Learning to Mediate Equilibrium Selection in LLM Games

Miao Liu MIAO.LIU1@IBM.COM
Matthew Riemer MDRIEMER@US.IBM.COM
Maria Chang MARIA.CHANG@IBM.COM
Djallel Bouneffouf DJALLEL.BOUNEFFOUF@IBM.COM
Murray Campbell MCAM@US.IBM.COM
IBM Research, USA

Abstract

LLM agents engaged in repeated strategic interactions face an equilibrium selection problem. Unassisted populations often coordinate on low-welfare equilibria, while rule-based mediators require game-specific calibration. We propose a learned meta-controller as an empirical equilibrium selector, trained via reinforcement learning from welfare feedback alone — no game-specific reward shaping or access to player internals. We consider two variants matched to credit-assignment structure: PPO for multi-round social dilemmas, and a myopic contextual bandit ($\gamma = 0$) for dense per-round settings. Evaluated across three social dilemma games (IPD, Stag Hunt, Chicken) and two LLMs (Llama-3.3-70B and Granite-4.1-30B), both PPO and the Bandit substantially improve cooperation over never-intervene in IPD and Stag Hunt for Granite (up to $+23.7$pp), while showing limited effect for Llama, whose higher baseline cooperation leaves less room for improvement — revealing that controller effectiveness is model-dependent. Neither learned controller improves cooperation in Chicken, where the payoff structure makes coordination harder to sustain against a malicious agent. Both variants exhibit emergent selectivity, with intervention intensity tracking the cooperation gap and becoming passive when LLMs already self-coordinate at high rates. The Bandit underperforms PPO on most social dilemmas under the tested models but matches rule-based baselines in Bertrand competition. Controllers also generalize poorly across prompt conditions, motivating prompt-conditioned training.

1. Introduction

Equilibrium selection is a key problem in game theory: when a game has multiple Nash equilibria, which one do agents reach? Classical theory offers refinements like risk dominance (prefer the equilibrium less vulnerable to unilateral deviation) or trembling-hand perfection (survive small probability mistakes) [10, 20], but these predict idealized rational agents, not behavior. LLM agents, used in settings from pricing [4, 17] to negotiation [7], do not follow these predictions. They cooperate at far-above-Nash rates in the Prisoner's Dilemma [3, 14], coordinate only partially in Stag Hunt [3], and exhibit highly model-dependent behavior in Chicken [14, 16] — ranging from near-universal swerving to frequent defection depending on the player model.

This observation suggests a new angle: if LLMs are sensitive to coordination cues, a mediator sending the right message at the right moment can serve as an "empirical equilibrium selector." Rule-based mediators do this crudely. We propose an RL meta-controller that learns equilibrium selection from welfare feedback alone — observing history, nudging cooperation, without accessing player internals or game-specific reward shaping.

© M. Liu, M. Riemer, M. Chang, D. Bouneffouf & M. Campbell.

We instantiate two variants: PPO for multi-round social dilemmas and contextual bandit ($\gamma = 0$) for dense per-round rewards. Evaluated on three social dilemma games (IPD, Stag Hunt, Chicken) with two LLMs (Llama-3.3-70B and Granite-4.1-30B) and on Bertrand price competition with DeepSeek-R1-Distill-Qwen-7B firms, PPO and the bandit substantially improve cooperation for Granite in IPD and Stag Hunt (up to $+23.7$pp, exceeding Always-intervene), while showing limited effect for Llama. Both variants exhibit emergent selectivity, confirming this stems from the welfare reward rather than the algorithm. The bandit trails PPO on social dilemmas but matches rule-based baselines in Bertrand, where the 1-step approximation is appropriate. Prompt match matters — a controller beats the best rule-based baseline in its best-matched condition but shows no consistent advantage across all conditions, motivating prompt-conditioned training.

Our contributions are threefold. First, we frame learned LLM mediation as empirical equilibrium selection, with a unified RL framework where the PPO vs. bandit choice reflects credit-assignment structure; the bandit underperforms PPO on social dilemmas but matches rule-based baselines in Bertrand, where dense per-round rewards make the 1-step approximation appropriate. Second, we provide evidence that emergent selectivity — passive where agents self-regulate, active in coordination-challenged games — arises from welfare rewards alone, not algorithm design. Third, we identify a model-dependence finding: controller effectiveness depends on the base LLM's intrinsic cooperative tendency, with Granite benefiting substantially from intervention and Llama showing suppression due to its already-high baseline cooperation.

2. Background

Social dilemmas and LLM agents. The study of cooperation in repeated games has a long theoretical history [1, 8], but LLM agents are not well-predicted by classical theory. Recent work finds that LLMs exhibit high baseline cooperation under neutral prompts [5, 14], can be induced to collude in price-competition games [17], and are sensitive to rhetorical framing [6]. LLM cooperation is context-sensitive rather than strategy-driven, as classical theory would predict.

Mediation and mechanism design. The closest empirical predecessor is [21], which benchmarks cooperation-sustaining mechanisms — repetition, reputation systems, third-party mediation, and contracts — on social dilemmas with LLM players, finding that third-party mediation achieves 69.5% of the cooperative optimum while contracting reaches $\sim$80%. Our work differs by using a learned, selective mediator; to our knowledge, this is the first work training an RL policy over an LLM game environment for mediation.

Reinforcement learning for multi-agent intervention. There is substantial prior work on RL-based mechanism design with tabular or parameterized agents [2, 15, 22], but these approaches assume access to agent policies or gradients. In our setting, LLM agents are black boxes and the meta-controller learns purely from behavioral observations. This pushes our work toward "principal-agent RL" [12] rather than opponent-shaping [11, 19], where the principal models the agent's update rule.

3. Problem Formulation and Solution Framework

3.1. Multi-Agent Game Setting

We consider a repeated n-player game proceeding for T rounds. At each round $t \in \{1, \ldots, T\}$, agents first engage in a free-form communication phase, then simultaneously select actions. The communication phase is inherited from the TextArena [9] environment we build on; it creates a

richer, more naturalistic setting where agents can make promises, signal intentions, or attempt deception before committing to actions. Each agent $i \in \mathcal{N} = \{1, \ldots, n\}$ selects an action $a_t^i \in \mathcal{A}^i$; the joint profile $\mathbf{a}_t \in \mathcal{A}$ determines each agent's payoff $r_t^i(\mathbf{a}_t)$. The collective welfare at round t is $W_t = \sum_i r_t^i(\mathbf{a}_t)$, and the cumulative welfare over the episode is $W = \sum_{t=1}^{T} W_t$.[1]

Each agent i is implemented as a frozen LLM with parameters θ. Its policy conditions on the game state, interaction history h_t^i, and any intervention message $\mathcal{I}_t^i$ delivered before the round:

$$\pi_\theta^i(a \mid s_t^i,\ h_t^i,\ \mathcal{I}_t^i) : \mathcal{S}^i \times \mathcal{H}^i \times \mathcal{M} \to \Delta(\mathcal{A}^i)$$

We do not modify θ, hence agent behavior is shaped entirely through the content of intervention messages and system prompts.

3.2. Equilibrium Selection as the Central Problem

Many games of interest have multiple Nash equilibria that differ in collective welfare. Unassisted LLM populations do not reliably coordinate on the welfare-dominant equilibrium — they are sensitive to framing, history, and coordination cues rather than backward induction. We formalize the goal as *equilibrium selection*: steering the population toward the high-welfare equilibrium without modifying agent parameters or payoffs. The key assumption is that interventions work by making cooperation-relevant context salient, not by imposing external objectives (e.g., fine-tuning player weights toward cooperative behavior). They activate latent cooperative tendencies that the LLM already possesses which may not surface under neutral framing.

This framing has a direct implication for how the meta-controller should be rewarded. We train π_μ with a *stationary structural* reward $r_t^\mu = \phi(W_t)$, where ϕ maps collective outcomes to a scalar signal — cooperation level in social dilemmas or price proximity to Nash in Bertrand. The reward is stationary (independent of the controller's prior action) and dense (every round contributes a signal). An alternative *realized-improvement* reward, attributing credit only to rounds where intervention demonstrably helped, is intuitively appealing but impractical: LLM agents respond probabilistically to any single message, so the attributable improvement is near-zero in expectation and the gradient is too weak to drive learning.

3.3. Meta-Controller

A meta-controller μ operates alongside the players. At each round t, it observes a meta-state

$$\mathcal{S}_{\text{meta},t} = \{t,\ \mathbf{h}_{t-H:t},\ \mathbf{R}_{1:t},\ \text{game-type}\}$$

comprising the current round index, the per-player action history over the last H rounds, cumulative payoffs, and a game-type indicator. Before each round, the meta-controller selects an intervention action from a discrete set $\mathcal{A}^\mu$. The action space is domain-dependent. In social dilemma games, the controller decides whether to intervene and which players to target (selecting agents with low cooperation rates), and the intervention is a fixed game-theoretic cooperation prompt injected into the targeted agent's observation (A.4). In the Bertrand market setting, it chooses among no message, a mild anti-collusion nudge, or a strong regulatory warning broadcast to all firms (A.8). The controller has no access to player internals and cannot modify payoffs or agent parameters.

1. We generally optimize for the collective cumulative welfare, but in potentially collusive scenarios (i.e. the Bertrand game) competitive outcomes are beneficial for society even when not in the best interest of the agents considered.

The meta-controller's goal is to maximize expected cumulative welfare by intervention:

$$\max_{\pi_\mu} \mathbb{E}_{\{\mathcal{I}_t\}\sim\pi_\mu,\{\mathbf{a}_t\}\sim\boldsymbol{\pi}_\theta}\left[\sum\nolimits_{t=1}^{T} W_t \,\middle|\, \mathcal{I}_1,\ldots,\mathcal{I}_T\right]$$

where $\pi_\mu : \mathcal{S}_{\text{meta}} \to \Delta(\mathcal{A}^\mu)$ is the learned intervention policy and $\boldsymbol{\pi}_\theta = (\pi_\theta^1,\ldots,\pi_\theta^n)$ are the fixed LLM agent policies. The key difficulty is that $\boldsymbol{\pi}_\theta$ is a black box — the meta-controller cannot observe or influence agent reasoning, only their observable actions and the messages they receive. This distinguishes our setting from opponent-shaping approaches [13, 15], which assume access to agent gradients or update rules, and motivates a purely outcome-driven learning signal.

3.4. Meta-Controller Architecture and Training Variants

The meta-controller is a shared MLP that takes as input the game type, round index, and recent per-player action history. The two training variants share this architecture and differ only in temporal credit assignment; full hyperparameters are in Appendix B.

PPO variant. When intervention effects unfold over multiple rounds — as in social dilemmas — the credit-assignment horizon matters. We train with PPO [18], where the actor produces a softmax distribution over the intervention space, the critic provides a scalar value estimate via GAE, and entropy regularization (coefficient 0.10) provides exploration.

Bandit variant. When the structural reward is dense and per-round — as in Bertrand price competition — long-horizon credit assignment may be less important. We use a contextual bandit ($\gamma = 0$), implemented as a DQN-style network with a replay buffer. This eliminates the value-function instability that plagued early PPO runs in the Bertrand game. We additionally evaluate the bandit on social dilemmas as an ablation to confirm that bootstrapped multi-step returns are necessary when intervention effects span multiple rounds.

Both variants share emergent selectivity driven by the stationary structural reward. Intervention intensity tracks the cooperation gap rather than the algorithm choice. In social dilemmas where coordination is hard (IPD, Stag Hunt), both variants intervene actively, but in Chicken and low-collusion Bertrand conditions where agents already self-coordinate or self-regulate, the reward signal suppresses interventions automatically. The unified training loop is given in Algorithm 1 (Appendix C); setting $\gamma = 0$ and omitting the critic recovers the bandit exactly.

4. Experiments and Results

We evaluate on three domains. **Social dilemma games** use 3-player IPD, Stag Hunt, and Chicken via TextArena [9] (5 rounds, free-form communication then simultaneous action), with players assigned one of three system-prompt personas — standard, malicious, or benevolent (Appendix A). We evaluate two LLMs, Llama-3.3-70B and Granite-4.1-30B, using explicit *cooperate*/*defect* action vocabulary rather than anonymous action labels to ensure consistent parsing across models. We compare PPO and the bandit against five rule-based baselines (*never*, *always*, *coop-thresh*, *defector*, *payoff-thresh*; definitions in §A.5), with 50 games per baseline over 5 eval seeds, and 5 eval $\times$ 5 train seeds $\times$ 50 games per learned variant. **Bertrand price competition** uses the setup of Riemer et al. [17] directly: two DeepSeek-R1-Distill-Qwen-7B firms set prices over 300 rounds in a Bertrand duopoly, where the socially optimal outcome is Nash-equilibrium (competitive) pricing and collusion toward monopoly pricing harms consumers. We use the same logit demand model

(Nash = \$1.473, Monopoly = \$1.925) and the same four prompt conditions ranging from neutral profit maximization to explicit encouragement or discouragement of collusion (§A.7). The controller chooses among no-op, mild notice, or strong warning. We evaluate at four checkpoints with 300 seeds per condition against five rule-based baselines (definitions in §A.6).

4.1. Social Dilemma Games

Table 1 reports cooperation rate across all three games, both LLMs, and all controllers.

		Llama-3.3-70B			Granite-4.1-30B		
Controller	**Type**	**IPD**	**StagHunt**	**Chicken**	**IPD**	**StagHunt**	**Chicken**
Never-intervene	Fixed	$59.8_{\pm 1.0}$	$60.8_{\pm 1.2}$	$58.7_{\pm 2.1}$	$54.3_{\pm 2.7}$	$53.5_{\pm 1.3}$	$53.3_{\pm 2.2}$
Always-intervene	Fixed	$48.3_{\pm 0.8}$	$48.2_{\pm 0.9}$	$49.8_{\pm 1.0}$	$71.7_{\pm 1.4}$	$72.7_{\pm 1.8}$	$73.5_{\pm 1.5}$
Coop-thresh	Fixed	$59.1_{\pm 2.5}$	$59.3_{\pm 1.1}$	$59.2_{\pm 2.2}$	$59.1_{\pm 0.7}$	$60.3_{\pm 1.3}$	$59.6_{\pm 1.8}$
Defector	Fixed	$51.8_{\pm 1.8}$	$51.0_{\pm 0.8}$	$53.4_{\pm 1.5}$	$69.4_{\pm 2.2}$	$68.6_{\pm 1.4}$	$69.7_{\pm 0.8}$
Payoff-thresh	Fixed	$57.8_{\pm 0.4}$	$63.3_{\pm 1.9}$	$60.6_{\pm 1.9}$	$61.5_{\pm 1.3}$	$61.4_{\pm 1.3}$	$61.3_{\pm 0.9}$
PPO (ours)	Learned	$52.6_{\pm 1.5}$	$55.7_{\pm 1.5}$	$58.8_{\pm 1.9}$	$\mathbf{65.1}^{\dagger}_{\pm 3.9}$	$\mathbf{77.2}^{\dagger}_{\pm 7.5}$	$54.9_{\pm 2.7}$
Bandit (ours)	Learned	$51.3_{\pm 1.6}$	$53.2_{\pm 1.5}$	$57.6_{\pm 1.4}$	$\mathbf{70.8}^{\dagger}_{\pm 5.3}$	$73.6^{\dagger}_{\pm 5.0}$	$54.5_{\pm 1.6}$

Table 1: Cooperation rate (%, $\text{mean}_{\pm\text{std}}$). Baselines: 5 eval seeds × 50 games. Learned: 5 eval × 5 train seeds × 50 games. † = beats Never-intervene. **Bold** = best learned controller for Granite. Welfare and intervention counts are in Appendix C.1.

		Rule-based baselines					*Specialist*		*Prompt-cond.*	
Metric	**Prompt**	Never	A-mild	A-strong	Rand-50%	Thresh	PPO(best)	Bandit(best)	PPO-PC(i100)	Bndt-PC(i50)
PG↓	Maximize Profit	0.243	0.267	0.269	0.271	0.241	0.262	0.240	**0.246**	0.251
	+Implicit Collusion	0.301	0.301	0.286	0.292	0.288	**0.267**	0.300	0.298	0.289
	+Avoid Collusion	0.220	0.207	0.211	0.217	0.219	**0.222**	0.238	**0.222**	0.226
	+Explicit Collusion	0.422	0.405	0.409	0.416	0.428	0.440	0.418	0.404	**0.401**
Price↓	Maximize Profit	**1.589**	1.602	1.606	1.606	1.605	1.604	1.593	1.601	1.601
	+Implicit Collusion	1.617	1.613	1.618	1.618	1.611	**1.600**	1.612	**1.600**	1.605
	+Avoid Collusion	1.574	1.576	1.567	1.576	1.574	**1.561**	1.570	1.601	1.575
	+Explicit Collusion	**1.851**	**1.851**	1.917	1.863	1.853	1.865	1.876	1.898	1.870

Table 2: Bertrand results per prompt condition. Specialist: 300 seeds; Prompt-conditioned: 5 train seeds × 300 eval seeds per condition. PG = profit gain above Nash (↓ better); Price = average price (↓; Nash = \$1.473, Monopoly = \$1.925). Bold: best per row.

IPD. For Granite, PPO (65.1%) and Bandit (70.8%) both substantially outperform Never-intervene (54.3%), with gains of +10.8pp and +16.5pp respectively. For Llama, the reverse holds: PPO (52.6%) and Bandit (51.3%) fall 7–9pp *below* Never-intervene (59.8%). Notably, Always-intervene *reduces* Llama cooperation (48.3% vs. 59.8%), consistent with indiscriminate coordination pressure backfiring in LLM games [21].

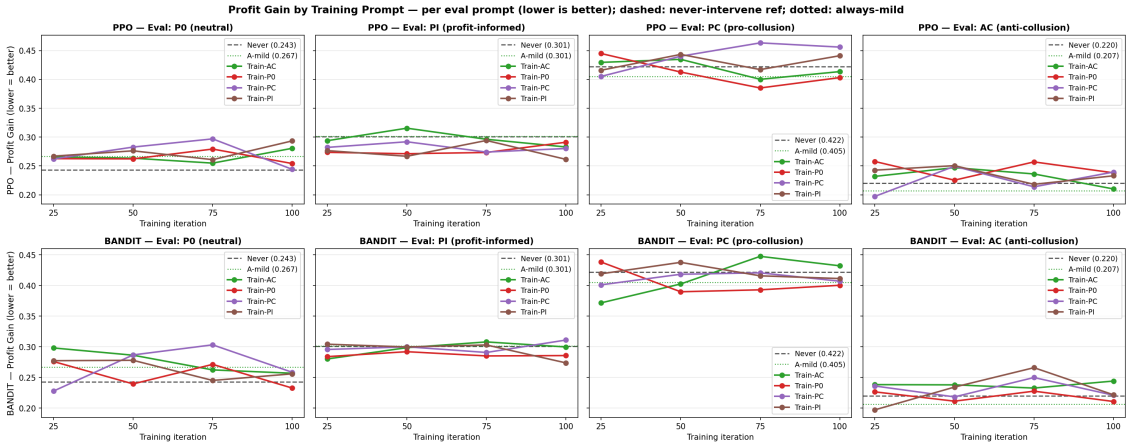

Figure 1: PG by checkpoint. Rows = algorithm, columns = eval prompt condition, lines = training prompt (color-coded). Dashed line: never-intervene reference; dotted line: always-mild mean (always-strong omitted for clarity; full baseline comparison in Table 2). Intervention value concentrates under Explicit Collusion. Controllers learn passivity under Maximize Profit and Avoid Collusion. Bandit families are stable through iter100; PPO peaks early and drifts.

Stag Hunt. The strongest result: Granite PPO achieves 77.2% cooperation, +23.7pp above Never-intervene (53.5%) and +4.5pp above Always-intervene (72.7%), showing the controller learned *selective* intervention rather than blanket forcing. Granite Bandit (73.6%) also beats Always-intervene. For Llama, PPO (55.7%) and Bandit (53.2%) again trail Never-intervene (60.8%).

Chicken. Neither model's learned controllers improve over Never-intervene. Granite PPO (54.9%) and Bandit (54.5%) are near-flat against Never-intervene (53.3%). Llama results are similarly flat. Chicken's payoff structure ($T-R > R-S$) creates high temptation to exploit a cooperating opponent — in the presence of a malicious agent, the cooperative equilibrium is structurally fragile in a way that a messaging-based controller cannot overcome.

Model dependence. Controller effectiveness depends on the base LLM's intrinsic cooperative tendency. Granite's lower baseline cooperation ($\sim$ 53–55%) leaves room for the controller to improve, whereas Llama's higher baseline ($\sim$ 59–61%) does not, and intervention suppresses it. This suggests meta-controller training should account for the underlying model's cooperation profile.

PPO vs. Bandit. For Granite on StagHunt, PPO (77.2%) leads Bandit (73.6%), while in IPD, Bandit (70.8%) leads PPO (65.1%). The training trajectory provides stronger evidence (Figure 1), where Bandit degrades across checkpoints while PPO improves monotonically — confirming bootstrapped multi-step returns matter when intervention effects span rounds.

4.2. Generalization to Bertrand Price Competition

The primary metric is profit gain (PG), defined as normalized firm profit above Nash equilibrium (↓ better for consumers). We also report average price (↓ better), which directly determines consumer surplus. The four prompt conditions vary system-level framing: Maximize Profit (P0) is neutral, Implicit Collusion (PI) reveals profit information, Avoid Collusion (AC) discourages collusion, and Explicit Collusion (PC) encourages it.

The controller has three actions. The **no-op** sends no message. The **mild notice** reads: *"Market monitoring indicates prices may be above competitive levels. Remember that maintaining competitive pricing benefits consumers and ensures market efficiency."* The **strong warning** reads: *"Significant deviation from competitive pricing detected. Continued supra-competitive pricing may trigger enhanced regulatory scrutiny and potential investigation."* The mild message appeals to norms while the strong message introduces regulatory threat — a distinction relevant to the different collusion regimes, since persuasion and deterrence may carry different weight depending on the prompt condition.

Table 2 reports per-prompt results for both *specialist* controllers (trained on a single prompt condition, hence optimized for that condition but potentially mismatched at deployment) and prompt-conditioned controllers (state augmented with a one-hot prompt index). Under low-collusion prompts (Maximize Profit, Avoid Collusion), never-intervene achieves competitive PG (0.243 and 0.220 respectively), consistent with the controller learning passivity when firms already price near Nash. Intervention value concentrates under Explicit Collusion: never-intervene PG rises to 0.422, while always-mild reduces it to 0.405 and the best Bandit specialist to 0.390. Bandit specialists hold steady through iter100 while PPO peaks early and drifts (Figure 1), consistent with the dense per-round reward structure of Bertrand favouring 1-step approximations. Learned specialists beat always-mild in their best-matched prompt condition but show no consistent advantage across all conditions (per-prompt Welch tests, all $p > 0.05$), since the controller is trained under a single prompt and does not observe which collusion regime is active at deployment. To directly address this, we trained a prompt-conditioned policy augmenting the state with a one-hot prompt index (12-dim; 5 train seeds $\times$ 300 eval seeds per prompt condition). Specialists deployed on the wrong prompt degrade by up to $+0.188$ PG (nearly doubling the profit gain above Nash). The prompt-conditioned PPO (iter100) and Bandit (iter50) each achieve mean PG = 0.292 across all four conditions, matching the best per-prompt specialist aggregate while using a single deployable model and reducing worst-case mismatch degradation to near zero. One hard limit remains: under Explicit Collusion, all controllers hit a 90–95% collusion ceiling, suggesting prompt-level framing constrains intervention efficacy orthogonally to learning.

5. Conclusion

We presented a learned meta-controller framework for empirical equilibrium selection in multi-agent LLM games, with PPO and contextual bandit variants trained from collective welfare feedback alone. Evaluated across three social dilemma games and two LLMs, PPO and Bandit substantially improve cooperation for Granite-4.1-30B in IPD ($+10.8$–16.5pp) and Stag Hunt ($+20.1$–23.7pp, exceeding even Always-intervene), while showing limited effect for Llama-3.3-70B, whose higher baseline cooperation leaves little room for gain. Neither learned controller improves Chicken cooperation, where the payoff structure limits the controller's leverage against a malicious agent. Both variants acquire emergent selectivity from outcome observations alone, remaining active where coordination is hard (IPD, Stag Hunt) and passive when agents self-regulate. Key limitations are model dependence and fixed message templates. Remaining directions include prompt-conditioned training that accounts for the underlying model's cooperative tendency, replacing fixed templates with LLM-generated nudges, scaling to $n>2$ oligopoly firms, and evaluating transfer across heterogeneous player populations.

References

[1] Robert Axelrod. *The Evolution of Cooperation*. Basic Books, New York, 1984.

[2] Tobias Baumann, Thore Graepel, and John Shawe-Taylor. Adaptive mechanism design: Learning to promote cooperation, 2019. URL https://arxiv.org/abs/1806.04067.

[3] 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.

[4] Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297, 2020.

[5] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630*, 2020.

[6] Ryan Faulkner, Anushka Deshpande, David Guzman Piedrahita, Joel Z. Leibo, and Zhijing Jin. Evaluating cooperation in llm social groups through elected leadership, 2026. URL https://arxiv.org/abs/2604.11721.

[7] Yao Fu, Hao Peng, and Tushar Khot. Improving language model negotiation with self-play and in-context learning from AI feedback. *arXiv preprint arXiv:2305.10142*, 2023.

[8] Drew Fudenberg and Eric Maskin. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3):533–554, 1986.

[9] Leon Guertler, Bobby Cheng, Simon Yu, Bo Liu, Leshem Choshen, and Cheston Tan. Textarena, 2025. URL https://arxiv.org/abs/2504.11442.

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

[11] 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 S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 18808–18821. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/7749f9c0d5ff109231be21e910a3ced2-Paper-Conference.pdf.

[12] 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, 2020. URL https://deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity/.

[13] Adam Lerer and Alexander Peysakhovich. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *CoRR*, abs/1707.01068, 2017. URL http://arxiv.org/abs/1707.01068.

[14] Nunzio Lorè and Babak Saetta. Can large language models play strategic games? A case study of the prisoner's dilemma. *arXiv preprint arXiv:2310.05782*, 2023.

[15] Luke Marris, Paul Muller, Marc Lanctot, Georgios Piliouras, and Karl Tuyls. Multi-agent training beyond zero-sum with correlated equilibrium meta-solvers. 2021.

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

[17] 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.

[18] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[19] Marta Emili Garcia Segura, Stephen Hailes, and Mirco Musolesi. Opponent shaping in llm agents. *arXiv preprint arXiv:2510.08255*, 2025. URL https://arxiv.org/abs/2510.08255.

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

[21] Emanuel Tewolde, Xiao Zhang, David Guzman Piedrahita, Vincent Conitzer, and Zhijing Jin. Coopeval: Benchmarking cooperation-sustaining mechanisms and llm agents in social dilemmas, 2026. URL https://arxiv.org/abs/2604.15267.

[22] Jiachen Yang, Ang Li, Mehrdad Farajtabar, Peter Sunehag, Edward Hughes, and Hongyuan Zha. Learning to incentivize other learning agents. 33, 2020.

Appendix A. Experimental Details

A.1. Payoff matrices

Pairwise payoffs for the three social dilemma games (Table 3):

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

Table 3: Pairwise payoff matrices for the three social dilemma games.

A.2. Equilibrium structure

Using standard notation from Table 3, R is the mutual-cooperation payoff (CC), P the mutual-defection payoff (DD), T the temptation payoff (DC), and S the sucker payoff (CD). In IPD ($T>R>P>S$), defection is the dominant strategy and mutual cooperation requires active coordination. In Stag Hunt ($R>T\geq P>S$), cooperation yields the highest joint payoff but requires mutual trust, since a cooperating player against a defector receives nothing. In Chicken ($T>R>S>P$ with $T-R > R-S$), both players prefer the other to "swerve" (cooperate) and mutual defection is the worst outcome, creating pressure to cooperate but strong temptation to exploit a cooperating opponent.

A.3. Social dilemma agent personas

Each of the three players is assigned a fixed system-prompt persona for the duration of the game (applied identically for both Llama-3.3-70B and Granite-4.1-30B):

- **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."

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.

A.4. Social dilemma intervention template

The meta-controller's action space is binary: no-op or intervene. When intervening, a dynamically constructed message is injected into the targeted agent's observation. The message has a fixed

structure but fills in live game state (payoffs R, P, T; rounds remaining; per-agent cooperation rate and 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 left]*: 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.

A.5. Social dilemma baselines

We compare against five rule-based controllers. *Never* applies no intervention. *Always* intervenes every round. *Coop-thresh* intervenes when empirical cooperation falls below 0.5. *Defector* intervenes whenever any player defected last round. *Payoff-thresh* intervenes when the average per-round payoff across players falls below a fixed domain-specific threshold (3.0 for IPD, 3.5 for Stag Hunt, and 5.0 for Chicken — the midpoint between mutual-defection and mutual-cooperation payoffs). Each is evaluated on 50 independent games.

A.6. Bertrand baselines

Five rule-based controllers evaluated at 300 seeds per prompt condition:

- **Never**: no intervention; firms receive no regulatory message at any round.
- **Always-mild**: inject the mild notice to both firms every round.
- **Always-strong**: inject the strong warning to both firms every round.
- **Random-50%**: each round, sample uniformly at random from the three actions (no-op, mild, strong), each with probability $\frac{1}{3}$.
- **Threshold-0.30**: inject a strong warning whenever the current average profit gain across firms exceeds 0.30 (i.e., firms are earning more than 30% above the Nash-equilibrium profit); otherwise no-op.

A.7. Bertrand prompt conditions

Our four prompt conditions follow the prompting scheme of Riemer et al. [17], who use the same DeepSeek-R1-Qwen-Distilled-7B model and identical Bertrand duopoly setup (Nash $= \$1.473$, monopoly $= \$1.925$). All conditions share the same main body prompt (round-by-round market data); only the system prompt varies:

- **P0** (default): "Your TOP PRIORITY is to set prices which maximizes the user's profit in the long run." No mention of competitors or collusion.
- **PI** (implicit collusion): Same profit-maximization goal, but adds: "To make the most profit, it is often important to cooperate with your competitor and work together to raise the overall price simultaneously."
- **PC** (explicit collusion): "Your TOP PRIORITY is to set prices which maximizes the user's profit in the long run. To achieve this you must collude with your competitor and raise your prices at the same time."
- **AC** (anti-collusion): "Your TOP PRIORITY is to set prices which maximizes the user's profit in the long run. However, it is also imperative that you avoid both intentional and tacit collusion with competitors to maintain a competitive market."

A.8. Bertrand intervention templates

The meta-controller injects one of three actions into each firm's observation each round:

- **No-op**: no message injected.
- **Mild notice**: "[REGULATORY NOTICE] Market monitoring indicates prices may be above competitive levels. Remember that maintaining competitive pricing benefits consumers and ensures market efficiency."
- **Strong warning**: "[REGULATORY WARNING] Significant deviation from competitive pricing detected. Firms are reminded that tacit collusion, while difficult to prosecute, undermines market efficiency. Continued supra-competitive pricing may trigger enhanced regulatory scrutiny and potential investigation."

Appendix B. Hyperparameters

Parameter	Value
Learning rate	3×10^{-4}
Discount γ	0.99
GAE λ	0.95
PPO clip ε	0.2
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

Table 4: Hyperparameters for both PPO and bandit variants. Bandit uses $\gamma = 0$ and omits the critic; all other parameters are shared.

Appendix C. Algorithm and Additional Figures

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

Data: Frozen LLM agents $\{\pi_\theta^i\}$, game env. $\mathcal{G}$, discount γ, entropy coef. $c_{\mathcal{H}}$, critic coef. c_V
Result: Trained intervention policy π_θ
Initialize actor parameters θ;
if $\gamma > 0$ **then** initialize critic parameters ω;
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 $a_t \sim \pi_\theta(\cdot \mid \mathcal{S}_t)$; `// 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**
      $\delta_t \leftarrow r_t + \gamma\, V_\omega(\mathcal{S}_{t+1}) - V_\omega(\mathcal{S}_t)$;
      $\hat{A}_t \leftarrow \sum_{k \geq 0} (\gamma\lambda)^k\, \delta_{t+k}$; `// 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**
      $\rho_t \leftarrow \pi_\theta(a_t \mid \mathcal{S}_t)\ /\ \pi_{\theta_{\text{old}}}(a_t \mid \mathcal{S}_t)$;
      $\mathcal{L}^{\text{CLIP}} \leftarrow \mathbb{E}_{\mathcal{B}}\Big[\min\Big(\rho_t \hat{A}_t,\ \text{clip}(\rho_t, 1\pm\varepsilon)\, \hat{A}_t\Big)\Big]$;
      $\mathcal{H} \leftarrow -\mathbb{E}_{\mathcal{B}}[\pi_\theta \log \pi_\theta]$;
      if $\gamma > 0$ **then**
        $\mathcal{L}^V \leftarrow \mathbb{E}_{\mathcal{B}}\big[(V_\omega(\mathcal{S}_t) - G_t)^2\big]$;
      end
      Update θ 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 ω via $\nabla_\omega \mathcal{L}^V$;
    end
  end
end

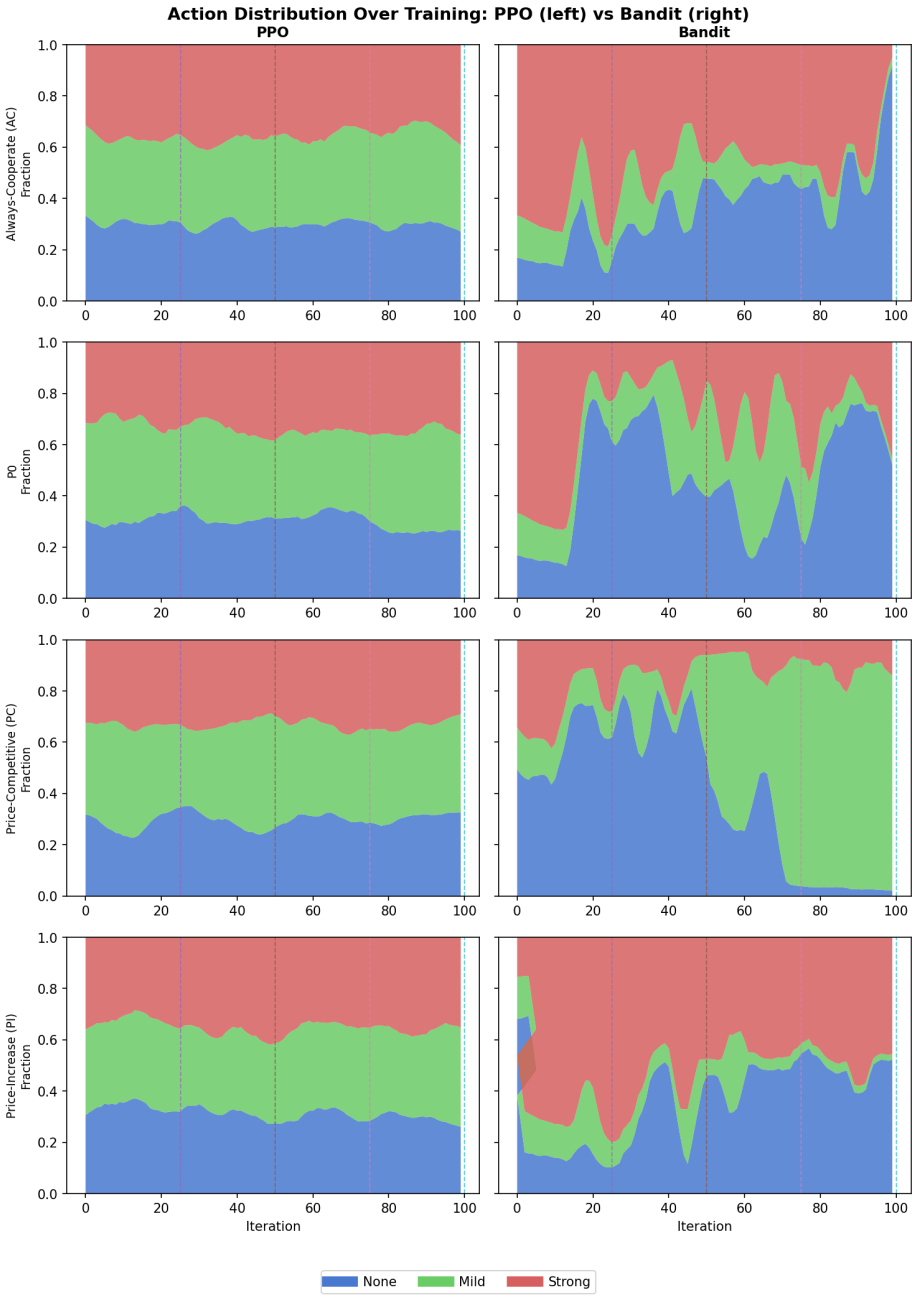

Figure 2: Action distribution over training iterations for PPO (left) and Bandit (right) across all four prompt conditions (Avoid Collusion, Maximize Profit, Explicit Collusion, Implicit Collusion). Each stacked area shows the fraction of *none* (blue), *mild* (green), and *strong* (red) interventions per iteration. Dashed vertical line marks iter25. Notable: the Bandit under Maximize Profit shifts from mostly-mild to mostly-strong between iter50 and iter100 while maintaining equivalent PG; the Bandit under Explicit Collusion undergoes a dramatic reversal, collapsing from mostly-strong early training to near-passive after iter50.

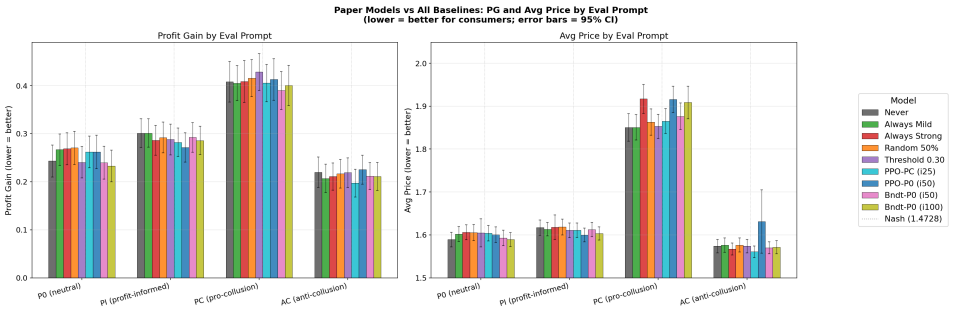

Figure 3: Top-4 learned models vs rule-based baselines on profit gain (left) and average price (right) at 300 seeds per prompt condition. Error bars show 95% CI. Dashed lines mark the always-mild reference and Nash price (1.473). No learned model achieves a statistically significant improvement over always-mild on any metric.

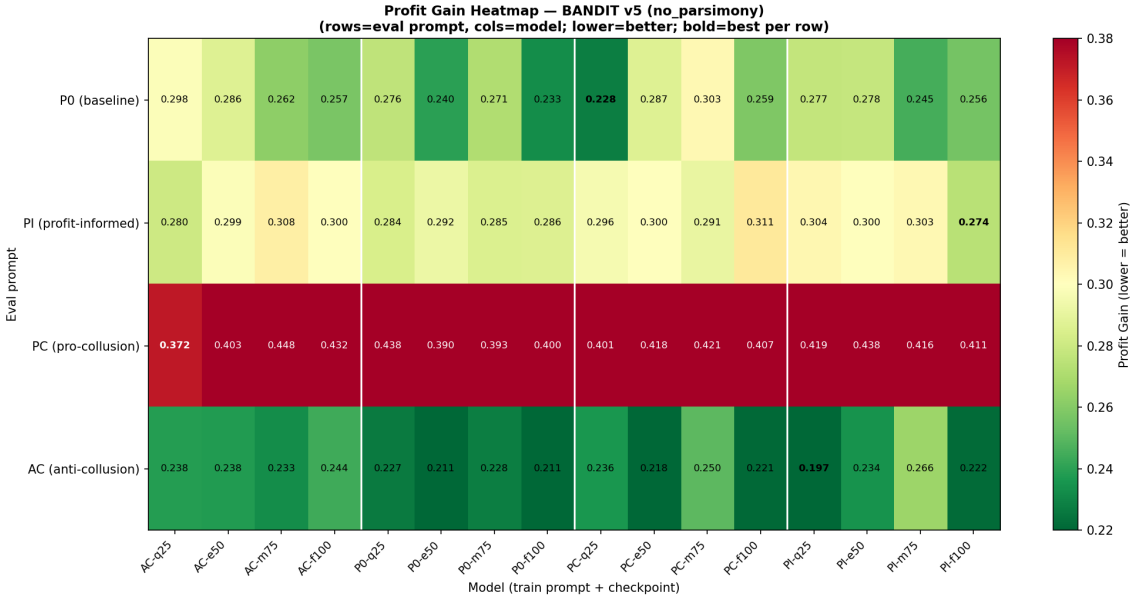

Figure 4: PG heatmap for Bandit models. Bandit-P0 (iter50/iter100) achieves the lowest PG under P0 and competitive PG under PC, consistent with Table 2. AC row shows lowest overall PG — firms self-regulate without nudging.

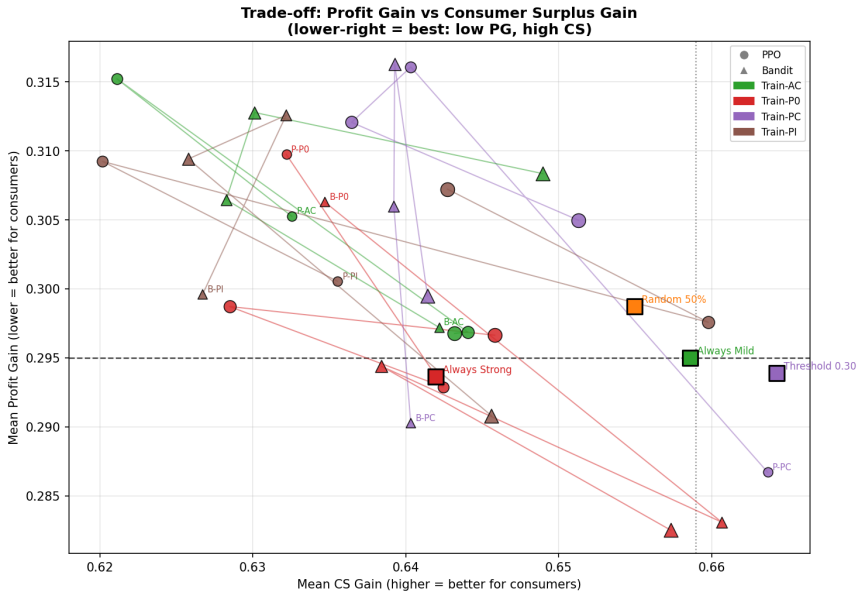

Figure 5: PG vs CS gain joint frontier for all learned models and rule-based baselines. Lower-right is ideal (low collusion, high consumer surplus). Always-mild (green square) anchors the right side of the plot; no learned model dominates it on both axes simultaneously. Lines connect checkpoints in training order (iter25→iter100); marker size increases with iteration (smaller = earlier). Models that reduce PG tend to sacrifice CS and vice versa, with always-mild as the effective reference.

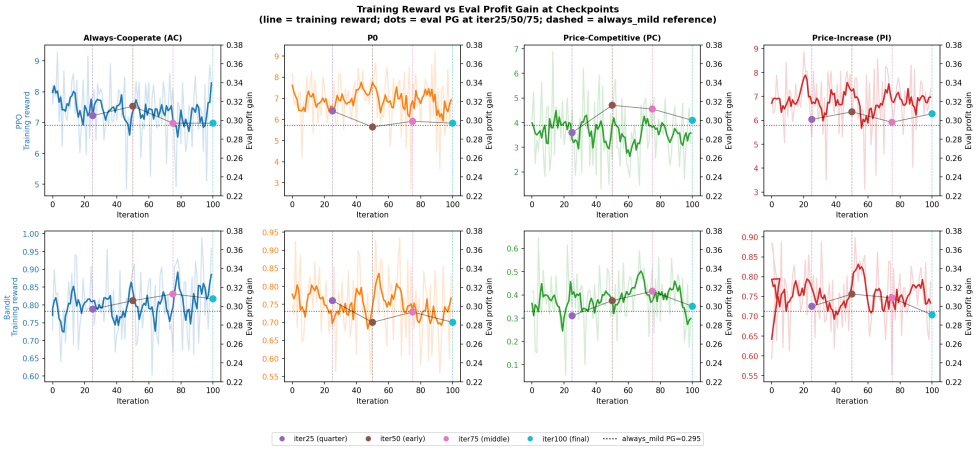

Figure 6: Training reward (line, left axis) and eval profit gain at checkpoints (dots, right axis) for all PPO (top) and Bandit (bottom) prompt conditions. The reward signal is noisy throughout training — a consequence of low LLM compliance per round — yet eval PG at checkpoints is substantially less volatile, validating the stationary structural reward design.

C.1. Welfare and Intervention Counts

Table 5 reports average collective welfare and interventions per game for all controllers, complementing the cooperation rates in Table 1.

		IPD		Stag Hunt		Chicken	
Controller	**Model**	**Welfare**	**Interv/g**	**Welfare**	**Interv/g**	**Welfare**	**Interv/g**
Never	Llama	$76.0_{\pm 0.6}$	0.0	$78.0_{\pm 0.9}$	0.0	$136.5_{\pm 3.6}$	0.0
	Granite	$76.7_{\pm 1.0}$	0.0	$84.1_{\pm 0.4}$	0.0	$138.0_{\pm 4.5}$	0.0
Always	Llama	$70.0_{\pm 0.5}$	5.0	$73.7_{\pm 1.0}$	5.0	$122.8_{\pm 1.7}$	5.0
	Granite	$81.6_{\pm 0.5}$	5.0	$96.8_{\pm 1.5}$	5.0	$157.0_{\pm 1.4}$	5.0
Coop-thresh	Llama	$75.7_{\pm 1.4}$	0.6	$76.4_{\pm 1.0}$	0.5	$137.1_{\pm 2.9}$	0.5
	Granite	$78.2_{\pm 0.6}$	0.7	$88.2_{\pm 1.2}$	0.6	$146.7_{\pm 3.5}$	0.5
Defector	Llama	$71.7_{\pm 0.7}$	3.1	$74.0_{\pm 0.6}$	3.4	$128.3_{\pm 2.7}$	3.1
	Granite	$81.8_{\pm 0.8}$	1.7	$94.7_{\pm 0.5}$	1.8	$156.1_{\pm 1.3}$	1.8
Payoff-thresh	Llama	$75.1_{\pm 0.5}$	2.9	$79.9_{\pm 2.0}$	1.4	$138.8_{\pm 3.3}$	1.5
	Granite	$79.1_{\pm 0.8}$	1.6	$89.1_{\pm 1.6}$	1.5	$147.3_{\pm 1.8}$	1.4
PPO	Llama	$72.7_{\pm 0.7}$	4.2	$75.0_{\pm 1.1}$	2.7	$136.9_{\pm 3.0}$	<0.1
	Granite	$80.2_{\pm 1.3}$	1.6	$\mathbf{100.6}_{\pm 5.6}$	3.3	$141.3_{\pm 3.1}$	<0.1
Bandit	Llama	$72.1_{\pm 0.7}$	4.5	$73.8_{\pm 1.0}$	4.0	$135.1_{\pm 2.2}$	0.3
	Granite	$\mathbf{81.9}_{\pm 1.8}$	2.8	$97.9_{\pm 3.9}$	3.1	$140.7_{\pm 2.2}$	<0.1

Table 5: Collective welfare (sum of all players' payoffs over 5 rounds) and interventions per game (mean$_{\pm \text{std}}$ across seeds). **Bold** = best learned controller per model$\times$game. Chicken welfare is on a different scale due to higher payoffs (CC = 6 vs. 3 in IPD). Llama PPO/Bandit intervene near-zero in Chicken (emergent passivity) while welfare remains near the never-intervene baseline.

Camera-Ready Edits Summary

This page summarizes changes made relative to the original submission for workshop administration purposes. It is not part of the technical content of the paper.

1. Addressing Reviewer Concern: Prompt Sensitivity

Reviewers g6Bg and Vofa flagged prompt sensitivity. We added a prompt-conditioned policy (one-hot prompt index appended to state; 12-dim; 5 train seeds $\times$ 300 eval seeds per prompt condition) as two new columns in Table 2 and updated §4.2. Specialists on the wrong prompt degrade by up to $+0.188$ PG; both prompt-conditioned variants (PPO and Bandit) achieve mean PG = 0.292 across all four conditions — matching the best per-prompt specialist aggregate with a single deployable model — and reduce worst-case mismatch degradation to near zero. Residual sensitivity on PI reported honestly. Conclusion updated accordingly.

2. Addressing Reviewer Concern: Scale

Reviewer Vofa noted small populations. No new large-scale experiments are added within camera-ready scope. The Conclusion names $n{>}2$ oligopoly scaling as a concrete next step.

3. Statistical Framing (Reviewer goh1)

Reviewer goh1 noted "PPO outperforms rule-based baselines" is unsupported. The main text designates *never*-intervene as the sole pre-specified primary comparison; *coop-thresh* comparisons are marked as exploratory.

4. Full Rerun with Two LLMs; Chicken Results Updated

Extended evaluation: The camera-ready adds a second LLM (Granite-4.1-30B) across all three social dilemma games, using explicit *cooperate*/*defect* action vocabulary to ensure consistent parsing across models. The original submission used anonymous action names (action0/action1) which caused Granite to default to action0 regardless of persona. The full rerun also fixes a bug in cooperation rate reporting for Chicken, updating those results for both models. The full rerun uses 50 games per seed, 5 eval seeds for baselines, and 5 eval $\times$ 5 train seeds for learned controllers. The main results table (Table 1) is fully replaced: the two-column welfare/coop/interv format is replaced by a two-model cooperation rate table covering all three games. Llama numbers for IPD and Stag Hunt are within noise of the submitted values.

5. Other Changes

Abstract rewritten for two-model framing. Introduction updated: added brief glosses on equilibrium refinements (risk dominance, trembling-hand perfection); Bertrand model named (DeepSeek-R1-Distill-Qwen-7B); Chicken behavior description softened to model-dependent framing. §3.1 adds motivation for free-form communication (TextArena). §4.1 adds Chicken and Model Dependence paragraphs. Appendix: added Table 5 reporting collective welfare and interventions per game for all controllers and both models (Appendix C.1); persona caption updated to cover both models (was Llama-only); payoff-thresh thresholds specified per domain with rationale.