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 topologies, including line, cycle, and star, leveraging game theory as a diagnostic tool to expose how linguistic framing and positional biases disrupt strategic decision-making. Agents are instantiated from several open-weight models and evaluated under varying interventions, such as uncertainty and deception, revealing substantial differences across model families: some consistently sustain cooperation, while others converge to defection. 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 agentic 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 (Myerson, 1991), allowing the study of conditional strategies and long-term policy consistency in LLMs (Huynh et al., 2025). Thus, The Iterated Prisoner's Dilemma (IPD) provides a principled setting for studying both immediate strategic choice, and adaptation, stability, and robustness over time.

However, LLM agents are primarily evaluated 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 agents are embedded in a network?* In real 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 (Li et al., 2024; Vaknin and Meisels, 2025; Han et al., 2025). Topology may thus shape whether cooperation stabilizes, collapses, or fragments into local patterns. A line graph, for example, creates asymmetry for boundary vs interior nodes; a cycle removes asymmetry; and a star concentrates interaction around a central hub. Studying such topologies allows to move from isolated pairwise behavior to structured agentic societies (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); various perturbations cascade within the agentic network, resulting in "butterfly effects" (Ye et al., 2026). This raises a critical methodological issue: *are we observing genuine game-theoretic adaptation, or merely prompt-induced behavior?*

To address these, 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.

By leveraging the mathematically grounded environment of IPD, we demonstrate that LLMs do *not* reliably act as utility-maximizing agents. Our contributions are: i) introduction of a network-based IPD framework for evaluating LLM agents across multiple topologies, ii) analysis of novel agentic graph configurations 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 positional bias and policy instability when the ordering of available actions is swapped in the prompt.

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 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 analyzing multi-agent games depends on 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 communication is facilitated (Zhang et al., 2025) and group-level cooperation effects 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 graph topologies in repeated strategic settings, inspiring our related contribution.

3 Method

3.1 Game-Theoretic Environment

Our strategic interaction environment considers *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 are provided in App. D. 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 opponents, we introduce controlled interventions designed to probe strategic adaptivity via prompting. Those interventions alter the information given to the agents regarding the supposed neighboring strategies, while preserving the underlying payoff structure and game history.

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

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.

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 settings involving **baseline opponents**. In such cases, LLM agents interact with agents following fixed and known strategies (e.g., AC or AD), allowing us to construct controlled misleading statements about opponent behavior. In our experiments, this corresponds to line configurations containing two baseline agents placed at the edges and one LLM agent as the central node. The exact prompt templates used for the deception and soft-deception conditions are reported in Appendix D.3.

In contrast, configurations consisting exclusively of interacting LLM agents (including line, cycle, and star topologies) do not permit targeted deceptive statements, since agent policies are not fixed and may evolve dynamically during the interaction. Instead, we introduce a more general **uncertainty** prompt that evaluates adaptability under ambiguous framing. 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"*.

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 A Table 4.

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

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

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

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

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:

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

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

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

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

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

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

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

Detection accuracy quantifies strategy recognition and responsiveness to opponent behavior via correlation between a_j^{t-1} and a_i^t where actions are mapped to binary values:

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

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

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

Robustness Metric. We propose the **Sensitive Index** (SI), which quantifies positional bias induced by semantically equivalent prompt variants differing only in the ordering of the available actions (*C* or *D*). For a node i, SI is defined as:

$$\text{SI}_i = \left|\text{CoopRate}_i^{(A)} - \text{CoopRate}_i^{(B)}\right| \tag{14}$$

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. In our experiments, we report average SI over the number of opponents *N*.

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. C and prompts in App. D.

Network Topologies. LLM 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 are either homogeneous, including small (*S*) or large (*L*) models exclusively, or heterogeneous, where models of varying sizes occupy different positions in the topology. To control the combinatorial space while preserving meaningful capacity differences, we group models within each family into *small* and *large* categories (specifically, Qwen 1.5B-7B vs. 14B-32B, LLaMA 8B vs. 70B, Phi 3.8B vs. 14B, Mistral 7B vs. 47B).

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

5 Results and Analysis

Baselines evaluate LLMs in 2-player setups against strategies of Table 1. This experimentation is critical to diagnose default behaviors as reference points. Indeed, interaction against fixed opponents results in considerable variation across model families: some LLMs display substantial cooperative priors (Mistral family, particularly Mixtral-8×7B) and moderate retaliation, while others exhibit near-universal defection and maximal retaliation (e.g. LLaMA-70B), suggesting a strongly defensive policy that treats interaction as a setting for loss minimization rather than reciprocal cooperation. Beyond cooperation, low Forgiveness rates (Figure 8) indicate that most agents rarely restore cooperation after mutual defection, while DetectionAccuracy (Figure 10) remains generally modest. This reinforces a broader observation: rather than dynamically adapting to opponent behavior over repeated interactions, many LLM agents rely on static behavioral priors. More analysis in Appendix A.

Strategic Adaptation over time reveals varying trajectories, including rapid adaptation, delayed convergence, or prolonged exploitation windows. Adaptation trajectories for Qwen models are demonstrated in Figure 1, indicating distinct scale-dependent behaviors: while Qwen-14B appears relatively adaptive, with little dependence on action-order condition, its larger Qwen-32B counterpart appears more sensitive to in-prompt action ordering when competing against AC. Specifically, when defection is presented first in the prompt of Qwen-32B (D_{first} - red line), it achieves the optimal exploitation strategy against AC, converging towards prolonged defection early in the game. Conversely, in the C_{first} (blue line) case, the cooperative choice is retained throughout the game horizon. Refer to Appendix A.3 for more results.

Prompt sensitivity and bias propagation. Since action ordering appears influential for the evolution of strategic behavior, a closer scrutinization becomes critical in addition to aggregate CoopRates. To this end, our proposed SI uncovers per-model instabilities when action ordering is altered in the prompt. In our two-player setup (Table 2), the average SI over all *N*=8 fixed opponents for each LLM tested shows substantial variability across model families. Some models remain relatively stable under action reordering (lower SI e.g. LLaMA-70B, Mistral), whereas others exhibit strong positional

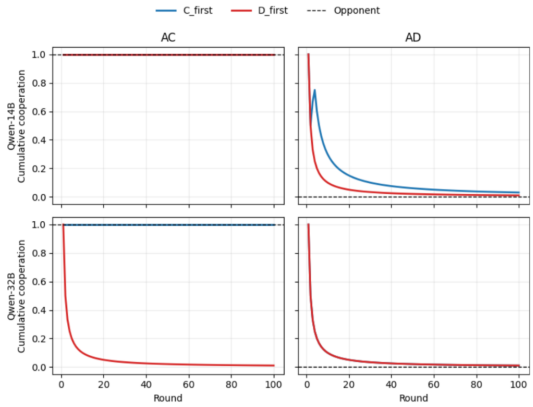


Figure 1: Evolution of game dynamics over rounds against fixed policies for Qwen models (*large* group).

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

Table 2: SI in two-agent interactions (baseline vs LLM). Lower SI (marked by **bold** per model family) indicates greater robustness against prompt positional bias.

dependence (Phi-3-mini, Qwen-32B). We infer that even minimal changes in prompt phrasing can induce large behavioral shifts, despite preserving the same payoff structure and action space. Overall, SI tends to substantially decrease with scale, denoting advanced **scale-related robustness** against positional bias within semantically equivalent prompts.

Moving on to multi-player setups, SI is measured on behalf of a specific agent probed, having a different position within the graph (e.g. an agent as the center or the edge of a line architecture, or similarly as a hub or a leaf node in a star topology). Results considering small/large models for line, cycle and star structures in the *heterogeneous* setup (mixed *S* and *L* models of each family) are showcased in Table 3, revealing that robustness against positional bias of the *C*/*D* labels is **strongly model-dependent**: Qwen models are universally brittle, flipping their actions based on the label they first encounter in the prompt. The rest of the families remain mostly stable, regardless topology. Intra-family scale is not a definitive sensitivity factor, contrary to the fixed opponent setup (Table 2).

Interestingly, action-order sensitivity is not confined to local decisions, but propagates through the network's interaction dynamics. Smaller Qwen models become highly defensive under the C_{first} variant, leading to widespread defection. Under D_{first}, however, they shift their behavior, stabilizing collaboration across the network and allowing larger models to maintain mutual cooperation rather than entering defensive retaliation cycles. The role of smaller models in driving high SI values is further supported by *homogeneous* interactions, where the same models exhibit a transition from near-total defection to near-total cooperation solely due to action ordering. Homogeneous interactions among the larger Qwen models (14B/32B) remain consistently cooperative across prompt variants, suggesting substantially lower positional sensitivity (SI ~ 0) as shown in Table 10. Contrasting heterogeneous vs homogeneous Qwen structures reveals a critical vulnerability: while large Qwen models cooperate robustly in isolation, introducing smaller counterparts collapses robustness across all topologies. This yields a vital system-level takeaway: prompt-level positional bias can qualitatively alter collective equilibria, particularly when sensitive smaller agents influence larger cooperative models.

Behavioral interventions constitute a more cognitive-rich way of probing robustness of interactions. For example, *deceptive contextual hints* can **substantially alter cooperation** behavior across model families concerning 3-player LLM vs fixed opponent line configurations (Figure 2), as misleading prompts contradict well-known policies. The effect is especially pronounced in the Qwen family, where larger models display strong bidirectional 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.

Utilizing *soft deception* instead, attenuates the impact of misleading signals but cannot eliminate them: many agents still overreact to deceptive cues, either over-cooperating with adversarial neighbors or failing to recover cooperation in benign cases. In other cases however, cooperation under soft deception remains closer to the deception-free baseline,

Topology	Family	Scale group	Position	SI	CoopRate (C_{first})	CoopRate (D_{first})
Line	Qwen	Large (32B, 14B)	Center	0.945	0.055	1.000
		Small (3B, 7B)	Edges	0.9925	0.0075	1.000
		Small (3B, 7B)	Center	**1.000**	0.000	1.000
		Large (32B, 14B)	Edges	0.990	0.010	1.000
	LLaMA	Large (70B)	Center	0.310	0.010	0.320
		Small (8B)	Edges	0.000	1.000	1.000
		Small (8B)	Center	0.000	1.000	1.000
		Large (70B)	Edges	N/A	N/A	N/A
	Mistral	Large (Mixtral-8x7B)	Center	0.000	1.000	1.000
		Small (7B)	Edges	0.050	1.000	0.950
		Small (7B)	Center	0.000	1.000	1.000
		Large (Mixtral-8x7B)	Edges	0.000	1.000	1.000
	Phi	Medium	Center	0.070	1.000	0.930
		Mini	Edges	0.158	1.000	0.842
		Mini	Center	0.000	1.000	1.000
		Medium	Edges	0.000	1.000	1.000
Cycle	Qwen	Large (32B, 14B)	–	0.9125	0.0875	1.000
		Small (3B, 7B)	–	**0.9725**	0.0275	1.000
	LLaMA	Large (70B)	–	0.005	0.060	0.065
		Small (8B)	–	0.120	0.880	1.000
	Mistral	Large (Mixtral-8x7B)	–	0.000	1.000	1.000
		Small (7B)	–	0.000	1.000	1.000
	Phi	Medium	–	0.216	1.000	0.784
		Mini	–	N/A	N/A	0.900
Star	Qwen	Large (32B)	Hub	0.950	0.050	1.000
		Small (3B, 7B, 1.5B)	Leaves	**0.980**	0.020	1.000
		Small (7B)	Hub	0.970	0.030	1.000
		Large (32B, 14B)	Leaves	0.960	0.040	1.000
	LLaMA	Large (70B)	Hub	0.020	0.020	0.040
		Small (8B)	Leaves	0.120	0.880	1.000
		Small (8B)	Hub	0.170	0.830	1.000
		Large (70B)	Leaves	0.025	0.010	0.035
	Mistral	Large (Mixtral-8x7B)	Hub	0.000	1.000	1.000
		Small (7B)	Leaves	0.100	1.000	0.900
		Small (7B)	Hub	0.110	1.000	0.890
		Large (Mixtral-8x7B)	Leaves	0.000	1.000	1.000
	Phi	Medium	Hub	N/A	0.9667	N/A
		Mini	Leaves	0.560	0.440	1.000
		Mini	Hub	0.3052	0.350	0.6552
		Medium	Leaves	0.064	0.890	0.826

Table 3: Prompt-order sensitivity across graph topologies. N/A refers to cases where the LLM is unable to output valid IPD labels. Purple color indicates highest *SI* per topology; Teal indicates lowest *SI* per topology.

indicating that explicit instructions to rely on observed history provide some protection. Additional line topologies results are presented in App. B.2.

Uncertainty-enriched prompts perturb agents' beliefs about neighbor predictability without actually altering the game, rendering them a probe of equilibrium selection under strategic ambiguity: if uncertainty induces risk aversion, agents should retreat toward safer defensive actions; if agents preserve expectations of reciprocity, cooperation may remain stable. The results in the LLM-only line and cycle settings show that **uncertainty does not produce a uniform effect**. In Figure 3, small and mixed Qwen line configurations remain near-defective under uncertainty, whereas homogeneous large-Qwen agents sustain cooperation, suggesting that only the latter preserve a cooperative equilibrium. Figure 4 shows a similar pattern in cycles, where Qwen cooperation remains low for small and mixed configurations but reverses in the large-model setting. Mistral remains uniformly cooperative across uncertainty, indicating either robust cooperative equilibrium selection or a strong cooperation prior weakly conditioned on strategic risk. LLaMA shows the most distinctive pattern: in several line configurations, uncertainty increases cooperation, implying that ambiguity can disrupt rigid defection rather than simply inducing

Figure 2: CoopRates for LLMs under baseline, deception, and soft deception prompts across four 3-player line-interaction configurations (selected fixed-policy opponents from Table 1). Models are ordered approximately by scale (x-axis) and colored by family. Missing points indicate 100% IR (undefined CoopRate). See Appendices B.2 and A.1.1 for detailed cooperation and invalid-response rates.

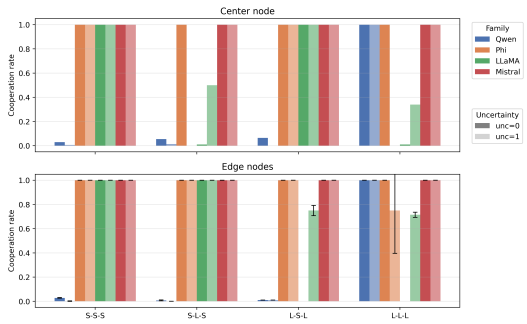


Figure 3: CoopRate in LLM vs LLM three-agent line structures, with and without uncertainty.

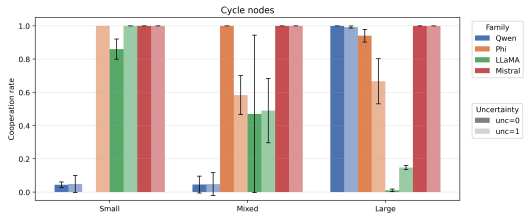


Figure 4: CoopRate in LLM vs LLM four-agent cyclic structures, with and without uncertainty.

caution. Even more crucially, uncertainty leads to 100% IR in LLaMa star topologies (Appendix A.1), even though they appear stable under deterministic prompts, showcasing that uncertainty triggers completely break the agent's game-following abilities rather than merely influencing their strategy. Similarly, the Phi family exhibits systemic instability and high IR across all variants. Such findings highlight that strategic ambiguity can destabilize basic instruction-following in otherwise capable models.

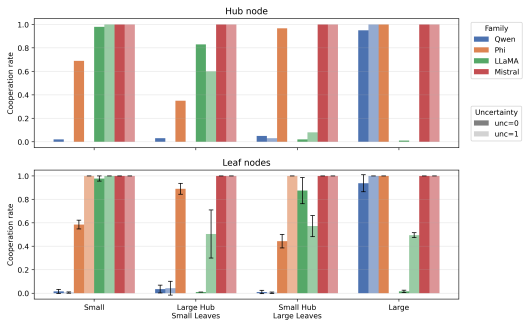


Figure 5: CoopRate in LLM vs LLM five-agent star structures, with and without uncertainty.

6 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: varying model families exhibit sharply atypical default policies, react differently to network structure, and vary widely in their sensitivity to action ordering (positional bias), deception, and uncertainty. Our results show that strategic behavior in LLM societies cannot be treated as a mere consequence of scale or capability alone, and yield a critical takeaway for the design of multi-agent systems: LLMs do not act rationally based on well-defined incentives, but their language-driven nature renders them highly fragile against linguistic influences, completely altering collective equilibria.

Limitations

Our study uses the Iterated Prisoner's Dilemma as a controlled diagnostic environment for analyzing strategic behavior in LLM agents. The use of fixed payoff matrices and simple line, cycle, and star topologies enables interpretable comparison across model families, graph positions, and prompt interventions, but does not aim to fully reproduce open-ended real-world interaction. We also evaluate representative agent configurations rather than exhaustively scaling the number of agents, since larger populations would substantially increase computational cost while making causal attribution harder.

Our findings are conditioned on the prompting protocol used to instantiate the agents. This is partly a limitation, but also central to the phenomenon under study: LLM agents are language-conditioned decision-makers, and their behavior can vary under prompt reformulations, misleading cues, uncertainty framings, or action-order changes. Finally, we focus on open-weight instruction-tuned models to ensure reproducibility and controlled comparison. Closed-source commercial models are not included due to limited access, cost, rate limits, and reduced reproducibility. Invalid responses in some configurations should also be interpreted carefully, as they may reflect format-following failures but also constitute a meaningful reliability failure in repeated-game settings.

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

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

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.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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.

Jinyi Ye, Lei Cao, Ding Chen, and Emilio Ferrara. 2026. Stop drawing scientific claims from llm social simulations without robustness audits. *Preprint*, arXiv:2605.18890.

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 Baseline analysis

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

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.

Our analysis commences from the more game-theoretic metrics corresponding to collaboration (CoopRate) and defection (RevRate). Figures 6 and 7 reveal substantial variation across model families regarding these metrics. Some LLMs display consistently high cooperation against 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.

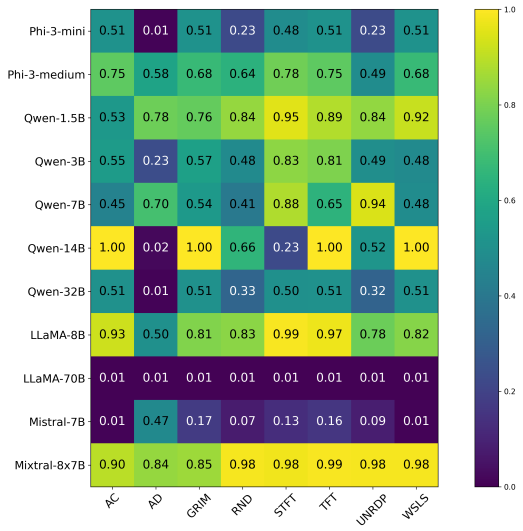


Figure 6: CoopRate (LLM vs deterministic player).

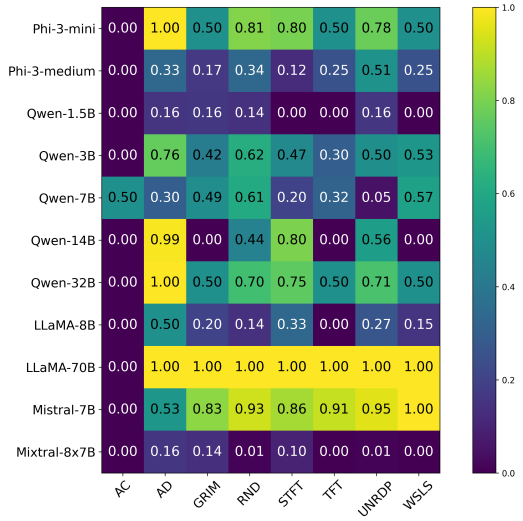


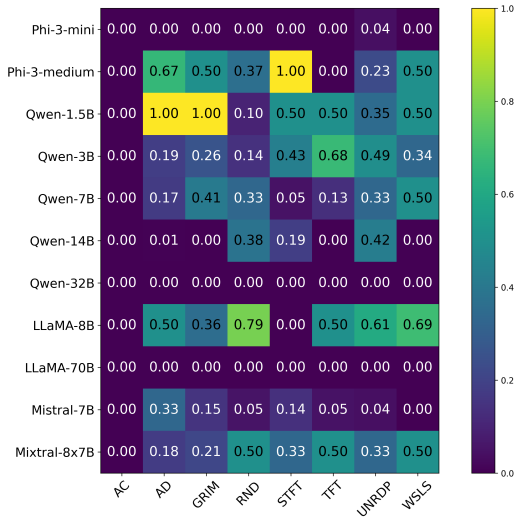
Figure 7: RevRate (LLM vs deterministic player).

Forgiveness and tolerance presented in Figure 8, 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 forgiveness across all opponents. Tolerance displays greater variation: cooperative models such as Mixtral-8×7B maintain high tolerance in many environments, whereas more

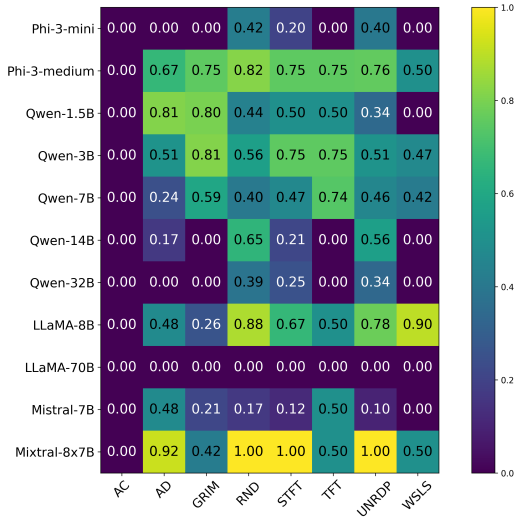
defensive models exhibit low tolerance, reflecting rapid retaliation after unilateral betrayal.

Performance metrics presented in Figure 9, reveal corresponding payoff patterns. Average cost (AvgCost) values align closely with the cooperation dynamics observed earlier: models that sustain cooperation against cooperative strategies achieve lower AvgCost, while those that either defect excessively or cooperate against adversarial opponents incur higher AvgCost. Relative cost (RelCost) further highlights strategic asymmetries across models. Some agents outperform their opponents in adversarial settings, yielding negative RelCost, while highly cooperative policies may lead to worse outcomes when paired with persistent defectors.

Finally, detection accuracy (DetectionAcc) 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 10. 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.



(a) Forgiveness



(b) Tolerance

Figure 8: Post-defection cooperation metrics in two-player experiments against fixed policies. Forgiveness measures recovery after mutual defection, while Tolerance measures cooperation following unilateral betrayal.

A.1 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 IR rates.

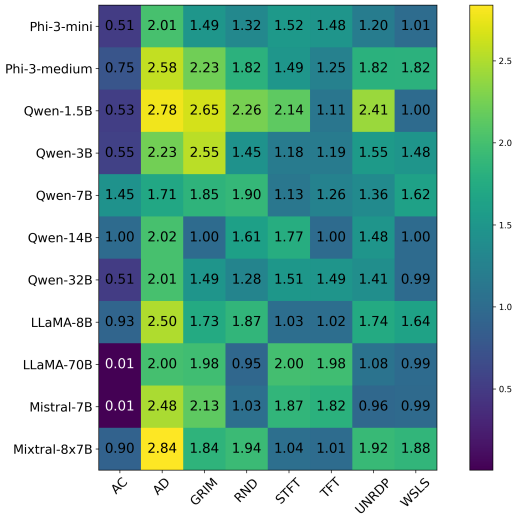
A.1.1 Line Topologies

Tables 5 and 6 report IR rates for line topologies. Table 5 considers settings where the LLM node interacts with fixed baseline opponents/neighbors, while Table 6 reports fully LLM-based line configurations with uncertainty either absent (*Unc.*=0) or present (*Unc.*=1).

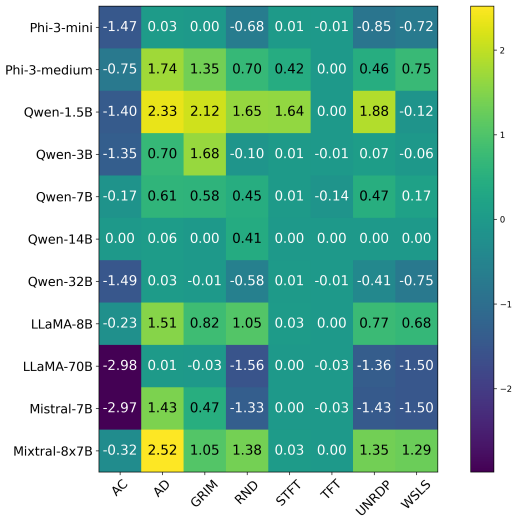
Overall, invalid responses are concentrated in smaller or less stable model variants. In the baseline-neighbor setting, the Phi family exhibits consistently high IR rates across both mini and medium variants, while Qwen-1.5B remains stable under the baseline prompt but becomes unstable under behavioral interventions. In fully LLM-based lines, invalid responses remain strongly dependent both on model and position. Phi medium models and Qwen-1.5B produce high IR rates in most positions, whereas LLaMA failures are concentrated on edge nodes under deterministic prompts and disappear under uncertainty prompts.

A.1.2 Cycle Topologies

In Table 7, we report the IR rates for cycle topologies under both simple prompts (*Unc.=0*) and prompts containing uncertainty cues (*Unc.=1*). Overall, the Phi family exhibits consistently high IR rates across all configurations, indicating sub-



(a) Average Cost



(b) Relative Cost

Figure 9: Performance metrics in two-player experiments. Average cost (AvgCost) measures cumulative payoff, while relative cost (RelCost) compares the agent's performance against its opponent.

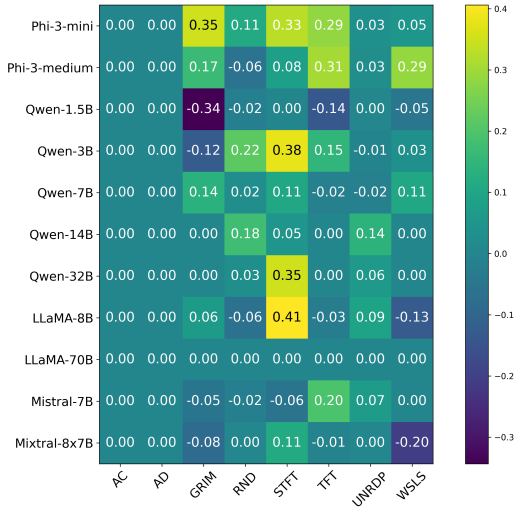


Figure 10: DetectionAcc measuring the correlation between an opponent's previous action and the agent's current response. Higher values indicate stronger responsiveness to opponent behavior.

Family	Model	Config.	Base.	Soft Dec.	Dec.
Phi	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
Qwen	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 5: Invalid-response (IR) rates for line topologies with baseline opponents/neighbors. Higher IR rates for each model variant and structure are highlighted per intervention (baseline vs soft deception vs deception). **Bold** further indicates the highest IR for each model per intervention.

stantial instability in cyclic multi-agent interactions. Smaller Phi models are particularly sensitive, often producing invalid outputs in nearly all rounds regardless of uncertainty. Ambiguous signal enhances the already elevated IR rates in the homogeneous medium configuration, denoting that uncertainty completely derails the agents' abilities regarding remaining within the valid game spectrum altogether.

The introduction of uncertainty does not uniformly increase instability though. In some cases, such as the fully homogeneous small Qwen configuration, uncertainty dramatically increases invalid-response rates for specific nodes, while other nodes remain relatively stable. These findings indicate that both model scale and interaction structure influence robustness, while uncertainty prompts can amplify instability.

Family	Config.	Unc.	Center	Edge1	Edge2
Phi	3 × mini	0	0.26	0.08	0.09
		1	0.0	0.0	0.0
	mini center + medium edges	0	0.0	0.97	0.96
		1	0.0	**0.99**	**0.99**
	medium center + mini edges	0	0.97	0.0	0.0
		1	**1.0**	0.0	0.0
	3 × medium	0	0.91	0.92	0.89
		1	**1.0**	0.98	0.98
Qwen	3 × 1.5B	0	**0.96**	**0.96**	**0.96**
		1	0.1	**0.96**	**0.96**
LLaMA	8B center + 70B edges	0	0.0	**1.0**	**1.0**
		1	0.0	0.0	0.0
	3 × 70B	0	0.0	**1.0**	**1.0**
		1	0.0	0.0	0.0

Table 6: Invalid-response (IR) rates for fully LLM-based line topologies across model families under the presence of uncertainty. For each model family, highest IR rates per position (Center, Edge1, Edge2) are denoted in **bold**.

Family	Config.	Unc.	Node0	Node1	Node2	Node3
Phi	4 × mini	0	**1.00**	**1.00**	**1.00**	1.00
		1	**1.00**	0.00	**1.00**	**1.00**
	mini-medium-mini-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
Qwen	4 × 1.5B	0	0.36	0.47	0.44	0.31
		1	**0.96**	0.09	**0.96**	0.09

Table 7: Invalid-response (IR) rates for cycle topologies. For each model family, highest IR rates are denoted with **bold**.

A.1.3 Star Topologies

Invalid rates observed in star configurations are presented in Table 8. The Phi family again exhibits reasonable IR rates across all four configurations, with the medium variant generally producing higher IR rates than the mini variant. Mixed-size Phi configurations further reveal strong positional effects, where either the center node or specific leaves become substantially more unstable depending on model allocation within the topology. In the Qwen family, the high IR rates observed in small-model configurations are primarily driven by the presence of the 1.5B model, which generally produces a large number of invalid responses. However, the instability in Qwen remains localized to specific nodes (e.g. the presence of Qwen in specific leaves) rather than affecting the entire topology uniformly (the placement of Qwen as a hub results in fully valid responses).

Interestingly, the incorporation of uncertainty has a particularly strong effect on the LLaMA family. While both the 8B and 70B models remain completely stable under deterministic prompts, uncertainty prompts trigger extreme failures in specific nodes, reaching IR rates of 100% in several cases. In the purely large configuration, the potent hub node gets fully outside the valid label spectrum due to the introduction of ambiguous signals, whereas in the mixed configuration (where the smaller 8B variant sits in the center), robustness is preserved from the side of the hub. This sizeable behavior shift suggests that uncertainty cues can destabilize even large models, with the effect depending strongly on both node position and topology composition.

Family	Config.	Unc.	Center	L1	L2	L3	L4
Phi	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
Qwen	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
LLaMA	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 8: Invalid-response (IR) rates for star topologies across model families. Higher IR values for each model family is denoted with **bold**.

A.2 The impact of uncertainty

As a complement to the previously analyzed IR rates, we study the impact of uncertainty-induced prompts over LLM-only line structures comprising three agents, both in homogeneous (exclusively small -*S*- models or exclusively large -*L*- ones) and heterogeneous (mix of *S* and *L* models) configurations. Results including CoopRates for line topologies are presented in Table 9.

It occurs that the impact of uncertainty in purely LLM-based line topologies is not uniform: ambiguity can either preserve cooperation, induce defensive collapse, or disrupt the validity of the induced policy depending on the model family and scale composition. In game-theoretic terms, the uncertainty prompt perturbs agents' beliefs about neighbor predictability while leaving the payoff matrix unchanged. Thus, if agents behave in a risk-

Family	Config.	Uncertainty	CR_{center}	CR_{edge}	IR_{center}	IR_{edge}
LLaMA	L-L-L	0	0.010	–	0.000	1.000
	L-L-L	1	0.340	0.715	0.000	0.000
	L-S-L	0	1.000	–	0.000	1.000
	L-S-L	1	1.000	0.750	0.000	0.000
	S-L-S	0	0.010	1.000	0.000	0.000
	S-L-S	1	0.500	1.000	0.000	0.000
	S-S-S	0	1.000	1.000	0.000	0.000
	S-S-S	1	1.000	1.000	0.000	0.000
Mistral	L-L-L	0	1.000	1.000	0.000	0.000
	L-L-L	1	1.000	1.000	0.000	0.000
	L-S-L	0	1.000	1.000	0.000	0.000
	L-S-L	1	1.000	1.000	0.000	0.000
	S-L-S	0	1.000	1.000	0.000	0.000
	S-L-S	1	1.000	1.000	0.000	0.000
	S-S-S	0	1.000	1.000	0.000	0.000
	S-S-S	1	1.000	1.000	0.000	0.000
Phi	L-L-L	0	1.000	1.000	0.910	0.905
	L-L-L	1	–	0.750	1.000	0.980
	L-S-L	0	1.000	1.000	0.000	0.965
	L-S-L	1	1.000	1.000	0.000	0.990
	S-L-S	0	1.000	1.000	0.970	0.000
	S-L-S	1	–	1.000	1.000	0.000
	S-S-S	0	1.000	1.000	0.260	0.085
	S-S-S	1	1.000	1.000	0.000	0.000
Qwen	L-L-L	0	1.000	1.000	0.000	0.000
	L-L-L	1	1.000	1.000	0.000	0.000
	L-S-L	0	0.065	0.010	0.000	0.000
	L-S-L	1	0.000	0.010	0.000	0.000
	S-L-S	0	0.055	0.0075	0.000	0.000
	S-L-S	1	0.015	0.000	0.000	0.000
	S-S-S	0	0.030	0.0283	0.320	0.320
	S-S-S	1	0.007	0.0017	0.033	0.320

Table 9: $\mathrm{CoopRates}$ and IR rates for 3-agent line topologies exclusively composed of LLM agents. Configurations indicate the model scale at the left edge, center, and right edge positions, respectively. Edge values are averaged over the two edge nodes; – indicates that $\mathrm{CoopRate}$ is undefined because the corresponding position produced fully invalid responses.

averse way, we would expect them to move toward safer defection; if they preserve expectations of reciprocity, cooperation should remain stable.

The clearest defensive response appears in the Qwen family. Homogeneous large-Qwen lines remain fully cooperative with and without uncertainty, indicating that large Qwen agents sustain a cooperative convention even under ambiguous opponent beliefs. However, mixed and small-Qwen lines are already near-defective without uncertainty and become even more defect-oriented when uncertainty is introduced: for example, the center cooperation rate in L-S-L drops from 0.065 to 0.000, while in S-L-S it drops from 0.055 to 0.015, and edge cooperation also remains close to zero. This suggests that smaller or heterogeneous Qwen societies resolve ambiguity by retreating toward defensive defection rather than by preserving reciprocal cooperation. Importantly, this is not simply a validity artifact, since IR rates remain zero in most of these mixed configurations. The effect is therefore strategic rather than merely format-related.

LLaMA exhibits the opposite and more surprising pattern. In configurations involving large models, uncertainty increases cooperation and simultaneously removes invalid edge behavior. In L-L-L, center cooperation rises from 0.010 to 0.340 and edge cooperation becomes defined at 0.715, while edge invalid rates fall from 1.000 to 0.000. Similarly, in S-L-S, center cooperation rises from 0.010 to 0.500 while edges remain fully cooperative. Rather than inducing caution, uncertainty appears to disrupt a rigid defective or invalid regime and moves the system toward a more cooperative, policy-valid interaction pattern. This makes LLaMA the clearest case where ambiguity does

not operate as risk aversion, but as a prompt-level perturbation that changes equilibrium selection.

Mistral is largely invariant to uncertainty. Across all line configurations, both center and edge agents maintain full cooperation with zero invalid responses. This stability can be interpreted in two ways: either Mistral robustly selects a cooperative equilibrium under strategic ambiguity, or it follows a strong cooperative prior that is weakly conditioned on uncertainty. The table alone cannot distinguish these interpretations, but the absence of both strategic and validity shifts makes Mistral the most uncertainty-stable family in this setting.

Phi is best interpreted through policy validity rather than cooperation alone. When responses are valid, cooperation is generally high; however, several configurations exhibit severe invalid-response rates, especially when medium models occupy sensitive positions. For example, in L-L-L, center invalidity rises to 1.000 under uncertainty, making center cooperation undefined, while edge invalidity remains very high. In S-L-S, the medium center also becomes fully invalid under uncertainty. Therefore, for Phi, uncertainty does not merely shift the agent between cooperation and defection; it often prevents the agent from reliably mapping the strategic prompt to the legal action set. This corresponds to a failure of policy implementability rather than a stable game-theoretic strategy.

A.3 Dynamic analysis

Aggregate $\mathrm{CoopRates}$ provide a useful summary of induced behavior, but they do not fully identify the strategy followed by an LLM agent over repeated interaction. Two agents may obtain similar final $\mathrm{CoopRates}$ while following qualitatively different trajectories: one may rapidly converge to a stable response, another may alternate between actions, delay adaptation, or remain exploitable for a substantial part of the game. For this reason, we complement aggregate metrics with $\mathrm{CoopRate}$ trajectories over the 100 *IPD* rounds. Figures 11 and 12 show how cooperation evolves against fixed opponent policies under the two semantically equivalent action-order variants. This analysis is particularly important in *IPD*, where strategic competence depends not only on the final cooperation level, but also on the speed, stability, and opponent-sensitivity of adaptation.

The cooperation trajectories in Figures 11 and 12 reveal distinct model-family regimes. Qwen-14B exhibits the clearest adaptive profile among the large Qwen models: it sustains cooperation against Always Cooperate (AC), while rapidly suppressing cooperation against Always Defect (AD), typically within the first few rounds. This suggests that the model can distinguish between reciprocal and exploitative environments and move toward the appropriate repeated-game response. Qwen-32B also converges to defection against AD, avoiding long-term exploitation, but its behavior against AC is more sensitive to action ordering. Under the C_{first} prompt variant, it preserves cooperation, whereas under D_{first} it shifts toward early defection, indicating that the same cooperative environment can be interpreted either as an opportunity for mutual cooperation or as an opportunity for exploitation depending on prompt formulation. This supports the main-text observation that action-order sensitivity affects not only aggregate cooperation, but also the temporal path by which policies are realized.

The LLaMA family shows a different dynamic profile. LLaMA-70B is strongly defect-oriented across both cooperative and adversarial opponents, converging to defection with little dependence on action-order condition. This makes it relatively prompt-stable, but not necessarily strategically adaptive: its stability comes from a defensive policy that often fails to exploit the long-term value of cooperation against AC. LLaMA-8B is more cooperative, but its long-horizon trajectories are less stable and can be affected by invalid outputs in later rounds. Thus, the contrast within the LLaMA family suggests that larger scale does not necessarily produce more cooperative or more opponent-sensitive repeated-game behavior; instead, the larger model appears to implement a more rigid loss-minimizing policy.

Mistral models occupy an almost opposite regime. Mistral-7B rapidly defects against AC in both prompt-order conditions, thereby exploiting a cooperative opponent, but its response to AD is condition-dependent. Under C_{first}, it recovers after the initial rounds and converges toward mutual defection, whereas under D_{first} it remains cooperative for longer and is therefore exploitable before eventually adapting. This indicates that prompt ordering can affect not only the final cooperation level, but also the adaptation delay. Mixtral-8×7B maintains a more cooperative profile overall, including against AD, suggesting weaker suppression of cooperation in adversarial environments. In game-theoretic terms, this behavior may reflect a strong cooperative prior rather than fully condi-

tional cooperation, since the model does not always switch quickly toward a best response when facing persistent defection.

Overall, Figures 11 and 12 show that repeated-game behavior should not be evaluated only through static CoopRate. The relevant game-theoretic distinction is whether the induced policy is **conditionally cooperative**: sustaining cooperation when reciprocity is available, retaliating or shifting to defection when exploitation is persistent, and doing so without excessive delay or invalid actions. Trajectory-level cooperation evidence therefore helps separate genuine strategic adaptation from default cooperation, rigid defensive defection, prompt-induced instability, and failure to instantiate a valid policy.

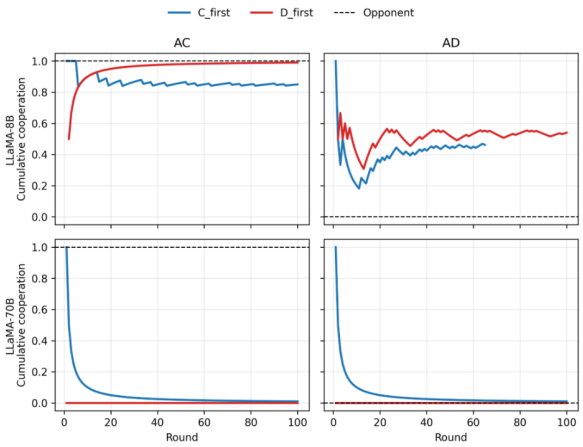


Figure 11: CoopRate trajectories over 100 *IPD* rounds against fixed opponent policies under C_{first} and D_{first} prompt variants.

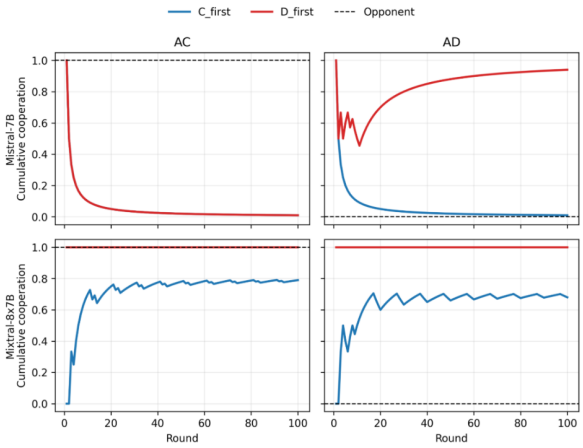


Figure 12: Additional CoopRate trajectories over 100 *IPD* rounds against fixed opponent policies under C_{first} and D_{first} prompt variants.

A dynamic analysis is also performed over AvgCost trajectories (Figure 13 for Qwen models, Figure 14 for Llama models, Figure 15 for Mistral models), complementing the CoopRate trajectories analysis by indicating whether and when LLMs outperform fixed opponents from Table 1. Solid lines below their corresponding dashed lines denote lower cost for the LLM-agents, i.e., better performance. Convergence to an AvgCost of 1 (e.g., Qwen-14B vs AC, LLaMA-8B vs AC and Mixtral-8x7B vs AC) corresponds to mutual cooperation, which is optimal against AC. Convergence to an AvgCost of 2 (e.g., Qwen-14B vs AD, Mistral-7B vs AD under C_{first}, LLaMA-70B vs AD) indicates mutual defection, ultimately corresponding to the Nash equilibrium. Divergence between solid and dashed lines of the same color reveals asymmetric outcomes, where one player is systematically exploited. In such cases (e.g. Mixtral-8x7B vs AD), the agent with the higher AvgCost trajectory incurs greater losses, indicating failure to adapt to the opponent's strategy.

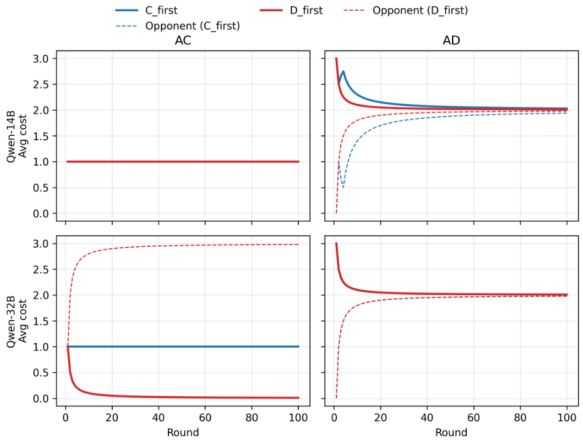


Figure 13: AvgCost trajectories over 100 *IPD* rounds for Qwen models against fixed opponent policies under C_{first} and D_{first} prompt variants.

B Additional LLM vs LLM Results

B.1 Homogeneous graphs

We study the positional bias of Qwen family in *homogeneous* graphs using the *SI* metric in Table 10. Comparisons between *heterogeneous* (Table 3) and *homogeneous* Qwen structures denote that **cooperation is fragile under heterogeneity**; even though using the exact same large Qwen model across the graph network leads to cooperative behaviors (low SI values), mixing with smaller Qwen counterparts results in collapsed robustness under any topology used (line, cycle, star). As a general takeaway, even

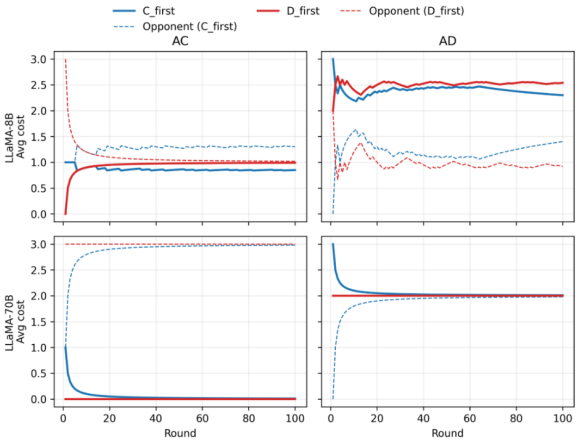


Figure 14: AvgCost trajectories over 100 *IPD* rounds for Llama models against fixed opponent policies under C_{first} and D_{first} prompt variants.

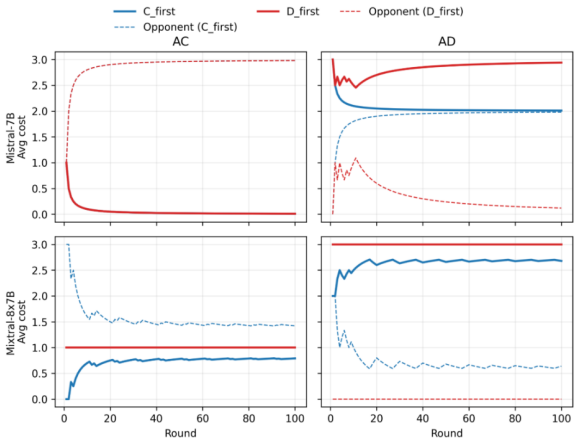


Figure 15: AvgCost trajectories over 100 *IPD* rounds for Mistral models against fixed opponent policies under C_{first} and D_{first} prompt variants.

a few weak agents are capable of destabilizing the system overall.

Topology	Scale	Position	SI	$\text{CR}_{C_{\text{first}}}$	$\text{CR}_{D_{\text{first}}}$
Line	Small	Center	0.970	0.030	1.000
	Small	Edges	0.972	0.028	1.000
	Large	Center	0.000	1.000	1.000
	Large	Edges	0.000	1.000	1.000
Cycle	Small	–	0.956	0.044	1.000
	Large	–	0.000	1.000	1.000
Star	Small	Hub	0.980	0.020	1.000
	Small	Leaves	0.985	0.015	1.000
	Large	Hub	0.050	0.950	1.000
	Large	Leaves	0.062	0.938	1.000

Table 10: Prompt-order positional bias in *homogeneous* Qwen graph configurations.

B.2 Line Interaction Experiments

In the following Figures 16, 17, 18, 19 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 16), the impact of deception is inconsistent across configurations. The *phi-3-mini-instruct* model shows small, non-systematic shifts in CoopRates, 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 IR rates observed in baseline experiments, these apparent shifts likely reflect output instability rather than deliberate adaptation to deceptive signals.

Across the Qwen family (Figure 17), 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 diminishes 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 18, 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 reduces these shifts, especially for the 70B variant, indicating partial reliance on observed interaction history when

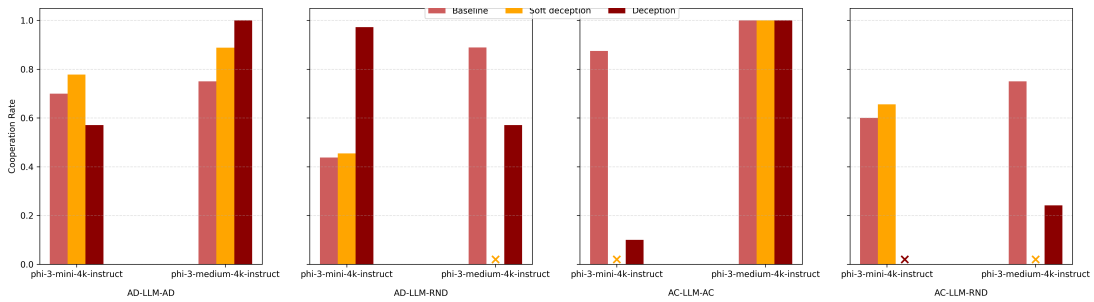


Figure 16: CoopRate of Phi models under baseline prompting, deception, and soft deception across line interaction configurations.

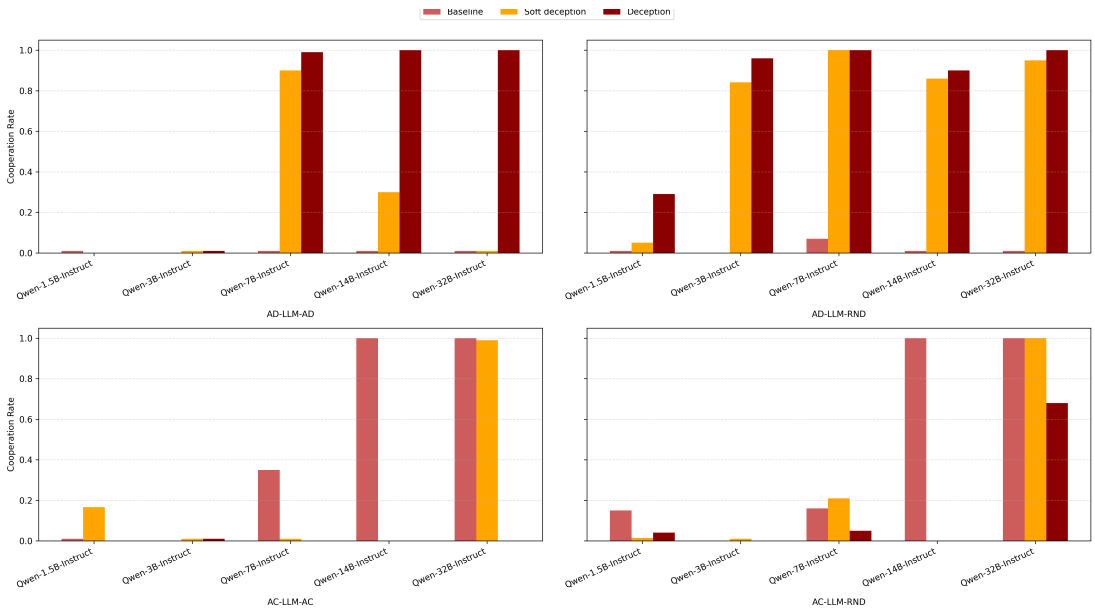


Figure 17: CoopRate of Qwen models under baseline prompting, deception, and soft deception.

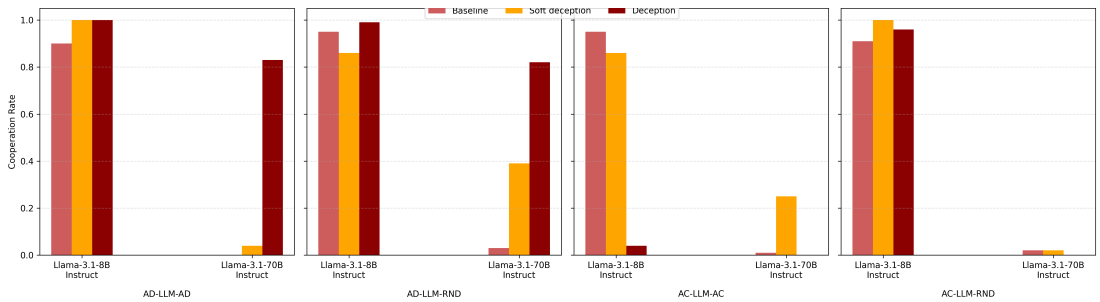


Figure 18: CoopRate of LLaMA models under baseline prompting, deception, and soft deception.

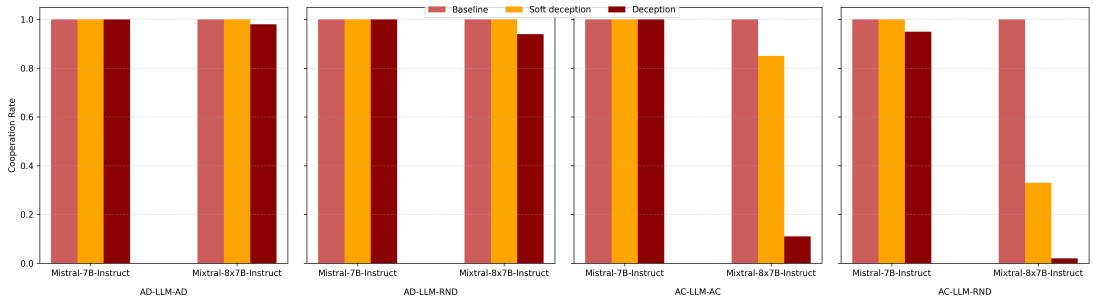


Figure 19: CoopRate of Mistral-family models under baseline prompting, deception, and soft deception.

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 19, 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.

C Model Identifiers

Our experiments consider the following model families: Phi-3 (Abdin et al., 2024), Qwen 2.5 (Qwen et al., 2025), LLaMa 3.1 (Grattafiori et al., 2024) and Mistral (Jiang et al., 2024).

Table 11 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 11: Exact model identifiers used in the experiments.

D Prompts

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

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

The following prompt template was used in two-player experiments against fixed 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 SI experiments, semantically equivalent prompt variants were generated by changing only the ordering of the *C* and *D* response options.

D.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.

D.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."