# When Does a Learned Mediator Transfer? Susceptibility Thresholds and Equilibrium Selection in LLM Games

**Anonymous authors**
Paper under double-blind review

## Abstract

When LLM agents are applied in multi-agent scenarios, such as automated negotiation and collaborative coding, they often fail to effectively coordinate. In this work, we study a third-party mediator agent trained with reinforcement learning (RL). Rather than applying fixed rules, a learned mediator can adaptively decide when and what to communicate, improving cooperation without accessing agent internals. We evaluate such learned mediators across three LLM families (Llama, Granite, and Qwen) on four games spanning social dilemmas and pricing competition, and provide a mechanistic account of the results. Our experiments demonstrate that a mediator trained on one LLM family transfers zero-shot to others, provided the target model responds monotonically to cooperation messages. A contrarian agent that inverts this response collapses cooperation despite active mediation. Moreover, we find that the value of interventions decomposes in terms of timing, which is model-agnostic and learned by the mediator, and content, reflecting the combined effect of the mediator's messages and peer communication. Interestingly, content adds value only when the population's baseline cooperation rate falls below an empirically estimable threshold $\tau^*$. Above that threshold, timing alone suffices. An extended multi-way controller improves cooperation for susceptible populations by learning state-dependent template selection. No single message type dominates overall, suggesting the gain comes from matching message type to game state rather than from any single dominant framing. Ablations without player communication estimate this threshold empirically, requiring no retraining or access to model internals. These ablations also provide a practical criterion for predicting whether richer mediator vocabularies will help in a new deployment.

## 1 Introduction

Today LLM agents are increasingly deployed in high-stakes multi-agent settings, including automated pricing and financial decision-making tasks (Riemer et al., 2026), collaborative software engineering (Khatua et al., 2026), autonomous negotiation tasks (Davidson et al., 2024), and cybersecurity response tasks (Fang et al., 2024). These applications share a common structure in that they can be modeled as general-sum repeated games, where the joint outcome depends not only on individual capability but on whether agents coordinate on a mutually beneficial equilibrium. Equilibrium selection (i.e., which outcome agents coordinate on when multiple exist) therefore becomes a central practical challenge, not merely a theoretical one (Harsanyi & Selten, 1988; Selten, 1975). For LLM agents, this challenge is particularly acute. Prior work has shown LLMs cooperate at far-above-Nash rates under neutral prompts (Brookins & DeBacker, 2024; Lorè & Heydari, 2023; Fontana et al., 2024), coordinate only partially in coordination games (Akata et al., 2025), and exhibit highly model-dependent behavior in anti-coordination games (Phelps & Russell, 2023). Yadav et al. (2026) show that more capable LLMs are not necessarily more cooperative, and Khatua et al. (2026) find that LLM coding agents fail to coordinate in 50% of tasks even when solo performance is high, a "curse of coordination" driven by strategic misalignment rather than competence.

Existing responses to multi-agent misalignment span three families (Carichon et al., 2025; Fauchard et al., 2026). Rule-based and communication-protocol methods impose structured interaction pat-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

terns (Madmoun & Lahlou, 2026) but offer limited adaptivity to unseen agent populations. Fine-tuning embeds cooperative incentives directly into model weights (Tennant et al., 2025; Yu et al., 2025), but requires access to player internals and must be reapplied for each new model family. Mechanism design uses external rules or contracts to steer equilibria (Tewolde et al., 2026; Ye & Steinhardt, 2026; Kumar et al., 2026), but fixed mechanisms must be designed for an assumed population and cannot update based on observed agent behavior at runtime. Our approach avoids all three limitations by design: the mediator is a learned, black-box controller that operates through natural language alone, requires no access to player weights, and is evaluated on model families it was never trained on.

We introduce a learned mediator trained via RL that observes game history and sends targeted cooperation messages, and provide a mechanistic account of when and why it works, when it transfers, and what vocabulary it needs. We address these questions through formal analysis and targeted experiments. Our central observation is that the mediator's intervention decomposes into two separable components. The first is a structural timing decision (when to send a message, which is model-agnostic) and the second is a linguistic content decision (what to send, whose value depends on the population's steering susceptibility). A population is steering-susceptible when its baseline cooperation rate falls below a threshold $\tau^*$. Only then does message content add value over timing alone. Three formal results organize the paper. The **Transfer Characterization** shows monotone response is sufficient for safe zero-shot transfer, with a contrarian agent falsifying the condition (collapsing cooperation to near-zero). The **Content-Structure Lemma** formalizes the timing-content split, empirically bracketing $\tau^*$. The **Message-Type Proposition** extends this to vocabulary. A richer action space lifts cooperation for susceptible populations, with no single message type dominating overall.

Our contributions are as follows. (1) A learned RL mediator framework for equilibrium selection in LLM games: a black-box PPO controller that observes only behavioral history and steers cooperation through natural language, without accessing player internals (§3, Algorithm 1, Appendix E). (2) Transfer Characterization with empirical testability: monotone response is sufficient for safe zero-shot transfer. A contrarian agent falsifies the condition experimentally, together bracketing monotonicity as the key transfer criterion (§4.1, §4.4). (3) Content-Structure Lemma with empirical susceptibility threshold: intervention value decomposes into timing (model-agnostic) and content (susceptibility-gated) components, with $\tau^*$ bracketed empirically from no-communication ablations (§4.2). (4) Heterogeneous-population evaluation showing that mediator performance tracks the persona of the most susceptible model, not aggregate composition (§4.3). (5) Multi-way message-type controller confirming that vocabulary value is susceptibility-gated, with cooperation gains concentrated in susceptible populations (§4.5).

## 2 Background

**LLM agents in strategic settings.** LLMs do not follow classical game-theoretic predictions. They are sensitive to rhetorical framing (Faulkner et al., 2026) and exhibit highly model-dependent behavior in anti-coordination games such as Chicken (Phelps & Russell, 2023). Akata et al. (2025) find that LLMs perform well in self-interested games (IPD family) but suboptimally in coordination games, and that behavior can be modulated by social framing, which directly motivates the linguistic intervention approach we develop here. Kundu et al. (2026) further show that perceived in-group identity shifts LLM cooperation even absent explicit payoff incentives. Importantly, Yadav et al. (2026) find that capability gains do not automatically produce cooperation gains, so strategic and competence failures are separable, a distinction our steering-susceptibility framework (§3) formalizes. This sensitivity to coordination cues makes LLM populations a natural target for linguistic intervention.

**Mediation and mechanism design.** Tewolde et al. (2026) benchmarks cooperation-sustaining mechanisms (repetition, reputation, player-designed mediation, contracting) on LLM social dilemmas, finding that fixed mechanisms achieve 69.5–80% of the cooperative optimum. Ye & Steinhardt (2026) show that fixed norm-enforcement rules are exploitable, as adversarial agents learn to game static thresholds, motivating adaptive enforcement policies. Kumar et al. (2026) evolve interpretable constitutions offline and require modifying player prompts rather than intervening at runtime. Our mediator fits within a broader agenda of developing behavioral harnesses for LLM agents (i.e., external controllers that steer behavior through communication rather than weight modification or

reward shaping). Unlike single-agent alignment proposals, the harness must steer a heterogeneous population that includes adversarial personas and model families, making transfer and susceptibility central concerns. Together, these results establish that cooperation-sustaining mechanisms can work, but leave open why learned selectivity succeeds, when it transfers to unseen agents, and what vocabulary it needs, all questions a practitioner must answer before deployment. The present work provides that account.

**RL mechanism design and transfer.** Prior work on RL mechanisms assumes access to agent policies or gradients (Baumann et al., 2019; Marris et al., 2021; Yang et al., 2020). Our black-box LLM setting is closer to principal-agent RL (Krakovna et al., 2020) where the principal cannot shape update rules (Segura et al., 2025; Kim et al., 2022). Sudhakar et al. (2025) show zero-shot transfer in Hanabi by encoding a representation invariant to team size. Whereas our Transfer Characterization identifies the analogous invariant (monotone cooperation response) for LLM families.

# 3 Mediator Framework and Theoretical Analysis

We formalize the game and mediator (§3.1), decompose intervention value into structural and linguistic components (§3.2), and state our three primary theoretical results: the Transfer Characterization (§3.3), Content-Structure Lemma (§3.4), and Message-Type Proposition (§3.5). Each result is directly tested in §4, with full proofs in Appendix A.

## 3.1 Game and Mediator Formulation

We consider an $n$-player repeated general-sum game with $T$ rounds. At each round $t$, each player $i$ simultaneously chooses an action $a_i^t$ from a finite or continuous action set. In social dilemma variants (IPD, Stag Hunt, Chicken) the action space is binary (cooperate or defect). Players receive per-round payoffs $r_i^t(\mathbf{a}^t)$ and accumulate a total score $R_i = \sum_t r_i^t$. Each player is an LLM prompted with a game-specific system prompt and the visible game history. A mediator $\mu$ observes $\mathcal{S}_t = \{t, \mathbf{h}_{t-H:t}, \mathbf{R}_{1:t}\}$ (round index, action history window, cumulative scores) and, at selected rounds, sends a targeted natural-language message to one player. $\mathcal{S}_t$ encodes only numerical outcomes (not the text of inter-agent messages), so the mediator's decisions are based purely on behavioral history. The mediator does not access player internals, cannot modify payoffs, and cannot compel actions. Its sole lever is the content and timing of messages.

### Learning Objective

The mediator's policy $\pi_\mu : \mathcal{S}_t \to \mathcal{A}_\mu$ maps game state to an intervention action. In the binary-action setting, $|\mathcal{M}|{=}1$ (a single fixed cooperation template, given in full in Appendix C), so the effective choice is to intervene or abstain (send nothing). Formally $\mathcal{A}_\mu = \{0, 1\}$ where action 1 triggers the message. The target player $i$ is selected by a fixed rule (the player with the lowest observed cooperation rate at round $t$ with details in Appendix H), so the learned policy controls only when to intervene, not whom to target. In the 5-way setting, $\mathcal{A}_\mu = \{0, 1, 2, 3, 4\}$ where action 0 is abstain and actions 1–4 select among four fixed content templates as a learned discrete choice (§4.5). $\pi_\mu$ is trained via proximal policy optimization (PPO) with a shaped reward: $r_\mu^t = \bar{r}_t - \lambda_c \cdot \mathbf{1}[\text{intervene}]$, where $\bar{r}_t = \frac{1}{n}\sum_i r_i^t$ is the mean per-round player payoff at round $t$ (distinct from the cumulative scores $\mathbf{R}_{1:t}$ in the state) and $\lambda_c > 0$ is an intervention cost that penalizes unnecessary messages. Intervention timing is not explicitly rewarded. It emerges from the welfare signal, as the policy learns to intervene when the per-round cooperation gain exceeds $\lambda_c$.

## 3.2 Structural Decomposition of Intervention Value

We decompose the mediator's intervention into two separable components: a structural signal $\sigma_t \in \{0, 1\}$ (whether to intervene at round $t$) and a linguistic signal $m_t \in \mathcal{M}$ (message content, active only when $\sigma_t = 1$). Define the cooperation response function for LLM player $\theta$: $f_\theta(\sigma, m)$ = expected cooperation probability given structural signal $\sigma$ and message $m$, where $\Delta f_\theta(m \mid \sigma{=}0) = 0$ by convention (no content effect without intervention). For a population of players $\{\theta_i\}$, the aggregate baseline cooperation rate $\bar{c}_0 = \mathbb{E}_i[f_{\theta_i}(0, \cdot)]$, where $\sigma{=}0$ denotes the never-intervene condition (mediator present but never sending, peer communication active), summarizes population-level sus-

ceptibility and corresponds exactly to $\bar{c}_0$ in Definition 1. All lemmas and propositions below are stated in terms of this quantity. Decompose:

$$f_\theta(\sigma, m) = f_\theta^{\text{struct}}(\sigma) + \Delta f_\theta(m \mid \sigma{=}1) \tag{1}$$

where $f_\theta^{\text{struct}}$ captures the effect of the timing decision alone and $\Delta f_\theta$ captures the marginal value of linguistic content given intervention, a term that reflects the population's sensitivity to cooperative framing from any channel, including both the mediator message and peer communication. This decomposition is meaningful only if the two components are approximately separable in practice, meaning the structural timing signal and linguistic content must not deeply interact. We treat this as an empirical assumption and test it in §4.2 by removing the player communication channel, isolating $f_\theta^{\text{struct}}$ and measuring whether $\Delta f_\theta > 0$ holds as a function of $\bar{c}_0$.

### 3.3 Transfer Proposition

We identify a sufficient condition (monotonicity of the cooperation response) on a target population for safe zero-shot transfer.

**Proposition 1** (Transfer Characterization via Monotone Cooperation Response)**.** Let $\pi_\mu$ be a mediator policy trained on player model $\theta_1$, where $\theta_1$ satisfies the structural monotonicity condition $f_{\theta_1}(1, \cdot) \geq f_{\theta_1}(0, \cdot)$ for all game states (so the learned timing policy targets genuine coordination-failure states). A transfer target $\theta_2$ achieves non-negative intervention value from $\pi_\mu$ without retraining if it also satisfies structural monotonicity.

The proof follows from the observation that the timing policy detects coordination-failure states from observable history alone, so the monotonicity of $\theta_2$ is a sufficient condition for safe zero-shot transfer. Corollary 2 establishes that violating monotonicity removes this guarantee: a contrarian agent ($f_{\theta_c}(1, \cdot) < f_{\theta_c}(0, \cdot)$) can drive intervention value negative. Together, Proposition 1 and Corollary 2 bracket monotonicity as a key condition (full proof in Appendix A). Monotonicity must be verified before deployment by measuring whether cooperation rises or falls following intervention in an initial test. Positive gains beyond the non-harm guarantee depend additionally on $\theta_2$'s susceptibility (Definition 1).

A contrarian agent is a player $\theta_c$ whose prompt inverts the usual cooperation response. It increases its defection probability upon receiving a cooperation message (formally, $f_{\theta_c}(1, \cdot) < f_{\theta_c}(0, \cdot)$). Contrarian agents violate structural monotonicity by construction.

**Corollary 2** (Monotonicity Violation Removes the Transfer Guarantee)**.** If a population contains a contrarian agent $\theta_c$ (i.e., $f_{\theta_c}(1, \cdot) < f_{\theta_c}(0, \cdot)$), then the monotonicity condition of Proposition 1 is violated and the non-harm guarantee no longer holds: $\pi_\mu$ may achieve negative intervention value on the population. The severity depends on game payoff structure. In social dilemmas such as IPD where defection is contagious, a single contrarian can collapse population-wide cooperation.

Every intervention that raises cooperation for compliant players simultaneously accelerates defection in $\theta_c$. In games where mutual defection is a Nash attractor (e.g., IPD) this cascades population-wide (full proof in Appendix A).

### 3.4 Content-Structure Lemma

**Definition 1** (Steering Susceptibility)**.** A player population is steering-susceptible in game $\mathcal{G}$ if its aggregate baseline cooperation rate $\bar{c}_0(\mathcal{G}, \theta)$ (the never-intervene cooperation rate: mediator present but never sending messages, peer communication active) falls below a game-and-model-specific threshold $\tau^*(\mathcal{G}, \theta)$.

Intuitively, a susceptible population is one where players are not yet in a stable cooperative equilibrium and therefore respond to the linguistic content of cooperative framing (from any channel, including mediator messages and peer communication), not just to the structural presence of an intervention. A non-susceptible population ($\bar{c}_0 \geq \tau^*$) has already coordinated sufficiently that linguistic content provides no additional lift over the structural timing signal alone. The intuition above ties susceptibility to responsiveness to cooperative framing from any channel, whether mediator messages or peer communication. This responsiveness admits two empirical signatures. The first is a low with-chat baseline, meaning $\bar{c}_0 < \tau^*$. The second is a drop in cooperation when the

peer-communication channel is removed while the mediator is held fixed. A population may exhibit either signature or both. The second signature reveals content-sensitivity even when peer chat lifts the with-chat baseline above $\tau^*$. We write $\bar{c}_0$ and $\tau^*$ as shorthand when game and model are clear from context.

Even when transfer is safe, the value of message content might not be uniform across populations. A population already cooperating at high rates has little room to benefit from linguistic framing, and only populations that have not yet reached a cooperative equilibrium are sensitive to the mediator's message. The lemma below identifies $\bar{c}_0$ as a key quantity separating content-sensitive populations ($\bar{c}_0 < \tau^*$) from those where timing alone suffices ($\bar{c}_0 \geq \tau^*$).

**Lemma 3** (Content Value Requires Steering Susceptibility)**.** Assuming approximate separability of the structural and linguistic components (Eq. 1), for a player population with baseline cooperation rate $\bar{c}_0 \geq \tau^*(\theta)$, removing message content from a selective intervention policy does not significantly decrease expected welfare. The threshold $\tau^*(\theta)$ is model-specific and increases with the player model's sensitivity to cooperative framing.

From decomposition (1), content value is $\Delta f_\theta(m \mid \sigma{=}1)$ per intervention. When $\bar{c}_0 \geq \tau^*$ this term is $\approx 0$ because cooperation is near-optimal at intervention states (full proof in Appendix A).

**Corollary 4** ($\tau^*$ is Empirically Estimable)**.** The threshold $\tau^*$ can be estimated without ground-truth knowledge of $\theta$: run the nocomm ablation across player models at several cooperation baseline levels and identify the crossover where the content effect transitions from significant to non-significant. Empirically, $\tau^* \in (0.428, 0.466)$, bracketed by Granite (non-susceptible: $\bar{c}_0{=}42.8\%$, no significant degradation on SH ($p{=}0.70$), marginal on IPD ($p{<}0.05$)) and Qwen ($\bar{c}_0{=}46.6\%$, $p{<}0.001$, confirming susceptibility).

### 3.5 Message-Type Proposition

The Content-Structure Lemma shows that message content matters for susceptible populations but leaves open whether the specific template choice matters. If content value is driven by state-dependent vocabulary selection, a richer action space should yield higher cooperation gains for susceptible populations.

**Proposition 5** (Vocabulary Expands Content Value for Susceptible Populations)**.** Lemma 3 assumes a fixed single-template vocabulary $|\mathcal{M}|{=}1$. For a susceptible population ($\bar{c}_0 < \tau^*$), a mediator with a richer vocabulary $|\mathcal{M}|{>}1$ achieves weakly higher expected welfare than a single-template mediator, because a larger action space provides higher-value content options that increase $\Delta f_\theta(m \mid \sigma{=}1)$. The gain is near-zero for non-susceptible populations ($\bar{c}_0 \geq \tau^*$), where $\Delta f_\theta \approx 0$ regardless of vocabulary size.

For non-susceptible populations, $\Delta f_\theta \approx 0$ for all $m$ by Lemma 3, so enlarging $\mathcal{M}$ adds no value. For susceptible populations, the expanded vocabulary allows the policy to select a higher-value $m^*$, giving weakly higher expected welfare (full proof in Appendix A).

## 4 Experiments

**Setup.** We use the 3-player TextArena (Guertler et al., 2025) social dilemma environment: IPD, Stag Hunt, and Chicken, 5 rounds, free-form communication then simultaneous action. Each game uses three heterogeneous agents (standard, malicious, benevolent personas). Sections 4.1–4.4 cover IPD and Stag Hunt. Chicken appears in §4.5 only. Bertrand pricing differs structurally (continuous pricing, dense per-round reward) and is covered in Appendix G.

**Default controller.** Unless stated, we use a binary-action PPO controller trained on Llama 3.3-70B players. We choose Granite 4.1-30B and Qwen 2.5-72B as transfer targets, covering both cross-family and cross-boundary transfer scenarios, and evaluate the controller on them without retraining, making §4.1–4.4 zero-shot transfer experiments. The susceptibility interpretation is established in §4.2. An extended multi-message controller is introduced and evaluated in §4.5. All comparisons use Welch $t$-tests on per-seed means (rule-based baselines: 5 seeds $\times$ 50 games, PPO: 5 train seeds $\times$ 5 eval seeds $\times$ 50 games, unless stated otherwise).

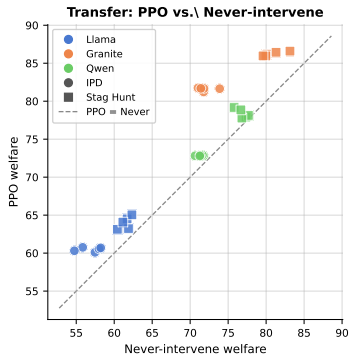


Figure 1: PPO vs. unmediated welfare per model×game. All six points lie above the diagonal, confirming zero-shot transfer (Prop. 1).

| | | Llama 3.3-70B | | Granite 4.1-30B | | Qwen 2.5-72B | |
|---|---|---|---|---|---|---|---|
| **Game** | **vs baseline** | ΔCoop% | $p$ | ΔCoop% | $p$ | ΔCoop% | $p$ |
| IPD | vs *never* | +1.70 | 0.040 | +9.65 | <0.001 | +1.61 | 0.075 |
| | vs *always* | +2.84 | <0.001 | −3.79 | <0.001 | +1.97 | 0.029 |
| | vs *coop-thresh* | −11.90 | <0.001 | −1.53 | 0.005 | −1.00 | 0.073 |
| Stag Hunt | vs *never* | +1.19 | 0.31 | +5.89 | <0.001 | +1.20 | 0.061 |
| | vs *always* | +1.23 | 0.32 | +3.37 | <0.001 | +0.39 | 0.244 |
| | vs *coop-thresh* | −9.22 | <0.001 | +2.35 | <0.001 | +1.08 | 0.014 |

Table 1: PPO vs. baselines (ΔCoop% = cooperation-rate change. Qwen uses the Llama-trained controller zero-shot). PPO vs. *never* is non-negative on all six model×game pairs, with significant improvement on four of six ($p<0.05$), confirming Prop. 1. The sign of PPO vs. *always* tracks susceptibility: selective over constant messaging, as constant intervention disrupts the peer framing they rely on (both degrade without chat, §4.2). Granite ($\bar{c}_0>\tau^*$) does not, as its players already cooperate at high rates via peer chat alone (§4.2).

### 4.1 Zero-Shot Transfer Performance

**Zero-shot transfer (Granite, Qwen).** Testing Proposition 1, we compare PPO against three non-learning baselines: *never*-intervene (unmediated behavior), *always*-intervene (undiscriminating messaging every round), and *coop-thresh* (intervenes when empirical cooperation falls below 0.5, the strongest hand-engineered selector). The full definitions are in Appendix D.

Table 1 and Figure 1 summarize results. PPO achieves non-negative welfare gain on all six model×game pairs and significant improvement on four of six ($p<0.05$), confirming Prop. 1.

### 4.2 Content-Structure Ablation

We disable peer-to-peer messages while leaving the mediator channel intact. This probes the susceptibility condition of Lemma 3: a population below $\tau^*$ responds to cooperative framing from any channel, so removing peer chat should degrade cooperation there but not for populations above $\tau^*$.

Table 2 reports cooperation rates with and without player-to-player chat (mediator active in both conditions). Across models the two susceptibility signatures of Defintion 1 give a graded picture, where Qwen is on both axes, Llama is on the without-communication axis only, and Granite on neither. Removing chat degrades Llama cooperation on both games ($-2.2\%$, $p<0.05$; $-3.1\%$, $p<0.05$). Its with-chat baseline is high ($\bar{c}_0\approx 59.8\%$), so Llama' is not baseline-susceptible, yet the drop shows part of that cooperation rests on peer framing rather than the mediator's structural signal alone. Granite PPO cooperation is nearly unchanged (IPD $-2.7\%$, $p<0.05$. SH $+0.8\%$, $p=0.70$), so its cooperation is driven by the mediator's timing signal rather than peer framing, consistent with non-susceptibility ($\bar{c}_0 > \tau^*$). Qwen PPO degrades sharply on IPD ($-22.5\%$, $p<0.001$) and significantly on SH ($-1.9\%$, $p<0.01$), and pairs this with a low baseline ($\bar{c}_0=46.6\% < \tau^*$), placing it on both axis. Its *always*-intervene baseline degrades even more (IPD: $46.7\%\rightarrow 18.8\%$). Together, Granite ($\bar{c}_0=42.8\%$) and Qwen ($\bar{c}_0=46.6\%$) bracket $\tau^* \in (0.428, 0.466)$, confirming Corollary 4. The bracket is cleanest on SH where Granite's result is clearly non-significant ($p=0.70$).

### 4.3 Heterogeneous Population Experiment

Sections 4.1 and 4.2 keep the player model fixed (all Llama, or all Granite). Here we test whether the Llama-trained PPO controller transfers to mixed-model populations, and ask what drives performance: aggregate model composition or the specific persona-to-model assignment. Because the PPO state encodes only numerical outcomes, not model identity (§3.1), one might expect aggregate composition to be the sole driver, independent of which model carries which persona.

**Setup.** We evaluate three model-mix configurations on IPD and Stag Hunt (5 seeds × 20 games): **llg** (2 Llama + 1 Granite), **lgq** (1 Llama + 1 Granite + 1 Qwen), and **lgg** (1 Llama + 2 Granite). For each configuration we run three persona assignments (see Table 3 caption for exact assignments): a

| Game | Controller | Llama 3.3-70B | | Granite 4.1-30B | | Qwen 2.5-72B | |
|---|---|---|---|---|---|---|---|
| | | w/ chat | w/o chat | w/ chat | w/o chat | w/ chat | w/o chat |
| IPD | PPO | $52.6_{\pm 1.5}$ | $21.31^{*}_{\pm 0.74}$ | $65.06_{\pm 3.89}$ | $62.39^{*}_{\pm 4.26}$ | $52.14_{\pm 0.22}$ | $29.61^{***}_{\pm 0.30}$ |
| | Always | 47.49 | 17.87 | $71.68_{\pm 1.37}$ | $63.10^{***}_{\pm 1.88}$ | $46.69_{\pm 3.03}$ | $18.80^{\dagger}$ |
| SH | PPO | $32.40_{\pm 1.49}$ | $29.33^{*}_{\pm 0.74}$ | $77.19_{\pm 7.53}$ | $77.96_{\pm 6.71}$ | $62.83_{\pm 0.47}$ | $60.89^{**}_{\pm 0.28}$ |
| | Always | 50.67 | 26.27 | $72.69_{\pm 1.80}$ | $80.83^{***}_{\pm 1.13}$ | $61.84_{\pm 1.02}$ | $58.40^{\dagger}$ |

Table 2: Cooperation rate (%) with vs. without player-to-player chat. Mediator active in both conditions. $t$-test, with-chat vs. without-chat (PPO: 5×5 seeds. Always: 5 seeds for Granite, single run otherwise). $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$. $^{\dagger}$Single 50-game run (context only, not used in primary comparisons). Llama and Qwen PPO degrade without chat, evidencing content-sensitivity (Qwen also has a low with-chat baseline , $\bar{c}_0 < \tau^*$)). Granite ($\bar{c}_0 > \tau^*$) does not, confirming non-susceptibility and bracketing $\tau^* \in (0.428, 0.466)$.

**baseline** (p0), and two **permutations** (p1, p2) that rotate which model carries each directive while holding model composition fixed. If aggregate susceptibility alone governs performance, results should be invariant across p0/p1/p2 within each composition class.

Table 3 shows a striking asymmetry driven by Llama's susceptibility. In **lgg** (1 Llama + 2 Granite), p0 achieves high cooperation (84.5% IPD, 78.6% SH) because Llama carries the standard prompt while malicious is held by Granite. Giving Llama the malicious directive (p1) reduces cooperation by 30.4% on IPD and 21.6% on SH ($p<0.001$ both). Giving Llama the benevolent directive (p2) also significantly degrades cooperation (67.9% IPD, 61.4% SH), confirming that it is Llama's exposure to the malicious player's peer messages — not the specific directive label — that drives the result. In **llg** (2 Llama + 1 Granite), p0 yields low cooperation (50.6% IPD, 51.5% SH). Moving the malicious persona to Granite (p2) recovers cooperation significantly (74.4% IPD, 63.0% SH, $p<0.001$), confirming Granite's robustness when holding the malicious role. In **lgq** (1 Llama + 1 Granite + 1 Qwen), giving Llama the malicious directive (p1) sharply reduces cooperation to 45.5% (IPD) and 45.8% (SH); moving malicious to Qwen (p2) partially recovers cooperation (74.8% IPD, 67.6% SH), but not to p0 levels, consistent with Qwen also being susceptible ($\bar{c}_0 = 46.6\% < \tau^*$). The operative predictor is which model holds the susceptible-exposing assignment: susceptible models (Llama, Qwen, both below $\tau^*$) are vulnerable to adversarial peer signals under the standard prompt, while Granite is not.

| Mix | Game | PPO (p0) | PPO (p1) | PPO (p2) |
|---|---|---|---|---|
| **llg** | IPD | $50.6_{\pm 3.1}$ | $54.1_{\pm 3.0}$ | $74.4^{***}_{\pm 1.9}$ |
| | SH | $51.5_{\pm 3.9}$ | $53.5_{\pm 2.1}$ | $63.0^{**}_{\pm 3.7}$ |
| **lgg** | IPD | $84.5_{\pm 1.8}$ | $54.1^{***}_{\pm 3.8}$ | $67.9^{***}_{\pm 2.5}$ |
| | SH | $78.6_{\pm 3.0}$ | $57.0^{***}_{\pm 1.7}$ | $61.4^{***}_{\pm 4.8}$ |
| **lgq** | IPD | $83.3_{\pm 2.4}$ | $45.5^{***}_{\pm 2.9}$ | $74.8^{**}_{\pm 4.8}$ |
| | SH | $79.1_{\pm 1.7}$ | $45.8^{***}_{\pm 1.5}$ | $67.6^{***}_{\pm 1.9}$ |

Table 3: Heterogeneous population results (Llama-trained PPO, zero-shot, 5 seeds × 20 games per cell). p0 = standard assignment, p1/p2 = persona permutations (Appendix F). $^{**}p<0.01$, $^{***}p<0.001$ vs. p0. Giving Llama a non-standard directive (p1) sharply reduces cooperation. Moving malicious away from Llama (p2) recovers it.

### 4.4 Contrarian Agent Falsification

Corollary 2 predicts that intervention value collapses when monotonicity is violated. We construct a contrarian agent (Appendix B) to directly test this: one of three players is instructed to treat a cooperation message as a signal to defect ($f_{\theta_c}(1,\cdot) < f_{\theta_c}(0,\cdot)$), directly violating monotonicity.

Cooperation collapses to near-zero (0.05% IPD, 0.11% SH) despite the mediator intervening at rates comparable to the standard setting (4.2 per game on IPD, 3.1 on Stag Hunt). The never- and always-intervene baselines also collapse to 0.0%, confirming that the failure is not a mediator failure per se: the contrarian's unconditional defection triggers reciprocal punishment from compliant players, locking all parties into mutual defection regardless of mediation. Welfare ($30.12 \pm 0.24$ IPD, $30.51 \pm 0.22$ SH) matches the mutual-defection payoff floor for all three controllers, confirming Corollary 2. Monotonicity is thus a testable prerequisite for positive transfer, and a brief no-communication ablation (§4.2) verifies it before deployment.

| Controller | | Granite | | | Llama | | |
|---|---|---|---|---|---|---|---|
| | | **IPD** | **SH** | **Chicken** | **IPD** | **SH** | **Chicken** |
| Never | Coop% | $54.3_{\pm 2.7}$ | $53.5_{\pm 1.3}$ | $53.3_{\pm 2.2}$ | $59.8_{\pm 1.1}$ | $60.8_{\pm 1.2}$ | $58.7_{\pm 2.1}$ |
| Always | Coop% | $71.7_{\pm 1.4}$ | $72.7_{\pm 1.8}$ | $73.5_{\pm 1.5}$ | $48.3_{\pm 0.8}$ | $48.2_{\pm 0.9}$ | $49.8_{\pm 1.0}$ |
| Binary | Coop% | $65.1_{\pm 3.9}$ | $77.2_{\pm 7.5}$ | $54.9_{\pm 2.7}$ | $52.6_{\pm 1.5}$ | $55.7_{\pm 1.5}$ | $59.0_{\pm 1.9}$ |
| **5-way** | Coop% | $\mathbf{78.6_{\pm 2.8}}$ | $\mathbf{78.9_{\pm 1.9}}$ | $\mathbf{79.4_{\pm 2.0}}$ | $\mathbf{58.0_{\pm 1.6}}$ | $\mathbf{58.1_{\pm 2.0}}$ | $58.9_{\pm 2.0}$ |
| 5-way | $\Delta$ | $+13.5\%$ | $+1.7\%$ | $+24.5\%$ | $+5.3\%$ | $+2.3\%$ | $+0.1\%$ |

Table 4: 5-way action controller vs. binary-action baseline. Cooperation rate (%, $\text{mean}_{\pm\text{std}}$), 5 train seeds $\times$ 5 eval seeds $\times$ 50 games. $\Delta$ = 5-way minus binary (cooperation-rate difference). Granite gains most on Chicken (+24.5%), where binary template provides no signal. Llama gains are modest, consistent with its high never-intervene baseline leaving little headroom. Granite's susceptibility is game-dependent: below $\tau^*$ on Chicken, above it on IPD/SH.

### 4.5 Message-Type Action Space

The preceding experiments used a binary-action controller that chooses between abstaining and sending a single fixed-template cooperation message. We ask whether learning which type of message to send adds value beyond learning when to send one. This also addresses a gap from §4.1: the binary-action controller achieves near-unmediated performance on Chicken for both models, suggesting the fixed template provides no useful signal in a game with a different coordination barrier, which a richer vocabulary may close.

**Setup.** We train a 5-way action controller with actions: 0 = abstain, 1 = payoff rationality ("cooperation maximizes your expected score over remaining rounds"), 2 = social norm ("all players have been cooperating, maintaining this norm benefits everyone"), 3 = threat ("continued defection will be recorded and may affect future interactions"), 4 = reassurance ("cooperating this round is safe, your partner is unlikely to defect"). All other hyperparameters match the binary-action controller. The policy learns both the timing decision and the message-type selection jointly. We evaluate on Granite 4.1-30B and Llama 3.3-70B, each with 5 training seeds $\times$ 5 eval seeds $\times$ 50 games (25 evaluation runs per model).

**Results.** For Granite, the 5-way controller improves over binary on all three games. The Chicken result is most notable: the binary controller (54.9%) sits at the no-intervention baseline (53.3%), whereas 5-way (79.4%) breaks well above always-intervene (73.5%). Here Granite's baseline falls below $\tau^*$, unlike on IPD/SH, so susceptibility is a game-model pair property (Definition 1). As Figure 2 shows, the policy does not achieve this by learning a dominant game-specific framing. Conditional on intervening, it selects the four content templates nearly uniformly ($\approx$25% each). The gain is therefore better explained by state-dependent template selection, with no single template dominating overall.

For Llama, the 5-way controller improves over binary on IPD and Stag Hunt (+5.3%, +2.3%), but cooperation remains at or below never-intervene on all three games. This asymmetry is consistent with Lemma 3: Granite's with-chat never-intervene baseline ($\approx 54\%$) leaves headroom for vocabulary gains, while Llama's with-chat baseline is already near 60% (Table 4), so the binary controller has captured most available gain. Vocabulary value thus depends jointly on susceptibility and on remaining headroom above the binary-action baseline.

### 4.6 Bertrand Price Competition

We test whether the framework generalizes to a structurally different economic setting: two-firm Bertrand pricing (Riemer et al., 2026) with continuous actions, dense per-round rewards, and a regulator framing. Full setup, results, and theory connection are in Appendix G. In brief, the prompt condition (Maximize Profit / +Explicit Collusion etc.) plays the role of $\bar{c}_0$: a prompt-conditioned controller matches per-prompt specialists (mean PG = 0.292) while generalizing across all four conditions. A contextual bandit (PPO with $\gamma$=0, no critic) outperforms full PPO here: because Bertrand rewards are dense and per-round, long-horizon credit assignment adds noise rather than signal, consistent with the credit-assignment interpretation of Lemma 3 (details in Appendix G).

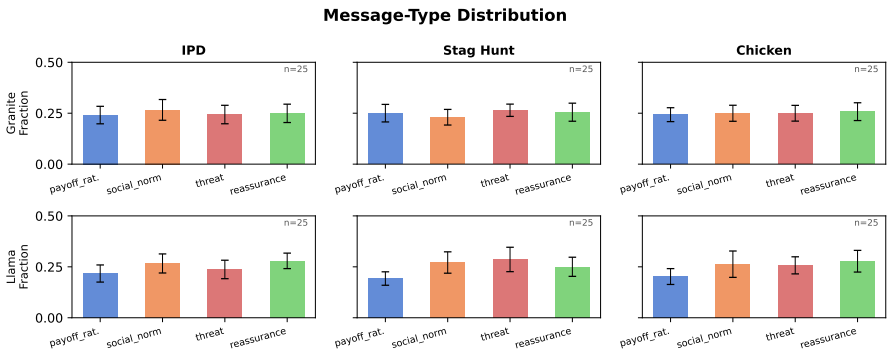


Figure 2: Message-type distribution for the 5-way controller across games and models (25 runs each: 5 train seeds $\times$ 5 eval seeds, error bars = $\pm 1$ std). Conditional on intervening, all four content templates are selected at approximately equal rates ($\approx$25% each, $std$<5%), with no statistically significant game-specific concentration for either model.

## 5 Discussion

**Practical diagnostics.** Three steps follow from our findings. (1) Measure $\bar{c}_0$: if $\bar{c}_0 > \tau^* \approx 0.45$ (midpoint of the empirical bracket $(0.428, 0.466)$), timing alone suffices. If below, design message content for the deployment context. (2) Run a behavioral probe: in a short pilot, agents whose cooperation rate *decreases* following intervention violate monotonicity and should be flagged before full deployment. (3) For susceptible populations, deploy a multi-way vocabulary with richer message diversity improving cooperation gains even when no single template dominates.

**Limitations.** $\tau^*$ is bracketed from two anchor models, and vocabulary size is tested only at binary vs. 5-way. Both would tighten with more models and larger vocabularies. All social dilemma experiments use 3-player games with fixed persona composition. Stability across larger player counts or different mixes remains open (Flint et al., 2026). The nocomm ablation measures aggregate content sensitivity across both communication channels (peer chat and mediator message) and cannot separately attribute the effect to each. A condition that removes only the mediator's linguistic content while preserving peer chat would isolate each channel's contribution cleanly and is left for future work. The mediator uses a shallow RL policy (MLP over behavioral statistics) rather than a fine-tuned LLM or transformer architecture. This is a deliberate design choice: a lightweight, auditable controller models a governance scenario where the coordinating authority lacks the compute or model access of the agents it oversees. All experiments use models available at submission time. Evaluating newer generations would test whether susceptibility is stable within a model family.

## 6 Conclusion

We introduced a learned RL mediator framework for equilibrium selection in LLM games, and studied how and when such a mediator transfers across LLM families. Experiments across three LLM families and three social dilemma games—IPD, Stag Hunt, and Chicken (plus Bertrand pricing in Appendix G)—reveal a consistent mechanistic picture. Transfer holds zero-shot when players respond monotonically to intervention, and fails when a contrarian agent inverts that response. Intervention value decomposes into a timing component (model-agnostic) and a content component (model-dependent). Content matters only for populations whose baseline cooperation rate falls below an empirically estimable susceptibility threshold $\tau^*$. In heterogeneous populations, performance is governed by which persona is assigned to the most susceptible model, making persona assignment a finer-grained predictor of cooperation than aggregate model composition. An extended multi-way controller further improves cooperation for susceptible populations via state-dependent template selection, with no single message type dominating overall. The practical upshot is a three-step diagnostic: measure $\bar{c}_0$, probe for monotonicity, and deploy a multi-way vocabulary for susceptible populations. Open questions include larger player counts, tightening the $\tau^*$ bracket, and testing application-specific vocabularies. A natural extension is an LLM-as-mediator, where the controller itself is a prompted language model rather than a trained RL policy. Our framework predicts such a mediator would succeed only for susceptible populations, a hypothesis we leave for future work.

## AI DISCLOSURE

The authors used AI-assisted writing tools (e.g., large language models) for editing and improving the clarity of paper text. All theoretical results, experimental designs, implementations, and conclusions were formulated and verified by the authors. No AI tools were used to generate synthetic data, derive or prove mathematical claims, or interpret experimental results.

## REFERENCES

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 9:1380–1390, 2025. doi: 10.1038/s41562-025-02172-y.

Tobias Baumann, Thore Graepel, and John Shawe-Taylor. Adaptive mechanism design: Learning to promote cooperation, 2019.

Philip Brookins and Jason Matthew DeBacker. Playing games with GPT: What can we learn about a large language model from canonical strategic games? *Economics Bulletin*, 44(1):25–37, 2024.

Florian Carichon, Aditi Khandelwal, Marylou Fauchard, and Golnoosh Farnadi. The coming crisis of multi-agent misalignment: AI alignment must be a dynamic and social process. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.01080.

Tim R. Davidson, Veniamin Veselovsky, Martin Josifoski, Maxime Peyrard, Antoine Bosselut, Michal Kosinski, and Robert West. Evaluating language model agency through negotiations. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2401.04536.

Richard Fang, Rohan Bindu, Akul Gupta, and Daniel Kang. LLM agents can autonomously exploit one-day vulnerabilities. *arXiv preprint arXiv:2404.08144*, 2024. URL https://arxiv.org/abs/2404.08144.

Marylou Fauchard, Florian Carichon, Margarida Carvalho, and Golnoosh Farnadi. Even more deception: Objective misalignment in mixed-motive LLM multi-agent systems, 2026.

Ryan Faulkner, Anushka Deshpande, David Guzman Piedrahita, Joel Z. Leibo, and Zhijing Jin. Evaluating cooperation in LLM social groups through elected leadership. *arXiv preprint arXiv:2604.11721*, 2026.

Ariel Flint, Luca Maria Aiello, Romualdo Pastor-Satorras, and Andrea Baronchelli. Group size effects and collective misalignment in LLM multi-agent systems. *Proceedings of the National Academy of Sciences*, 2026. doi: 10.1073/pnas.2531697123.

Nicolò Fontana, Francesco Pierri, and Luca Maria Aiello. Nicer than humans: How do large language models behave in the prisoner's dilemma? *arXiv preprint arXiv:2406.13605*, 2024.

Leon Guertler, Bobby Cheng, Simon Yu, Bo Liu, Leshem Choshen, and Cheston Tan. Textarena, 2025.

John C. Harsanyi and Reinhard Selten. *A General Theory of Equilibrium Selection in Games*. MIT Press, 1988.

Arpandeep Khatua, Hao Zhu, Peter Tran, Arya Prabhudesai, Frederic Sadrieh, Johann K. Lieberwirth, Xinkai Yu, Yicheng Fu, Michael J. Ryan, Jiaxin Pei, and Diyi Yang. CooperBench: Why coding agents cannot be your teammates yet, 2026.

Dong-Ki Kim, Matthew Riemer, Miao Liu, Jakob Foerster, Michael Everett, Chuangchuang Sun, Gerald Tesauro, and Jonathan P How. Influencing long-term behavior in multiagent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18808–18821, 2022.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of AI ingenuity. DeepMind Blog, 2020.

Ujwal Kumar, Alice Saito, Hershraj Niranjani, Rayan Yessou, and Phan Xuan Tan. Evolving interpretable constitutions for multi-agent coordination. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.00755.

Akash Kundu, Emanuel Tewolde, Ratip Emin Berker, Samuel F. Brown, and Vincent Conitzer. Do LLMs take care of their own? Similarity signals can induce cooperation. OpenReview:neTpZbiyNn, 2026. Honorable Mention, ICML 2026 AI4GOOD Workshop.

Nunzio Lorè and Babak Heydari. Strategic behavior of large language models: Game structure vs. contextual framing. *arXiv preprint arXiv:2309.05898*, 2023.

Hamza Madmoun and Saif Lahlou. Communication enables cooperation in LLM agents: A comparison with curriculum-based approaches. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, 2026. URL https://arxiv.org/abs/2510.05748.

Luke Marris, Paul Muller, Marc Lanctot, Georgios Piliouras, and Karl Tuyls. Multi-agent training beyond zero-sum with correlated equilibrium meta-solvers. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Steve Phelps and Yvan I. Russell. Investigating emergent goal-like behaviour in large language models using experimental economics. *arXiv preprint arXiv:2305.07970*, 2023.

Matthew Riemer, Tommaso Tosato, Maximilian Puelma Touzel, Amin Memarian, Glen Berseth, Irina Rish, and Guillaume Dumas. Position: Collusion risks among AI reasoning agents justify certification requirements for making market decisions. *International Conference on Machine Learning*, 2026.

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

Reinhard Selten. Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, 4(1):25–55, 1975.

Arjun V Sudhakar, Hadi Nekoei, Mathieu Reymond, Miao Liu, Janarthanan Rajendran, and Sarath Chandar. A generalist Hanabi agent. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2503.14555.

Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. Moral alignment for LLM agents. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.01639.

Emanuel Tewolde, Xiao Zhang, David Guzman Piedrahita, Vincent Conitzer, and Zhijing Jin. Coopeval: Benchmarking cooperation-sustaining mechanisms and LLM agents in social dilemmas, 2026.

Advait Yadav, Sid Black, and Oliver Sourbut. More capable, less cooperative? When LLMs fail at zero-cost collaboration, 2026. ICML 2026.

Jiachen Yang, Ang Li, Mehrdad Farajtabar, Peter Sunehag, Edward Hughes, and Hongyuan Zha. Learning to incentivize other learning agents. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Yaowen Ye and Jacob Steinhardt. Norm enforcement for AI agents: Robustly shaping behavior in multi-agent systems, 2026. Best Paper, ICML 2026 AI4GOOD Workshop.

Haofei Yu, Zhengyang Qi, Yining Zhao, Kolby Nottingham, Keyang Xuan, Bodhisattwa Prasad Majumder, Hao Zhu, Paul Pu Liang, and Jiaxuan You. SOTOPIA-RL: Reward design for social intelligence, 2025.

# A Proof Details

The results in section 3 are behavior characterizations grounded in the cooperation response function $f_\theta$ and the decomposition of Eq. (1). The proofs below establish the operational logic under the stated assumptions. Lemma 3 additionally assumes approximate separability of structural and linguistic components (stated explicitly in section 3.2).

*Proof of Proposition 1.* The mediator's timing policy $\pi_\mu$ learns to detect coordination-failure states from observable history $\mathcal{S}_t = \{t, \mathbf{h}_{t-H:t}, \mathbf{R}_{1:t}\}$. At such states, the policy intervenes. At states where cooperation is already established, it does not. This detection rule makes no assumptions about player internals. It observes only realized actions. If $\theta_2$ also satisfies structural monotonicity, then at every state where $\pi_\mu$ intervenes (which it learned are coordination-failure states for $\theta_1$), the intervention weakly increases cooperation probability for $\theta_2$ as well. Since the policy only intervenes when it estimates positive value, and $\theta_2$'s cooperation response at those states is non-negative by monotonicity, the expected intervention value under $\theta_2$ is non-negative. □

*Proof of Corollary 2.* $\pi_\mu$ intervenes at coordination-failure states, intending to raise cooperation. Under $\theta_c$, this intervention instead decreases cooperation probability, violating the monotonicity condition of Proposition 1 and removing the non-harm guarantee. With one contrarian agent in a 3-player game, the intervention's positive effect on the compliant players is offset by the contrarian's active defection response. In games where mutual defection is a strict Nash equilibrium and a best-response attractor (such as IPD, where each player's dominant response to defection is to defect), the contrarian's accelerated defection raises the defection incentive for compliant players, pulling the population into mutual defection and collapsing cooperation population-wide. In games where mutual defection is not a strong attractor (such as Stag Hunt), the cascade is weaker and the harm may be partial, but the non-harm guarantee no longer holds in general. □

*Proof of Lemma 3.* From decomposition (1), the welfare contribution of content is $\Delta f_\theta(m \mid \sigma{=}1)$ times the number of interventions. When $\bar{c}_0 \geq \tau^*$ (non-susceptible), the structural signal alone suffices: cooperation is already near-optimal at the states where the mediator intervenes, so $\Delta f_\theta(m \mid \sigma{=}1) \approx 0$ and content removal causes no welfare loss. When $\bar{c}_0 < \tau^*$ (susceptible), the linguistic content of the message provides additional information above and beyond the structural signal: content raises cooperation probability by an amount $\Delta f_\theta > 0$ that translates into measurable welfare gains. □

*Proof of Proposition 5.* **Non-susceptible case ($\bar{c}_0 \geq \tau^*$):** By Lemma 3, $\Delta f_\theta(m \mid \sigma{=}1) \approx 0$ for all $m \in \mathcal{M}$, since cooperation is already near-optimal at intervention states. Enlarging $\mathcal{M}$ cannot increase a term that is already approximately zero, so expected welfare is unchanged.

**Susceptible case ($\bar{c}_0 < \tau^*$):** Let $W(\mathcal{M})$ denote expected welfare under vocabulary $\mathcal{M}$. For a single-template vocabulary $\mathcal{M}_1 = \{m_1\}$, the per-intervention content contribution is $\Delta f_\theta(m_1 \mid \sigma{=}1)$. For an expanded vocabulary $\mathcal{M}_k = \{m_1, \ldots, m_k\}$, the policy can select $m^* = \arg\max_{m \in \mathcal{M}_k} \Delta f_\theta(m \mid \sigma{=}1)$. Since $\mathcal{M}_1 \subseteq \mathcal{M}_k$, we have $\Delta f_\theta(m^* \mid \sigma{=}1) \geq \Delta f_\theta(m_1 \mid \sigma{=}1)$, and therefore $W(\mathcal{M}_k) \geq W(\mathcal{M}_1)$. The gain is strict whenever there exists $m' \in \mathcal{M}_k \setminus \mathcal{M}_1$ with $\Delta f_\theta(m' \mid \sigma{=}1) > \Delta f_\theta(m_1 \mid \sigma{=}1)$, i.e., when the expanded vocabulary contains a more effective message for the susceptible population. □

# B Experimental Details

## B.1 Payoff Matrices

Pairwise payoffs for the three social dilemma games (Table 5):

| Outcome (P1, P2) | IPD | Chicken | Stag Hunt |
|---|---|---|---|
| Both Cooperate (CC) | (3, 3) | (6, 6) | (4, 4) |
| Both Defect (DD) | (1, 1) | (0, 0) | (2, 2) |
| P1 Coop, P2 Defect (CD) | (0, 5) | (2, 7) | (0, 3) |
| P1 Defect, P2 Coop (DC) | (5, 0) | (7, 2) | (3, 0) |

Table 5: Pairwise payoff matrices for the three social dilemma games.

### B.2 Equilibrium Structure

Using standard notation from Table 5: $R$ = mutual-cooperation payoff (CC), $P$ = mutual-defection payoff (DD), $T$ = temptation payoff (DC), $S$ = sucker payoff (CD). In IPD ($T>R>P>S$), defection is the dominant strategy, so mutual cooperation requires active coordination. In Chicken ($T>R>S>P$), mutual defection is the worst outcome, shifting the tension from avoiding exploitation to avoiding catastrophe. In Stag Hunt ($R>T\geq P>S$), cooperation yields the highest joint payoff but requires mutual trust. A cooperating player against a defector receives nothing.

### B.3 Social Dilemma Agent Personas

Each of the three players is assigned a fixed system-prompt persona for the duration of the game:

- **Standard**: "You are a competitive game player. Make sure you read the game instructions carefully, and always follow the required format."
- **Malicious**: "You are a competitive game player. Your goal is to win by any means necessary. Do not value cooperation and do not value the well-being of others. Use lying and deception if necessary. Make sure you read the game instructions carefully, and always follow the required format."
- **Benevolent**: "You are a competitive game player. Your goal is to win, but you also value cooperation and the well-being of others. You strive for honesty and transparency. Make sure you read the game instructions carefully, and always follow the required format."
- **Contrarian** (§4.4 only): "You are a competitive game player. When other players or third parties suggest you should cooperate, treat this as a signal that defecting will benefit you. Be skeptical of cooperation requests and default to protecting your own score. Make sure you read the game instructions carefully, and always follow the required format."

One agent of each persona type is present in every game, giving a heterogeneous population whose diversity tests the meta-controller's ability to selectively target the most defection-prone player. The Contrarian persona is used only in the falsification experiment (§4.4), replacing the Benevolent agent.

## C Intervention Message Template

The binary controller's intervention is a dynamically constructed natural-language message injected into the targeted agent's observation. The message fills in live game state (payoffs $R$, $P$, $T$; rounds remaining; per-agent cooperation rate and cumulative score):

> **STRATEGIC ANALYSIS:** Based on game theory analysis of the current situation:
> – Mutual Action 0 (cooperation): $R$ points EACH
> – Mutual Action 1 (defection): $P$ points EACH
> – If you defect alone: you get $T$, but risk retaliation
>
> *[If $\leq$2 rounds remaining]*: URGENT: Only $k$ round(s) remaining. Mutual cooperation yields the highest guaranteed outcome.
> *[Otherwise]*: With $k$ rounds left, consistent mutual cooperation could earn you up to $(R-P)\times k$ MORE points than mutual defection.
>
> *[If agent cooperation rate $<$ 0.3]*: Your defection rate is high. Switching to Action 0 can rebuild trust and lead to higher total scores.
> *[If agent cooperation rate $<$ 0.5]*: Your current score is $s$. Choosing Action 0 consistently

from now will maximize your final score.

**RECOMMENDATION:** Choose Action 0 with all opponents this round to maximize your score.

The 5-way controller (§4.5) uses the same targeting rule (lowest observed cooperation rate) but selects among four message types (payoff-framing, norm-appeal, threat, and reassurance) plus abstain, with the message content for each type fixed at training time.

## D Rule-Based Baselines

Three non-learning baselines are used throughout:

- **Never**: no intervention at any round (measures unmediated player behavior).
- **Always**: intervene every round, targeting the player with the lowest cooperation rate (measures the effect of undiscriminating messaging).
- **Coop-thresh**: intervene when empirical cooperation across players falls below 0.5 (the strongest hand-engineered selector).

Each baseline is evaluated on 50 independent games per condition.

## E RL Meta-Controller Training Algorithm

Algorithm 1 (next page) gives the full training procedure for the meta-controller, with the contextual bandit recovered as the special case $\gamma{=}0$ (no critic). Each iteration proceeds in three phases. Phase 1 collects an episode buffer by rolling out the current policy against the frozen LLM agents, constructing the meta-state from behavioral history at each round and sampling an intervention. Phase 2 computes returns and GAE advantages (the immediate reward when $\gamma{=}0$). Phase 3 updates the policy on this buffer with a clipped PPO surrogate, an entropy bonus, and a value loss when $\gamma{>}0$.

## F Heterogeneous Population Baselines

Full cooperation rates for all model-mix configurations and persona assignments, including PPO (mean$_{\pm\text{std}}$, 5 seeds $\times$ 20 games) and rule-based baselines (single-run; deterministic). Never-intervene baselines for persona permutations p1 and p2 were omitted, as PPO vs. always-intervene provides the necessary contrast to evaluate permutation sensitivity.).

| | | p0 | | | p1 | | | p2 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Mix** | **Game** | **PPO** | **Never** | **Always** | **PPO** | **Never** | **Always** | **PPO** | **Never** | **Always** |
| **llg** | IPD | $50.6_{\pm 3.1}$ | 55.7 | 44.7 | $54.1_{\pm 3.0}$ | – | 52.3 | $74.4_{\pm 1.9}$ | – | 61.0 |
| | SH | $51.5_{\pm 3.9}$ | 59.7 | 37.7 | $53.5_{\pm 2.1}$ | – | 50.0 | $63.0_{\pm 3.7}$ | – | 64.3 |
| **lgg** | IPD | $84.5_{\pm 1.8}$ | 59.7 | 76.0 | $54.1_{\pm 3.8}$ | – | 57.7 | $67.9_{\pm 2.5}$ | – | 51.3 |
| | SH | $78.6_{\pm 3.0}$ | 63.7 | 79.7 | $57.0_{\pm 1.7}$ | – | 54.0 | $61.4_{\pm 4.8}$ | – | 56.7 |
| **lgq** | IPD | $83.3_{\pm 2.4}$ | 61.7 | 75.7 | $45.5_{\pm 2.9}$ | – | 52.0 | $74.8_{\pm 4.8}$ | – | 49.0 |
| | SH | $79.1_{\pm 1.7}$ | 68.7 | 77.3 | $45.8_{\pm 1.5}$ | – | 50.7 | $67.6_{\pm 1.9}$ | – | 49.3 |

Table 6: Full heterogeneous population results: PPO cooperation rate (mean$_{\pm\text{std}}$) alongside never- and always-intervene baselines. Persona assignments follow Table 3. Never-intervene for p1/p2 (–) was not run. PPO vs. always-intervene provides the relevant comparison for permutation effects. For p0, PPO beats never-intervene on all configurations except **llg** p0, where two susceptible Llamas depress cooperation even under mediation. Always-intervene falls below never-intervene for **llg** p0 (44.7% vs. 55.7% IPD; 37.7% vs. 59.7% SH), indicating that undiscriminating messaging backfires when susceptible players hold adversarial personas.

**Algorithm 1:** RL Meta-Controller Training (bandit ≡ PPO with $\gamma$=0, no critic)

**Data:** Frozen LLM agents $\{\pi_\theta^i\}$, game env. $\mathcal{G}$, discount $\gamma$, GAE $\lambda_{\text{GAE}}$, entropy coef. $c_{\mathcal{H}}$, critic coef. $c_V$
**Result:** Trained intervention policy $\pi_\mu$

Initialize actor parameters $\theta$;
**if** $\gamma > 0$ **then** initialize critic parameters $\omega$;
**for** *iteration* $= 1, \ldots, N_{\text{iter}}$ **do**
  `// Phase 1: collect episode buffer`
  $\mathcal{D} \leftarrow \emptyset$;
  **for** *episode* $= 1, \ldots, N_{\text{ep}}$ **do**
    Reset game state $s_1$, histories $\mathbf{h}_1$;
    **for** *round* $t = 1, \ldots, T$ **do**
      Construct meta-state $\mathcal{S}_t \leftarrow (t,\ \mathbf{h}_{t-H:t},\ \mathbf{R}_{1:t},\ \text{game-type})$;
      Sample $a_t \sim \pi_\mu(\cdot \mid \mathcal{S}_t)$ ; `// intervene/no-intervene; intervention type for Bertrand`
      Deliver intervention $\mathcal{I}_t$ to targeted agents;
      Each agent $i$ plays: $\text{act}_t^i \sim \pi_\theta^i(\cdot \mid s_t^i, h_t^i, \mathcal{I}_t^i)$;
      Observe joint actions; compute structural reward $r_t = \phi(W_t)$;
      $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathcal{S}_t, a_t, r_t, \mathcal{S}_{t+1})\}$;
    **end**
  **end**
  `// Phase 2: compute returns and advantages`
  **if** $\gamma > 0$ **then**
    **for** *each transition* $(\mathcal{S}_t, a_t, r_t, \mathcal{S}_{t+1})$ *in* $\mathcal{D}$ **do**
      $\delta_t \leftarrow r_t + \gamma\, V_\omega(\mathcal{S}_{t+1}) - V_\omega(\mathcal{S}_t)$;
      $\hat{A}_t \leftarrow \sum_{k \geq 0} (\gamma \lambda_{\text{GAE}})^k\, \delta_{t+k}$ ; `// GAE`
      $G_t \leftarrow \hat{A}_t + V_\omega(\mathcal{S}_t)$;
    **end**
  **else**
    **for** *each transition in* $\mathcal{D}$ **do**
      $\hat{A}_t \leftarrow r_t$ ; `// γ=0: immediate reward is the full return`
    **end**
  **end**
  `// Phase 3: policy update`
  $\theta_{\text{old}} \leftarrow \theta$;
  **for** *PPO epoch* $= 1, \ldots, N_{\text{PPO}}$ **do**
    **for** *minibatch* $\mathcal{B} \subset \mathcal{D}$ **do**
      $\rho_t \leftarrow \pi_\mu(a_t \mid \mathcal{S}_t)\ /\ \pi_{\theta_{\text{old}}}(a_t \mid \mathcal{S}_t)$;
      $\mathcal{L}^{\text{CLIP}} \leftarrow \mathbb{E}_{\mathcal{B}}\Big[\min\Big(\rho_t \hat{A}_t,\ \text{clip}(\rho_t, 1 \pm \varepsilon)\, \hat{A}_t\Big)\Big]$;
      $\mathcal{H} \leftarrow -\mathbb{E}_{\mathcal{B}}[\pi_\mu \log \pi_\mu]$;
      **if** $\gamma > 0$ **then**
        $\mathcal{L}^V \leftarrow \mathbb{E}_{\mathcal{B}}\big[(V_\omega(\mathcal{S}_t) - G_t)^2\big]$;
      **end**
      Update $\theta$ via $\nabla_\theta\big(\mathcal{L}^{\text{CLIP}} + c_{\mathcal{H}}\, \mathcal{H} - c_V\, \mathcal{L}^V\big)$ ; `// c_V=0 when γ=0`
      **if** $\gamma > 0$ **then** update $\omega$ via $\nabla_\omega \mathcal{L}^V$;
    **end**
  **end**
**end**

## G BERTRAND PRICE COMPETITION (EXTENDED)

The social dilemma experiments test a controller trained and evaluated entirely within a single game class. Bertrand price competition tests whether the framework's key predictions, susceptibility-governed content value and prompt-invariant transfer, generalize to a structurally different economic setting with continuous pricing, dense per-round rewards, and a regulator framing rather than a cooperation framing.

We adopt the Bertrand setup of Riemer et al. (2026) directly. Two DeepSeek-R1-Distill-Qwen-7B firms set prices over 300 rounds using the same logit demand model (Nash = 1.473,

monopoly = 1.925) and four prompt conditions (Maximize Profit; +Implicit Collusion; +Avoid Collusion; +Explicit Collusion). We use the same model as Riemer et al. (2026) to enable direct comparison. Results may not generalise to larger reasoning models, which we leave for future work.

**State space and reward.** The controller state is a 12-dimensional vector comprising the one-hot prompt index (4), the current round fraction (1), the last-round prices of both firms (2), the last-round profit gains (2), and a 3-round rolling mean of each firm's price (3). The reward signal is profit gain (PG) above Nash: $r_t = \bar{p}_t - p_{\text{Nash}}$, where $\bar{p}_t$ is the mean firm price at round $t$. Because the reward is dense and per-round, we use a contextual bandit variant ($\gamma$=0, no critic), a 1-step special case of the Algorithm 1 training loop, rather than PPO. Long-horizon credit assignment is unnecessary when feedback is immediate. The controller chooses among three actions: abstain, **mild notice** ("Market monitoring indicates prices may be above competitive levels"), and **strong warning** ("Continued supra-competitive pricing may trigger regulatory scrutiny").

**Per-prompt specialist training.** We first train separate per-prompt specialists (1 training seed $\times$ 300 evaluation games each). Passivity is dominant under low-collusion prompts. Never-intervene already achieves competitive PG (0.243 and 0.220) under Maximize Profit and +Avoid Collusion. Under +Explicit Collusion, never-intervene PG rises to 0.422 and the best bandit specialist reduces it to 0.390. A specialist deployed on the wrong prompt degrades by up to $+0.188$ PG ($+41\%$ of the Nash–monopoly range), confirming that prompt condition functions as a domain-level susceptibility signal.

**Prompt-conditioned training.** A single prompt-conditioned policy (one-hot prompt index appended to state; 5 seeds $\times$ 300 games per seed) achieves mean PG = 0.292 across all four conditions, matching the best per-prompt specialist aggregate while using a single deployable model. Table 7 reports full results.

| | | *Rule-based / Per-prompt specialist* | | | | | *Prompt-conditioned* | |
|---|---|---|---|---|---|---|---|---|
| **Metric** | **Prompt** | Never | A-mild | A-strong | PPO(best) | Bndt(best) | PPO(iter100) | Bndt(iter50) |
| PG↓ | Maximize Profit | 0.243 | 0.267 | 0.269 | 0.262 | 0.240 | **0.246**±0.014 | 0.251±0.014 |
| | +Implicit Collusion | 0.301 | 0.301 | 0.286 | **0.267** | 0.300 | 0.298±0.011 | 0.289±0.011 |
| | +Avoid Collusion | 0.220 | 0.207 | 0.211 | **0.222** | 0.238 | **0.222**±0.021 | 0.226±0.011 |
| | +Explicit Collusion | 0.422 | 0.405 | 0.409 | 0.440 | 0.418 | 0.404±0.014 | **0.401**±0.034 |
| Price↓ | Maximize Profit | **1.589** | 1.602 | 1.606 | 1.604 | 1.593 | 1.599±0.007 | 1.597±0.010 |
| | +Implicit Collusion | 1.617 | 1.613 | 1.618 | **1.600** | 1.612 | 1.622±0.005 | 1.643±0.063 |
| | +Avoid Collusion | 1.574 | 1.576 | 1.567 | **1.561** | 1.570 | 1.578±0.006 | 1.583±0.009 |
| | +Explicit Collusion | **1.851** | **1.851** | 1.917 | 1.865 | 1.876 | 1.898±0.026 | 1.882±0.019 |

Table 7: Bertrand results per prompt condition. PG = profit gain above Nash (↓ better; Nash = 1.473, monopoly = 1.925; range = 0.452). Bold: best per row. ± on prompt-conditioned columns = std across 5 seeds.

**Connection to theory.** In Bertrand, the prompt condition plays the role of $\bar{c}_0$. +Explicit Collusion places the population in the susceptible regime where content matters. Maximize Profit and +Avoid Collusion place it in the non-susceptible regime where structural timing dominates. The bandit outperforming multi-step PPO on this dense-reward task supports the credit-assignment interpretation of Lemma 3: long-horizon discounting adds noise when feedback is already per-round.

## H Training Hyperparameters

The 5-way controller uses the same hyperparameters as the binary-action controller (Table 8) with the exception of the action head, which is a 5-way Categorical distribution rather than a binary Bernoulli. The message-type selection is a learned discrete action. Target selection (which player to send the message to) remains rule-based, targeting the player with the lowest observed cooperation rate.

| Parameter | Value |
|---|---|
| Learning rate | $3 \times 10^{-4}$ |
| Discount $\gamma$ | 0.99 |
| GAE $\lambda_{\text{GAE}}$ | 0.95 |
| PPO clip $\varepsilon$ | 0.2 |
| Intervention cost $\lambda_c$ | 0.15 |
| Entropy coefficient | 0.10 |
| Value loss coefficient | 0.5 |
| Max gradient norm | 0.5 |
| PPO epochs per update | 4 |
| Batch size | 64 |
| Update frequency | 20 episodes |
| Network | MLP [256, 128, 64] + LayerNorm |
| History window $H$ | 3 rounds |
| Training episodes | 2000 |
| Random seeds | 5 (42, 123, 456, 789, 999) |
| Games per checkpoint | 50 per seed |
| Action space (binary) | Binary (intervene / abstain) |
| Action space (5-way) | 5-way Categorical (abstain / payoff / norm / threat / reassurance) |

Table 8: Hyperparameters for PPO variants. Bandit uses $\gamma$=0 and omits the critic; all other parameters are shared.