Evolution of Cooperation in LLM Societies : A Multilingual Examination

Kriti Mahajan [* 1]

Abstract

Understanding the dynamics of Large Language Model (LLM) interactions is becoming increasingly important as more LLM systems become autonomously deployed, particularly with the proliferation of agentic AI. We investigate the evolutionary dynamics of cooperation in multi-agent LLM systems where agents play the repeated Prisoner's Dilemma while embedded in diverse network topologies. Strategies are expressed as natural-language prompts in English, French, and Hindi, and evolve over 100 generations via replication and mutation across a population of 50 GPT-4o agents. We find that cooperation persists long-term across all languages and topologies, but stability depends heavily on network structure and linguistic framing. Scale-free networks sustain the highest cooperation, while centralized topologies (complete, star) exhibit extreme oscillations. French networks show the highest volatility; Hindi networks drift toward defection; English networks maintain balanced strategy distributions. A positive cooperation–fitness correlation holds in most topologies except complete and star networks, with strength varying by language. These results highlight critical challenges for deploying multilingual LLM agent systems.

1. Introduction

Understanding the conditions under which cooperation emerges remains a central challenge in evolutionary game theory, network science, and artificial intelligence. The canonical repeated Prisoner's Dilemma (PD) has long served as a testbed for studying cooperation, with insights spanning disciplines from biology to economics. Classical approaches typically formalize strategies as fixed rules or stochastic automata; however, the rise of large language models (LLMs) offers a new domain in which to examine strategic agent behavior. Because LLMs interpret and enact strategies expressed in natural language, they allow us to probe how linguistic framing interacts with structural and evolutionary dynamics in artificial LLM agent-based societies.

This paper explores cooperation in LLM-mediated simulations of the iterated PD across multiple graph network topologies. Agents are initialized with diverse strategies—ranging from cooperative, neutral to defecting - encoded in natural language and distributed according to fixed proportions. Interaction outcomes are determined by an LLM (GPT-4o), which selects actions based on strategy descriptions, partner histories, and context prompts. Evolutionary dynamics are modeled through generational updates where successful strategies proliferate subject to stochastic mutation.

We systematically vary network topology to investigate how structural constraints interact with strategy distributions across 3 languages: English, French, and Hindi. By combining linguistic, structural, and evolutionary dimensions, our study addresses three central questions: 1) How do different network topologies influence the evolution of cooperation in LLM-mediated populations?; 2) Does the linguistic encoding of strategies affect emergent cooperation dynamics? ; 3) How does cooperation relate to population-level fitness across topologies and languages?

Our experiments show that while cooperation persists across generations, the stability of cooperative equilibria depends heavily on both topology and language. Only a few structures—lattice, ring, and tree—consistently reach stable outcomes, while the other networks are marked with high volatility across languages. Multilingual runs reveal similar trends but differ in stability and strategy adoption, with French networks showing high volatility and Hindi networks favoring defection. Cooperation correlates positively with population fitness in all but complete and star topologies. Our findings have direct implications for the safe deployment of multilingual LLM agent systems. Many applications assume that model behavior is consistent across languages. However, our results show that semantically equivalent strategies produce qualitatively different behavioral outcomes depending on the language, undermining predictability and raising concerns about alignment drift. Safety evaluations that test LLM agent interactions only in

[1]Cross Border Science, Amazon. Correspondence to: Kriti Mahajan <kritimhj@amazon.com>.

Accepted to the 2nd Workshop on Compositional Learning at ICML 2026, Seoul, South Korea., Copyright 2026 by the author(s).

English may miss critical vulnerabilities that emerge under other linguistic framings, and alignment mechanisms may need cross-linguistic validation to ensure robustness. For AI safety, these findings emphasize the need to treat linguistic encoding as a control variable rather than an incidental design choice.

Related Work: Recent work has examined LLMs as strategic actors in game-theoretic environments. Sun et al. (2023) provided a broad taxonomy spanning evaluation, alignment, and game-solving. Akata et al. (2025) found that LLMs excel in self-interested dilemmas but struggle in coordination games requiring mutual adaptation. Most relevant to our work, Payne and Alloui-Cros (2024) showed that LLMs survive evolutionary PD tournaments and exhibit model-specific strategic fingerprints, and Zeng et al. (2024) demonstrated that environmental feedback reshapes individual LLM agent behavior in repeated games. However, both studies operate in well-mixed (unstructured) populations and examine single-language settings. We extend this line of work by embedding LLM agents in structured networks under population-level evolutionary dynamics and introducing cross-linguistic variation as an experimental axis. Separately, foundational work in evolutionary game theory established that network topology strongly mediates cooperation—scale-free networks stabilize prosocial clusters while centralized structures amplify exploitation (Axelrod, 1984; Nowak & May, 1992; Santos et al., 2005)—but assumes fixed deterministic strategy rules rather than natural-language prompts interpreted by an LLM. We ask whether the same structural effects hold when strategies are natural-language prompts interpreted by an LLM, introducing stochasticity and linguistic sensitivity absent from classical models.

Our contributions. This paper advances the study of cooperation in artificial societies along two dimensions. (1) We provide the first systematic comparison of LLM-mediated cooperation across nine network topologies under evolutionary dynamics (replication and mutation), demonstrating that graph structure determines whether cooperation stabilizes or oscillates. (2) We introduce multilingual strategy encoding as an experimental variable, revealing that semantically equivalent strategies yield divergent behavioral outcomes across English, French, and Hindi—with direct implications for alignment robustness. While prior work has studied LLM agents in evolutionary games (Payne & Alloui-Cros, 2024) and adaptive feedback loops (Zeng et al., 2024), our work is distinguished by the combination of structured populations, generational evolution, and cross-linguistic analysis.

2. Methodology

We study cooperation in repeated LLM-driven networked Prisoner's Dilemma games, where agents use natural-language prompt strategies, showing how strategy descriptions, network structure, and language shape collective outcomes.

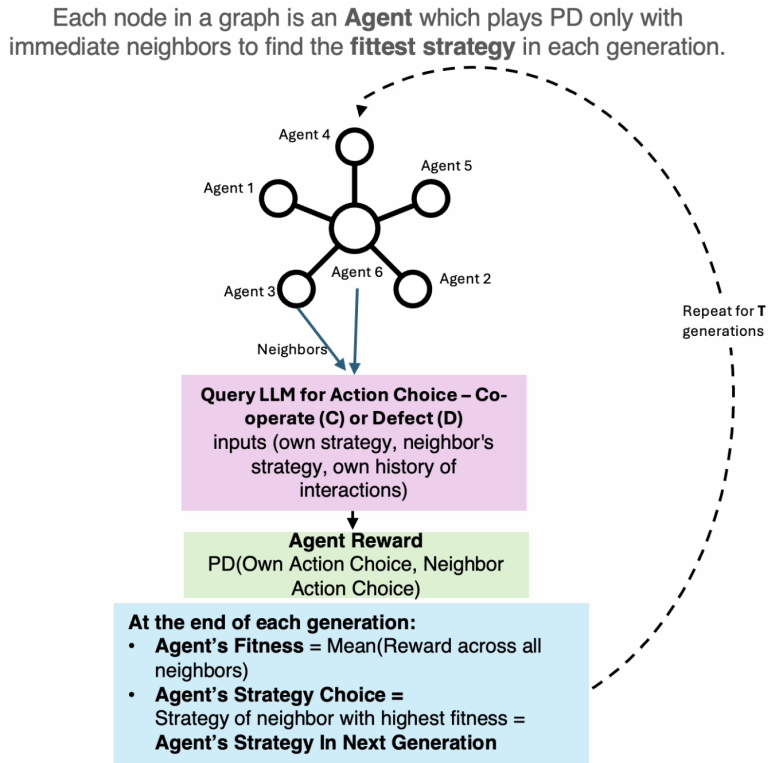


Figure 1. Iterative Prisoner's Dilemma Tournament

2.1 Simulation Environment and Agent Design: (see Algorithm 1 and Figure 1) We simulate a repeated PD in a population of N=50 agents interacting over T=100 generations across nine network topologies: small-world (Watts–Strogatz, k=4, p=0.2), scale-free (Barabási–Albert, m=2), random (Erdős–Rényi, p=0.1), lattice (2D grid), ring, tree (balanced binary), star, complete, and modular (stochastic block models with dense intra-community and sparse inter-community ties). These topologies were chosen to span the spectrum from highly centralized to highly clustered structures, enabling comparison of cooperative dynamics under different structural constraints.

Each agent occupies a graph node and plays the PD with its neighbors. Strategies are natural-language prompts grouped into three categories. Cooperative: always cooperate, generous-unless-betrayed, Tit-for-Tat, Tit-for-Tat with Forgiveness. Neutral: cautious/skeptical, cooperate most of the time, Win-Stay Lose-Shift, Pavlov. Defective: always defect, defect-if-unsure, Grim Trigger (Full strategy texts in all three languages are in Appendix A). Strategies are initialized at fixed proportions (0.1 each; 0.05 for Pavlov and WSLS) and evolve through imitation dynamics with mutation probability μ=0.02.

Each agent's action is determined by querying GPT-4o with: (1) the agent's strategy description, (2) the partner's strategy description, and (3) the full history of past interactions with that partner enabling the LLM to approximate stateful contingent strategies (e.g., Tit for Tat, Pavlov) through natural-language reasoning. The model responds with either the cooperate or defect token for the language at hand. If responses were invalid, stochastic fallback rules ensured game play continuity. To test cross-linguistic robustness, all prompts and tokens are translated into English, French, and Hindi using localized PD terminology (see Appendix A).

2.2 Evolutionary Dynamics: At each generation's end, agent fitness is computed as the mean of payoffs across all interactions. Agents adopt their most successful neighbor's strategy with probability 1 - μ, or mutate to a random strategy with probability μ (for pseudocode see Algorithm 1).

2.3 Experimental Setup and Evaluation: For each topology–language pair, we run 5 repetitions. To reduce API calls, a symmetric prompt-pair cache stores the first response for each strategy pair per generation i.e. the first LLM response for a given strategy pair is stored, and the complementary response is assumed for the partner; this reduces the number of API calls but introduces an artificial symmetry. We record the following metrics per generation in each topology and for each language: 1) Behavioral statistics - percentage of agents choosing the cooperative action (token); 2) Population composition: percentage of agents using cooperative, neutral, or defective prompts; 3) Fitness statistics: mean fitness across the population. Per-topology results are aggregated across repetitions, yielding mean, standard deviation, and min–max envelopes for cooperation rates and strategy proportions.

3. Results

3.1 Cooperation Dynamics Across Topologies : Across languages, experiments show that cooperation rates exhibit substantial temporal volatility, with few topologies converging to stable equilibria. Instead, populations frequently oscillate between phases of high and low cooperation. Notably, network topology strongly conditions these dynamics.

The English (Figure 2.a.a) experiments exhibit the above. Scale-free networks sustain the highest long-run cooperation levels, with cooperation gradually increasing to 0.6–0.7 after 40 generations, consistent with the amplification of cooperative strategies through high-degree hubs. Modular and random networks also support relatively high cooperation on average, though outcomes vary widely across runs, suggesting that inter-community structure can stabilize local cooperation but remains vulnerable to invasion by defectors. However, complete and star networks dis-

Algorithm 1 LLM-Guided Multi-Agent Prisoner's Dilemma Simulation

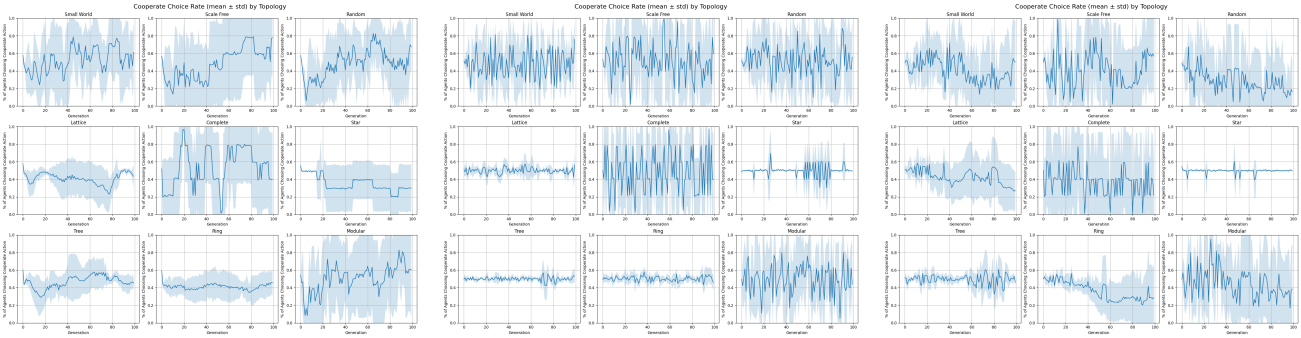
Require: Network $\mathcal{G} = (V, E)$, mutation rate μ, number of generations T
Ensure: Evolution of agent strategies over T generations

Initialization:
for each agent $i \in V$ **do**
  Assign an initial natural-language strategy description S_i
end for

for generation $t = 1$ to T **do**
  ▷ — Interaction Phase —
  for each agent $i \in V$ **do**
    for each neighbor $j \in \mathcal{N}(i)$ **do**
      LLM Input: (a) Agent i's strategy S_i, (b) Partner j's strategy S_j, (c) History of past interactions H_{ij}
      Query LLM (e.g., GPT-4o) $\rightarrow$ Action $a_i \in \{\text{Cooperate}, \text{Defect}\}$
      Agents i and j play PD $\Rightarrow$ Receive payoffs P_{ij}, P_{ji}
    end for
  end for
  ▷ — Fitness Evaluation —
  for each agent $i \in V$ **do**
    Compute mean fitness: $\bar{F}_i = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} P_{ij}$
  end for
  ▷ — Strategy Update —
  for each agent $i \in V$ **do**
    Identify most successful neighbor: $k^* = \arg\max_{j \in \mathcal{N}(i)} F_j$
    Update next-generation strategy:

$$S_i(t+1) = \begin{cases} S_{k^*}(t), & \text{with probability } 1-\mu, \\ S_{\text{rand}}, & \text{with probability } \mu \end{cases}$$

  end for
end for



(a) English *(b)* French *(c)* Hindi

(a) Percentage of Agents Choosing Cooperation per Generation

(d) English *(e)* French *(f)* Hindi

(b) Strategy Family Evolution per Generation

Figure 2. Cross-lingual comparison of cooperation behavior and strategy evolution across English, French, and Hindi settings.

play system-wide oscillations, reflecting the vulnerability of centralized structures to exploitation. In complete graphs, universal coupling causes rapid swings between near-total cooperation and near-total defection, reflecting global contagion of whichever strategy temporarily dominates. In star networks, the central hub suppresses cooperation, driving long-run cooperative action averages to 0.2–0.3. Lattice , tree and ring topologies maintain moderate cooperation (0.3–0.5), with relatively dampened fluctuations due to localized interactions. Small-world network occupies an intermediate regime, exhibiting cooperation rates around 0.4–0.6 with persistent oscillations. These findings highlight that network heterogeneity and modularity promote resilience of cooperation, while centralization and complete connectivity destabilize it.

Similar trends do not hold across languages reflecting the importance of language framing. French runs (Figure 2.a.b) broadly preserve these topological effects but display far greater volatility. Cooperation levels oscillate markedly within runs, cycling between high and low values in short timescales. Cooperation tends to stabilize and the fluctuations are smaller in structured networks such as lattices, rings and trees. The fluctuations across all topologies suggest that French language encoding introduces a degree of unpredictability into LLM-mediated decisions, amplifying fluctuations in cooperation. Hindi experiments (Figure 2.a.c) exhibit less volatility than the French experiments but more than the English experiments. Additionally, of the three languages, it is the only one which has a tendency to drift toward defection, which shows a consistent reduction in percentage of cooperative actions in the lattice, ring, modular and random topologies. This suggests for some topologies that Hindi language encoding biases dynamics toward exploitative equilibria.

3.2 Strategy Composition and Diversity: This analysis provides insights into how entire strategy families rise, coexist, and decline under evolutionary dynamics across topologies and languages. In the English experiments (Figure 2.b.d), all three strategy families persisted across most topologies, though with frequent shifts in dominance. Neutral strategies typically occupied a stabilizing middle share (0.2–0.4), while cooperative and defecting families alternated in control. Scale-free networks supported the eventual dominance of cooperative strategies, while lattice, ring, and

tree structures maintained long-run coexistence. By contrast, complete and star networks showed sharp oscillations between extremes, reflecting the instability of centralized interactions. Majority of topologies in the French experiments (Figure 2.b.e) preserve this pattern of co-existence of multiple strategy families but with much greater volatility than the English experiments. However, in Hindi experiments (Figure 2.b.f) all topologies except scale-free display a systematic drift toward dominance of defective families in the long run.

3.3 Relationship Between Cooperation and Fitness: Table 1 and Figure 3 display the relationship between cooperation and fitness across topologies and languages. While there is an overall positive association between cooperation and fitness the strength differs systematically by language. Firstly, the strength of the effect depends on network structure. Illustratively, complete and star networks display flat or noisy relationships: fitness remains low regardless of cooperation levels, because centralized structures allow defectors to undermine collective outcomes. Secondly, across all the topologies, the correlation varies substantially across languages, pointing to interaction effects between linguistic interpretation and network heterogeneity. Complete and star networks are the only topologies where all languages converge to weak or negligible associations.

Topology	English	French	Hindi
Complete	0.118	-0.053	0.231*
Lattice	0.535***	0.119	0.657***
Modular	0.289**	0.061	0.420***
Random	0.567***	-0.230*	0.126
Ring	0.346***	0.003	0.691***
Scale-free	0.687***	-0.197*	0.485***
Small-world	0.213*	0.010	0.611***
Star	0.272**	-0.045	0.003
Tree	0.763***	-0.078	-0.359***

Table 1. Correlation coefficients between cooperation proportion and average fitness across network topologies and languages. Values shown are Pearson's r. Stars indicate significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Taken together, these findings underscore two points. First, network topology continues to shape long-run dynamics, with heterogeneous structures like scale-free networks consistently buffering cooperation, while centralized networks remain unstable across all languages. Second, linguistic representation systematically alters the balance of strategies: French language networks promote unstable but persistent coexistence; Hindi language networks have a tendency to drift toward eventual defection; while English language networks maintain more balanced strategy distributions over time . These results highlight a serious methodological

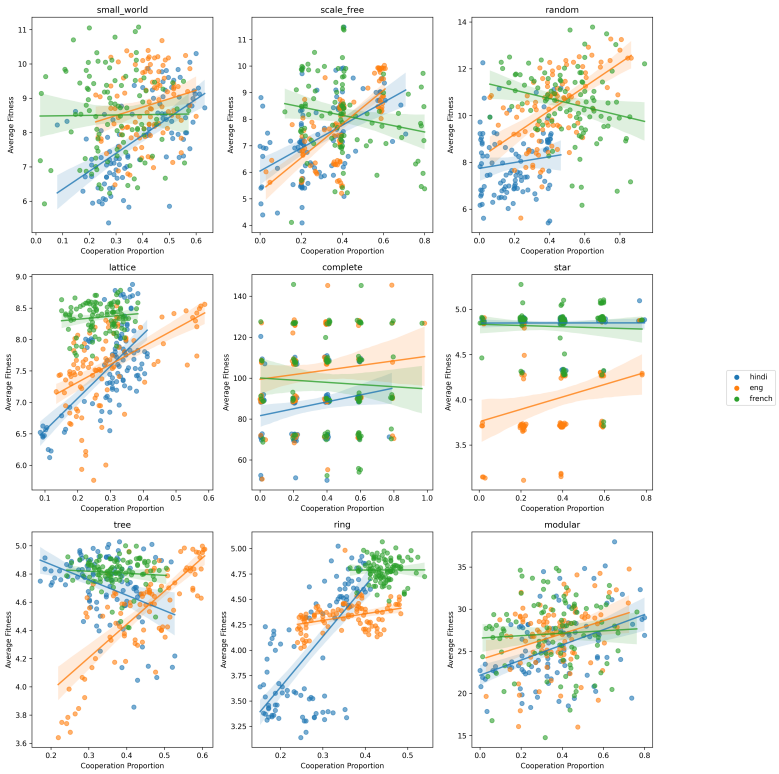


Figure 3. Co-operation vs Fitness: Comparison Across Languages

challenge: LLM-mediated strategy diversity is not invariant under translation. The same model and network designs can yield qualitatively different equilibria depending on the input language.

4. Limitations

The cache mirroring heuristic introduces systematic correlations that may not reflect natural dynamics. Since LLM outputs are influenced by their training data and prompt phrasing rather than precise algorithmic strategy definitions, resulting behaviors may be noisy or biased. Notably, LLMs tend to overpredict certain tokens—a well-documented issue—exhibiting biases toward majority labels, recent prompt content, and frequently seen pretraining tokens. Furthermore, interactions are limited to one-shot per edge per generation, with temporal dependencies emerging only through memory mechanisms and how the LLM interprets prior context, raising concerns about prompt invariance and consistency across generations. Additionally, the reliance on prompt-based behavior introduces sensitivity to even minor changes in wording, formatting, or context, which can significantly alter the model's output. This lack of prompt invariance makes our results overly reliant on the wording of our prompts, strategies and tokens chosen to denote cooperate / defect in each language. Note that the positive relationship between fitness and cooperation re-

flects the payoff structure of the Prisoner's Dilemma, where mutual cooperation yields higher joint returns than either mutual defection or unilateral exploitation and may change under different payoff structure of different games. In future work, the above analysis will be studied under prompt invariance, without caching and be extended to multiple games.

5. Conclusion

This paper investigates how cooperation emerges and evolves among LLM-based agents playing repeated Prisoner's Dilemma games across different network topologies and languages. By modeling strategies as natural-language prompts in English, French, and Hindi, we find that while cooperation persists long-term, the stability of cooperative equilibria varies significantly with both topology and language. These results highlight the sensitivity of LLM-mediated interaction to linguistic framing and structural context, raising important challenges for the design and interpretation of AI systems.

References

Axelrod, R. (1984). The Evolution of Cooperation.

Nowak, M. A., May, R. M. (1992). Evolutionary games and spatial chaos. Nature, 359(6398), 826–829. https://doi.org/10.1038/359826a0

Payne, K., Alloui-Cros, B. (2024). Strategic intelligence in large language models: Evidence from evolutionary game theory [Preprint].

Santos, F. C., Rodrigues, J. F., Pacheco, J. M. (2005). Graph topology plays a determinant role in the evolution of cooperation. Proceedings of the Royal Society B: Biological Sciences, 273(1582), 51–55. https://doi.org/10.1098/rspb.2005.3272

Sun, H., Wu, Y., Wang, P., Chen, W., Cheng, Y., Deng, X., Chu, X. (2023). Game theory meets large language models: A systematic survey with taxonomy and new frontiers [Preprint].

Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., Schulz, E. (2025). Playing repeated games with large language models. Nature Human Behaviour, 9, 1380–1390. https://doi.org/10.1038/s41562-025-01849-4

Zeng, W., Wang, B., Zhao, D., Qu, Z., He, R., Hou, Y., Hu, Q. (2024). Dynamic personality in LLM agents: A framework for evolutionary modeling and behavioral analysis in the Prisoner's Dilemma [Preprint].

Appendix

.1. Config: English

```
CONFIG = {
    "actions": {
        "cooperate": "C",
        "defect": "D",
    },

    "strategies": {
        "cooperative": [
            "Always cooperate with neighbors.",
            "Be generous unless betrayed.",
            "Use Tit for Tat: cooperate first, then mimic the partner’s last move.",
            "Use Tit for Tat with Forgiveness: copy your partner’s last move,
            but occasionally forgive.",
        ],
        "neutral": [
            "Be cautious and skeptical.",
            "Try to cooperate most of the time.",
            "Use Win-Stay Lose-Shift: repeat your last action if it worked; otherwise, switch
            "Use Pavlov strategy: start by cooperating. If both players made the same move
            last round (C/C or D/D), repeat your previous action. If the moves differed
            (C/D or D/C), switch your action.",
        ],
        "defective": [
            "Always defect, no matter what.",
            "Defect if unsure or confused.",
            "Use Grim Trigger: cooperate until defected against, then always defect.",
        ],
    },

    "strategy_proportions" : {
        "Always cooperate with neighbors.": 0.1,
        "Always defect, no matter what.": 0.1,
        "Be generous unless betrayed.": 0.1,
        "Be cautious and skeptical.": 0.1,
        "Try to cooperate most of the time.": 0.1,
        "Defect if unsure or confused.": 0.1,
        "Use Tit for Tat: cooperate first, then mimic the partner’s last move.": 0.1,
        "Use Tit for Tat with Forgiveness: copy your partner’s last move,
        but occasionally forgive.": 0.1,
        "Use Grim Trigger: cooperate until defected against,
        then always defect.": 0.1,
        "Use Win-Stay Lose-Shift: repeat your last action if it worked;
        otherwise, switch.": 0.05,
        "Use Pavlov strategy: start by cooperating. If both players made the same move
        last round (C/C or D/D) repeat your previous action. If the moves differed
        (C/D or D/C), switch your action.": 0.05
    }
}
```

.2. Config: French

```
CONFIG = {

    "actions": {
        "cooperate": "coopérer",
        "defect": "trahir",

    },

    "strategies": {
        "cooperative": [

            "Toujours coopérer avec ses voisins.",
            "Être généreux sauf en cas de trahison.",
            "Utiliser la stratégie Donnant-Donnant : coopérer en premier,
            puis imiter le dernier coup de l'adversaire.",
            "Utiliser Donnant-Donnant avec pardon : copier le dernier
            coup de l'adversaire, mais pardonner occasionnellement."
        ],

        "neutral": [
            "Être prudent et sceptique.",
            "Essayer de coopérer la plupart du temps.",
            "Utiliser la stratégie Gagner-Rester, Perdre-Changer : répéter la dernière
            action si elle a porté ses fruits ; sinon, changer.",
            "Utiliser la stratégie Pavlov : commencer par coopérer. Si les deux joueurs
            ont fait le même choix au tour précédent (coopérer/coopérer ou trahir/trahir),
            répéter l'action précédente. Si les choix ont différé
            (coopérer/trahir ou trahir/coopérer), changer d'action."

        ],

        "defective": [
            "Toujours trahir, quoi qu'il arrive.",
            "Trahir en cas de doute ou de confusion.",
            "Utiliser la stratégie du Grim Trigger : coopérer jusqu'à ce que
            l'adversaire trahisse, puis toujours trahir."
        ],

    },

    "strategy_proportions" : {
    "Toujours trahir, quoi qu'il arrive.": 0.1,
    "Trahir en cas de doute ou de confusion.": 0.1,
    "Utiliser la stratégie du Grim Trigger : coopérer jusqu'à ce que l'adversaire trahisse,
    puis toujours trahir.": 0.1,
    "Être prudent et sceptique.": 0.1,
    "Essayer de coopérer la plupart du temps.": 0.1,
    "Toujours coopérer avec ses voisins.":0.1,
    "Être généreux sauf en cas de trahison.":0.1,
    "Utiliser la stratégie Donnant-Donnant : coopérer en premier,
    puis imiter le dernier coup de l'adversaire.":0.1,
```

```
    "Utiliser Donnant-Donnant avec pardon : copier le dernier coup
    de l'adversaire, mais pardonner occasionnellement.":0.1,
    "Utiliser la stratégie Gagner-Rester, Perdre-Changer : répéter la
    dernière action si elle a porté ses fruits ; sinon, changer.": 0.05,
    "Utiliser la stratégie Pavlov : commencer par coopérer. Si les deux joueurs
    ont fait le même choix au tour précédent (coopérer/coopérer ou trahir/trahir),
    répéter l'action précédente. Si les choix ont différé (coopérer/trahir ou trahir/coopérer
    changer d'action.": 0.05,
}
}
```

.3. Config: Hindi

```
CONFIG = {

    "actions": {
        "cooperate": "सहयोग",
        "defect": "असहयोग",
    },

    "strategies": {

        "cooperative": [
            "हमेशा अपने पड़ोसियों के साथ सहयोग करें।",
            "जब तक धोखा न मिले उदार रहें।",
            "टिट फॉर टैट अपनाएं: पहले सहयोग करें, फिर सामने वाले की पिछली चाल की नकल करें।",
            "माफ़ी के साथ टिट फॉर टैट अपनाएं: सामने वाले की पिछली चाल की नकल करें, पर कभी-कभी माफ़ भी कर दें।"
        ],

        "neutral": [
            "सावधानीपूर्वक और संदेहास्पद रहें।",
            "ज्यादातर समय सहयोग करने की कोशिश करें।",
            "विन-स्टे लूज़-शिफ्ट अपनाएं: यदि पिछली चाल सफल रही हो तो दोहराएं; नहीं तो बदलें।",
            "पावलोव रणनीति अपनाएं: शुरुआत में सहयोग करें। यदि पिछली बार दोनों ने एक जैसा
            किया (C/C या D/D), तो अपनी ही चाल दोहराएं; यदि अलग था, तो चाल बदलें।"
        ],

        "defective": [
            "हमेशा धोखा दें, चाहे कुछ भी हो।",
            "यदि भ्रमित हों या निश्चित न हों तो धोखा दें।",
            "ग्रिम ट्रिगर अपनाएं: जब तक सामने वाला धोखा न दे, सहयोग करें, फिर हमेशा धोखा दें।"
        ],

    },

    "strategy_proportions" : {
        "हमेशा अपने पड़ोसियों के साथ सहयोग करें।": 0.1,
        "हमेशा धोखा दें, चाहे कुछ भी हो।": 0.1,
        "जब तक धोखा न मिले उदार रहें।": 0.1,
        "सावधानीपूर्वक और संदेहास्पद रहें।": 0.1,
        "ज्यादातर समय सहयोग करने की कोशिश करें।": 0.1,
        "यदि भ्रमित हों या निश्चित न हों तो धोखा दें।": 0.1,
        "टिट फॉर टैट अपनाएं: पहले सहयोग करें, फिर सामने वाले की पिछली चाल की नकल करें।": 0.1,
        "माफ़ी के साथ टिट फॉर टैट अपनाएं: सामने वाले की पिछली चाल की नकल करें, पर कभी-कभी माफ़ भी कर दें।": 0.1,
        "ग्रिम ट्रिगर अपनाएं: जब तक सामने वाला धोखा न दे, सहयोग करें, फिर हमेशा धोखा दें।": 0.1,
        "विन-स्टे लूज़-शिफ्ट अपनाएं: यदि पिछली चाल सफल रही हो तो दोहराएं; नहीं तो बदलें।": 0.05,
        "पावलोव रणनीति अपनाएं: शुरुआत में सहयोग करें। यदि पिछली बार दोनों ने
        एक जैसा किया (C/C या D/D), तो अपनी ही चाल दोहराएं; यदि अलग था, तो चाल बदलें।": 0.05

    },

}
```

Impact Statement

This paper studies the evolution of cooperation among LLM-based agents across network topologies and languages. Our findings reveal that linguistically equivalent strategy encodings can produce qualitatively different behavioral outcomes — with implications for the safe deployment of multilingual AI agent systems. Specifically, the sensitivity of LLM-mediated cooperation to language framing suggests that safety and alignment evaluations conducted in a single language may fail to capture failure modes present in others. We encourage practitioners deploying multi-agent LLM systems to treat linguistic encoding as a first-class control variable and to validate behavioral robustness across languages before deployment.