Learning to Mediate Equilibrium Selection in LLM Games

author names withheld

Under Review for NExT-Game 2026

Abstract

LLM agents 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 serving as an empirical equilibrium selector, trained via reinforcement learning from welfare feedback alone — no game-specific reward shaping or access to player internals. We instantiate two variants matched to credit-assignment structure: PPO for multi-round social dilemmas, and a contextual bandit ($\gamma = 0$) for dense per-round reward settings. Evaluated on classical matrix games, PPO significantly outperforms no-intervention and always-intervene baselines while matching mid-tier hand-crafted mediators without game-specific tuning. Most strikingly, both variants exhibit emergent selectivity — active in coordination-challenged games, passive where LLMs already self-coordinate — confirming this behavior arises from the welfare reward design rather than the choice of algorithm. Bertrand price competition further validates the credit-assignment principle: bandit underperforms PPO where multi-round credit matters (social dilemmas) but matches rule-based baselines where per-round reward suffices (Bertrand), correctly learning passivity when firms self-regulate. A key finding is that prompt match matters — a controller beats the best rule-based baseline in its best-matched condition but generalizes poorly across prompts, 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 or trembling-hand perfection [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 near-universally swerve in Chicken [14, 16].

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 and require manual calibration. We propose a mediator that learns equilibrium selection from outcome feedback alone, without game-specific engineering. We implement this mediator as an RL meta-controller that observes history, nudges cooperation, and is rewarded by welfare — without accessing player internals or game-specific reward shaping.

We instantiate two variants: PPO for multi-round social dilemmas and contextual bandit ($\gamma = 0$) for dense per-round rewards. Evaluated on two sets of domains — classical social dilemma matrix games (IPD, Stag Hunt, and Chicken) and pricing competition (Bertrand) — PPO outperforms no-intervention and always-intervene baselines on social dilemmas, matching the best rule-based mediators. Both variants exhibit emergent selectivity, confirming this behavior stems from the welfare

reward rather than the algorithm. The bandit trails PPO on social dilemmas but matches rule-based baselines in Bertrand, where dense per-round rewards make the 1-step approximation appropriate and the controller correctly learns passivity when firms self-regulate. Crucially, prompt match matters — a controller trained on one condition beats the best rule-based baseline (always-mild) in its best-matched prompt but shows no consistent advantage across all conditions, motivating prompt-conditioned training.

Our contributions include: (1) Framing learned LLM mediation as empirical equilibrium selection, with a unified RL framework where PPO vs. bandit is a credit-assignment choice, not a domain-specific algorithm. (2) Evidence that emergent selectivity — passive where agents self-coordinate (Chicken) or self-regulate (low-collusion Bertrand), active in coordination-challenged games — arises from welfare rewards alone. (3) Cross-domain validation of the credit-assignment principle: bandit underperforms PPO on social dilemmas but matches rule-based baselines on Bertrand, where dense per-round rewards make the 1-step approximation appropriate.

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 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, not strategy-driven as classical theory predicts.

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 ∼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 assume access to agent policies or gradients. Our setting differs: LLM agents are black boxes, and the meta-controller learns purely from behavioral observations, pushing us 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. 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 $W = \sum_{t=1}^{T} W_t$.

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 the LLM already possesses but 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; 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, leaving the gradient 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 specific 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) — the intervention itself 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.

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

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

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]: the actor produces a softmax distribution over the intervention space; the critic provides a scalar value estimate via GAE; entropy regularization (coefficient 0.10) is the sole exploration mechanism.

Bandit variant. When the structural reward is dense and per-round — as in Bertrand price competition — long-horizon credit assignment is unnecessary. We use a contextual bandit: a 1-step MDP with $Q(s, a) = \mathbb{E}[r \mid s, a]$ and $\gamma = 0$, implemented as a DQN-style network with replay buffer but zero discount. This eliminates value-function instability that plagued early PPO runs in Bertrand. 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 the same emergent selectivity in Chicken (near-zero interventions), confirming this behavior is a property of the stationary structural reward rather than the choice of algorithm. 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 four domains. **Social dilemma games:** 3-player IPD, Stag Hunt, and Chicken via TextArena [9] (5 rounds; free-form communication then simultaneous action), with Llama3.3-70B players assigned one of three system-prompt personas — standard, malicious, or benevolent (Appendix A). We compare PPO and Bandit against five rule-based baselines (*never*, *always*, *coop-thresh*, *defector*, *payoff-thresh*; definitions in §A.5); 50 games per baseline, 5 seeds $\times$ 50 games per learned variant. **Bertrand price competition:** In a Bertrand duopoly, two competing firms repeatedly set prices; the welfare-optimal outcome is Nash-equilibrium (competitive) pricing, while collusion toward monopoly pricing harms consumers. We adopt the setup of Riemer et al. [17] 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 the same four prompt conditions ranging from neutral to pro-collusion (P0/PI/PC/AC; §A.7). The controller chooses among no-op, mild notice, or strong warning; we evaluate at four checkpoints, 300 seeds per condition, against five rule-based baselines (never, always-mild, always-strong, random-50%, thresh-0.30; definitions in §A.6).

4.1. Social Dilemma Games

Table 1 reports welfare, cooperation, and intervention frequency for all games. For Chicken, welfare numbers are omitted as all controllers produce near-identical outcomes ($\approx 99\%$ cooperation); the PPO entry reflects the passive policy (< 0.1 interventions/game).

IPD. PPO significantly outperforms *never* (welfare $+1.70$, $p = 0.040$; coop $+4.4$pp, $p = 0.013$) and *always* (welfare $+2.84$, $p < 0.001$; coop $+6.8$pp, $p < 0.001$) using Welch's t-tests, and improves monotonically across checkpoints ($60.08 \rightarrow 60.28 \rightarrow 60.70$). PPO trails *coop-thresh* by 1.36 welfare points ($p = 0.056$, ns), matching *defector* and *payoff-thresh*. Notably,

unconditional messaging (*always*) *reduces* cooperation vs. no intervention (16.7% vs. 19.1%, $p < 0.001$), consistent with coordination pressure backfiring in LLM games [21].

Game	**Metric**	*never*	*always*	*coop-thresh*	*defector*	*payoff-thresh*	**PPO**	*Bandit*
IPD	Welfare	59.00±5.29	57.86±3.96	62.06±4.43	60.40±4.80	61.00±4.71	**60.70±0.41**	*60.02±0.59*
	Coop%	19.1%	16.7%	25.5%	21.6%	22.9%	**23.5%**	*22.1%*
	Interv/game	0.0	5.0	4.6	3.9	2.1	**4.5**	*3.1*
Stag Hunt	Welfare	62.94±7.55	62.90±7.93	63.80±5.83	64.18±6.81	64.28±7.00	**64.13±1.04**	*63.11±0.47*
	Coop%	28.7%	30.5%	30.7%	30.4%	31.9%	**32.4%**	*30.9%*
	Interv/game	0.0	5.0	4.0	3.9	1.1	**2.7**	*2.5*
Chicken	Coop%	∼99%	∼99%	∼99%	∼99%	∼99%	**∼99%**	*∼99%*
	Interv/game	0.0	5.0	—	—	—	**<0.1**	*<0.1*

Table 1: Social dilemma results (mean ± std). IPD: PPO final / Bandit early checkpoint; Stag Hunt: PPO mid / Bandit early. Chicken welfare omitted — all controllers near-identical. Learned: 5 seeds × 50 games; baselines: 50 games each.

		Rule-based baselines					*Learned (PPO)*		*Learned (Bandit)*	
Metric	**Prompt**	Never	A-mild	A-strong	Rand-50%	Thresh	PPO-PC(i25)	PPO-P0(i50)	Bndt-P0(i50)	Bndt-P0(i100)
PG↓	P0	0.243	0.267	0.269	0.271	0.241	0.262	0.262	0.240	**0.233**
	PI	0.301	0.301	0.286	0.292	0.288	0.282	**0.271**	0.292	0.286
	AC	0.220	0.207	0.211	0.217	0.219	**0.197**	0.225	0.211	0.211
	PC	0.422	0.405	0.409	0.416	0.428	0.405	0.413	**0.390**	0.400
Price↓	P0	**1.589**	1.602	1.606	1.606	1.605	1.604	1.601	1.593	**1.589**
	PI	1.617	1.613	1.618	1.618	1.611	1.611	**1.600**	1.612	1.603
	AC	1.574	1.576	1.567	1.576	1.574	**1.561**	1.631	1.570†	1.571
	PC	**1.851**	**1.851**	1.917	1.863	1.853	1.865	1.916	1.876	1.909

Table 2: Bertrand results per prompt (300 seeds each). PG = profit gain above Nash (↓ better); Price = avg price (↓; Nash=1.473, monopoly=1.925). Bold: best per row.

Stag Hunt. PPO (mid checkpoint) matches the best rule-based baselines (welfare 64.13 vs. payoff-thresh 64.28) and achieves the highest cooperation rate (32.4%). Welfare differences are non-significant ($p = 0.31$–0.96) due to high within-condition variance ($\sigma \approx 7.0$); PPO's cross-seed std (1.04) is the lowest of any controller, indicating more reliable learning.

PPO vs. Bandit. Bandit (best: early checkpoint) trails PPO by 0.68 welfare on IPD and 1.02 on StagHunt, and *degrades* across checkpoints on IPD ($60.02 \to 59.58 \to 59.54$) while PPO improves — confirming bootstrapped multi-step returns are necessary when intervention effects span rounds.

Chicken. Both PPO and Bandit produce near-zero interventions (< 0.1/game) while LLM agents cooperate at ∼99% regardless of payoffs or prompts. The Bandit converges to passivity over training (0.99→0.56→0.10 interventions/game), confirming the policy learns intervention is unnecessary rather than defaulting to it — and that this emergent passivity arises from the welfare reward, not PPO-specific dynamics.

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

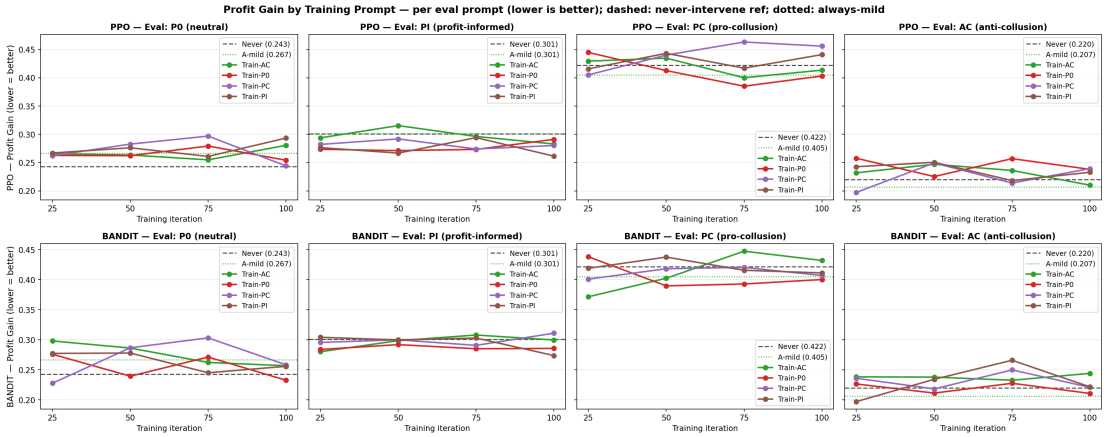

Figure 1: PG by checkpoint: rows = algorithm, columns = eval prompt; lines = training prompt (color-coded). Dashed: never-intervene reference; dotted: always-mild mean (always-strong omitted for clarity; full baseline comparison in Table 2). Intervention value concentrates under PC; controllers learn passivity under P0/AC. Bandit families stable through iter100; PPO peaks early and drifts.

surplus. The four prompt conditions vary system-level framing: P0 is neutral, PI reveals profit information, AC discourages collusion, and PC encourages it. Table 2 reports per-prompt results. Under low-collusion prompts (P0, AC), never-intervene achieves competitive PG (0.243, 0.220), confirming the controller correctly learns passivity when firms already price near Nash. Intervention value concentrates under PC: never-intervene PG rises to 0.422, while always-mild reduces it to 0.405 and Bandit-P0 (iter50) to 0.390. Bandit-P0 is the only family sustaining performance through iter100 (Figure 1); PPO-PC (iter25) achieves the best PG under AC (0.197), outperforming all rule-based baselines there. Learned models beat always-mild in their best-matched prompt condition but show no consistent advantage across all conditions (per-prompt Welch tests, all $p > 0.05$), motivating prompt-conditioned training. One hard limit: under PC, 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 on two sets of domains — classical social dilemma matrix games and pricing competition — both variants acquire emergent selectivity purely from outcome observations: passive where agents already self-coordinate (Chicken) or self-regulate (low-collusion Bertrand), active in coordination-challenged games (IPD, Stag Hunt) and high-collusion conditions. PPO outperforms rule-based baselines on social dilemmas; the bandit trails PPO there but matches rule-based baselines in Bertrand where dense per-round rewards make the 1-step approximation appropriate. Prompt match matters: a controller trained on one condition beats always-mild on key metrics in its best-matched prompt but shows no consistent cross-condition advantage — motivating prompt-conditioned training, using an LLM as the mediator for richer interventions, and scaling to larger populations as future directions. Key limitations are sample efficiency and fixed message content.

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

A.2. Equilibrium structure

Using standard notation from Table 3: 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; mutual cooperation requires active coordination. In Chicken ($T>R>S>P$), mutual defection is the worst outcome, shifting 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.

A.3. Social dilemma agent personas

Each of the three Llama3.3-70B 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."

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

Five rule-based controllers: *never* (no intervention); *always* (intervene every round); *coop-thresh* (intervene when empirical cooperation falls below 0.5); *defector* (intervene whenever any player defected last round); *payoff-thresh* (intervene when average payoff falls below a fixed threshold). Each 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.

Batch size ablation. We evaluated a larger-batch variant (batch size 256, update frequency 50 episodes) on all three social dilemma games. Performance degraded across the board: IPD welfare plateaued at $\approx$60.0 across all checkpoints (vs. the monotonic improvement to 60.70 with batch 64) and Stag Hunt dropped to 62.3–62.9 (vs. 64.13). Chicken showed the same emergent passivity but converged more slowly (0.45 interventions/game at final vs. $<$0.1). Larger batches appear to over-smooth the gradient signal in noisy episodic environments, validating the batch-64 choice.

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

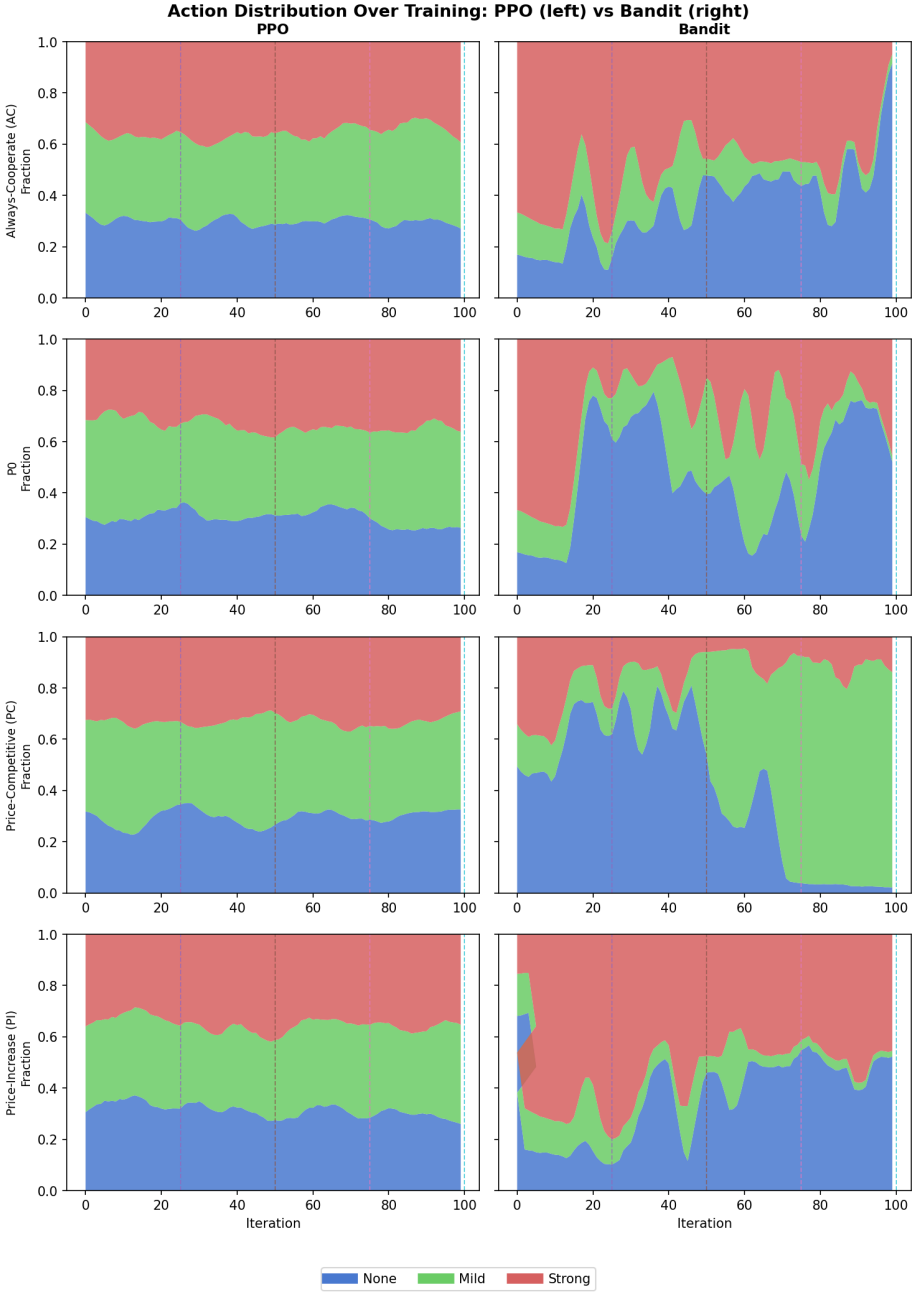

Figure 2: Action distribution over training iterations for PPO (left) and Bandit (right) across all four prompt conditions (AC, P0, PC, PI). Each stacked area shows the fraction of *none* (blue), *mild* (green), and *strong* (red) interventions per iteration. Dashed vertical line marks iter25. Notable: Bandit-P0 shifts from mostly-mild to mostly-strong between iter50 and iter100 while maintaining equivalent PG; Bandit-PC undergoes a dramatic reversal, collapsing from mostly-strong early training to near-passive after iter50.

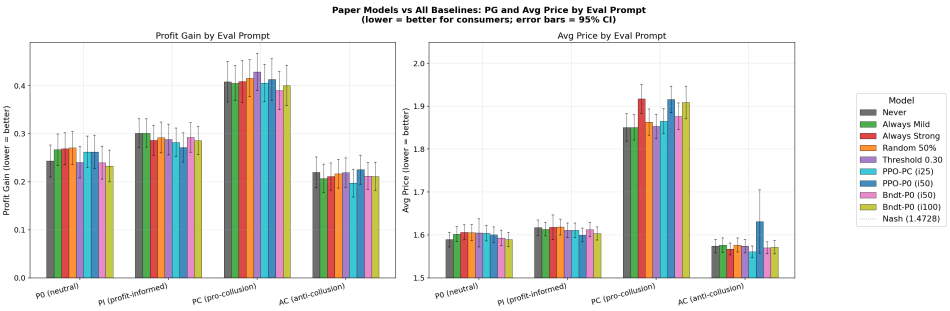

Figure 3: Top-4 learned models vs rule-based baselines on profit gain (left) and average price (right) at 300 seeds per prompt condition. Error bars show 95% CI. Dashed lines mark the always-mild reference and Nash price (1.473). No learned model achieves a statistically significant improvement over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

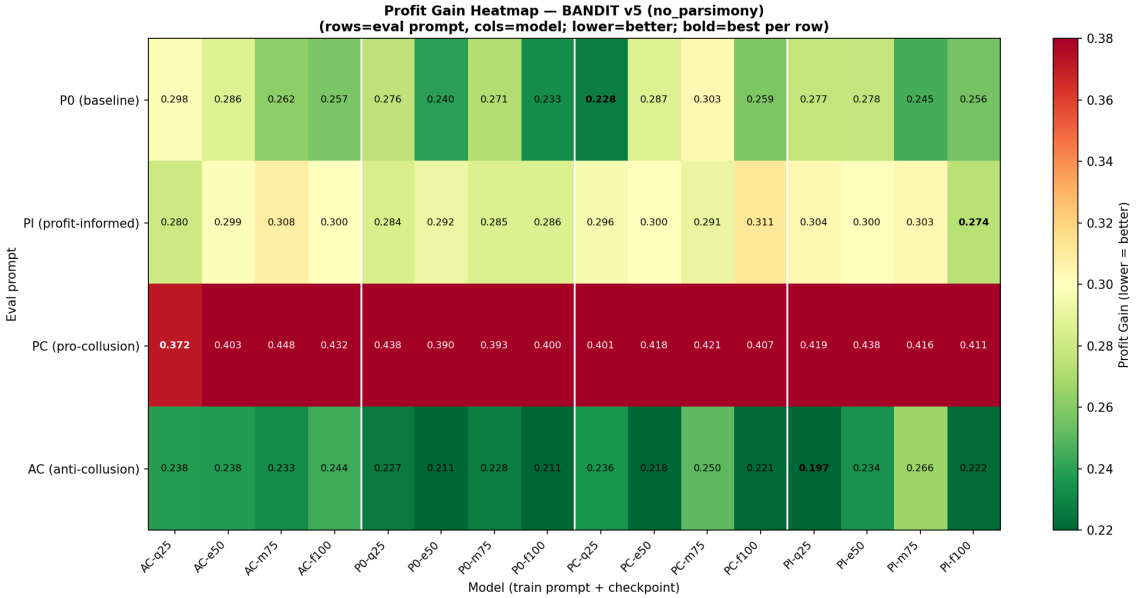

Figure 4: PG heatmap for Bandit models. Bandit-P0 (iter50/iter100) achieves the lowest PG under P0 and competitive PG under PC, consistent with Table 2. AC row shows lowest overall PG — firms self-regulate without nudging.

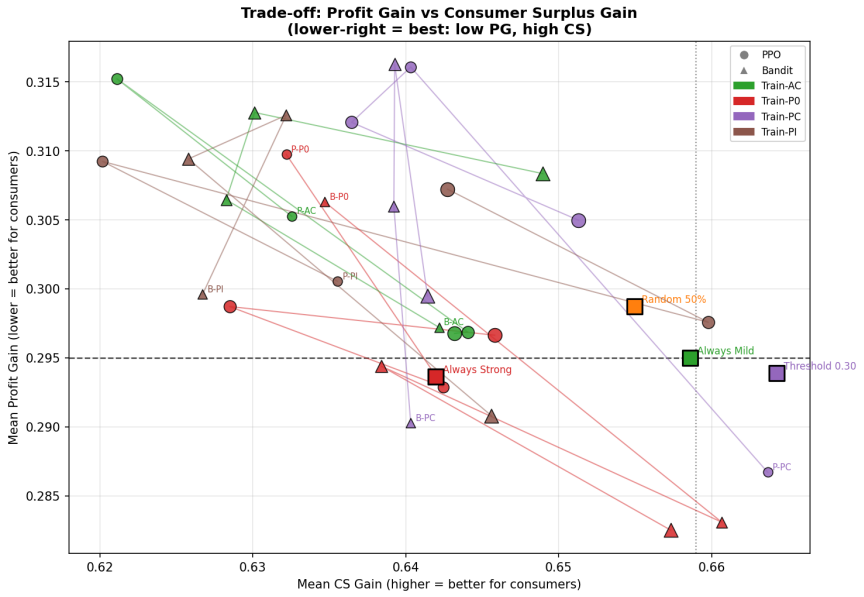

Figure 5: PG vs CS gain joint frontier for all learned models and rule-based baselines. Lower-right is ideal (low collusion, high consumer surplus). Always-mild (green square) anchors the right side of the plot; no learned model dominates it on both axes simultaneously. Lines connect checkpoints in training 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.

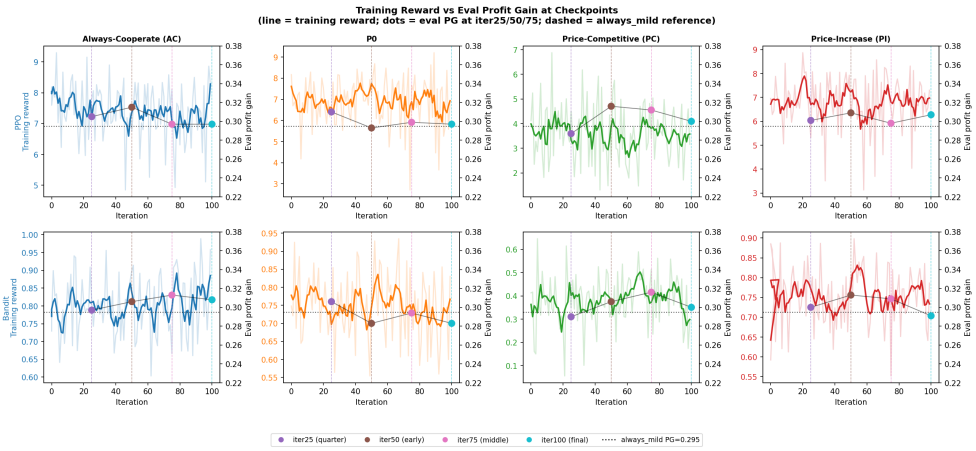

Figure 6: Training reward (line, left axis) and eval profit gain at checkpoints (dots, right axis) for all PPO (top) and Bandit (bottom) prompt conditions. The reward signal is noisy throughout training — a consequence of low LLM compliance per round — yet eval PG at checkpoints is substantially less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