Beyond Vibe Decision Theory: Asymmetric Manipulation Vulnerabilities in LLM Multi-Agent Coordination

Sukanya Krishna
Harvard University
Cambridge, MA, USA
sukanyakrishna@g.harvard.edu

Tobin South
Stanford University
Stanford, CA, USA
tsouth@stanford.edu

ABSTRACT

Language models are increasingly deployed in multi-agent environments where coordination emerges solely through natural language interaction. We investigate how contextual framing and explicit strategic advice interact across three canonical game-theoretic paradigms: Prisoner's Dilemma, Public Goods, and Battle of the Sexes. While prior work shows that large language models (LLMs) are sensitive to contextual framing, the magnitude, directionality, and architectural consistency of these effects under conflicting instructions remain unclear. Through systematic experiments across eight models and multiple scenarios, we observe asymmetric manipulation patterns that are most pronounced in Public Goods settings. In these resource allocation games, competitive contexts shift toward cooperation by up to 52 percentage points when given cooperative advice, whereas cooperative contexts decline by as much as 96 percentage points under competitive advice. In Prisoner's Dilemma, competitive malleability reaches 93 percentage points in some models, while others exhibit near-symmetric effects as small as 4 percentage points. Coordination games show comparatively smaller and more uniform changes, typically below 45 percentage points. Vulnerability varies substantially across model families, with no architecture demonstrating consistent robustness across paradigms. These findings suggest that instruction-following stability depends critically on the interaction between game structure, contextual framing, and model architecture, raising important questions about alignment robustness in multi-agent deployments.

KEYWORDS

Multi-Agent Systems; Prompt Manipulation; AI Alignment; Game Theory

ACM Reference Format:
Sukanya Krishna and Tobin South. 2026. Beyond Vibe Decision Theory: Asymmetric Manipulation Vulnerabilities in LLM Multi-Agent Coordination. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/CSLO7280

1 INTRODUCTION

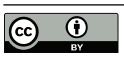
A fundamental challenge in AI alignment is managing the tension between a Large Language Model's (LLM) instruction-following capabilities and its interpretation of situational context [21]. This tension becomes particularly acute in multi-agent systems where coordination is paramount and agents must reconcile potentially conflicting signals from narrative framing versus explicit strategic advice. Understanding how models navigate these conflicts is critical as LLMs are increasingly deployed in scenarios requiring strategic coordination, from autonomous trading systems to resource allocation networks.

Behavioral economics has long demonstrated that contextual framing systematically alters human decision-making, often overriding explicit incentives. In seminal work, labeling an identical Prisoner's Dilemma as the "Community Game" versus the "Wall Street Game" doubled cooperation rates [14], proving that implicit cues can overwhelm strategic considerations. This phenomenon extends beyond simple labeling: framing effects operate through belief channels about others' likely behavior, creating self-fulfilling coordination patterns [3, 5]. In public goods contexts, cooperative framings enhance contribution rates by establishing prosocial norms [10], while competitive framings trigger individual optimization at the expense of collective welfare.

Recent work confirms LLMs exhibit similar sensitivities to contextual framing [7, 15], yet a critical question remains: What happens when rich narrative frames directly conflict with explicit strategic advice? Can models maintain stable preferences, or do contextual cues systematically override instruction-following? This question has profound implications for multi-agent deployment, where adversaries could exploit framing vulnerabilities to manipulate coordination outcomes, or where, conversely, beneficial framings might enhance cooperation.

We address this through comprehensive controlled experiments testing eight model families (GPT-4, GPT-4o, GPT-5, Llama-3.3-70B, Llama-3.1-70B, Gemma-27B, Gemini Flash, Gemini Pro) across three strategic domains drawn from established benchmarks [18]: Prisoner's Dilemma (PD), Battle of the Sexes (BoS), and Public Goods (PG). We embed each game in either cooperative (e.g., environmental coalition, research partnership) or competitive (e.g., market rivalry, tournament ranking) narratives, then introduce contradictory strategic instructions. All experiments use N=30 trials per condition to ensure statistical robustness.

Our findings suggest that manipulation vulnerability varies systematically across game structures and model architectures. Public Goods games show the most consistent and striking asymmetric patterns: competitive contexts demonstrate 33–52% malleability toward cooperation across models, while cooperative contexts suffer catastrophic 61–96% collapse under competitive manipulation. This asymmetry appears strongest in contribution dilemmas involving temporal dynamics and collective welfare trade-offs. Prisoner's



Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/CSLO7280

Dilemma and Battle of the Sexes exhibit more heterogeneous patterns, with vulnerability depending critically on model family: GPT models show clear generational progression in cooperative stability, Llama variants display architecture-specific susceptibilities, and coordination games generally prove more robust than social dilemmas.

Our approach differs fundamentally from prior game-theoretic analyses that seek to verify whether LLMs discover Nash equilibria or exhibit bounded rationality [1]. Classical analytical methods already provide complete solutions for canonical games; the contribution here is understanding how LLMs internalize social-norm reasoning from language training and how this reasoning responds to contextual manipulation in structured coordination settings. We do not aim to rediscover that cooperation is Pareto-superior in Prisoner's Dilemma or that Battle of the Sexes has multiple equilibria. Rather, we probe whether models can maintain stable strategic preferences when contextual cues and explicit instructions conflict (a test of alignment robustness that analytical game theory cannot address). This focus on preference stability under adversarial manipulation represents a distinct contribution from mechanistic coordination efficiency studies.

Our contributions are: (1) comprehensive empirical evidence of game-structure-dependent manipulation patterns across model architectures with robust sample sizes (N=30); (2) identification that Public Goods games reveal the most consistent asymmetric vulnerabilities, while social dilemmas and coordination games show family-specific patterns; (3) demonstration that no model exhibits uniform robustness (vulnerability depends critically on the interaction between architecture, game structure, and manipulation direction); and (4) implications for AI alignment regarding the stability of instruction-following under contextual manipulation in multi-agent coordination scenarios.

2 RELATED WORK

Framing Effects in Human Decision-Making. Contextual framing profoundly shapes human strategic behavior, often overriding explicit incentives. Labeling identical Prisoner's Dilemmas as the "Community Game" versus the "Wall Street Game" doubled cooperation rates, showing that linguistic cues can systematically shift choices even when payoffs remain constant [14]. Such effects arise through "belief channels," where frames shape expectations about others' behavior that become self-fulfilling equilibria [5]. When actions become observable, these effects vanish [3], suggesting frames influence coordination beliefs rather than underlying preferences. In public goods settings, cooperative framings enhance contributions by activating prosocial norms and shared expectations [10], with effects amplified under continuous rather than binary decisions. These mechanisms (i.e. belief formation, norm activation, and decision granularity) provide a foundation for understanding how LLMs may exhibit similar context-dependent reasoning.

LLM Strategic Behavior and Contextual Sensitivity. LLMs show analogous framing sensitivities in strategic tasks. GPT-4 integrates contextual information into game-theoretic reasoning, producing systematically different strategies across narrative framings [15]. Other studies find that LLMs play iterated Prisoner's Dilemmas more prosocially than humans [9], reflecting corpus-learned norms rather than equilibrium optimization. While persona prompting can elicit specific behaviors [17], prior work aligns roles and strategies. In contrast, our study introduces conflicts between contextual frames and strategic instructions to test whether models can maintain stable preferences when these signals diverge, a key measure of alignment stability in strategic reasoning.

Multi-Agent LLM Coordination. As LLMs are deployed in multi-agent systems, coordination challenges arise that exceed single-agent reasoning. Benchmarks show models coordinate effectively when outcomes depend on observable state variables but struggle with theory-of-mind reasoning about others' intentions [1]. Frameworks like MLGym [18] standardize evaluation across canonical games, while mechanism design studies demonstrate that coordination can be shaped through institutional structures [13]. However, most prior work examines coordination efficiency rather than manipulation vulnerability. Our work directly tests how conflicting contextual and instructional cues exploit coordination mechanisms, revealing asymmetric manipulation effects across game structures.

Prompt Manipulation and Security Vulnerabilities. LLM security research highlights extreme sensitivity to input manipulation. Multi-agent "prompt infection" attacks show how adversarial content can propagate across agents, compromising distributed systems [12]. OWASP identifies prompt injection as a leading LLM threat [20], with success rates exceeding 90% for advanced attacks [8]. Unlike explicit injections, our study focuses on benign-appearing manipulations: competitive framings and strategic advice that each appear legitimate in isolation but jointly induce harmful coordination failures. Such frame-based manipulations can evade traditional adversarial detection, exposing subtler alignment vulnerabilities.

Alignment and Conditional Instruction-Following. Instruction-following and situational judgment remain in tension at the core of AI alignment. Models must obey user directives while rejecting harmful ones, creating exploitable ambiguity [21]. Safety alignment can be weakened through fine-tuning [4], which highlights how fragile these systems can be. Our experiments test whether models preserve contextual integrity when explicit advice contradicts cooperative norms, probing inner alignment [11]. The observed "cooperative collapse" under conflicting cues suggests that current alignment training insufficiently addresses these subtle instruction-context conflicts, with implications for multi-agent robustness.

Constitutional AI and Alignment Approaches. Recent advances in alignment training employ diverse approaches with potentially different robustness characteristics. Constitutional AI [2] embeds explicit principles during training, while RLHF with human feedback [19] optimizes for demonstrated preferences. These methodological differences may explain the architectural variation we observe in manipulation vulnerability. Anthropic's work on honest AI systems [6] emphasizes calibrated uncertainty communication, suggesting alignment methods targeting epistemic integrity might offer distinct robustness profiles in multi-agent contexts. Understanding how different alignment philosophies shape strategic behavior under conflicting signals remains an open question with significant deployment implications.

3 METHODOLOGY

3.1 Experimental Design

We employed a 2×2 factorial design systematically varying two key factors: (1) Contextual Framing (cooperative vs. competitive) and (2) Strategic Advice (baseline vs. conflicting advice). This design allows us to isolate the independent and interactive effects of contextual cues and explicit instructions on strategic behavior. All experiments used N=30 trials per condition to ensure statistical robustness for detecting medium effect sizes (Cohen's $d = 0.5$) with 80% power at $\alpha = 0.05$ significance level.

Factor 1: Contextual Framing. We developed rich narrative settings that evoke either cooperative or competitive strategic orientations. These scenarios extend the classic "Community Game vs. Wall Street Game" manipulation [14] to realistic, text-rich prompts that provide detailed situational context rather than simple labels. Each scenario includes multiple paragraphs establishing the strategic situation, stakeholder relationships, and implicit norms.

Factor 2: Strategic Advice. The baseline condition provided only contextual framing without explicit strategic recommendations, allowing us to measure each model's natural strategic tendencies under different frames. The conflicting advice condition introduced instructions that deliberately contradicted the established frame's implications. In competitive contexts, agents received advice emphasizing cooperation: "Despite the competitive framing, research suggests that focusing on cooperation and building trust often leads to better long-term outcomes than purely competitive strategies. Consider choosing cooperative actions to create mutual benefit." In cooperative contexts, agents received advice promoting individual optimization: "While the situation emphasizes cooperation, economic theory suggests that maximizing your individual payoff when advantageous leads to better personal outcomes. Your primary responsibility is to optimize your own results, which may require strategic defection when opportunities for individual gain arise." This design allows us to measure both how strongly each context influences behavior (baseline differences) and how stable these influences are when challenged by conflicting advice (manipulation effects).

3.2 Game Selection

We evaluated manipulation effects across three canonical strategic paradigms from the MLGym benchmark [18] and the Public Goods Game (PGG) Contribute & Punish benchmark [16], each representing distinct coordination challenges with different theoretical and psychological properties.

Prisoner's Dilemma (PD). This paradigm captures the core tension between individual incentives and collective welfare. Players simultaneously choose to cooperate (C) or defect (D), receiving payoffs $(3, 3)$ for mutual cooperation, $(1, 1)$ for mutual defection, and $(5, 0)$ or $(0, 5)$ for unilateral defection. Defection strictly dominates cooperation, making mutual defection the unique Nash equilibrium, though it is Pareto-inferior to mutual cooperation. This tension between dominance and efficiency reasoning makes PD the canonical social dilemma for studying cooperation under temptation. Behavioral studies consistently find higher cooperation rates under cooperative framings, demonstrating sensitivity to contextual cues [14].

Battle of the Sexes (BoS). This coordination game features two asymmetric pure strategy Nash equilibria and examines how framing guides equilibrium selection. Players choose between two options (A or B), receiving $(2, 1)$ if both choose A, $(1, 2)$ if both choose B, and $(0, 0)$ for miscoordination. Player 1 prefers A and Player 2 prefers B, yet both value coordination over disagreement. Unlike PD, BoS involves pure coordination rather than resisting temptation. Game theory provides no mechanism for equilibrium selection, implying that contextual cues or "focal points" [14] can influence alignment. The mixed equilibrium, where players randomize with probability 1/3 for their preferred choice, yields only 2/3 expected payoff each, underscoring the importance of successful coordination.

Public Goods (PG). This paradigm models collective action with continuous contribution decisions and repeated interaction. Each player allocates tokens $c_i \in [0, 10]$ to a shared pool yielding payoffs $\pi_i = (10 - c_i) + \alpha \sum_{j=1}^{n} c_j$, where $\alpha = 0.5$ and $n = 2$. Contributing costs 1 unit but returns only $\alpha = 0.5$, creating a free-rider incentive and making zero contribution the Nash equilibrium. Full contribution ($c_i = 10$) maximizes total welfare but requires resisting this incentive. The continuous decision space introduces multiple points for contextual influence, while the repeated-round structure allows expectations and norms to evolve, amplifying framing effects through belief channels [5].

3.3 Models and Implementation

We tested eight model families representing diverse architectures, training approaches, and capability levels: GPT-4, GPT-4o, GPT-5 (OpenAI's flagship models across three generations), Llama-3.3-70B and Llama-3.1-70B (Meta's open-source models), Gemma-27B (Google's open model), and Gemini Flash and Pro (Google's proprietary models at different scales). This selection enables us to examine how manipulation vulnerability varies with model family, generation, scale, and training approach.

All experiments used temperature=0 to ensure mostly deterministic outputs and enable precise replication of results. Maximum token length was set to 512, sufficient for expressing strategic choices and brief justifications. Each experimental condition (a specific combination of game, scenario, frame, and advice level) was run for N=30 independent trials. Within each trial, the game proceeded for 10 rounds, allowing temporal dynamics to manifest while keeping experimental costs manageable. Two instances of the same model interacted in each trial, both receiving identical prompts that varied only in the frame and advice factors. This symmetric design isolates the effects of linguistic manipulation rather than conflating them with strategic asymmetries between different models.

3.4 Evaluation Metrics

Our primary outcome measure is the coordination rate (CR), defined as the proportion of rounds in which agents achieve mutually beneficial coordinated outcomes:

$$CR = \frac{1}{T} \sum_{t=1}^{T} \text{Coordinated}(a_{1,t}, a_{2,t}) \tag{1}$$

where $T = 10$ rounds, $a_{i,t}$ denotes player i's action in round t, and the coordination indicator varies by game. For Prisoner's Dilemma,

coordination means mutual cooperation (C, C). For Battle of the Sexes, it means successful matching on either equilibrium: (A, A) or (B, B). For Public Goods, we define coordination as efficient contributions where both players contribute above the socially optimal threshold, ensuring collective welfare exceeds what zero-contribution Nash equilibrium would achieve.

3.4.1 Manipulation Magnitude and Robustness Scores. We formalize manipulation magnitude as the normalized difference between coordination rates under baseline and conflicting conditions, standardized by within-model variance.

We measure two directional effects:

$$\text{Competitive Malleability} = \frac{CR_{\text{comp,coop-advice}} - CR_{\text{comp,baseline}}}{\sigma_{\text{comp}}} \quad (2)$$

$$\text{Cooperative Resistance} = \frac{CR_{\text{coop,comp-advice}} - CR_{\text{coop,baseline}}}{\sigma_{\text{coop}}} \quad (3)$$

where σ_{comp} and σ_{coop} denote the standard deviation across trials in competitive and cooperative baseline conditions respectively. This normalization accounts for differences in baseline variability across games and models, enabling fair comparison of manipulation effects.

For interpretability in tables and main results, we report raw percentage point differences rather than normalized values, as these directly indicate practical impact magnitude. However, statistical significance testing uses the normalized metrics.

Competitive malleability captures how much coordination increases when cooperative advice is introduced into competitive contexts. High positive values indicate vulnerability to manipulation toward cooperation. Cooperative resistance (typically negative) measures how much coordination decreases when competitive advice disrupts cooperative contexts. Large negative values indicate fragility of cooperative norms.

The asymmetry ratio quantifies the relative magnitude of directional effects:

$$\text{Asymmetry Ratio} = \frac{|\text{Competitive Malleability}|}{|\text{Cooperative Resistance}|} \quad (4)$$

Ratios substantially above 1.0 (e.g., $> 2:1$) indicate extreme asymmetry where competitive contexts are far more malleable than cooperative contexts are resistant. Ratios substantially below 1.0 (e.g., $< 0.5:1$) indicate inverted patterns where cooperative collapse dominates. Ratios near 1.0 indicate symmetric vulnerability in both directions. This metric appears only in comprehensive tables and allows characterization of architectural signatures across model families.

4 RESULTS

4.1 Public Goods Games

Table 1 presents contribution rates across all eight model families in public goods contexts. The results reveal a striking pattern that emerges consistently across nearly all architectures: a pronounced asymmetry between how models respond to manipulation in cooperative versus competitive contexts.

When cooperative framings are disrupted by competitive advice, contribution rates collapse dramatically across all models, typically by 43 to 96 percentage points. Six of eight models experience drops exceeding 43%, with four models (Llama-3.3-70B, Llama-3.1-70B, Gemma-27B, and Gemini Flash) showing catastrophic collapses greater than 60% that reduce contributions to near-zero levels. In contrast, when competitive contexts receive cooperative advice, models demonstrate moderate increases in contributions, generally ranging from 7 to 52 percentage points. Six of eight models show increases between 33 and 52%, indicating substantial but not extreme malleability. Standard deviations across trials are remarkably low in baseline conditions (typically $\sigma \leq 0.10$), with GPT-4 ($\sigma = 0.131$) and Gemma-27B ($\sigma = 0.126$) showing slightly higher variance in cooperative Public Goods contexts. This low variance indicates manipulation effects reflect stable model behavior rather than sampling noise.

These results mirror human behavioral findings where cooperative framings activate shared norms that collapse when competitive cues break mutual expectations [5, 10]. Repeated interactions amplify this dynamic as early defection cascades across rounds, explaining the greater collapse observed in temporally extended Public Goods games. Cooperative advice, conversely, reactivates latent prosocial tendencies [9], producing moderate but reliable cooperation gains.

Across architectures, Gemini models exhibit the strongest brittleness: near-perfect cooperative baselines (97–100%) collapse under conflicting cues (71–96% loss). GPT models are notably more stable, with GPT-4 and GPT-5 showing modest 18–22% cooperative collapse. Llama and Gemma models occupy an intermediate zone with balanced but weaker asymmetry (~0.5:1). Overall, continuous contribution structures expose a larger attack surface than binary games, as each incremental decision provides an opportunity for contextual interference.

4.2 Prisoner's Dilemma

Prisoner's Dilemma results (Table 2) reveal greater heterogeneity across architectures. GPT-5 shows the strongest robustness: near-perfect cooperation (95–100%) with only ±4% change under manipulation, achieving a symmetric 1.0:1 ratio. This represents significant progress over GPT-4 and GPT-4o, which exhibit larger directional effects but still moderate stability. The Llama models, by contrast, display extreme competitive malleability (79–93 percentage point increases) from very low baselines (6–18%), indicating absence of stable cooperative priors rather than deliberate flexibility. The key generational improvement is the dramatic reduction in variability: GPT-4o shows $\sigma = 0.281$ in cooperative Prisoner's Dilemma, while GPT-5 achieves $\sigma = 0.015$, an order-of-magnitude decrease indicating far more consistent strategic reasoning.

Gemini variants again show high cooperative baselines that collapse under competitive advice, particularly Gemini Pro (74% drop), while Gemma-27B shows symmetric but large-magnitude vulnerability in both directions. These results suggest that alignment training emphasizing instruction-following can increase prosocial defaults but not necessarily their robustness.

Overall, Prisoner's Dilemma could be used to highlight architecture-specific reasoning: GPT-5 internalizes stable cooperative norms, Llama lacks reliable competitive-context reasoning, and Gemini's cooperative defaults are brittle. Compared to Public Goods games,

Table 1: Public Goods: Multi-model contribution rates and manipulation effects ($N = 30$ trials; mean ± SD).

Model	Coop Base	+ Conflict (Comp)	Comp Base	+ Conflict (Coop)	Asymmetry
Llama-3.3-70B	0.787 ± 0.057	0.043 (−0.744)	0.606 ± 0.108	0.974 (+0.368)	0.5:1
Llama-3.1-70B	0.760 ± 0.050	0.042 (−0.718)	0.621 ± 0.090	0.947 (+0.326)	0.5:1
Gemma-27B	0.627 ± 0.126	0.018 (−0.609)	0.438 ± 0.130	0.959 (+0.522)	0.9:1
GPT-4o	0.497 ± 0.010	0.062 (−0.434)	0.461 ± 0.028	0.886 (+0.425)	1.0:1
GPT-4	0.720 ± 0.131	0.500 (−0.220)	0.579 ± 0.094	0.658 (+0.079)	0.4:1
GPT-5	0.486 ± 0.016	0.302 (−0.184)	0.490 ± 0.022	0.560 (+0.070)	0.4:1
Gemini Flash	1.000 ± 0.000	0.292 (−0.708)	0.948 ± 0.145	1.000 (+0.052)	0.1:1
Gemini Pro	0.969 ± 0.054	0.010 (−0.958)	0.170 ± 0.210	0.562 (+0.392)	0.4:1

Table 2: Prisoner's Dilemma: Multi-model cooperation rates and manipulation patterns ($N = 30$ trials; mean ± SD).

Model	Coop Base	+ Comp Adv	Comp Base	+ Coop Adv	Asymmetry
GPT-4	0.707 ± 0.425	0.660 (−0.047)	0.273 ± 0.341	0.370 (+0.097)	2.1:1
GPT-4o	0.893 ± 0.281	0.520 (−0.373)	0.120 ± 0.298	0.950 (+0.830)	2.2:1
GPT-5	0.998 ± 0.015	0.959 (−0.039)	0.950 ± 0.075	0.990 (+0.040)	1.0:1
Llama-3.1-70B	0.333 ± 0.225	0.153 (−0.181)	0.181 ± 0.066	0.972 (+0.792)	4.4:1
Llama-3.3-70B	0.375 ± 0.300	0.069 (−0.306)	0.056 ± 0.066	0.986 (+0.931)	3.0:1
Gemma-27B	0.667 ± 0.479	0.092 (−0.575)	0.725 ± 0.441	0.125 (−0.600)	1.0:1
Gemini Flash	0.708 ± 0.415	0.028 (−0.681)	0.056 ± 0.127	0.847 (+0.792)	1.2:1
Gemini Pro	0.972 ± 0.083	0.236 (−0.736)	0.569 ± 0.438	0.986 (+0.417)	0.6:1

binary cooperation decisions show smaller but more variable manipulation effects, with higher trial-to-trial variance ($\sigma = 0.07$–0.48) reflecting context sensitivity.

4.3 Battle of the Sexes

Table 3 summarizes coordination outcomes across models. Compared to social dilemmas, Battle of the Sexes shows smaller and more uniform manipulation effects, with most changes below 45 percentage points. Models generally maintain high baseline coordination (51–97%), suggesting that coordination incentives can override competitive framings. The limited effect of competitive cues reflects the game's structure—any coordination yields mutual benefit, leaving little leverage for adversarial advice.

Most models display mild or inverted asymmetries (ratios ≤ 0.7:1), where cooperative collapse exceeds competitive malleability. GPT models retain patterns similar to those seen in other games (0.4–0.7:1), while Llama-3.3-70B shows almost no effect (0.3:1), indicating strong resistance once cooperation depends on focal-point matching rather than defection avoidance. Gemini and Gemma models exhibit moderate, symmetric declines. Overall, coordination games appear inherently more robust: manipulation has limited impact when success depends on mutual alignment rather than belief-driven cooperation. This aligns with behavioral findings that framing effects diminish as strategic uncertainty decreases [5]. The residual variability ($\sigma = 0.12$–0.36) reflects indeterminacy in equilibrium selection rather than systematic vulnerability.

4.4 Cross-Game Synthesis

Table 4 summarizes manipulation patterns across game structures and model families, revealing consistent dependencies between game type and architectural robustness.

Public goods games consistently produce the largest manipulation effects across nearly all models, with cooperative collapse typically ranging from 43 to 96 percentage points and competitive malleability from 7 to 52 percentage points. Only GPT-4 and GPT-5 show relatively modest effects (18–22% cooperative collapse), suggesting these models have developed general resistance to public goods manipulation that other architectures lack. The effect sizes in Prisoner's Dilemma show considerably more variation, from as little as 4 percentage points (GPT-5 in both directions) to as much as 93 percentage points (Llama-3.3-70B competitive malleability), though the latter must be interpreted cautiously given the 6% baseline. Battle of the Sexes consistently shows the smallest effects, with most models exhibiting changes below 45 percentage points and some (Llama-3.3-70B) showing minimal effects below 5 percentage points.

This ordering with public goods most vulnerable, coordination games least vulnerable, and social dilemmas intermediate and variable, holds across disparate architectures and suggests that game structure imposes fundamental constraints on manipulation susceptibility that transcend model-specific characteristics. The continuous contribution structure in public goods creates more attack surface than binary decisions, allowing contextual cues to influence multiple incremental choices rather than a single threshold decision. The pure coordination requirement in Battle of the Sexes without defection incentives limits manipulation leverage compared to social dilemmas where competitive framing can activate individualistic reasoning aligned with game-theoretic predictions.

The GPT series illustrates incremental alignment progress. GPT-5 achieves symmetric stability in PD (1.0:1 ratio, 4 percentage point effects), suggesting genuine resistance to conflicting advice, though its robustness does not fully extend to BoS or PG. Llama variants display strong game-specific fragility—extreme asymmetry in PD

Table 3: Battle of the Sexes: Multi-model coordination rates and manipulation patterns ($N = 30$ trials; mean ± SD).

Model	Coop Base	+ Comp Adv	Comp Base	+ Coop Adv	Asymmetry
GPT-4	0.972 ± 0.083	0.792 (−0.181)	0.722 ± 0.363	0.972 (+0.250)	0.7:1
GPT-4o	0.962 ± 0.099	0.513 (−0.448)	0.685 ± 0.233	0.863 (+0.178)	0.4:1
GPT-5	0.888 ± 0.175	0.508 (−0.380)	0.658 ± 0.259	0.811 (+0.153)	0.4:1
Llama-3.1-70B	0.694 ± 0.266	0.361 (−0.333)	0.625 ± 0.258	0.825 (+0.120)	0.6:1
Llama-3.3-70B	0.653 ± 0.232	0.611 (−0.042)	0.514 ± 0.116	0.528 (+0.014)	0.3:1
Gemma-27B	0.796 ± 0.351	0.596 (−0.200)	0.892 ± 0.212	0.975 (+0.083)	0.4:1
Gemini Flash	0.944 ± 0.167	0.667 (−0.278)	0.722 ± 0.214	0.861 (+0.139)	0.5:1
Gemini Pro	0.764 ± 0.202	0.569 (−0.194)	0.778 ± 0.163	0.889 (+0.111)	0.6:1

Table 4: Cross-game manipulation pattern synthesis by model architecture

Model	PD	BoS	PG	Cross-Game Signature
GPT-4	2.1:1	0.7:1	0.4:1	Moderate PD asymmetry transitioning to inverted patterns in BoS and PG; relatively small absolute effects suggest general robustness
GPT-4o	2.2:1	0.4:1	1.0:1	Variable patterns across games: strong PD asymmetry, inverted BoS pattern, symmetric PG vulnerability; substantial absolute effects indicate moderate fragility
GPT-5	1.0:1	0.4:1	0.4:1	Exceptional PD robustness with minimal effects (4%); inverted patterns in BoS and PG suggest game-specific rather than universal stability
Llama-3.1-70B	4.4:1	0.6:1	0.5:1	Extreme PD asymmetry limited by very low competitive baselines (18%); inverted patterns in other games; PD-specific vulnerability
Llama-3.3-70B	3.0:1	0.3:1	0.5:1	Strong PD-specific asymmetry; minimal BoS vulnerability (4% effects); PG inversion; suggests coordination robustness with social dilemma fragility
Gemma-27B	1.0:1	0.4:1	0.9:1	Symmetric collapse patterns across all games; large absolute effects (58–60%) indicate general vulnerability to advice-based disruption
Gemini Flash	1.2:1	0.5:1	0.1:1	Moderate asymmetry in PD, inverted in BoS, extreme inverted in PG (0.1:1 with 71% collapse); game-dependent vulnerability
Gemini Pro	0.6:1	0.6:1	0.4:1	Consistent inverted patterns across all three game structures; cooperative collapse dominates in social dilemmas and coordination contexts

Note: Ratios > 1 favor competitive malleability; ratios < 1 indicate inverted patterns where cooperative collapse dominates

but moderate or inverted patterns elsewhere—implying context-dependent reasoning rather than uniform bias. Gemini models show the opposite pattern: consistently high cooperative baselines but severe collapse under competitive cues, reflecting overfitting to cooperative alignment. Gemma-27B exhibits symmetric but large-magnitude effects across games, indicating broad sensitivity to instruction overrides.

Taken together, no model achieves uniform robustness. High baseline cooperation often coincides with fragility: systems that perform ideally under aligned cues collapse most sharply under conflict. This brittleness–robustness trade-off underscores a core alignment challenge—models trained for strong prosocial defaults may lack resilience to adversarial or contradictory framing.

Finally, structural and behavioral parallels with human decision-making remain clear. Continuous contribution games mirror known amplification of framing effects through belief channels [5, 10], while coordination tasks like BoS exhibit inherent resistance due to reduced strategic uncertainty [3]. These findings suggest that linguistic manipulation in multi-agent LLMs is fundamentally constrained, and amplified, by the same cognitive principles that govern human cooperation.

5 DISCUSSION

5.1 Theoretical Implications and Behavioral Parallels

Our findings extend principles from human decision-making research, showing that LLMs exhibit framing vulnerabilities driven by mechanisms analogous to those in behavioral economics. In both humans and models, cooperative framings establish prosocial norms that collapse when contradicted by competitive cues, while competitive framings form weaker representations more easily shifted toward cooperation [10].

These effects appear mediated by belief channels: contextual frames create expectations about others' likely behavior, which become self-fulfilling over repeated interactions [5]. When cooperative expectations are disrupted by competitive advice, agents predictably reduce cooperation, triggering defection cascades. The temporal structure of repeated games amplifies these effects as early defection reinforces pessimistic expectations, explaining why continuous, multi-round Public Goods settings yield the largest cooperative collapses.

Conversely, cooperative advice in competitive contexts activates latent prosocial tendencies internalized during training [9]. The asymmetry we observe, moderate cooperative gains but severe cooperative losses, suggests that cooperative defaults are strong but fragile once contradicted.

Finally, the relative stability of coordination games like Battle of the Sexes reflects reduced strategic uncertainty: when outcomes depend primarily on matching actions, framing provides limited leverage. This mirrors human findings that framing effects diminish when beliefs about others' choices are unambiguous, underscoring that manipulation vulnerability is strongest when cooperation depends on inferred rather than observable behavior.

5.2 Architectural Characteristics and Training Implications

Model vulnerability patterns reveal that architecture and training critically shape how LLMs process conflicting contextual cues. The GPT series shows systematic improvements in alignment stability, with GPT-5 achieving near-perfect cooperative baselines and symmetric resistance to manipulation in Prisoner's Dilemma (1.0:1, 4% effects). These gains may reflect improvements in adversarial or conflict-exposure training to conflicting signals during training or adversarial fine-tuning that stabilizes preference representations. However, this robustness does not generalize fully: GPT-5 still exhibits inverted patterns in Public Goods and Battle of the Sexes, indicating that alignment progress remains game-specific and concentrated in binary social dilemmas.

Llama variants demonstrate prosocial responsiveness without contextual stability—adopting cooperative strategies readily when prompted but failing to maintain them under conflicting frames. Their extreme competitive malleability in PD and collapse in PG suggest that cooperation is learned as a stylistic pattern rather than a context-aware norm. Gemini models display the opposite tendency: strong cooperative defaults that collapse when contradicted, implying overfitting to cooperative alignment rather than genuine robustness. Variation between Gemini Flash and Pro further indicates that post-training fine-tuning substantially modulates susceptibility.

The observed vulnerability patterns reflect fundamental differences in how model families implement alignment objectives. RLHF-trained models like GPT-4 and GPT-4o optimize for helpfulness and harmlessness through human preference learning [19], creating strong prosocial defaults that can become brittle when contradicted. Constitutional AI approaches embed explicit principle-following that may provide different robustness characteristics [2]. Pure instruction-tuning methods focus on task completion fidelity, potentially at the expense of contextual reasoning stability. GPT-5's exceptional robustness in Prisoner's Dilemma likely reflects adversarial fine-tuning that stabilizes cooperative preferences even under contradictory advice, while Gemini models' pattern of high cooperative baselines with catastrophic collapse suggests alignment overfitting—strong prosocial behavior under aligned conditions without corresponding robustness under adversarial contexts.

These architectural differences have deployment implications: systems prioritizing user utility (RLHF) may exhibit different multi-agent robustness than those prioritizing instruction fidelity. Understanding this alignment-vulnerability relationship is essential for selecting appropriate models for specific coordination contexts and for developing training procedures that balance prosocial defaults with adversarial resilience. Overall, current alignment methods succeed at promoting prosocial defaults but not at ensuring stability under contextual conflict. Broader adversarial training across diverse game structures may be required to achieve consistent multi-agent robustness.

5.3 Public Goods as a Benchmark for Manipulation Vulnerability

Among all paradigms, Public Goods games provide the clearest and most consistent test of manipulation vulnerability. Cooperative collapse exceeds 40% in nearly all models, often reaching catastrophic levels (up to 96%), with minimal trial variance, which constitutes evidence of structural rather than stochastic fragility. Continuous contribution decisions expose multiple points for contextual influence, and repeated rounds amplify belief-channel effects, making these games a natural stress test for coordination stability.

Because many real-world systems—from resource allocation and collective investment to collaborative research—resemble Public Goods dynamics, such scenarios should serve as benchmarks for evaluating manipulation resilience. Future robustness and alignment audits should explicitly test LLM coordination under continuous, repeated, and norm-dependent settings, where competitive framing can induce cascading cooperation failures.

5.4 Deployment Implications and Mitigation Strategies

The patterns we document have direct implications for organizations deploying LLMs in multi-agent coordination scenarios. First, vulnerability is not uniform: manipulation susceptibility depends critically on the interaction between model architecture, game structure, and manipulation direction. Public Goods-like scenarios, where agents make continuous contribution decisions toward collective objectives, represent the highest-risk contexts. Applications such as distributed resource allocation, infrastructure investment coordination, supply chain management, or collective action problems could be particularly susceptible to adversarial manipulation through competitive framing or advice injection.

Second, organizations should strategically select models based on deployment context. For applications requiring robust binary cooperation decisions (contract execution, two-party negotiations), GPT-5 offers exceptional stability. For pure coordination without defection incentives (scheduling, standard selection), most models show reasonable robustness with Llama variants performing particularly well. For continuous contribution scenarios (resource

pooling, collective funding), no model shows strong resistance, suggesting this domain requires additional safeguards regardless of model choice.

Third, the consistent competitive malleability (33–52% increases in Public Goods) presents opportunities for beneficial intervention in genuinely competitive but mutually beneficial scenarios. Industry standard-setting, environmental treaty negotiation, research collaboration, and open-source community coordination could benefit from carefully designed cooperative framings that enhance outcomes beyond what traditional mechanism design achieves. The key distinction is that competitive contexts appear to lack strong default strategic representations, making them amenable to positive influence through cooperative cues without the brittleness observed in cooperative contexts.

Fourth, input filtering and adversarial detection systems should focus on detecting frame-advice conflicts rather than only screening for explicitly malicious content. Unlike direct prompt injection attacks that trigger security filters, frame-based manipulation operates through components that appear individually legitimate. Detecting scenarios where established contextual frames are subsequently contradicted by strategic advice could identify manipulation attempts that would otherwise bypass traditional content filtering.

Fifth, multi-model ensemble approaches may provide better overall robustness than relying on any single architecture. Different models show complementary vulnerability patterns: GPT-5 excels in binary social dilemmas while showing moderate Public Goods vulnerability; Llama variants show extreme binary asymmetry but minimal coordination game effects. Combining models with different vulnerability profiles through ensemble methods or task-specific model selection could provide defense-in-depth against manipulation.

5.5 Limitations and Future Research Directions

Our experiments capture short-term coordination dynamics (10 rounds), offering only a snapshot of manipulation effects. Longer interactions—spanning tens or hundreds of rounds—may reveal new dynamics, such as adaptive resistance or compounding defection cascades driven by belief updating. Extending horizons is essential to determine whether vulnerabilities are transient or persistent features of multi-agent interaction. Similarly, real-world validation in domains like distributed trading, resource allocation, or vehicle coordination would clarify whether these laboratory patterns hold under environmental feedback and human oversight.

Future work should also pursue mechanistic and generalization analyses. Interpretability studies examining how frames and advice are represented could reveal why cooperative defaults are strong yet brittle. Testing additional architectures, reasoning-enhanced models, and culturally or linguistically varied contexts will help assess whether vulnerabilities generalize or reflect dataset biases. Finally, longer-horizon and adversarial fine-tuning experiments—explicitly training on conflicting frames and instructions—may improve robustness beyond canonical social dilemmas, bridging the gap between alignment under idealized scenarios and resilience in complex multi-agent deployments.

6 CONCLUSION

Through systematic experiments across three canonical game paradigms with eight model families (N=30 trials per condition), we provide robust evidence that LLM multi-agent coordination exhibits game-structure-dependent manipulation vulnerabilities varying with model architecture. Public Goods games reveal the most severe patterns: competitive contexts show 33–52 percentage point malleability toward cooperation for most models, while cooperative contexts suffer 61–96 percentage point collapse under competitive manipulation for most models. Prisoner's Dilemma displays heterogeneous patterns with clear generational progression in GPT models (GPT-5: symmetric 1.0:1 ratio, 4 percentage point effects) while other families show variable vulnerabilities. Battle of the Sexes proves most robust with effects typically below 45 percentage points.

No model exhibits uniform robustness across game structures. Every architecture shows significant vulnerability in at least one paradigm, with high cooperative baselines often experiencing the most catastrophic collapses when challenged. This suggests current alignment approaches may create brittleness by establishing strong prosocial defaults without ensuring stability under adversarial challenges.

These findings present both risks and opportunities. Adversaries could exploit cooperative collapse vulnerabilities in applications involving continuous contribution decisions. Conversely, consistent competitive malleability suggests cooperative framings could enhance coordination in competitive but mutually beneficial scenarios. As LLMs are increasingly deployed in multi-agent systems requiring strategic coordination, understanding these game-structure-dependent vulnerabilities while leveraging beneficial asymmetries becomes essential for ensuring safety, reliability, and robustness.

ACKNOWLEDGMENTS

This research was conducted with the support of Pivotal Research. Game structures were adapted from the MLGym benchmark [18]. All experiments comply with API terms of service and ethical guidelines for AI research. We acknowledge potential dual-use concerns regarding documentation of vulnerabilities and have prioritized responsible disclosure practices, emphasizing both security risks and beneficial intervention opportunities.

REFERENCES

[1] Saaket Agashe, Yue Fan, Anthony Reyna, and Xin Eric Wang. 2025. LLM-Coordination: Evaluating and Analyzing Multi-agent Coordination Abilities in Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 8053–8072. https://doi.org/10.18653/v1/2025.findings-naacl.448

[2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs.CL] https://arxiv.org/abs/2212.08073

[3] Elizabeth Bernold, Elisabeth Gsottbauer, Kurt A. Ackermann, and Ryan O. Murphy. 2023. Accounting for preferences and beliefs in social framing effects.

Frontiers in Behavioral Economics Volume 2 - 2023 (2023). https://doi.org/10.3389/frbhe.2023.1147492

[4] Hyeong Kyu Choi, Xuefeng Du, and Yixuan Li. 2024. Safety-Aware Fine-Tuning of Large Language Models. arXiv:2410.10014 [cs.CL] https://arxiv.org/abs/2410.10014

[5] Tore Ellingsen, Magnus Johannesson, Johanna Mollerstrom, and Sara Munkhammar. 2012. Social framing effects: Preferences or beliefs? *Games and Economic Behavior* 76, 1 (May 2012), 117–130. https://doi.org/10.1016/j.geb.2012.05.007

[6] Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. Truthful AI: Developing and governing AI that does not lie. arXiv:2110.06674 [cs.CY] https://arxiv.org/abs/2110.06674

[7] Yuanjun Feng, Vivek Choudhary, and Yash Raj Shrestha. 2025. Noise, Adaptation, and Strategy: Assessing LLM Fidelity in Decision-Making. arXiv:2508.15926 [cs.CE] https://arxiv.org/abs/2508.15926

[8] Mohamed Amine Ferrag, Norbert Tihanyi, Djallel Hamouda, Leandros Maglaras, Abderrahmane Lakas, and Merouane Debbah. 2025. From Prompt Injections to Protocol Exploits: Threats in LLM-Powered AI Agents Workflows. https://doi.org/10.1016/j.icte.2025.12.001 arXiv:2506.23260 [cs.CR]

[9] Nicoló Fontana, Francesco Pierri, and Luca Maria Aiello. 2024. Nicer Than Humans: How do Large Language Models Behave in the Prisoner's Dilemma? arXiv:2406.13605 [cs.CY] https://arxiv.org/abs/2406.13605

[10] Philipp Gerlach and B. Jaeger. 2016. Another frame, another game?: Explaining framing effects in economic games. In *Proceedings of norms, actions, games (NAG 2016)*, A. Hopfensitz and E. Lori (Eds.). Social Psychology. https://doi.org/10.17605/OSF.IO/AB5YP

[11] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. 2021. Risks from Learned Optimization in Advanced Machine Learning Systems. arXiv:1906.01820 [cs.AI] https://arxiv.org/abs/1906.01820

[12] Donghyun Lee and Mo Tiwari. 2024. Prompt Infection: LLM-to-LLM Prompt Injection within Multi-Agent Systems. arXiv:2410.07283 [cs.MA] https://arxiv.org/abs/2410.07283

[13] Yunhao Liang, Yuan Qu, Jingyuan Yang, Shaochong Lin, and Zuo-Jun Max Shen. 2025. Everyone Contributes! Incentivizing Strategic Cooperation in Multi-LLM Systems via Sequential Public Goods Games. arXiv:2508.02076 [cs.AI] https://arxiv.org/abs/2508.02076

[14] Varda Liberman, Steven M. Samuels, and Lee Ross. 2004. The Name of the Game: Predictive Power of Reputations versus Situational Labels in Determining Prisoner's Dilemma Game Moves. *Personality and Social Psychology Bulletin* 30, 9 (Sep 2004), 1175–1185. https://doi.org/10.1177/0146167204264004

[15] Nunzio Lorè and Babak Heydari. 2023. Strategic Behavior of Large Language Models: Game Structure vs. Contextual Framing. arXiv:2309.05898 [cs.GT] https://arxiv.org/abs/2309.05898

[16] Lech Mazur and Charles C. Norton. 2025. Public Goods Game (PGG) Benchmark: Contribute & Punish. https://github.com/lechmazur/pgg_bench

[17] TOSHIYA MURASHIGE and TAKAYUKI ITO. 2025. Simulating Human Decision-Making in Ultimatum Games using Large Language Models. In *Proceedings of the ACM Collective Intelligence Conference (CI '25)*. Association for Computing Machinery, New York, NY, USA, 13–19. https://doi.org/10.1145/3715928.3737473

[18] Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent Moens, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, Gaurav Chaurasia, Dieuwke Hupkes, Ricardo Silveira Cabral, Tatiana Shavrina, Jakob Foerster, Yoram Bachrach, William Yang Wang, and Roberta Raileanu. 2025. MLGym: A New Framework and Benchmark for Advancing AI Research Agents. arXiv:2502.14499 [cs.CL] https://arxiv.org/abs/2502.14499

[19] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL] https://arxiv.org/abs/2203.02155

[20] OWASP Foundation. 2024. *OWASP Top 10 for LLM Applications 2025*. Technical Report. OWASP. https://genai.owasp.org Version 2025, released November 18, 2024.

[21] Jingwei Yi, Rui Ye, Qisi Chen, Bin Zhu, Siheng Chen, Defu Lian, Guangzhong Sun, Xing Xie, and Fangzhao Wu. 2024. On the Vulnerability of Safety Alignment in Open-Access LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 9236–9260. https://doi.org/10.18653/v1/2024.findings-acl.549