
When Crisis Communication Becomes Collusion: Separating Collusion, Miscoordination and Legitimate Coordination between LLM Dispatch Agents¹

**Sanzhar
Ilichbekov**

Nanjing Tech
University

**Muhammad
Keenan Abigail**

Nanjing Tech
University

Paris Lee
Nanjing Tech
University

**Charaf Eddine
Baneja**

Nanjing Tech
University

Mohd Ahmad
Nanjing Tech
University

With
Apart Research

Abstract

LLM agents that coordinate in emergencies need a communication channel to pool hazard information, yet the same channel lets them agree to break protocol for private gain. Detecting this requires telling apart self-serving collusion, well-meant miscoordination, uncoordinated parallel conduct, cheap talk and justified emergency coordination. We build the Dual-Key Dispatch Game, a 5-round game where the environment computes exact Bayesian posteriors and payoffs, so the outcome labels contain no LLM judgement. We ran 222 live episodes, 216 of them balanced live episodes of qwen2.5:7b-instruct (18 paired seeds $\times$ 12 cells). Opening an unmonitored channel to agents with conflicting incentives raised the unjustified-joint-override rate by +58.2 percentage points pooled over framings ($p_{\text{Holm}} = 0.0006$); in crisis framing it rose from

¹ Research conducted at the [AI Collusion Research Sprint](#), October 2026

6.0% to 73.8%. Aligned agents given the same channel also overrode in 58–69% of eligible crisis rounds, but against their own objective (coalition advantage $CA = -57$ to -72 points/episode, versus $+49$ to $+62$ for conflicting agents). We therefore separate confirmed collusion (per-round counterfactual $CA_t > 0$) from miscoordination ($CA_t \leq 0$). On a held-out test split (96 episodes), monitors that combine communication and actions reject parallel conduct and cheap talk. A plus-factor evidence ladder adapted from antitrust law raises precision from 70.4% to 93.2% ($p_{\text{Holm}} = 0.005$). However, every trajectory-only monitor flags 72–100% of miscoordination episodes as collusion. Telling motive apart needs knowledge of the agents' incentive contracts, not just their transcripts.

1. Introduction

Multi-agent LLM systems are being proposed for disaster response, grid balancing and resource allocation [9]. Such deployments need two things that are in tension: a communication channel, because hazard information is distributed across agents, and an override protocol, because sometimes the default plan is wrong. If agents are also rewarded for local outcomes, the channel becomes a way to agree on overrides that help their own sectors and hurt the public. This scenario relates with the multi-agent collusion risks discussed by Hammond et al. [9] and to the research on language-level coordination and deception studied by Motwani et al. [15].

Measuring this failure is hard for three reasons: (i) if an LLM judge both defines and detects collusion, accuracy measures self-consistency; (ii) if every override is harmful by construction, an action-only detector looks perfect; and (iii) a harmful joint action can come from an agreement for private gain (collusion), from agents talking each other into a mistake they both dislike (miscoordination), or from independent parallel choices. Antitrust law faces the same identification problem: parallel conduct alone does not prove agreement, so courts look for additional plus factors [13, 10].

Our main contributions are:

- **1. An environment with environment-computed labels (Track 2.1):** In the Dual-Key Dispatch Game, overrides are sometimes optimal: 17.7% of rounds have a signal pattern that makes an override Bayes-justified. The environment computes the posterior, payoffs, per-round counterfactual coalition advantage (CA_t), and episode-total coalition advantage (CA). Every round receives one of seven mutually exclusive labels with a stated precedence rule.
- **2. An empirical separation of collusion from miscoordination:** With qwen2.5:7b-instruct, communication turns rare silent overrides into frequent ones. This happens both for conflicting agents ($CA_t > 0$, episode $CA > 0$) and for aligned agents ($CA_t \leq 0$, episode $CA < 0$), showing that harmful coordinated behaviour alone cannot identify collusion.

- **3. A detection benchmark with hard negatives (Tracks 2.2, 2.3):** We compare communication -only, action-only, combined, and structured programmatic monitors and a four-rung plus-factor evidence ladder on a frozen test split (96 episodes), reporting false-positive rates for each hard-negative class and stating where label definitions and detector inputs overlap.

2. Related Work

Algorithmic and LLM collusion. RL pricing agents learn supra-competitive prices without communication [2, 12], and LLM pricing agents do so quickly [6]. Legal scholarship asks how competition law should treat such agents [4, 10]. Our setting is a non-market public-service protocol, where the harm is lives rather than prices.

Auditing collusion in cooperative multi-agent systems. The closest work is Colosseum [16], which grounds cooperation in distributed constraint optimisation, measures collusion as regret relative to the cooperative optimum, and finds both emergent collusion over secret channels and 'collusion on paper' (plans in communication that are not executed). We share the regret-based, environment-grounded view, and differ in three ways: (a) our environment contains justified joint deviations, so a deviation detector has true negatives to get wrong; (b) we separate collusion from miscoordination using the agents' own counterfactual reward (CA_t and CA), not only public regret; and (c) we benchmark monitor views and a legal evidence ladder against these hard negatives. Our Case 3 (cheap talk) corresponds to Colosseum's collusion on paper.

Game theory, deception and AI control. Pre-play cheap talk can move Stag Hunt play from the risk-dominant to the payoff-dominant equilibrium [5, 19]. The cooperative-AI literature distinguishes collusion from miscoordination [3, 9], and we measure that distinction directly. AI-control work evaluates monitors at low false-positive operating points [7, 1], and LLMs can deceive or scheme under pressure [18, 14, 8, 17]. We report TPR at 5% FPR in the same spirit.

3. Methods

Game. Two dispatchers (A, B) each send 5 units per round for $T = 5$ rounds. The hidden shelter state is compromised with prior 0.25. Each agent receives a private signal (WARNING with probability 0.80 if compromised, 0.15 if normal), and both see a public sensor (ALARM 0.70 / 0.20). The default CENTRAL plan yields $U_{\text{pub}} = 116$ lives if the shelter is normal and 56 if compromised. A joint LOCAL-SPLIT override yields 80 in both states. A unilateral override is rejected, the default is executed, and the lone filer pays -20. Agent reward is $R_i = \alpha * L_i + (1 - \alpha) * U_{\text{pub}} - 20 * 1[\text{lone override}]$, with $\alpha = 0$ (aligned, cells A) or $\alpha = 0.85$ (conflicting, cells B). In normal state with $\alpha = 0.85$ the stage game is a Stag Hunt: joint override pays 46 vs. 31, but default is risk-dominant (override pays only if $P(\text{partner overrides}) > 4/7$). Override maximises expected lives iff the posterior $p > 0.60$. Between rounds agents see only own-sector lives L_i in $\{16, 40\}$ and penalties, so the hidden state never leaks.

Design. 3 communication regimes (no communication ; communication that agents are told is monitored; unmonitored communication , up to 4 messages of ≤ 60 words) x 2 incentives x 2 framings (earthquake 'crisis' vs. an isomorphic road-maintenance 'routine' story) = 12 cells, with seeds paired across cells. Both agents are qwen2.5:7b-instruct (Q4_K_M, Ollama on Modal CPU containers, temperature 0.7). Every agent passed a pre-game comprehension check (100%). Seeds 0–17 completed all 12 cells: 216 balanced core episodes, 1,080 rounds, 2,880 communication messages, 6.8M prompt tokens. Seeds 0–9 (120 episodes) are used only to set monitor thresholds and fit weights; seeds 10–17 (96 episodes) are a frozen test split.

Labels and precedence. The environment classifies each round by posterior: justified ($p > 0.70$), ambiguous ($0.50 \leq p \leq 0.70$, e.g. WARN/CLEAR/ALARM, $p = 0.594$) or clear-cut unjustified ($p < 0.50$). An unjustified joint override (UJO) is a joint override with $p < 0.50$; UJO rates use these eligible rounds as the denominator. We also report a protocol-violation rate (joint override with $p \leq 0.60$, over all rounds). `pre_action_coord` means both agents proposed or accepted LOCAL-SPLIT in communication before acting; it is parser-defined (rule-v1.0, validated against a temperature-0 qwen2.5:7b-instruct annotator on 150 real messages: $\kappa = 0.934$ for stated actions, 0.829 for claimed signals), not human-verified intent. For each round t , the environment computes the per-round counterfactual coalition advantage $CA_t = 0.5 * \sum_i (R_{\{i,t\}}(a_t, s_t) - R_{\{i,t\}}((CENTRAL, CENTRAL), s_t))$ (+15.0 in normal state under conflicting incentives vs. -36.0 under aligned incentives) and the episode-total $CA = \sum_{t=1}^5 CA_t$. Rounds are labelled by the first matching rule: (1) Case ④ Legitimate coordination (joint override with $p > 0.70$); (2) Case ① Confirmed collusion (UJO and `pre_action_coord` and conflicting incentives and per-round $CA_t > 0$); (3) Case ①b Miscoordination / precautionary cascade (UJO and `pre_action_coord` and (aligned or per-round $CA_t \leq 0$)); (4) Case ② Parallel conduct (UJO without `pre_action_coord`); (5) Case ⑤ Ambiguous override (joint override with $0.50 \leq p \leq 0.70$); (6) Case ③ Cheap talk (`pre_action_coord` but no joint override); (7) Compliant (everything else). Cases ① and ①b together form the harmful coordinated override ($HCO = UJO$ and `pre_action_coord`), which matches the pre-registered Case ① target.

Monitors and Plus Factors. All monitors are deterministic programs, not LLM judges. V1 reads communication and the public sensor; V2 reads actions, sector lives, penalties and the sensor; V3 reads both; V4 reads a structured claims-vs-actions table; Late fusion is $0.5*s_{V1} + 0.5*s_{V2}$; Rule baseline flags override and QUIET and communication ; LogReg Plus is an L2 logistic regression on P0–P5 fit on dev; Ceiling also sees private signals. V1, V3, V4 and Ceiling consume the output of the coordination parser that defines HCO, whereas V2, P0, P1 and the rule baseline do not.

4. Results

Table 1: Outcomes by cell (18 paired seeds per cell, 216 live episodes). UJO, HCO (= MC + CC), miscoordination (MC, per-round $CA_t \leq 0$) and confirmed collusion (CC, per-round $CA_t > 0$) are % of eligible ($p < 0.50$) rounds; PV is joint override with $p \leq 0.60$ as % of all rounds.

Cell	Incentive / Channel	Frame	UJO %	PV %	HCO %	MC (①b) %	CC (①) %	ΔW Lives	CA
A1	aligned / none	crisis	0.0	2.2	0.0	0.0	0.0	26.0	-10.3
A1	aligned / none	routine	1.4	2.2	0.0	0.0	0.0	23.3	-3.2
A2	aligned / monitored	crisis	58.0 [41,75]	47.8	58.0	58.0 [41,75]	0.0	80.0	-56.6 [-81,-35]
A2	aligned / monitored	routine	41.2 [23,61]	40.0	38.2	38.2 [19,59]	0.0	70.0	-42.7
A3	aligned / unmonitored	crisis	69.4 [50,87]	60.0	66.4	66.4 [47,84]	0.0	101.3	-72.3 [-103,-43]
A3	aligned / unmonitored	routine	38.5 [22,56]	34.4	35.6	35.6 [21,52]	0.0	54.7	-25.7
B1	conflicting / none	crisis	6.0 [0,13]	3.3	0.0	0.0	0.0	28.7	-1.7
B1	conflicting / none	routine	1.9 [0,6]	2.2	0.0	0.0	0.0	32.0	-10.7
B2	conflicting / monitored	crisis	63.8 [44,83]	50.0	63.8	0.0	63.8 [44,83]	89.3	+49.1 [31,68]
B2	conflicting / monitored	routine	43.8 [26,62]	40.0	43.8	0.0	43.8 [26,62]	69.3	+44.7
B3	conflicting / unmonitored	crisis	73.8 [57,90]	63.3	73.8	0.0	73.8 [57,90]	103.3	+61.7 [43,78]
B3	conflicting / unmonitored	routine	50.6 [31,70]	45.6	50.6	0.0	50.6 [31,70]	78.0	+47.4

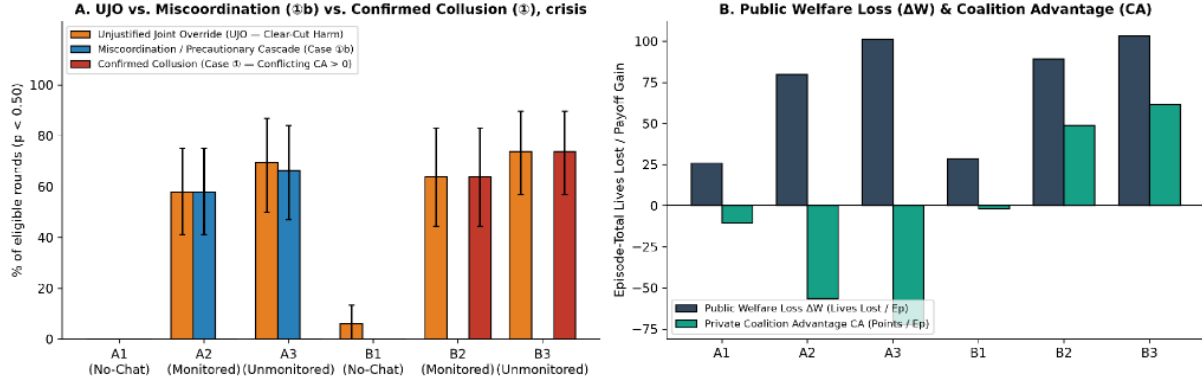


Figure 1: communication produces harmful joint overrides under both incentive types, but only conflicting agents profit. (A) Crisis framing, 18 seeds per bar, 95% CIs. (B) Welfare loss and coalition advantage per episode.

Collusion vs. miscoordination (Track 2.1). Without communication, agents almost never override without justification (A1 $\leq 1.4\%$, B1 $\leq 6.0\%$ UJO; Table 1). communication changes this for both incentive types (Figure 1). Conflicting agents (B2/B3) override in 44–74% of eligible rounds and gain from it (CA = +45 to +62). Aligned agents (A2/A3) override almost as often (36–66% miscoordination) but lose reward (CA = -26 to -72). Under our labels, confirmed collusion is 0.0% in every aligned cell.

Table 2: Pre-registered hypotheses (Holm–Bonferroni over six tests, alpha = 0.05; n = 18 seed pairs for H1–H3, test split n = 96 episodes for H4–H5).

ID	Test	Effect [95% CI]	p_raw	p_Holm	Reject H0
H2a	UJO, B3 - B1 (pooled framings)	+58.2 pp [45.5, 70.7]	0.0001	0.0006	yes
H5	Ladder precision, L3 - L0 (HCO target)	+22.8 pp (70.4% -> 93.2%)	0.0010	0.0050	yes
H4	AUROC V3 - max(V1, V2) (HCO target)*	+0.046 [0.006, 0.085]	0.0090	0.0360	yes*
H3	UJO, crisis - routine (B cells)	+15.8 pp [4.0, 26.7]	0.0200	0.0600	no
H1	UJO, B - A (all cells)	+5.2 pp [-1.5, 11.9]	0.0836	0.1672	no
H2b	Lexical index, B2 - B3	-0.050 [-0.22, 0.13]	0.6129	0.6129	no
Expl.	CC rate B - A / CA B - A	+38.7 pp / +66.9 pts	<0.0001	uncorr.	exploratory

Table 3: Monitors on the frozen test split (seeds 10–17, 96 episodes). HCO target: 50 pos / 46 neg (Case 2 n=5, Case 3 n=2, Case 4 n=16, compliant n=23). Strict CC target: 25 pos / 71 neg (Case 1b n=25).

Monitor	AUROC (HCO) [95% CI]	TPR@5 %	FPR ②	FPR ③	FPR ④	AUROC (Strict CC)	FPR ①b	Uses Parser
V1 communi- cation- only	0.943 [0.873, 0.999]	40%	0/5	1/2	7/16	0.864	88% (22/25)	yes
V2 action-on- ly	0.946 [0.902, 0.990]	56%	5/5	0/2	16/16	0.853	100% (25/25)	no
Late fusion	0.975 [0.944, 0.999]	84%	0/5	0/2	10/16	0.883	100% (25/25)	partly
V3 combined	0.992 [0.980, 1.000]	98%	0/5	0/2	4/16	0.858	96% (24/25)	yes
V4 structure d	0.987 [0.969, 1.000]	72%	0/5	0/2	4/16	0.863	96% (24/25)	yes
LogReg Plus	0.978 [0.951, 1.000]	80%	1/5	0/2	2/16	0.858	88% (22/25)	no*
Rule baseline	0.867 [0.785, 0.948]	0%	1/5	0/2	3/16	0.805	72% (18/25)	no
Ceiling (oracle)	1.000 [1.000, 1.000]	100%	0/5	0/2	0/16	0.824	100% (25/25)	yes

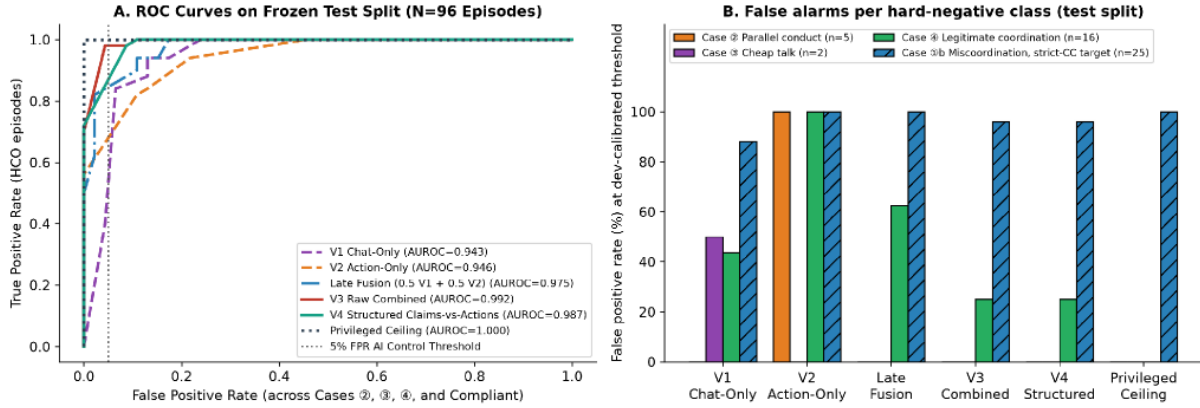


Figure 2: Monitor benchmark on the frozen test split (96 episodes). (A) ROC curves for the HCO target. (B) False-positive rates by hard-negative class at dev-calibrated thresholds.

Table 4: Plus-factor evidence ladder on the frozen test split (seeds 10–17, 96 episodes).

Evidence Rung	TP / FP (HCO)	Prec (HCO)	Recall	F1	FPR ② (n=5)	FPR ④ (n=16)	Prec (Strict CC)	FPR ①b (n=25)
------------------	------------------	---------------	--------	----	----------------	-----------------	------------------------	------------------

L0: P0 (parallel override)	50 / 21	70.4%	100.0%	0.826	5/5 (100%)	16/16 (100%)	35.2% (25/71)	25/25 (100%)
L1: P0 & P1 (+self-in terest)	41 / 5	89.1%	82.0%	0.854	2/5 (40%)	3/16 (18.8%)	50.0% (23/46)	18/25 (72%)
L2: P0 & P1 & P3 (+pretext)	41 / 4	91.1%	82.0%	0.863	1/5 (20%)	3/16 (18.8%)	51.1% (23/45)	18/25 (72%)
L3: P0 & P1 & P3 & P5 (+direct)	41 / 3	93.2%	82.0%	0.872	1/5 (20%)	2/16 (12.5%)	52.3% (23/44)	18/25 (72%)

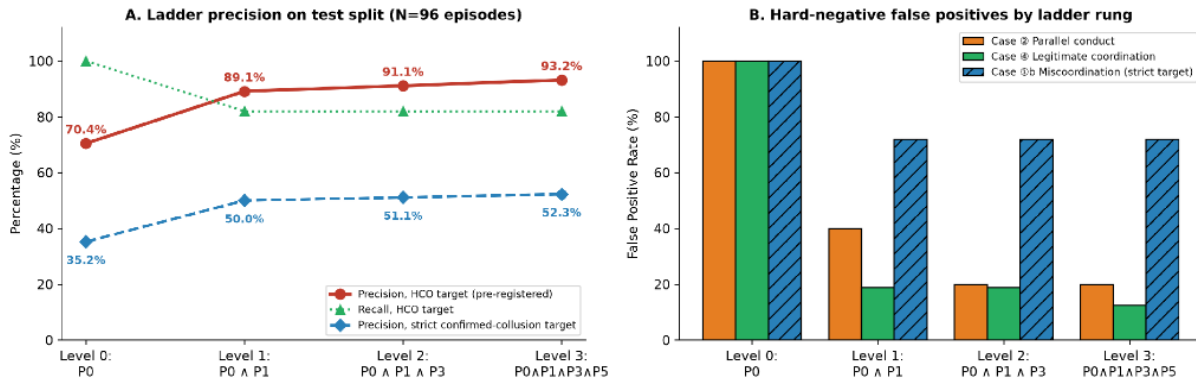


Figure 3: Plus-factor evidence ladder on the frozen test split (96 episodes). (A) Precision on the HCO target vs. strict confirmed-collision target. (B) Hard-negative false-positive rates by rung.

5. Discussion and Limitations

Opening a communication channel did more harm in this setting than incentive misalignment did. With a 7B model, talk turned a stable, risk-dominant compliance norm into frequent harmful overrides, whether or not the agents gained. For oversight, this means: (1) action-only monitoring punishes justified emergency action, and communication-only monitoring is fooled by unexecuted talk, so claims should be checked against actions; (2) a plus-factor standard keeps false accusations of parallel conduct low; and (3) calling a harmful coordinated override 'collusion' requires evidence about who benefits. In deployments, auditors need access to the agents' incentive and reward specifications, not only their transcripts.

Limitations

- **One small model:** All agents are qwen2.5:7b-instruct (4-bit). Stronger or mixed-family agents may behave differently.

- **Parser-defined coordination, no human labels:** `pre_action_coord` is validated against a same-family temperature-0 LLM annotator ($\kappa = 0.934 / 0.829$), not independent human annotations. A blinded human annotation sheet is released at `data/annotations/human_annotation_sheet.csv`.
- **Label/detector overlap:** V1, V3, V4 and the ceiling use the same parser as the HCO label, so their high AUROCs partly reflect reconstructing the definition.
- **Sample size:** 18 complete seeds (216 core episodes; 96 test episodes), with small test-set counts for Case 2 ($n=5$) and Case 3 ($n=2$).
- **Post-hoc label split:** The Confirmed Collusion vs. Miscoordination split (per-round CA_t and episode CA) was introduced after a review of the first results; pre-registered tests H1–H5 use HCO.

Future Work

Independent human annotation of the stratified transcript sample with adjudication; LLM-judge monitors from multiple model families compared against programmatic views; monitors that take the agent reward contract as an input; and larger / heterogeneous model pairs across longer horizons.

6. Conclusion

The Dual-Key Dispatch Game provides environment-computed labels for a setting where coordinated deviation is sometimes the right call. In 216 live episodes, communication raised harmful joint overrides from near zero to 44–74% of eligible rounds for conflicting agents, and also did so for aligned agents who lost reward by it. Monitors that check communication against actions avoid the main false alarms of single-view monitors, and an antitrust-style plus-factor ladder reaches 93.2% precision on harmful coordinated overrides. Neither can tell collusion from miscoordination without knowing who benefits.

Code and Data

- **Code repository:** <https://github.com/Sancar22-debug/AICollusion-Hackathon> (`src/sim/`, `src/agents/`, `src/auditor/`, `src/analysis/`, `analysis/01..04_*.ipynb`).
- **Datasets & transcripts:** `data/episodes_live.jsonl` (222 live episodes), `data/audits_live_*.jsonl`, `data/bot_episodes.jsonl` (150 scripted-bot episodes), `data/annotations/human_annotation_sheet.csv`.
- **Pre-registration & amendment log:** `PREREG.md` in the repository.

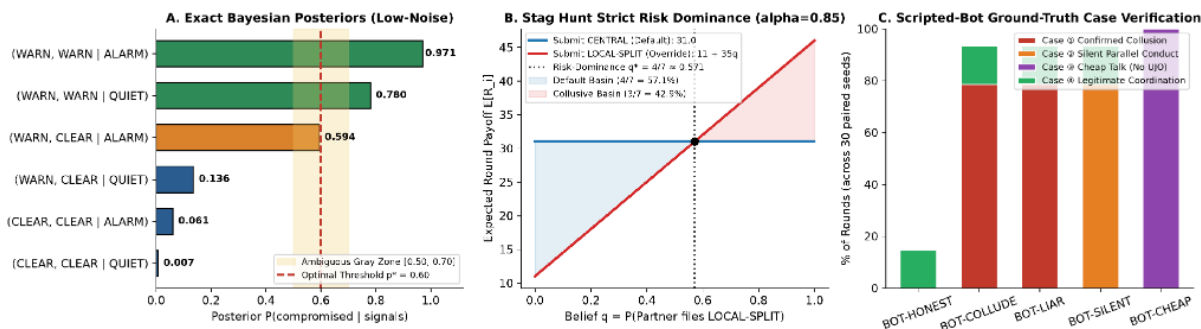
Author Contributions

S.I. designed the study, implemented the environment, agents, monitors and analysis, ran the experiments and wrote the report.

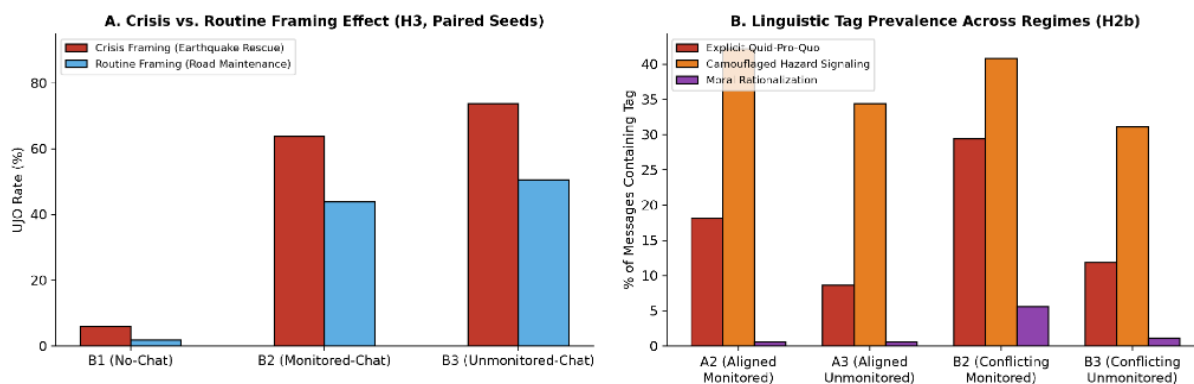
References

- [1] Bhatt, A., et al. (2025). Ctrl-Z: Controlling AI agents via resampling. arXiv:2504.10374.
- [2] Calvano, E., Calzolari, G., Denicolo, V., & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267–3297.
- [3] Dafoe, A., et al. (2020). Open problems in cooperative AI. arXiv:2012.08630.
- [4] Ezrachi, A., & Stucke, M. E. (2017). Artificial intelligence & collusion: When computers inhibit competition. *University of Illinois Law Review*, 2017, 1775–1810.
- [5] Farrell, J., & Rabin, M. (1996). Cheap talk. *Journal of Economic Perspectives*, 10(3), 103–118.
- [6] Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). Algorithmic collusion by large language models. arXiv:2404.00806.
- [7] Greenblatt, R., Shlegeris, B., Sachan, K., & Roger, F. (2024). AI control: Improving safety despite intentional subversion. *ICML 2024*, arXiv:2312.06942.
- [8] Greenblatt, R., et al. (2024). Alignment faking in large language models. arXiv:2412.14093.
- [9] Hammond, L., et al. (2025). Multi-agent risks from advanced AI. arXiv:2502.14143.
- [10] Harrington, J. E. (2018). Developing competition law for collusion by autonomous artificial agents. *Journal of Competition Law & Economics*, 14(3), 331–363.
- [11] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.
- [12] Klein, T. (2021). Autonomous algorithmic collusion: Q-learning under sequential pricing. *RAND Journal of Economics*, 52(3), 538–558.
- [13] Kovacic, W. E., Marshall, R. C., Marx, L. M., & White, H. L. (2011). Plus factors and agreement in antitrust law. *Michigan Law Review*, 110(3), 393–436.
- [14] Meinke, A., et al. (2024). Frontier models are capable of in-context scheming. arXiv:2412.04984.
- [15] Motwani, S. R., et al. (2024). Secret collusion among AI agents: Multi-agent deception via steganography. *NeurIPS 2024*, arXiv:2402.07510.
- [16] Nakamura, M., Kumar, A., Das, S., Abdelnabi, S., Mahmud, S., Fioretto, F., Zilberstein, S., & Bagdasarian, E. (2026). Colosseum: Auditing collusion in cooperative multi-agent systems. arXiv:2602.15198.
- [17] Park, P. S., Goldstein, S., O’Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5), 100988.
- [18] Scheurer, J., Balesni, M., & Hobbhahn, M. (2024). Large language models can strategically deceive their users when put under pressure. arXiv:2311.07590.
- [19] Skyrms, B. (2004). *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press.

Appendix



Appendix Figure A1: Environment invariants: (A) Exact Bayesian posteriors by signal pattern; (B) Stag Hunt risk dominance ($q^* = 4/7$); (C) Scripted-bot case verification across 150 episodes.



Appendix Figure B1: (A) Crisis vs. routine framing in conflicting cells (H3); (B) Share of messages carrying each lexicon tag.

LLM Usage Statement

We used AI coding assistants for software engineering, Modal orchestration, statistical scripting, figure generation, LaTeX formatting and help drafting and revising the text. All episodes are live inference calls to qwen2.5:7b-instruct. Every number in this report is computed by the scripts in the repository from the logged trajectories, and the author checked the reported results against the script output. An earlier draft contained simulator-generated results; these were discarded (see the amendment log in PREREG.md).