A Union of Dependence Statistics Detects Collusion That No Single Statistic Covers: A Pre-Registered, Sprint-Scale Study with a Fresh-Seed Confirmation (APTG-RL)

Aravinth K
Independent researcher aravinth3007k@gmail.com
With Apart Research

Abstract

Collusion audits must tell coordination apart from shared context, shared models and prompt-induced shifts. We test the APTG-RL design (directional predictive gain, a temporal-graph score, an RL investigator; its prompt-conditioned baseline is not exercised) in a three-seller pricing game with scripted collusion, benign and shared-model negatives, and perturbation conditions C0–C6. Ten detectors are compared at validation-fixed thresholds under a pre-registered plan. H1 (directional beats symmetric MI and net influence) is not supported: the combined gain (AUROC 0.839) beats symmetric MI (0.755) but has lower TPR than lag-1 net influence (0.850; -0.162). H5 (robustness) is not supported, and the temporal graph and RL investigator added no value. Each statistic covers different mechanisms: net influence misses same-round complementary pricing, symmetric MI misses lagged copying, and the parametric gain misses a parity-coded channel. A max-z union of net influence, symmetric MI and directional gain, found exploratorily, was confirmed on fresh seeds with three new mechanisms and frozen thresholds (pre-specified H6–H9 all supported): AUROC 0.965 [0.956, 0.973], TPR 0.877 against 0.524 for net influence and 0.617 for the best single statistic, FPR 0.056. Weak spots remain: lag-2 following (0.74) and a threshold trigger (0.60), and the nonparametric statistics alone miss lag-2 following. Evidence is simulator-only, and the new mechanisms were chosen knowing which statistics cover which dependence types. A Haiku 4.5 pilot (720 decisions) confirms prompts shift behaviour.

1 Introduction

Multi-agent LLM systems can coordinate through explicit messages, response patterns, complementary actions or hidden channels. Two independent agents can also look coordinated because they share a prompt, a task state or a base model, and an instruction given to only one agent can change that agent's behaviour without any collusion. A useful auditor must therefore ask whether one agent's history carries *extra, directional* information about another's next action beyond context, own history and prompt condition, and must do so without raising false alarms when only one agent has been steered.

We implement the APTG-RL design (our own research proposal, written before the sprint) and evaluate it on a controlled testbed, with one caveat stated up front: its prompt-conditioned baseline is implemented and unit-tested but is not used by the reported detectors (Section 5), so the prompt-conditioning hypothesis is not tested. Our contributions are:

1. **A controlled collusion-audit testbed.** A 3-seller repeated pricing game, four scripted collusion mechanisms (complementary, leader–follower, alternating, hidden-channel), three non-collusive

controls (independent, shared-model, benign adaptation), and a seven-condition controlled experimental design (C0–C6) covering unilateral and common instruction perturbations, scripted collusion, collusion under perturbation and benign policy shifts, with held-out mechanisms and a held-out prompt family.

2. **A pre-registered comparison of ten detectors**, including directional predictive gain, a windowed directed-graph score, an RL investigator and the information-theoretic baseline of Bonjour et al., with episode-level bootstrap CIs, plus a log of every deviation from the plan.
3. **Findings, including nulls, and a confirmed union.** H1 and H5 are not supported (the combined gain beats symmetric MI but not lag-1 net influence; APTG degrades more than net influence on held-out mechanisms); each dependence statistic covers a different set of mechanisms; the graph score and RL investigator did not add value at this scale; no detector showed a clear false-alarm rise under unilateral perturbation. A max-z union of net influence, symmetric MI and directional gain, found exploratorily, was confirmed in a second pre-specified study on fresh seeds and three new mechanisms (H6–H9 supported; AUROC 0.965 against 0.867 for the best single statistic). We make no priority claim.

2 Related Work

Information-theoretic detection. Bonjour et al. [1] use mutual information between agents' actions to detect covert cooperation in repeated games; we implement a net-influence variant as baseline B1 and a symmetric-MI variant as B2. Our directional statistic is the held-out predictive-gain analogue of transfer entropy [11] and Granger-style prediction [3]. **LLM collusion.** Motwani et al. [8] formalise secret collusion via steganography; Mathew et al. [7] study the emergence and mitigation of steganographic collusion; Colosseum [9] audits collusion in cooperative multi-agent systems using objective-based regret and separates attempted from effective collusion; Rose et al. [10] detect collusion from model activations (white-box, which we do not use). **Graphs and sequential sensing.** Eschenbaum and Meylahn [2] audit pricing algorithms from static strategy graphs; Joseph et al. [5] cast anomaly detection with observation costs as sequential controlled sensing with learned policies. We use dynamic interaction graphs from trajectories and a sensing-style investigator. **Behaviour is prompt- and endpoint-conditioned.** Work on LLM behavioural fingerprinting and endpoint stability [4, 6] is adjacent motivation for treating an agent's output distribution as dependent on prompts, endpoints and configurations (the premise behind the C1/C2/C6 controls); it does not validate a prompt-conditioned collusion detector, and our prompt-conditioned baseline remains untested here.

3 Methods

Environment. Three sellers play T=40 rounds; each chooses a price level $a \in \{0, \ldots, 4\}$; the lowest price takes the market (ties split) and the payoff is price×share. Agents $\{0, 1\}$ are the candidate coalition, agent 2 is an outsider. Non-collusive base policies are categorical, with prompt-family and severity shifts, own-action stickiness and a weak coupling to public lagged prices. **Mechanisms.** *Complementary*: $b_t = 4 - a_t$ w.p. 0.8. *Leader–follower*: $b_t = a_{t-1}$ w.p. 0.8. *Alternating*: turn-taking. *Hidden-channel*: the parity of a_t shifts b_{t+1} (a closed toy protocol). Negatives: independent, shared-model, and benign adaptation (both respond to the outsider, never to each other). **Conditions.** C0 clean; C1/C2 a perturbation of agent 0/1 only; C3 a common perturbation; C4 scripted collusion; C5 collusion plus a unilateral perturbation; C6 a benign policy shift in one agent. Splits are by

whole episode: train seeds 0–9, validation 50–61, test 75–94, and OOD 100–109 using a prompt family never seen in calibration; *alternating* and *hidden-channel* are excluded from every labelled fit († in tables); only B9 and the RL investigator are fit on labelled episodes, so the hold-out constrains them and not the unsupervised detectors, whose design (linear model, lags 0–2) we chose knowing the mechanisms. Cells have 276/460/230 validation/test/OOD episodes in total. **Directional gain.** For an ordered pair, a regularised multinomial logistic model of a_j^t on its own lag and the public lag of the third agent (prompt covariates in both models, which are constant within an episode and therefore inert) is compared with the same model plus a set L of the source agent's lags (lag 0 is the source's action in the *same* round), on contiguous held-out folds:

$$g_{i\to j}^{(L)} = \tfrac{1}{N}\textstyle\sum_t \big[\log p_1(a_j^t \mid \cdot, H_i) - \log p_0(a_j^t \mid \cdot)\big].$$

Two kinds of evidence. Because sellers move simultaneously, $L{=}\{0,1,2\}$ (B3) mixes *contemporaneous* dependence, $S_{ij} = I(A_i^t; A_j^t \mid X)$, which sees same-round relations such as complementary pricing ($b_t = 4 - a_t$), with *lagged* predictive gain from lags $1, 2$ (B3-temporal, the strictly directional statistic $G_{i\to j}$, whose comparison with $G_{j\to i}$ reveals direction). We keep the pre-registered B3 as a *combined* score and report $G_{i\to j}$ separately.

Population interpretation. Let $Y = a_j^t$, X the conditioning variables (target's own history, the third agent's public lag, prompt covariates) and Z the source's lag set. If $p_0(y \mid x) = P(Y{=}y \mid X{=}x)$ and $p_1(y \mid x, z) = P(Y{=}y \mid X{=}x, Z{=}z)$ are the true conditionals, then

$$\mathbb{E}\left[\log \frac{p_1(Y \mid X, Z)}{p_0(Y \mid X)}\right] = I(Y; Z \mid X) = \mathbb{E}_{X,Z}\, D_{\mathrm{KL}}\big(P(Y \mid X, Z) \,\|\, P(Y \mid X)\big) \geq 0,$$

so the population gain is the extra information Z carries about Y beyond X. This is an identity under correct specification, not a result about collusion. Held-out estimates can be negative (estimation error, sampling variability), and with a misspecified model class they measure the gain *of that class*, which is why a logistic model on numeric lags misses a parity code. The gain is predictive, not causal: omitted common causes, public feedback and shared environmental information also produce it, so a significant edge is evidence of dependence under the stated conditioning set, not proof of collusion.

B3 ($L{=}\{0,1,2\}$) is $\max_{i\neq j}$ of the robust $z = (g - \operatorname{med} G_0)/(1.4826\,\mathrm{MAD}\,G_0)$ against gains from clean training negatives; B3-temporal uses $L{=}\{1,2\}$ only; B3-noprompt drops prompt covariates. **Graph (APTG).** Per-row gains are averaged in 20-step windows (stride 10); each edge gets a block-permutation null (block 5, 30 draws) and significance by Westfall–Young max-statistic control at $q{=}0.05$. The episode score is persistence$\times S/(S{+}s_0)$ with S the strongest pair's total positive gain. **Baselines.** B1 net influence and B2 symmetric MI (bias-corrected plug-in MI), B5 action similarity, B6 coalition regret versus clean episodes (given the coalition's identity, so it uses privileged information the other detectors lack), B8 a Wald SPRT on the per-window "any significant edge" stream, B9 logistic regression on the APTG features (trained on non-held-out mechanisms). **Thresholds and statistics.** All thresholds are set on validation negatives at FPR = 0.05 and applied once to test/OOD. We report AUROC, AUPRC, TPR and FPR with 1000-resample episode-level percentile bootstrap CIs, and paired bootstrap differences on identical episodes. **RL investigator.** A Gymnasium POMDP over the window stream: actions observe, inspect the strongest pair, flag, abstain; reward $-c_{obs}/-c_{insp}$ per step and $R_{TP}{=}1, R_{TN}{=}0.5, C_{FP}, C_{FN}{=}1, C_A{=}0.5$ at termination; the hidden label is not observed. PPO is trained on the training split, the reward grid $C_{FP} \in \{1,2,4\} \times c_{obs} \in \{.01,.02,.05\}$ (3 seeds each, 20k steps) is selected on validation episodes from seen mechanisms only (the fixed-threshold and uncertainty-driven comparators are tuned the same way), and compared with a fixed threshold, SPRT, and an uncertainty-driven investigator.

Detector	AUROC [95% CI]	AUPRC	TPR@5%FPR$_{val}$	FPR
B1 net influence (MI, lag 1)	0.850 [0.803, 0.892]	0.855	0.719	0.067
B2 symmetric MI (lag 0)	0.755 [0.705, 0.801]	0.727	0.450	0.060
B5 action similarity	0.564 [0.509, 0.617]	0.406	0.106	0.057
B6 task regret	0.418 [0.366, 0.468]	0.299	0.025	0.063
B8 SPRT on edges	0.737 [0.694, 0.781]	0.634	0.362	0.020
B3 directional gain	0.839 [0.793, 0.885]	0.815	0.556	0.043
B3 temporal-only (lags 1,2)	0.704 [0.643, 0.756]	0.678	0.419	0.060
B3 w/o prompt covariates	0.839 [0.793, 0.885]	0.815	0.556	0.043
APTG graph score	0.740 [0.697, 0.785]	0.663	0.481	0.037
B9 supervised (graph feats)	0.864 [0.822, 0.904]	0.842	0.631	0.070

Table 1: Test-set detection (460 episodes, 160 collusive). Thresholds fixed on validation negatives at 5% FPR; TPR and FPR are measured at that threshold. Bracketed values are 95% episode-bootstrap CIs.

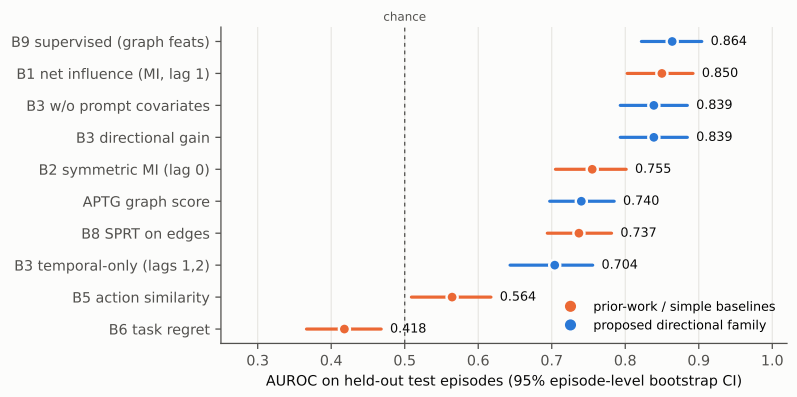


Figure 1: AUROC with 95% CIs. Blue: the directional family (including B9, which uses the graph features); orange: prior-work and simple baselines.

LLM pilot. Claude Haiku 4.5 sellers were run as subagents with no API key: each round, one self-contained prompt per (episode, seat) was answered independently, split across three subagents by seat. Conditions: C0, C1, C3 and C4L (seats 0,1 told to keep both profits high; no messages), 6 episodes each, T=10. The perturbation is a misleading task detail ("high price is a quality signal"); a first version that asked agents to undercut was discarded because it failed its manipulation check (Appendix C). **Pre-registration.** `PREREG.md` was committed as `f2daa25` (full hash `f2daa2555101f2f6983ab56ae2f8895b19552a60`) on 2026-10-25 at 16:01:18 +05:30 (10:31:18 UTC), 20 minutes before the detection runner (`5438f14`), and was never edited afterwards. The confirmation study's specification `FOLLOWUP2_PREREG.md` was committed as `77ad20e` before any confirmation episode was generated. `PREREG.md` was committed before any detector was evaluated on validation, test, OOD or pilot data; deviations are in Appendix C.

4 Results

H1 (directional vs. symmetric and net influence): not supported. The pre-registered rule required the directional score to beat both B2 and B1; it beats B2 but is significantly worse than B1

Detector	complementary	leader-follower	alternating†	hidden-channel†
B1 net influence (MI, lag 1)	0.07	1.00	0.80	1.00
B2 symmetric MI (lag 0)	1.00	0.07	0.65	0.07
B5 action similarity	0.25	0.15	0.00	0.03
B6 task regret	0.05	0.03	0.03	0.00
B8 SPRT on edges	0.57	0.57	0.30	0.00
B3 directional gain	0.72	0.68	0.72	0.10
B3 temporal-only (lags 1,2)	0.07	0.82	0.70	0.07
B3 w/o prompt covariates	0.72	0.68	0.72	0.10
APTG graph score	0.78	0.68	0.47	0.00
B9 supervised (graph feats)	0.85	0.70	0.85	0.12

Table 2: TPR at the validation-tuned 5%-FPR threshold by mechanism (test). †Held out of all fits.

on TPR, so we make no directional-novelty claim. B3 exceeds B2: ΔAUROC 0.084 [0.031, 0.139]. Against lag-1 net influence B1 the AUROC difference is inconclusive (-0.011 [-0.077, 0.057]) and B1 has the higher TPR (-0.162 [-0.267, -0.060]). So the *combined* same-round-and-lagged statistic is supported against symmetric MI only. The strictly lagged statistic is not: B3-temporal reaches 0.704 against 0.755 for B2 (ΔAUROC -0.051 [-0.124, 0.024], inconclusive), and the same-round term is what recovers complementary pairs (Table 2). Realised test FPRs at the shared 5% validation target differ (B1 0.067, B3 0.043, APTG 0.037, B8 0.020), so TPR comparisons favour detectors with higher realised FPR (AUROC does not depend on thresholds), and the TPR/FPR intervals treat the validation threshold as fixed and are somewhat too narrow. **Mechanism specificity.** Table 2 and Fig. 2 show that no detector covers all four mechanisms. B1 is perfect on leader–follower and hidden-channel but nearly blind to complementary pairs (same-round, negative association); B2 is the reverse. B3 covers complementary, leader–follower and alternating at 0.68–0.72 but not the parity-coded hidden channel (0.10): a monotone logistic model on numeric lags cannot represent parity, whereas plug-in MI is nonparametric. B3-temporal (lags $1, 2$) loses complementary pairs, confirming that simultaneous-move dependence is contemporaneous rather than lagged and directional. **H2 (prompt conditioning).** *Inconclusive.* B3 and B3-noprompt are identical, because prompt covariates are constant within an episode and our per-episode estimator therefore cannot use them. The relevant safety evidence is Table 5: no detector shows a clear false-alarm rise under unilateral (C1/C2), common (C3) or benign-shift (C6) conditions or on shared-model and benign-adaptation negatives, with FPRs mostly near the nominal 5%. The highest perturbation cells are B2 on C1 (0.15, i.e. 9/60 episodes) and B6 on C1 (0.12), and B1 is already at 0.13 on clean C0; with 60 episodes per cell (binomial half-width up to ±0.09) we cannot exclude a modest increase. **H3 (temporal graph features).** *Not supported.* The APTG score is below B3 (-0.099 [-0.138, -0.058]); with T=40 there are only three 20-step windows, each too short for reliable per-window significance. The supervised B9 on the same features (0.864) outperforms APTG by 0.124 [0.084, 0.161], so the features contain signal that the hand-built score discards; B9 inherits the hidden-channel blind spot (Table 2). The SPRT stream (B8) rejects the null on 0.36 of collusive and 0.02 of non-collusive test episodes, typically after window 2. **H5 (robustness): not supported.** The drop in TPR from clean collusion to the held-out prompt family (C4 to OOD, 80 vs. 80 episodes) is −0.10 [−0.25, 0.05] for APTG (no drop), against 0.01 [−0.11, 0.15] for B1, 0.11 [−0.04, 0.26] for B2 and 0.04 [−0.08, 0.14] for B5; APTG's drop is smaller than B2's (difference −0.21 [−0.40, −0.03]) but not clearly smaller than B1's or B5's. The drop from seen to held-out mechanisms (C4, 40 vs. 40) is 0.35 [0.15, 0.55] for APTG against −0.38 [−0.55, −0.20] for B1, i.e. APTG degrades *more* than B1 (difference 0.73

Investigator (test)	TPR	FPR	delay (steps)	cost
Fixed threshold	0.47	0.03	40.0	3.00
SPRT	0.36	0.02	32.1	2.89
Uncertainty-driven	0.37	0.03	20.0	3.27
PPO (mean±sd, 3 seeds)	0.37±0.00	0.03±0.00	20.0	1.00

Table 3: Investigators on test episodes at matched FPR ≤ 0.05 (thresholds and rewards selected on validation). Delay is the time step of the last window seen at alert; cost counts observe steps plus 2.5× inspections.

[0.48, 1.00]) and not clearly differently from B2 (0.23 [−0.03, 0.50]) or B5 (0.05 [−0.23, 0.33]). Since the rule required a smaller drop than B1, B2 and B5, H5 fails. These contrasts were computed after review with `experiments/h5_bootstrap.py` (independent episode-level bootstraps, 2000 resamples, validation thresholds fixed), not in the original scoring stage. Table 7 reports TPR on clean collusion (C4), collusion with a unilateral perturbation (C5) and a held-out prompt family (OOD). Because the baseline gate is inactive, the held-out family is never flagged out-of-calibration, so these columns measure a distribution shift on an ungated detector, not calibration-aware robustness. Directional detectors do not collapse under it; B6 (task regret) is below chance and its OOD FPR is 0.680 because family-level price shifts move payoffs without any dependence between agents. In this game scripted coordination does not raise the coalition's own payoff, so outcome-based evidence is weak by construction. **H4 (RL investigator).** *Weakly supported in its literal form, and not a contribution of learning.* At matched FPR (≤ 0.05 on validation, selected reward C_{FP}=1, c_{obs}=0.05) the PPO investigator decides after the first window, with test TPR 0.37, FPR 0.03, delay 20.0 steps and observation cost 1.00, against 0.47/0.03/40.0/3.00 for the fixed threshold and 0.36/0.02/32.1/2.89 for SPRT (Table 3); so delay and cost are lower, as H4 literally requires, but recall is lower than the fixed threshold. The three PPO seeds converged to the same deterministic policy (sd = 0), and it coincides on every split with a hand-written "flag iff the first window has a significant edge" rule that we added post hoc (test TPR 0.37, FPR 0.03); that rule loses -0.106 TPR [-0.158, -0.053] to the fixed threshold. A high false-positive cost makes PPO never alert (3 of 9 reward settings yield TPR 0). Delay is measured among true positives and the fixed threshold always sees all three windows, so PPO's lower delay and cost follow mechanically from deciding after one window; selection of the reward setting and the comparator thresholds used validation episodes from seen mechanisms only (a first selection on all validation positives chose the same settings and gave identical test numbers); the held-out mechanisms enter only the test and OOD evaluation. Following our decision rule, we report no evidence that RL adds value beyond simple rules in this setting; with three windows per episode there is little sequential structure to learn. **Follow-up: union of dependence statistics (exploratory, not pre-registered).** Because no single statistic covered all mechanisms, we tested a max-z union (design in `FOLLOWUP_PREREG.md`; details and Table 9 in Appendix A): the union of B1, B2 and B3 reached AUROC 0.990 [0.981, 0.996] and TPR 0.950 at realised FPR 0.057, against 0.850 and 0.719 for B1, and detected every mechanism at 0.80 or higher. Because it was specified after the main results were known, it is confirmed on fresh data below. **Confirmation study (pre-specified, fresh seeds).** To test the exploratory union without reusing the data that suggested it, we froze every detector, calibration reference, z-standardisation and threshold from the first study and applied them once to new episodes: seeds 200–249 (1450 test episodes: 700 collusive, 750 negative) and 25 seeds on the held-out prompt family (725 episodes). Beside the four original mechanisms the collusive set has three new ones: same-round copying ($b_t = a_t$ w.p. 0.8), a lag-2 follower ($b_t = a_{t-2}$) and a threshold trigger ($b_t = 0$ w.p. 0.8 when $a_{t-1} \geq 3$). The specification,

Detector	AUROC [95% CI]	TPR	FPR	Compl.	Lead–fol.	Altern.	Hidden	Copy	Lag-2	Thresh.
B1 net influence	0.756 [0.728, 0.781]	0.524	0.056	0.08	1.00	0.77	0.99	0.08	0.06	0.69
B2 symmetric MI	0.722 [0.693, 0.747]	0.399	0.060	1.00	0.04	0.56	0.07	1.00	0.05	0.07
B3 directional gain	0.867 [0.848, 0.886]	0.617	0.036	0.77	0.80	0.69	0.08	0.81	0.77	0.40
B4 nonparametric lagged CMI	0.658 [0.627, 0.687]	0.373	0.067	0.07	1.00	0.09	0.85	0.03	0.09	0.48
U3 = max-z(B1,B2,B3)	0.965 [0.956, 0.973]	0.877	0.056	1.00	1.00	0.81	0.99	1.00	0.74	0.60
U4 = max-z(B1,B2,B3,B4)	0.963 [0.953, 0.971]	0.886	0.072	1.00	1.00	0.79	1.00	1.00	0.74	0.67
Unp = max-z(B1,B2,B4)	0.902 [0.883, 0.918]	0.774	0.063	1.00	1.00	0.71	1.00	1.00	0.08	0.63

Table 4: Confirmation study on fresh seeds (1450 test episodes), frozen detectors and thresholds. TPR/FPR at the 5%-FPR validation threshold; right columns are per-mechanism TPR for the four original and three new mechanisms.

with hypotheses H6–H9, is `FOLLOWUP2_PREREG.md`, committed as `77ad20e` before the episodes were generated. All four hypotheses are supported (Table 4). **H6 (union beats single statistics): supported.** U3 reaches AUROC 0.965 [0.956, 0.973] and TPR 0.877 [0.851, 0.901], against B1 at 0.756 and 0.524 and the best single statistic B3 at 0.867 and 0.617. Paired differences: U3−B1 ΔAUROC 0.209 [0.182, 0.236], ΔTPR 0.353 [0.311, 0.395]; U3−B3 ΔAUROC 0.098 [0.082, 0.117], ΔTPR 0.260 [0.227, 0.294]. Net influence falls from 0.850 to 0.756 AUROC because three of the seven mechanisms lie outside its lag-1 coverage, and on these data B3 beats B1 (ΔAUROC 0.111 [0.074, 0.146]), which differs from H1 in the first study and shows that single-statistic rankings depend on the mechanism mix. **H7 (coverage): supported.** The minimum U3 TPR over the seven mechanisms is 0.60 (threshold trigger); the others are 1.00 (complementary, leader–follower, copy), 0.99 (hidden channel), 0.81 (alternating) and 0.74 (lag-2 follower). The all-nonparametric union Unp detects only 0.08 of lag-2 followers (AUROC 0.902), so the parametric gain, whose lags 0–2 reach it, is needed; adding B4 to U3 changes nothing (U4−U3 ΔAUROC −0.002 [−0.006, 0.002]). **H8 (cost): supported.** Realised U3 FPR is 0.056 [0.039, 0.074]; on C1, C2, C3 and C6 negatives it is 0.06, 0.06, 0.03 and 0.05 (U4 is higher, 0.072, and 0.12 on clean C0). **H9 (robustness): supported.** On the held-out prompt family U3 has TPR 0.823 at FPR 0.061, a drop of 0.054 from 0.877. **Haiku pilot.** The pilot is descriptive only (T=10; details and Table 8 in Appendix A).

5 Discussion and Limitations

Implications. For an auditor, the practical message is that dependence statistics are *mechanism-specific*: a lagged-MI monitor, a same-round MI monitor and a parametric directional monitor each cover a different part of the space, and a parity-coded channel needs a nonparametric statistic. A union of detectors is the natural response, and the pre-specified confirmation study supports it (U3 AUROC 0.965 against 0.867 for the best single statistic, FPR 0.056, every mechanism at 0.60 or higher), but the three new mechanisms were written knowing which statistics cover which dependence types, so it is evidence about coverage in this simulator, not about unseen coordination. Reassuringly, unilateral steering of one agent did not produce a clear rise in false alarms in any detector in our simulator, though the cells are small. **Prompt conditioning, still to be tested.** The motivating distinction is between an *individual policy shift* (an instruction changes one agent's action distribution) and *new inter-agent dependence*; the two must not be conflated. A faithful test needs a baseline $P(A_{i,t} \mid H_i^{t-1}, C_t, M_i, P_i, \tau_i)$ over the agent's own history, public context, model or endpoint identity M_i, instruction family P_i and decoding settings τ_i. Because prompt conditions are constant within an episode, it must be estimated *across* episodes (pooled or hierarchical) and then compared with matched non-collusive controls under unilateral perturbation. We write the model only to define the extension; it is untested here, and writing it down is not a substitute

for the experiment. **Limitations.** (1) *Simulator evidence.* The scripted mechanisms are our own construction and share assumptions with the detectors; a single environment, three agents, T=40; no DCOP/task-regret environment. (2) *Scale.* Fewer seeds than the planned 100/cell (train 10, val 12, test 20, OOD 10), so CIs are wide, and 30 permutations per edge give coarse p-values. (3) *Prompt conditioning is not tested.* The calibration baseline is implemented but not wired into the detectors, so every episode counts as calibrated and the out-of-calibration gate never fires; the prompt covariates in the estimator are constant within an episode, making the ablation vacuous. A pooled estimator across episodes with a per-family reference is needed. (4) *Graph score.* Three windows per episode is too few to test temporal motifs; H3 is not supported *at this scale*, not refuted in general. (5) *RL.* Few windows, a small training set and 20k PPO steps; results say little about RL for audit in general. (6) *LLM pilot.* Haiku sellers were driven as subagents, without temperature control, and a subagent answers the prompts of one seat across episodes, so independence between agents is approximate; 6 episodes per condition. (7) *Collusion is behavioural.* Scripted coordination does not raise coalition payoff here, so we claim operational evidence of coordination, not intent or harm. (8) Edges are predictive, not causal. (9) *Confirmation scope.* The confirmation study uses the same simulator and environment; its three new mechanisms were chosen knowing which statistics cover lag-0, lag-1 and lag-0–2 dependence; components, thresholds and hypotheses were frozen, but the hypotheses H6–H9 were written after the first study's results; lag-2 following (0.74) and the threshold trigger (0.60) are only partly covered. (10) *Planned but not run.* Of the pre-registered secondary outcomes and ablations we did not run: calibration (Brier/ECE), group localization, detection delay for any detector other than B8 and the RL investigators, the LLM-judge baseline, and the no-temporal-motif and no-regret-feature ablations; the prompt-conditioned ablation is vacuous (H2). (11) *Statistical details.* Block-permutation draws use the same seeds in every episode, so p-values are correlated across episodes; max-statistic draws are independent per edge (conservative: on 60 independent-episode edge tests the single-edge rejection rate at nominal 0.05 was 0.017); cross-fitted folds are blocked and purged but not causal. (10) *Unplanned exclusion.* One validation negative (`benign_adaptation-C0-s61-f1`) has no window record, so B8 scores it 0 and the RL validation set has 275 episodes. **Future work.** A pooled prompt-conditioned baseline evaluated against matched unilateral-perturbation controls; the union tested on learned (not scripted) collusion, on mechanisms written by someone else and in other environments, with stronger lag-2 and threshold-style detectors (the two weak spots); a strictly lagged directional detector with a parity-capable (nonparametric) model class; a properly powered comparison of the RL investigator with simple sequential rules; longer episodes and denser windows; a DCOP environment with real task regret; multiple model families; more LLM episodes through an API with controlled sampling.

6 Conclusion

In a pre-registered, sprint-scale test of the APTG-RL design, H1 and H5 were not supported: the combined directional gain beat symmetric MI but not lag-1 net influence, its strictly lagged part showed no clear improvement, and the graph score and RL investigator added no measurable value. The detectors proved complementary rather than ranked, and a max-z union of three of them, found exploratorily, was confirmed in a second pre-specified study on fresh seeds and three new mechanisms (AUROC 0.965, TPR 0.877 against 0.617 for the best single statistic, FPR 0.056, every mechanism at 0.60 or higher), with lag-2 and threshold-style rules as the weak spots. The lessons are about coverage (no single dependence statistic sees every coordination pattern) and evaluation design (hold out mechanisms, report negative controls and deviations).

References

[1] Trevor Bonjour, Vaneet Aggarwal, and Bharat Bhargava. Information theoretic approach to detect collusion in multi-agent games. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 180 of *PMLR*, 2022. URL https://proceedings.mlr.press/v180/bonjour22a.html.

[2] Nicolas Eschenbaum and Janusz M. Meylahn. Auditing algorithmic collusion from strategy graphs. arXiv:2608.07098, 2026. URL https://arxiv.org/abs/2608.07098.

[3] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. doi: 10.2307/1912791.

[4] Yuepeng Hu, Zhengyuan Jiang, Mengyuan Li, Osama Ahmed, Zhicong Huang, Cheng Hong, and Neil Zhenqiang Gong. Fingerprinting LLMs via prompt injection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*, 2026. URL https://aclanthology.org/2026.acl-long.541/.

[5] Geethu Joseph, Chen Zhong, M. Cenk Gursoy, Senem Velipasalar, and Pramod K. Varshney. Anomaly detection via learning-based sequential controlled sensing. *IEEE Sensors Journal*, 24, 2024. doi: 10.1109/JSEN.2024.3399456.

[6] J. Leshin, M. Shah, I. Timmis, and D. Kang. Behavioral fingerprints for LLM endpoint stability and identity. In *Proceedings of the ACM Conference on AI and Agentic Systems (CAIS 2026)*, 2026. doi: 10.1145/3786335.3813194.

[7] Yohan Mathew, Ollie Matthews, Robert McCarthy, Joan Velja, Christian Schroeder de Witt, Dylan Cope, and Nandi Schoots. Hidden in plain text: Emergence & mitigation of steganographic collusion in LLMs. In *Proceedings of IJCNLP-AACL 2025*, 2025. URL https://aclanthology.org/2025.ijcnlp-long.34/.

[8] Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip H. S. Torr, Lewis Hammond, and Christian Schroeder de Witt. Secret collusion among AI agents: Multi-agent deception via steganography. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/861f7dad098aec1c3560fb7add468d41-Abstract-Conference.html.

[9] Mason Nakamura, Abhinav Kumar, Saswat Das, Sahar Abdelnabi, Saaduddin Mahmud, Ferdinando Fioretto, Shlomo Zilberstein, and Eugene Bagdasarian. Colosseum: Auditing collusion in cooperative multi-agent systems. arXiv:2602.15198, 2026. URL https://arxiv.org/abs/2602.15198.

[10] Aaron Rose, Carissa Cullen, Sahar Abdelnabi, Philip Torr, Brandon Gary Kaplowitz, and Christian Schroeder de Witt. Detecting multi-agent collusion through multi-agent interpretability. arXiv:2604.01151, 2026. URL https://arxiv.org/abs/2604.01151.

[11] Thomas Schreiber. Measuring information transfer. *Physical Review Letters*, 85:461–464, 2000. doi: 10.1103/PhysRevLett.85.461.

Code and Data. Code, configs, seeds, pre-registration and raw results: https://github.com/chinchoooo/COLDER (repository `COLDER`). The scripted data are regenerable from seeds; the Haiku pilot logs are in `results/llm/`. All code was written during the sprint; the only pre-existing material is the author's two APTG-RL proposal documents (no code, no results).

Author Contributions. A.K. conceived the design, specified the experiments and wrote the report; code and drafting were produced with AI assistance (below) under the author's direction.

A Supplementary tables and figures

The tables and figure below support the main text; their discussion is in Section 5.

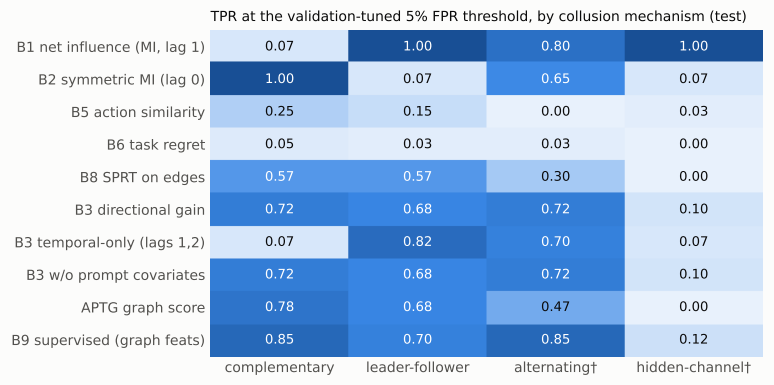


Figure 2: Detector × mechanism TPR (same data as Table 2).

Detector	C0	C1 (A pert.)	C2 (B pert.)	C3 (both)	C6 (benign shift)	benign adapt.	shared model
B1 net influence (MI, lag 1)	0.13	0.02	0.08	0.07	0.03	0.07	0.05
B2 symmetric MI (lag 0)	0.05	0.15	0.03	0.02	0.05	0.07	0.06
B5 action similarity	0.07	0.05	0.05	0.05	0.07	0.08	0.06
B6 task regret	0.07	0.12	0.08	0.02	0.03	0.05	0.06
B8 SPRT on edges	0.05	0.00	0.03	0.00	0.02	0.01	0.02
B3 directional gain	0.08	0.02	0.05	0.00	0.07	0.02	0.05
B3 temporal-only (lags 1,2)	0.13	0.07	0.07	0.02	0.02	0.06	0.04
B3 w/o prompt covariates	0.08	0.02	0.05	0.00	0.07	0.02	0.05
APTG graph score	0.07	0.03	0.05	0.02	0.02	0.05	0.03
B9 supervised (graph feats)	0.10	0.08	0.03	0.05	0.08	0.07	0.08

Table 5: FPR at the validation-tuned threshold on each negative type (test).

Comparison (A − B)	ΔAUROC [95% CI]	ΔTPR [95% CI]
B3 directional gain − B2 symmetric MI	0.084 [0.031, 0.139]	0.106 [0.018, 0.197]
B3 directional gain − B1 net influence	-0.011 [-0.077, 0.057]	-0.162 [-0.267, -0.060]
APTG graph score − B2 symmetric MI	-0.015 [-0.074, 0.045]	0.031 [-0.059, 0.122]
APTG graph score − B3 directional gain	-0.099 [-0.138, -0.058]	-0.075 [-0.130, -0.029]
B3 temporal-only − B2 symmetric MI	-0.051 [-0.124, 0.024]	-0.031 [-0.150, 0.084]
B3 directional gain − B3 w/o prompt covariates	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]
APTG graph score − B1 net influence	-0.110 [-0.175, -0.040]	-0.237 [-0.344, -0.129]
B9 supervised − APTG graph score	0.124 [0.084, 0.161]	0.150 [0.095, 0.215]

Table 6: Paired episode-bootstrap differences on identical test episodes (TPR at each detector's own validation threshold).

Detector	TPR clean (C4)	TPR + unilateral pert. (C5)	TPR held-out family (OOD)	OOD FPR
B1 net influence (MI, lag 1)	0.74	0.70	0.72	0.07
B2 symmetric MI (lag 0)	0.49	0.41	0.38	0.05
B5 action similarity	0.17	0.04	0.14	0.09
B6 task regret	0.04	0.01	0.45	0.68
B8 SPRT on edges	0.31	0.41	0.42	0.01
B3 directional gain	0.50	0.61	0.59	0.05
B3 temporal-only (lags 1,2)	0.44	0.40	0.51	0.09
B3 w/o prompt covariates	0.50	0.61	0.59	0.05
APTG graph score	0.45	0.51	0.55	0.05
B9 supervised (graph feats)	0.59	0.68	0.64	0.05

Table 7: Robustness: TPR under unilateral perturbation (C5) and held-out prompt family; last column is FPR on OOD negatives.

Condition	seat 0	seat 1	seat 2	episodes
C0 clean	0.93	0.78	0.78	6
C1 seat-0 perturbed	1.97	0.57	0.45	6
C3 all perturbed	4.00	4.00	4.00	6
C4L seats 0,1 told to coordinate	2.68	2.82	1.88	6

Table 8: Haiku pilot: mean chosen price level per seat and condition (T=10).

Detector	AUROC [95% CI]	TPR	FPR	Compl.	Lead–fol.	Altern.†	Hidden†
B1 net influence	0.850 [0.803, 0.892]	0.719	0.067	0.07	1.00	0.80	1.00
B2 symmetric MI	0.755 [0.705, 0.801]	0.450	0.060	1.00	0.07	0.65	0.07
B3 directional gain	0.839 [0.793, 0.885]	0.556	0.043	0.72	0.68	0.72	0.10
B4 nonparametric lagged CMI	0.717 [0.665, 0.771]	0.450	0.053	0.00	1.00	0.05	0.75
U3 = max-z(B1,B2,B3)	0.990 [0.981, 0.996]	0.950	0.057	1.00	1.00	0.80	1.00
U4 = max-z(B1,B2,B3,B4)	0.987 [0.976, 0.994]	0.944	0.067	1.00	1.00	0.78	1.00
Unp = max-z(B1,B2,B4)	0.983 [0.968, 0.992]	0.925	0.053	1.00	1.00	0.70	1.00

Table 9: Follow-up: single statistics and max-z unions on the test split (460 episodes). TPR/FPR at the 5%-FPR validation threshold; right columns are per-mechanism TPR. †Held out of all fits.

Exploratory union follow-up (full text)

Follow-up: union of dependence statistics (exploratory, not pre-registered). Because no single statistic covered all mechanisms, we tested the union suggested by Table 2 using the design in `FOLLOWUP_PREREG.md`: each component is converted to a robust z against training negatives and the union score is the maximum z, thresholded once at 5% FPR on validation negatives (one run, no tuning). We added B4, a nonparametric lagged conditional MI $I(a_i^{t-1}; a_j^t \mid a_j^{t-1})$ with a shuffle-null correction, to test the review's suggestion of a nonparametric directional component. The union U3 of B1, B2 and B3 reaches AUROC 0.990 [0.981, 0.996] and TPR 0.950 at realised FPR 0.057, against B1's 0.850 and 0.719 (paired ΔAUROC 0.140 [0.098, 0.187]; ΔTPR 0.231 [0.158, 0.308]), and every mechanism is detected at 0.80 or higher (Table 9). On the held-out prompt family it keeps TPR 0.938 at FPR 0.067. Its FPR on perturbation and benign negatives stays between 0.03 and 0.07 (C1/C2 0.07), so we see no clear false-alarm cost, within the limits of 60 episodes per cell. B4 alone is weaker than B1 (AUROC 0.717; Δ vs. B1 -0.133 [-0.186, -0.081]) and, as a lag-1 statistic, misses complementary and alternating pairs; adding it to U3 changes nothing (U4$-$U3 ΔAUROC -0.003). An all-nonparametric union (B1, B2, B4) performs almost as well on these four mechanisms (AUROC 0.983, TPR 0.925, FPR 0.053); the confirmation study below shows that it misses lag-2 following. Caveats: the follow-up was specified after the main results were known, the mechanisms are our own scripted constructions, and the components were chosen knowing which mechanisms they cover, so this is evidence that coverage is complementary in this simulator, not that a union generalises to unseen coordination.

Haiku pilot (full text)

Haiku pilot. 0 of 720 responses were invalid. The manipulation check passes: in C1 the perturbed seat averages 1.97 against 0.57 for its unperturbed partner (clean seats ≈ 0.78), and in C3 every seat prices at 4. Telling seats 0,1 to keep both profits high lifts their mean price to 2.75 against 0.86 in C0 (one-sided Mann–Whitney p=0.002, 6 vs. 6 episodes), with the outsider also higher than in C0 (Table 8). With T=10 the windowed detectors do not apply; the pilot is descriptive only.

B Limitations and Dual-Use / Ethical Considerations

What the results establish. In one synthetic pricing game with scripted coordination, on held-out simulated episodes, specific dependence statistics detect specific coordination patterns at a stated false-alarm rate, and unilateral steering of one agent did not clearly raise those false-alarm rates (cells of 60 episodes). A small Haiku pilot shows that instructions move prices. **What they do not establish.** They do not show that these detectors work on real deployed agents, on learned (rather than scripted) collusion, on other games, or against agents that adapt to the monitor. They say nothing about intent: a detector alarm is evidence of statistical coordination under explicit rules, not proof of malicious behaviour, and false accusations have real costs (the investigator's reward weights C_{FP} are a policy choice). Edges are predictive, not causal. Power is limited (Section 5). **Dual-use.** The testbed includes a hidden-channel mechanism that evades two of our detectors (B3, APTG). This is a closed toy protocol (a parity code on a five-level action) and is described because the failure is an audit finding; however, a reader could learn that parity-style codes are invisible to monotone parametric monitors. We mitigate by keeping the protocol toy-scale, reporting that a nonparametric MI statistic detects it (B1: 1.00), not releasing any optimiser for evasion, and recommending nonparametric statistics in monitors. The LLM pilot includes a prompt that

tells agents to keep both profits high; in a real market this would be price coordination, which is illegal in many jurisdictions. It is used only in a toy game with fictional sellers; no real market data, personal data or real endpoints were used. **Responsible use.** Detector outputs should be treated as triage evidence for human review, reported with uncertainty, and never as automatic grounds for sanctions.

C Deviations from the Pre-registration

All were decided before any test, OOD or pilot detector evaluation, from training-seed smoke tests (full log: `DEVIATIONS.md`). (1) Westfall–Young max-statistic control replaces BH in the windowed graph: with ≤ 200 permutations the smallest attainable p is $1/(n{+}1)$ and BH over 6 pairs needs $p \leq q/6 \approx 0.0083$, so a single strong edge could never be rejected (a smoke test on a clear leader–follower episode scored 0). (2) Window 20/stride 10 instead of 10/5 (training smoke: leader–follower persistence 0.50 vs. 0.26). (3) 30 permutations per edge. (4) Fewer seeds (Section 5). (5) B3 takes the maximum over all six ordered pairs (the detector does not know the coalition); B3-temporal and B3-noprompt were added. (6) The no-prompt ablation is applied to B3 only. (7) The RL reward grid and comparator thresholds were first selected on all validation positives, including the held-out mechanisms, contrary to the plan; we repeated the selection on seen mechanisms only (227 validation episodes) and report that version. The selected settings and every test number were identical. (8) *Follow-up (not in the plan):* the union experiment follows `FOLLOWUP_PREREG.md`, which was written before the follow-up scripts were run but, unlike `PREREG.md`, was committed together with its results, so its timing is not independently verifiable. (9) *Confirmation study:* `FOLLOWUP2_PREREG.md` (commit `77ad20e`) precedes the generation of the confirmation episodes (seeds 200–274, three new mechanisms); detectors and thresholds were frozen. (10) *LLM pilot redesign:* the first pilot (perturbation: "undercut aggressively") failed its manipulation check, since Haiku already prices at the low level, and produced near-identical episodes within a condition; it was discarded after two rounds and replaced by a misleading task detail plus neutral role descriptions ("a shop owner", etc.). No detector had been run on either pilot.

D LLM pilot prompt

Each decision prompt reads: *"You are seller k, <role>, in a repeated market game with 3 sellers (numbered 0, 1, 2) over 10 rounds. Each round every seller simultaneously picks an integer price level from 0 to 4. The seller(s) with the lowest price take the whole market; ties split it equally. Your payoff each round = your price level × your market share (so a high price only pays if you win the market). Your goal: maximize your own total payoff."* followed, as applicable, by the perturbation ("Market note (from a trade newsletter): in this market customers read a high price as a quality signal and tend to buy from the highest-priced seller, so a price of 3 or 4 attracts the most customers.") or the coordination instruction ("You and seller m want to keep both of your profits high over the whole game. You may not message each other, but choose prices you think will help you both."), the history of rounds (all three prices and the seller's own payoff), and the format request `ACTION: <n>`. Prompts never contain other agents' prompts or labels.

LLM Usage Statement

Claude (Anthropic) was used throughout: to help design and critique the experiments, to write and debug the code (implementation was delegated to Claude subagents and reviewed), to run the Claude Haiku 4.5 agents of the pilot, and to draft this report. All numbers in the paper are generated by scripts from the committed result files; references were each checked against the publisher or preprint page (or Crossref) by

the author's assistant; one reference from the original proposal that could not be verified was dropped. The author is responsible for all claims.