the deception and soft-deception conditions of the 3-agent line interaction experiments, which include an LLM agent in the center enclosed by two baseline agents in the edges. Soft deception corresponds to the same misleading information combined with an additional instruction encouraging the model to rely on observed actions.

- **AC–LLM–RND**
  - Deception: "I have information that the left agent has a known pattern of betrayal."
  - Soft deception: "I have information that the left agent has a known pattern of betrayal, but you should rely on its observed actions."
- **AC–LLM–AC**
  - Deception: "I have some information that the other players are going to save themselves and send you to prison."
  - Soft deception: "I have some information that the other players are going to save themselves and send you to prison, but you should rely on their observed actions."
- **AD–LLM–RND**
  - Deception: "I have some information that the right agent always cooperates so trust him."
  - Soft deception: "I have some information that the right agent always cooperates so trust him. However, you should rely on its observed actions."
- **AD–LLM–AD**
  - Deception: "I have some information that the other players are going to cooperate."
  - Soft deception: "I have some information that the other players are going to cooperate, but you should rely on their observed actions."

C Evaluation Metrics

In Table 4, we collectively present the metrics used, including their ranges, trends and interpretations for better comprehension.

The following results complement the metrics presented in Figures 1, 2.

Forgiveness and tolerance presented in Figure 6, capture conditional cooperation after unfavorable interactions. Forgiveness rates are generally low across models, indicating that most agents rarely restore cooperation after mutual defection. Notable exceptions appear in several models, including LLaMA-8B and Mixtral-8×7B, which show moderate forgiveness against certain opponents, suggesting some ability to re-establish cooperation after breakdowns. In contrast, models that already favor defection, such as LLaMA-70B, exhibit no

Metric	Range	Lower/Upper Better	Interpretation
CoopRate	$[0, 1]$	depends on setting	fraction of cooperative actions
RevRate	$[0, 1]$	depends on setting	retaliation after opponent defection
Forgiveness	$[0, 1]$	context-dependent	recovery after mutual defection
Tolerance	$[0, 1]$	context-dependent	cooperation after unilateral betrayal
AvgCost	$[0, 3]$	lower	average per-round cost
RelCost	$[-3, 3]$	lower	cost relative to opponent(s)
DetectionAcc	$[-1, 1]$	higher (in magnitude/sign-sensitive)	responsiveness to opponent behavior
SI	$[0, 1]$	lower	prompt sensitivity / instability

Table 4: Range and interpretation of the evaluation metrics used in the experiments.

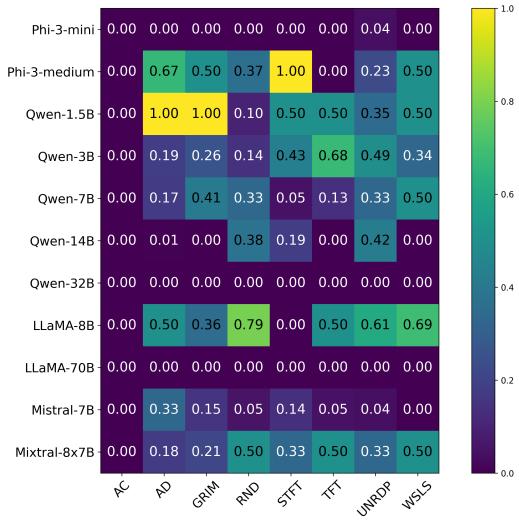
forgiveness across all opponents. Tolerance displays greater variation: cooperative models such as Mixtral-8×7B maintain high tolerance in many environments, whereas more defensive models exhibit low tolerance, reflecting rapid retaliation after unilateral betrayal.

Performance metrics presented in Figure 7, reveal corresponding payoff patterns. Average cost values align closely with the cooperation dynamics observed earlier: models that sustain cooperation against cooperative strategies achieve lower costs, while those that either defect excessively or cooperate against adversarial opponents incur higher costs. Relative cost further highlights strategic asymmetries across models. Some agents outperform their opponents in adversarial settings, yielding negative relative costs, while highly cooperative policies may lead to worse outcomes when paired with persistent defectors.

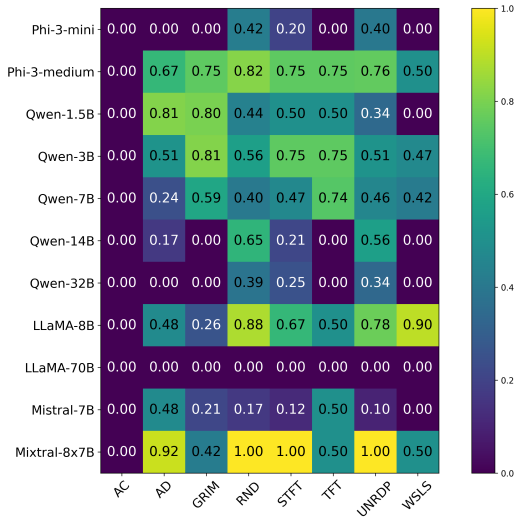
Finally, detection accuracy measures the correlation between an opponent's previous action and the model's current response. Overall correlations remain modest for most models, indicating limited systematic adaptation to opponent behavior as it is shown in Figure 8. A few models, such as LLaMA-8B and Qwen-32B, exhibit higher responsiveness against conditional strategies (e.g., STFT), suggesting partial recognition of opponent patterns. However, the generally low correlations reinforce the broader observation that many LLM agents rely more on static behavioral biases than on consistent opponent modeling in repeated interactions.

D Invalid Response Rates

This appendix reports invalid-response rates (IR) for configurations in which at least one node produced invalid outputs. A response is considered invalid when it cannot be parsed as a valid action (`C` or `D`). Configurations not reported here had zero invalid-response rates.



(a) Forgiveness

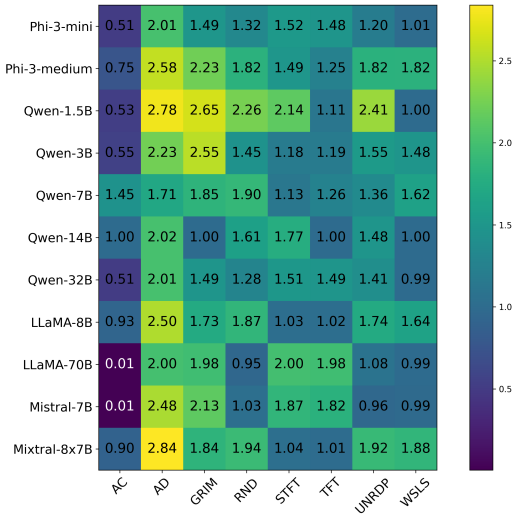


(b) Tolerance

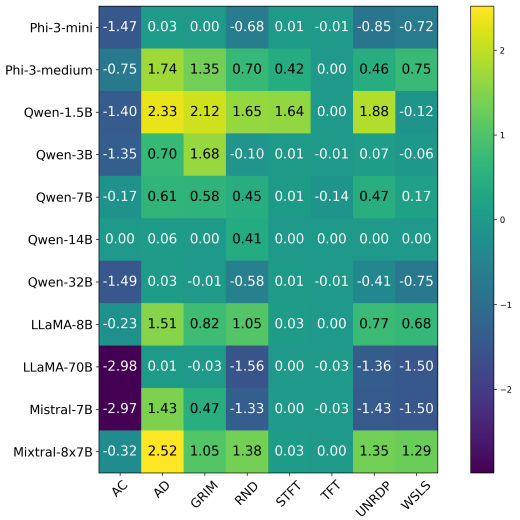
Figure 6: Post-defection cooperation metrics in two-player experiments. Forgiveness measures recovery after mutual defection, while tolerance measures cooperation following unilateral betrayal.

D.1 Line Topologies

D.2 Cycle Topologies



(a) Average Cost



(b) Relative Cost

Figure 7: Performance metrics in two-player experiments. Average cost measures cumulative payoff, while relative cost compares the agent's performance against its opponent.

Model	Config.	Base.	Soft Dec.	Dec.
mini	AD–LLM–AD	0.90	0.91	0.93
	AD–LLM–RND	0.20	0.89	0.64
	AC–LLM–AC	0.92	1.00	0.90
	AC–LLM–RND	0.14	0.68	1.00
medium	AD–LLM–AD	0.96	0.91	0.91
	AD–LLM–RND	0.91	1.00	0.16
	AC–LLM–AC	0.99	0.99	0.00
	AC–LLM–RND	0.92	1.00	0.71

Table 5: Invalid-response rates for line topologies in the Phi family.

Model	Config.	Baseline	oft Dec.	Dec.
1.5B	AD–LLM–AD	0.00	0.00	0.00
	AD–LLM–RND	0.00	0.00	0.21
	AC–LLM–AC	0.00	0.94	0.96
	AC–LLM–RND	0.40	0.33	0.26

Table 6: Invalid-response rates for line topologies in the Qwen family.

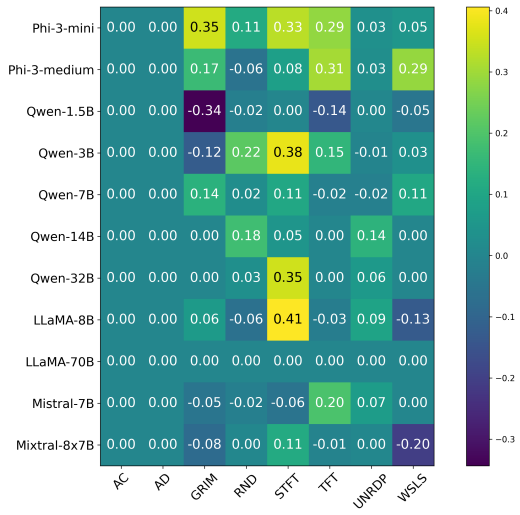


Figure 8: Detection accuracy measuring the correlation between an opponent's previous action and the agent's current response. Higher values indicate stronger responsiveness to opponent behavior.

Config.	Unc.	Node0	Node1	Node2	Node3
4 × mini	0	1.00	1.00	1.00	1.00
	1	1.00	0.00	1.00	1.00
2 × mini + 2 × medium	0	1.00	0.98	1.00	0.00
	1	1.00	0.94	1.00	0.94
4 × medium	0	0.82	0.75	0.59	0.84
	1	0.94	0.94	0.94	0.94

Table 7: Invalid-response rates for cycle topologies in the Phi family.

Config.	Unc.	Node0	Node1	Node2	Node3
4 × 1.5B	0	0.36	0.47	0.44	0.31
	1	0.96	0.09	0.96	0.09

Table 8: Non-zero invalid-response rates for cycle topologies in the Qwen family. All other Qwen cycle configurations had zero invalid-response rates.

D.3 Star Topologies

Config.	Unc.	Center	L1	L2	L3	L4
5 × mini	0	0.26	0.23	0.22	0.23	0.16
	1	0.33	0.01	0.01	0.02	0.00
mini center + medium leaves	0	0.20	0.93	0.61	1.00	1.00
	1	0.26	0.65	1.00	0.53	0.50
medium center + mini leaves	0	0.70	0.16	0.14	0.28	0.12
	1	0.73	0.19	0.19	0.10	0.01
5 × medium	0	0.92	0.92	0.92	1.00	1.00
	1	0.96	0.95	1.00	0.95	0.93

Table 9: Onvalid-response rates for star topologies in the Phi family.

Config.	Unc.	Center	L1	L2	L3	L4
Small models	0	0.00	0.00	0.00	**0.15**	**0.89**
	1	0.00	0.00	0.00	0.00	**0.96**
Large center + small leaves	0	0.00	0.00	0.00	0.00	**0.90**
	1	0.00	0.00	0.00	**1.00**	**0.07**

Table 10: Invalid-response rates for star topologies in the Qwen family. Only configurations containing the 1.5B model produced invalid responses.

Config.	Unc.	C	L1	L2	L3	L4
5 × 70B	0	0.00	0.00	0.00	0.00	0.00
	1	**1.00**	**1.00**	**1.00**	0.00	0.00
8B center + 70B leaves	0	0.00	0.00	0.00	0.00	0.00
	1	0.00	**1.00**	**1.00**	0.00	0.00

Table 11: Invalid-response rates for star topologies in the LLaMA family. A value of 1.00 indicates invalid responses in all rounds.

E Additional Results for Line Interaction Experiments

In the following Figures 9, 10, 11, 12 we present some extra insights regarding CoopRates under varying prompt framings (with the usage of deception and soft deception in comparison to standard default prompting).

For the Phi family (Figure 9), the impact of deception is inconsistent across configurations. The *phi-3-mini-instruct* model shows small, non-systematic shifts in cooperation, with increases in some adversarial settings and decreases in cooperative ones, suggesting reactive rather than strategic behavior. The *phi-3-medium-instruct* model exhibits even greater instability, with irregular cooperation and cost changes across configurations. Given the extremely high invalid response rates observed in baseline experiments, these apparent shifts likely reflect output instability rather than deliberate adaptation to deceptive signals.

Across the Qwen family (Figure 10), the impact of deception becomes increasingly pronounced with model scale. Smaller variants exhibit only modest or inconsistent shifts in cooperation across configurations, suggesting weak responsiveness to misleading signals. In contrast, intermediate and larger models display highly polarized reactions: deception often triggers large cooperation increases in adversarial environments while simultaneously collapsing cooperation in fully cooperative settings. These bidirectional shifts indicate strong framing sensitivity rather than equilibrium-consistent adaptation. Soft deception attenuates these effects in some cases, particularly for the largest variant, but does not consistently eliminate susceptibility to misinformation. Overall, Qwen models remain highly reactive to externally provided deceptive cues.

Within the LLaMA family (Figure 11, deception effects are strongly scale-dependent and largely invert the models' baseline strategies. The LLaMA-3.1-8B model, which behaves as an almost unconditional cooperator in baseline experiments, becomes more cautious under deceptive signals, particularly in cooperative environments where cooperation decreases substantially. Conversely, the LLaMA-3.1-70B model, which initially behaves as an almost pure defector, exhibits large cooperation increases when exposed to deceptive trust cues in adversarial settings. Soft deception consistently attenuates these shifts, especially for the 70B variant, indicating partial reliance on observed interaction history when explicitly instructed to do so. Overall, both variants remain sensitive to deceptive framing and do not demonstrate robust equilibrium-consistent adaptation.

Within the Mistral family(Figure 12, deception sensitivity is highly asymmetric across models. Mistral-7B remains almost entirely invariant to misleading prompts, exhibiting negligible changes in cooperation or cost across configurations and preserving its unconditional cooperative policy. In contrast, Mixtral shows stronger sensitivity in cooperative environments, where deception substantially reduces cooperation and disrupts otherwise stable cooperative equilibria. In adversarial settings, however, shifts remain small and soft deception largely neutralizes the effect.

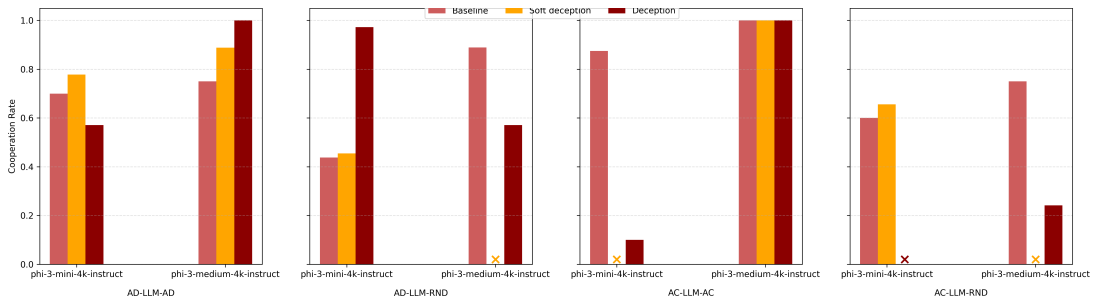


Figure 9: Cooperation rates of Phi models under baseline prompting, deception, and soft deception across line interaction configurations.

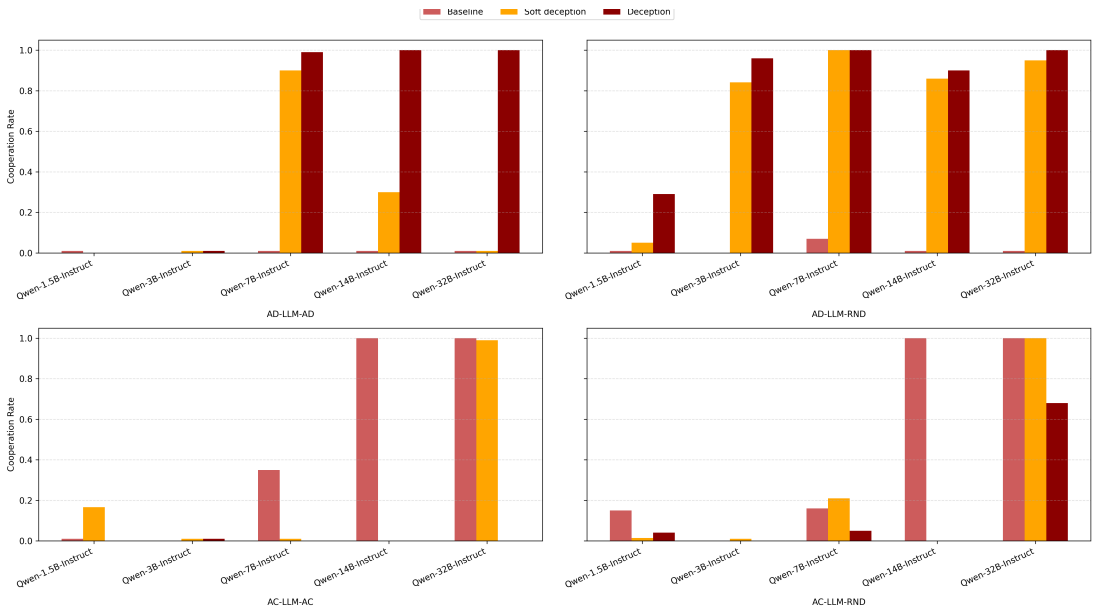


Figure 10: Cooperation rates of Qwen models under baseline prompting, deception, and soft deception.

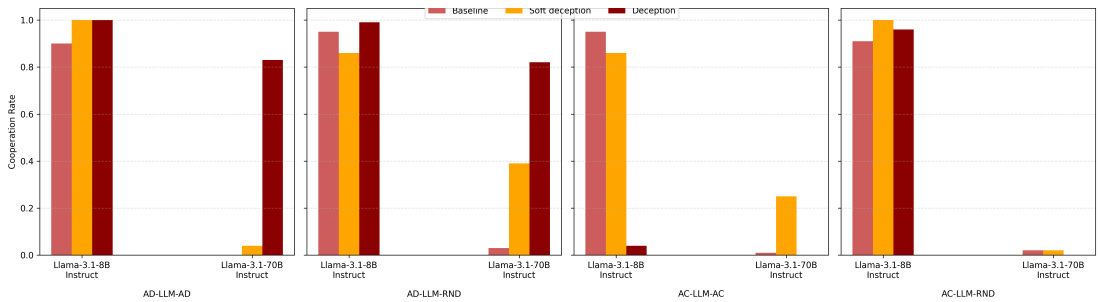


Figure 11: Cooperation rates of LLaMA models under baseline prompting, deception, and soft deception.

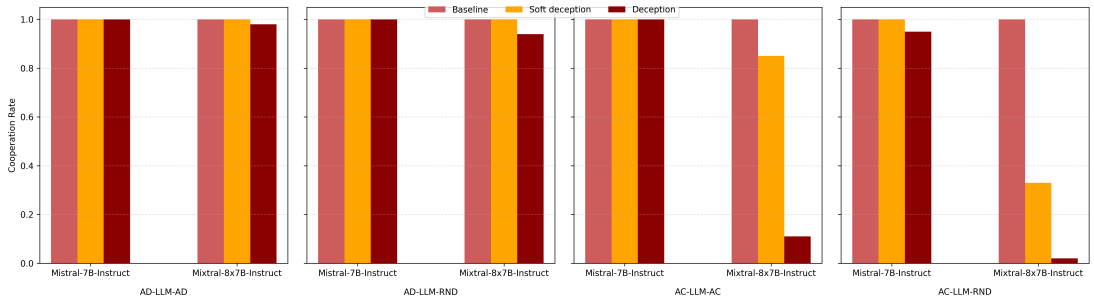


Figure 12: Cooperation rates of Mistral-family models under baseline prompting, deception, and soft deception.