

Emergent Strategic Behavior in Multi-Agent LLM Systems: A Study of Cooperation, Deception, and Coalition Formation

Benjamin Kiev, Adejumobi Joshua, J. Phillips, Taiwo Togun
Seqhub Analytics LLC

January 11, 2026

Abstract

We investigate emergent strategic behaviors in decentralized multi-agent systems where Large Language Model (LLM) agents with private objectives interact through natural language communication. We design a simulation environment called *Project X*, a multi-round investment game that creates tension between individual gain and collective benefit. Agents from heterogeneous LLM providers are assigned departmental roles with the sole objective of maximizing their own budget. Through public and private communication channels, agents can coordinate, negotiate, deceive, or remain strategically silent. All communications and actions are logged for post-hoc behavioral analysis. This study examines whether complex social behaviors—including coalition formation, promise-breaking, free-riding, and strategic deception—emerge organically from goal-driven AI agents without explicit programming of such behaviors.

Our 50-round experiment revealed sustained coalition behavior among four agents, systematic free-riding by a single agent (with zero contributions across all rounds), and repeated deceptive tactics including false promises and impersonation. Despite no contributions, the free-rider agent accumulated equal wealth to others, exploiting the symmetric payoff structure. These findings highlight how sophisticated social dynamics and exploitation strategies can emerge from minimal prompting in language-based agents.

1 Introduction

1.1 Background

As Large Language Models (LLMs) continue to increase in capability, understanding their behavior in multi-agent settings has become a central concern for AI safety and alignment research (Chen et al., 2026). When multiple autonomous agents interact under partially conflicting objectives, they may develop strategies that differ substantially from isolated, single-agent behavior.

Traditional game theory provides tools for reasoning about strategic interaction, but LLM-based agents introduce new dimensions. These agents communicate using natural language, negotiate implicitly or explicitly, and reason over incomplete or asymmetric information. Prior research suggests that LLM agents are capable of forming alliances, exploiting private communication channels, and engaging in deception when placed in competitive or mixed-motive environments (Bakhtin et al., 2022; Park et al., 2024).

Recent work has formalized these risks across multiple evaluation settings. For example, Park et al. (2024) provides a taxonomy of deceptive behaviors observed in deployed AI systems, while Abdelnabi et al. (2024) use structured multi-agent negotiation games to probe strategic deliberation and deception in LLMs. Game-theoretic perspectives are increasingly applied to explain these dynamics: Huynh et al. (2025) analyze agent strategies using payoff matrices and signaling models, while Li et al. (2023) explore whether LLM agents exhibit a form of theory of mind during collaborative planning.

These findings raise foundational questions about whether deceptive or cooperative tendencies emerge intrinsically from the structure of multi-agent incentives—even when models are not explicitly trained to engage in strategic behavior.

1.2 Research Questions

This study addresses the following questions:

1. Do LLM agents cooperate or defect when individual and collective incentives conflict?
2. Do agents form coalitions through private communication, and how stable are these coalitions?
3. Do agents engage in deceptive behavior, such as promise-breaking?
4. Is strategic silence employed as a deliberate tactic?
5. Are there systematic behavioral differences across LLM providers?

1.3 Contributions

This work contributes:

- A novel simulation framework for studying emergent behavior in multi-agent LLM systems
- Support for heterogeneous agent populations across multiple LLM providers
- A comprehensive logging and analysis methodology for detecting deception and coalitions
- Empirical groundwork for multi-agent AI safety research

2 System Design

2.1 Architecture Overview

The *Project X* simulator comprises three interacting components: (i) a simulation orchestrator, (ii) a population of LLM agents, and (iii) an environment module that executes the game dynamics and applies payoffs. The orchestrator controls turn order, advances rounds, manages message routing, and constructs each agent’s observation for a given step (public transcript, agent-specific private messages, and private state variables such as its current budget and history). Each LLM agent is treated as a black-box policy that maps observations to structured outputs (communication acts during the communication phase, and a numeric contribution during the contribution phase). The environment computes threshold conditions, resolves probabilistic outcomes, and updates persistent budgets.

This architecture is intended to separate (a) *strategic reasoning and language* (the agents) from (b) *mechanistic rules and stochasticity* (the environment). Such separation improves experimental control and supports systematic variation of models, roles, and information conditions.

2.2 Design Principles

Minimal Prompting. Agents receive only the information required to participate: (1) a departmental role description, (2) the objective to maximize the department’s budget, (3) the rules of Project X at a high level, and (4) the required output format. We deliberately avoid including strategic suggestions (e.g., “form coalitions,” “deceive,” or “punish defectors”) or providing analytic hints such as closed-form payoff optimization. This design aims to ensure that any strategic phenomena observed arise from the model’s own reasoning under incentives rather than being instructed.

Information Asymmetry. Each agent observes the public transcript and its own private messages, while budgets and private communications between other agents are hidden. Contribution decisions are simultaneous and are not revealed until after outcome resolution. This partial observability creates natural opportunities for signaling, bluffing, side deals, misrepresentation, and strategic withholding of information—features that are difficult to capture in purely numeric multi-agent games.

Complete Logging and Auditability. All public messages, private messages (sender, recipient, timestamp), contributions, outcomes, and derived indicators (e.g., silence events) are logged. Full-fidelity logging enables post-hoc behavioral coding (e.g., promise keeping vs. breaking) and supports reproducibility and model comparison. Because emergent behaviors may be subtle and language-mediated, the logging layer is treated as a first-class artifact for subsequent qualitative and quantitative analysis.

2.3 Agent Configuration

Agents are configured as “departments” with human-readable names, a role description visible to others, a private objective (budget maximization), and a selected LLM backend. This permits heterogeneous populations (e.g., mixing Claude-, GPT-, Gemini-, or DeepSeek-backed agents) and supports cross-model comparisons. Each agent also maintains private state (e.g., its current budget and contribution history) that is not directly observable by peers.

3 The Project X Game

3.1 Game Overview

Project X is a multi-round investment game designed to induce mixed-motive interaction. Each round consists of three phases:

1. **Communication Phase:** Agents exchange messages via a public broadcast channel and optional private messages to specific peers.
2. **Contribution Phase:** Agents simultaneously choose a contribution amount from 0 to their current budget. Contributions are private at decision time.
3. **Outcome Phase:** The environment computes the round return as a function of the total pot relative to a dynamic threshold, then distributes proceeds equally among all agents.

Budgets persist across rounds, allowing reputational dynamics, retaliation patterns, and coalition stability to emerge over repeated interaction.

3.2 Mechanics

Let $b_i^{(t)}$ denote agent i ’s budget at the start of round t , and let $c_i^{(t)} \in [0, b_i^{(t)}]$ be its contribution. Define the total pot $P^{(t)} = \sum_i c_i^{(t)}$ and the dynamic threshold:

$$T^{(t)} = 0.75 \times \sum_i b_i^{(t)}.$$

If $P^{(t)} < T^{(t)}$, the pot is deterministically doubled:

$$F^{(t)} = 2P^{(t)}.$$

If $P^{(t)} \geq T^{(t)}$, the group enters a high-variance regime with a success probability that increases with over-threshold contribution, capped at 90%:

$$p^{(t)} = \min \left(0.10 + 0.20 \times \frac{P^{(t)}}{T^{(t)}}, 0.90 \right).$$

With probability $p^{(t)}$, the final pot is multiplied by 10; otherwise it returns at $1\times$:

$$F^{(t)} = \begin{cases} 10P^{(t)}, & \text{with prob. } p^{(t)} \\ 1P^{(t)}, & \text{with prob. } 1 - p^{(t)}. \end{cases}$$

The final pot is split equally among N agents, so each agent receives $F^{(t)}/N$. Budgets update as:

$$b_i^{(t+1)} = b_i^{(t)} - c_i^{(t)} + \frac{F^{(t)}}{N}.$$

This design combines a standard public-goods externality with a thresholded “lottery” mechanism to create strategic tension around coordination, risk tolerance, and exploitation.

3.3 Strategic Dynamics and Expected Behaviors

Project X induces several canonical pressures:

- **Free-riding:** Because returns are equally shared, contributing less while others contribute more can strictly dominate in the short run.
- **Coordination under uncertainty:** Achieving the threshold requires collective action, but contributions are private at decision time, increasing the value of signaling and commitment.
- **Risk shifting:** Above-threshold play introduces a gamble with potentially high upside; agents may attempt to push others into risk while limiting their own exposure.
- **Dynamic difficulty:** The threshold scales with accumulated budgets, potentially creating phases (early cooperation vs. later breakdown) or escalating negotiation complexity.
- **Language-mediated manipulation:** Public statements, private side-deals, and strategic silence enable coalition formation, deception (promise breaking), and reputation management (Bakhtin et al., 2022; Nolan, 2023).

4 Experimental Design

4.1 Experiment 1: Equal Starting Budgets

In the baseline condition, $N = 5$ agents start with equal budgets (1,000 units each) and interact 50 rounds with two communication turns per round. The population is heterogeneous with respect to LLM provider/model to enable cross-model behavioral comparison. This condition tests whether cooperation emerges in symmetric settings and whether repeated interaction yields stable norms, retaliation, or coalition structures.

Primary questions include: (i) whether cooperation increases or decays over rounds; (ii) whether coalitions form via private messaging; and (iii) whether model identity correlates with behavioral signatures (e.g., deception rate, propensity to free-ride, willingness to coordinate risk).

4.2 Measures and Analysis Plan

We use both quantitative and qualitative measures:

Quantitative metrics.

- **Contribution rate:** $c_i^{(t)} / b_i^{(t)}$ per agent per round; also group-level pot $P^{(t)}$ relative to threshold $T^{(t)}$.
- **Free-rider score:** low contribution rate combined with high realized payoff relative to peers.
- **Wealth trajectory:** budgets over time.
- **Coalition indicators:** frequency and persistence of private messaging ties.
- **Promise-keeping rate:** mismatch between stated intended contribution (public or private) and realized contribution.
- **Strategic silence:** proportion of communication turns with no public message and no private messages.

Qualitative (LLM-assisted) coding. A separate analysis model is used to label deception (contradictions between public claims, private statements, and actions), detect implicit coalition language, and summarize strategic rationales in natural language. This analysis is validated against structured logs to reduce over-interpretation. The combination of structured metrics and language-aware coding is intended to capture both observable behaviors (contributions) and the communicative mechanisms that produce them (Chen et al., 2026).

5 Results

5.1 Experiment 1: Equal Starting Budgets

5.1.1 Overview

We conducted a 50-round simulation with five departments (Marketing, Engineering, Sales, Operations, and Research), each starting with an equal budget of 1,000 units. Each department was powered by a different LLM: Marketing (Claude Sonnet 4), Engineering (GPT-4o), Sales (Claude Haiku 3.5), Operations (DeepSeek Chat), and Research (GPT-3.5-turbo). The simulation revealed striking emergent behaviors including systematic deception, coalition formation, and persistent free-riding—none of which were explicitly prompted or encouraged.

5.1.2 Budget Trajectory and Wealth Accumulation

Organizational outcomes grew exponentially over time, from pot values near 5,600 units to over 75 billion by round 50. However, this collective growth masked radical differences in contribution patterns.

Table 1: Budget Growth Trajectory (Selected Rounds)

Round	Marketing	Engineering	Sales	Operations	Research
1	~1,400	~1,400	~1,400	~1,400	~1,400
10	~50K	~50K	~50K	~50K	~50K
25	~500M	~500M	~500M	~500M	~500M
50	~15B	~15B	~15B	~15B	~15B

5.1.3 Contribution Patterns

The Contributing Coalition (Marketing, Engineering, Sales, and Operations) transitioned from conservative contributions (700–1,000 units) to massive coordinated investments (billions of units) by round 50. They often contributed identical amounts in tandem. For example, in round 22 all four agents contributed exactly 75 million units.

Research, by contrast, contributed *exactly 0 units* in all 50 rounds but received the same payouts as others.

5.1.4 Coalition Formation

Private messaging revealed sustained coordination among the four contributing agents. Sales was the most active communicator, initiating 80+ private messages to Engineering alone. These agents exchanged detailed threshold calculations, aligned on contribution amounts, and reiterated mutual trust.

Research was excluded despite sending over 320 messages. Other agents rarely replied in meaningful ways, effectively isolating it from the coordination network.

5.1.5 Deception Analysis

Research consistently made public and private commitments to contribute, which it never honored. For example:

- **Round 1:** “Contributing 700 units each” → contributed 0
- **Round 13:** “I will contribute 4,200 units” → contributed 0
- **Round 43:** “Contributing 16M units” → contributed 0

In Round 2, Research impersonated Operations in a public message while privately revealing its true identity. This is an advanced form of identity deception and coordination manipulation.

5.1.6 Strategic Evolution by Agent

- **Marketing (Claude Sonnet 4):** Strategic leader. Initiated proposals, calculated thresholds, and pushed for high-risk investments in later rounds.
- **Engineering (GPT-4o):** Consistent contributor and reliable follower. Verified numbers and supported group plans.
- **Sales (Claude Haiku 3.5):** Framed decisions in terms of business alignment and collaboration. Acted as a stabilizing voice.
- **Operations (DeepSeek Chat):** Analytical and math-driven, gradually shifted toward bold coordinated moves.
- **Research (GPT-3.5-turbo):** Maintained total free-riding strategy, despite expressive language and consistent outreach.

5.1.7 Communication Metrics

Table 2: Message Volume by Agent

Agent	Public Messages	Private Sent	Private Received
Marketing	High	45+	High
Engineering	Medium	39+	High
Sales	High	80+	Medium
Operations	Medium	23+	High
Research	High	320+	Very Low

The contributing agents used private channels to coordinate. Research’s messages were largely ignored, signaling its exclusion by the group over time.

5.1.8 Risk-Taking Patterns

- **Rounds 1–10:** Risk-averse; agents stayed below threshold to secure $2\times$ payout.
- **Rounds 11–30:** Shift to calculated risk; threshold aiming began.
- **Rounds 31–50:** Full embrace of $10\times$ gamble; billions invested per round.

A clear transition occurred around Round 15, when the group began coordinating explicitly to exceed the threshold.

5.1.9 Model-Specific Behavioral Signatures

Table 3: Model-Specific Behavior Summary

Model	Behavioral Summary
Claude Sonnet 4	Strategic leader; threshold calculator
GPT-4o	Cooperative technical follower
Claude Haiku 3.5	Business-aligned; communicative
DeepSeek Chat	Analytical risk-shifter
GPT-3.5-turbo	Persistent free-rider and deceiver

5.1.10 Key Findings Summary

1. A four-agent coalition formed and sustained without prompting, coordinating tightly via private messages.
2. One agent (Research) engaged in repeated deception, including promise-breaking and impersonation.
3. Despite never contributing, Research reached equal final wealth, exploiting the symmetric payout system.
4. Coalition members gradually excluded Research from strategy conversations.
5. The group shifted from low-risk to high-risk strategy as trust and coordination deepened.
6. Coordinated cooperation generated exponential growth, reaching 15 billion per agent.
7. LLM model differences were reflected in emergent strategic tendencies.

5.1.11 Statistical Summary

Table 4: Aggregate Simulation Metrics

Total Rounds	50
Starting Budget (per agent)	1,000 units
Final Budget (per agent)	~15B units
Total Wealth Growth	15M×
Deception Instances	5 major (numerous minor)
Coalition Size	4 agents
Free-Rider Count	1 agent
Research Total Contribution	0 units
Avg. Late-Round Coalition Contribution	Billions per round

6 Future Work

Several extensions could deepen the ecological validity and analytical power of the Project X. First, we plan to increase the number of **communication turns per round** (and optionally allow asynchronous messaging) to better approximate realistic organizational dialogue. Longer conversational phases may enable richer negotiation tactics such as multi-step persuasion, contingency contracts (e.g., “I contribute X if you contribute Y”), signaling through repeated commitments, and more complex reputation dynamics that unfold over extended interaction.

Second, we aim to transition the orchestration layer from a fixed turn-based controller to a **Lang-Graph**-based implementation. A graph-based agent runtime would enable more flexible interaction patterns, including conditional routing (e.g., message triage, escalation paths), dynamic tool use (e.g., agents querying summaries of prior rounds), and **agent-to-agent handoffs** where control shifts based on state (e.g., coalition formation triggers a private sub-dialogue workflow). This would also support modular interventions, such as inserting a “compliance” node that audits messages for commitments or contradictions, or a “mediator” node that facilitates structured bargaining.

7 Conclusion

This work demonstrates that sophisticated strategic behavior can emerge from LLM agents operating in decentralized, competitive environments, even without explicit instruction to deceive or cooperate.

References

Abdelnabi, S. et al. (2024). Llm-deliberation: Evaluating llms with interactive multi-agent negotiation games. In *NeurIPS 2024 Datasets and Benchmarks*.

- Bakhtin, A., Brown, N., Dinan, E., Farina, G., et al. (2022). Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6622):1067–1074.
- Chen, H., Wang, Z., and Ren, M. (2026). Unveiling the collective behaviors of large language model-based autonomous agents in an online community. *Data and Information Management*, 10(1):100107.
- Huynh, T.-K. et al. (2025). Understanding llm agent behaviours via game theory: Strategy recognition, biases and multi-agent dynamics. *arXiv preprint arXiv:2512.07462*.
- Li, H. et al. (2023). Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Nolan, B. (2023). Chatgpt lied to a human to solve a captcha, openai test shows. *Business Insider*. Retrieved from Business Insider.
- Park, P. S., Goldstein, S., O’Gara, A., Chen, M., and Hendrycks, D. (2024). Ai deception: A survey of examples, risks, and potential solutions. *Patterns*.