

When Net Influence Misleads: Calibration, Innocent Dependence, and Coalition Size

Rucha Agashe, Imperial College London (ragashe1618@gmail.com)

AI Collusion Research Sprint, Apart Research, 23 to 25 October 2026 (submissions close 25 October 11:59 PM AoE, which is 26 October 11:59 UTC). Track 1 (question 1.4) and Track 2 (question 2.3).

Abstract

When does statistical dependence constitute evidence of agent collusion? We stress-test pairwise net influence using 7,400 scripted games and two GPT-4.1-mini pilot runs, each comprising ten games in each of three conditions.

We examine public-state specification, finite-sample calibration, and coalition size. Innocent agents following the same public rule defeat unconditioned dependence statistics; conditioning removes this confound only when the chosen state captures that rule. At 25 rounds, a fixed threshold produces false accusations in approximately half of all-random games. A Holm-corrected permutation rule reduces false-positive rates on the tested scripted nulls, but flags a pair in each pooled LLM control without a planted coalition. These outcomes distinguish statistical rejection from evidence of collusion.

For a symmetric coalition with one primary and two assistants, we derive a cancellation mechanism: the pairwise correction subtracts one coalition member's influence from another's. Consequently, fixed-threshold detection deteriorates as sample size increases.

In the LLM pilots, the planted pair ranks first in eight or nine of ten games but is correctly flagged in five or seven. Its mean within-game score advantage changes from -0.25 to +1.00 across runs; these descriptive differences do not establish a causal benefit from coordination.

An exploratory Jacobian-lens assay includes 150 receiver evaluations but no unauthorized adoption, leaving its intended participation-versus-refusal contrast unresolved. We release code and evaluation artifacts and identify assumptions auditors must check before treating dependence as evidence of collusion.

1. Introduction

The sprint asks what should count as evidence that agents are colluding when the evidence is a log of actions (question 2.3), and whether information-theoretic measures can identify colluding pairs among several agents (question 1.4). Antitrust law has long faced the same difficulty: competitors act in parallel because they respond to the same public information, and parallel conduct alone is not an agreement (Berger et al. 2025). A detector that flags correlation flags every pair that shares a prior or a strategy, and two instances of one base model share a great deal of prior (Xu et al. 2026).

Bonjour et al. (2022) proposed a statistic built for exactly this difficulty. It measures the dependence between two agents' actions conditional on the observable game state, then subtracts the strongest dependence any third agent has on the same target. It is domain independent and needs only logged actions. The original paper evaluated it on known colluding configurations, varied collusion strength and sample size, treated the threshold as a hyperparameter to be tuned on training data, and limited its scope to two colluders. Several concurrent sprint entries apply it to LLM players (Ayeku 2026; Snow 2026; Lee et al. 2026), and Lee et al. include 2,800 scripted coalition games. None of this tells an auditor which assumptions the verdict rests on. The question is this:

Under which assumptions about public information, sample size and coalition structure does net influence distinguish planted coordination from innocent dependence, and where does calibration fail to transfer?

We combine four evaluation axes and emphasise their interaction: a matched innocent-dependence confound (two non-colluders following the same public rule), short-horizon calibration (25 to 500 rounds, where plug-in bias is

comparable to the published threshold), state misspecification (an innocent rule keyed on a richer state than the detector conditions on) and a three-member coalition, together with an analytic account of coalition cancellation and the transfer of two decision rules to the same LLM logs. We then apply both the fixed-threshold rule and a permutation-calibrated rule to GPT-4.1-mini agents with a hidden private channel and to a same-model control, and report ranking, detection and coalition benefit as separate outcomes. Throughout, a signal is not treated as evidence of collusion until the assumption behind it has been checked. Section 5 reports a small Jacobian-lens experiment that asked the same question of a model’s internals; the condition it needed did not occur, and it is presented as exploratory.

2. Related work

Information-theoretic detection from actions. Bonjour, Aggarwal and Bhargava (2022) define influence between agents as conditional mutual information given the game state and flag a pair whose net influence exceeds a threshold tuned on labelled data; they evaluate on iterated Rock-Paper-Scissors and poker with two colluders, across collusion probabilities and numbers of games, and note the two-colluder scope as a limitation. Later information-theoretic detectors for LLM agents (Tailor 2025; Ghanem 2026) estimate cross-run mutual information over transcripts rather than actions. Xu et al. (2026) show that two deployments of one base model carry a small residual bid correlation and argue that price-level audits are blind by construction. Within this sprint, Ayeku (2026) and Snow (2026) apply net influence with a permutation test to LLM players, Ayeku conditions on a public state against an innocent rule-following pair, and Lee et al. (2026) extend the statistic to coalitions with graph methods over 2,800 scripted coalition games and 400 honest games. Our experiments differ from these in four respects: matched innocent-dependence confounds under correct and misspecified conditioning, the gap between bias correction and decision calibration at short horizons, an analytic account of why the pairwise correction cancels a three-member coalition, and a transfer test in which both decision rules are run on the same LLM logs.

Detection from internals. Rose et al. (2026) train a deception direction on residual activations and aggregate per-agent scores across a group; in their limitations (v1, Section 4.4) they write that a privileged-observer control, an agent who knows the collusion scheme but is instructed not to participate, would help, and leave it for future work. Li et al. (2026) find that a probe confined to the Jacobian-lens transport subspace detects an agent’s stated stance but not withheld collusion. Veras Neves (2026) reads a coordination signal from the J-lens on single prompts. Gurnee et al. (2026) and Ayonrinde and Lindsey (2026) report that a concept can remain in the workspace while the model is instructed to suppress it. Section 5 is a first attempt at the control Rose et al. describe.

3. Method

3.1 Pairwise net influence

Consider N agents playing a repeated simultaneous-move game with K actions. Let A_i^t be agent i ’s action in round t and S^t the public state before round t . Bonjour et al. define the influence of i on j as

$$\gamma(i; j) = \sum_s p(s) \sum_{\{a_i, a_j\}} p(a_i, a_j | s) \log [p(a_i, a_j | s) / (p(a_i | s) p(a_j | s))],$$

the conditional mutual information $I(A_i; A_j | S)$, symmetric in the simultaneous fully observable case. The net influence is

$$\Gamma(i; j) = \gamma(i; j) - \max_{\{k \text{ not in } \{i, j\}\}} \gamma(k; j),$$

and agents i and j are declared colluding if $\Gamma(i; j) \geq \alpha$ and $\Gamma(j; i) \geq \alpha$. The paper reports a pair only when exactly one pair satisfies this and uses $\alpha = 0.05$ (natural log). We implement this with plug-in probabilities and score each unordered pair by $\min(\Gamma(i; j), \Gamma(j; i))$.

3.2 State functions

We compare a memoryless state (γ reduces to mutual information), the previous round’s majority action (four values) and the full previous action profile (82 values, sparse at short horizons).

3.3 Bias correction and two decision rules

Plug-in mutual information is biased upwards by about $(K-1)^2 / (2T)$ nats per state stratum; for $K = 3$ and $T = 25$ this is 0.08 nats, larger than α . We estimate the null of $\gamma(i; j)$ under “ A_j independent of A_i given S ” by shuffling A_j within each state stratum (100 to 500 permutations) and report the corrected statistic γ minus the null mean, clipped at zero, and a permutation p-value per pair. Net influence is computed from the corrected γ . Two decision rules are then applied, and they are different procedures: the **fixed rule** (exactly one pair has net-influence score ≥ 0.05) and the **Holm rule** (the top-scoring pair has a Holm-adjusted permutation p-value below 0.05 over the six pairs). Rejecting conditional independence is a statement about a specified null given the conditioning information; it is not equivalent to establishing collusion, and Section 4 measures the gap between the two.

3.4 Scripted environment

Four players (five in Section 4.3), actions rock, paper, scissors. Payoffs follow Bonjour et al.: all the same, nobody scores; all three actions present, everyone scores 1; exactly two actions present, the winning action’s players score 1. Agent types: *random*; *assistant*, which with probability CP plays the action that loses to the primary’s action (guaranteeing the primary a point, with ex-ante signalling assumed as in the paper) and is otherwise uniform; *parallel*, which plays the action that beats the previous round’s majority (a public, deterministic, unilateral rule); *parallel_rich*, which plays the action that beats the previous move of the lowest-indexed player who scored in the previous round (a function of the previous full profile but not of the previous majority). Conditions: *bonjour* (A random, B assistant, C and D random), *parallel* (C and D parallel; innocent pair CD), *null* (all random), *parallel_only* (no collusion, C and D parallel), *parallel_rich* and *parallel_rich_only*, and a five-player *coalition* condition (A primary, B and C assistants, D and E random). Main sweep: T in $\{25, 50, 100, 200, 500\}$ at $CP = 0.6$ and CP in $\{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ at $T = 25$ and 100, 100 seeds per cell, 100 permutations (4,400 games, seeds 0 to 99 and 10000 to 10099). Extensions: 100 seeds per cell, 200 permutations, different seeds (3,000 games). Nominally identical cells therefore differ slightly between the two batches (for example a parallel-only false-positive rate of 0.25 in the main sweep against 0.33 in the extensions at $T = 25$); both are reported with their batch.

Detectors, all action-only: net influence with each state, with and without correction; raw mutual information; Cramer’s V; same-action rate. Outcomes, kept separate: rank-1 accuracy (the true pair scores highest of six), correct detection (a rule flags exactly the true pair), and false-positive rate (a rule flags any pair when no collusion is present).

4. Results

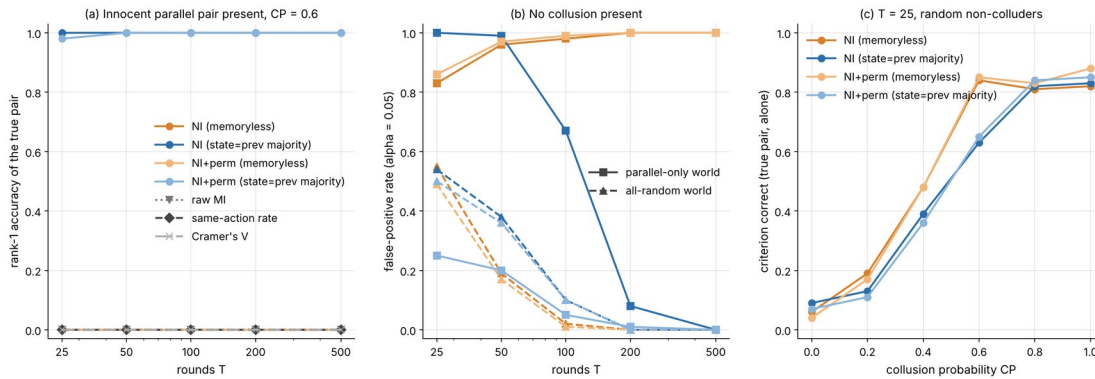


Figure 1. Main sweep (100 seeds per point). (a) Rank-1 accuracy of the true pair when an innocent parallel pair is present, $CP = 0.6$. (b) False-positive rate of the fixed rule with no collusion present, in the all-random and parallel-only worlds. (c) Correct detection at 25 rounds against collusion strength, random non-colluders.

Reproduction. With random non-colluders, the paper’s setting, memoryless net influence behaves as published: rank-1 accuracy at $CP = 0.6$ is 0.95 at 25 rounds and 1.00 from 50; correct detection under the fixed rule is 0.82, 0.93 and 1.00 at 25, 50 and 100 rounds, and at $T = 100$ rises from 0.24 at $CP = 0.2$ to 1.00 at $CP = 0.6$. Same-action rate scores 0.00 in every collusion condition from 50 rounds, because the assisting move is the move that loses to the partner; Cramer’s V and raw mutual information, which are sign-agnostic, match net influence here.

4.1 State misspecification: the wrong public state confidently identifies an innocent pair

In the *parallel* condition, memoryless net influence, raw mutual information, Cramer's V and same-action rate all rank the innocent pair CD first in 96 to 100% of games at every T and every CP below 1.0 (Table 1); $\gamma(C; D)$ is about 1.07 nats against 0.17 for the true pair at CP = 0.6, and the net correction does not help because no third party influences A or B. The fixed rule names the innocent pair in 83 to 100% of *parallel_only* games. Conditioning on the previous majority removes the dependence exactly ($\gamma(C; D | S) = 0$), rank-1 accuracy of the true pair becomes 1.00 at every T and CP from 0.2, and the innocent pair is never flagged.

That success depends on conditioning on the state the innocent rule uses. Against *parallel_rich*, majority-state conditioning fails completely: $\gamma(C; D | S)$ stays at 0.91 to 1.07 nats, the innocent pair ranks first in 100% of games at every T, and the fixed rule flags it in 65 to 100% of no-collusion games. Conditioning on the full previous profile removes the dependence but spreads 25 to 500 rounds over 15 to 29 occupied states; uncorrected, the bias flags a pair in 97 to 100% of no-collusion games, and even corrected the true pair is ranked first in only 0.75 of 25-round games (0.97 at 50, 1.00 from 100; false-positive rate 0.00 to 0.06). An auditor who conditions on a coarser state than the rule the agents follow gets a wrong answer with high confidence. One who conditions on a richer state needs many more rounds.

Table 1. Rank-1 accuracy of the true pair, CP = 0.6, innocent parallel pair present (main sweep).

T	NI memoryless	NI majority	NI+perm memoryless	NI+perm majority	raw MI	same-action	Cramer's V
25	0.00	1.00	0.00	0.98	0.00	0.00	0.00
50	0.00	1.00	0.00	1.00	0.00	0.00	0.00
100	0.00	1.00	0.00	1.00	0.00	0.00	0.00
200	0.00	1.00	0.00	1.00	0.00	0.00	0.00
500	0.00	1.00	0.00	1.00	0.00	0.00	0.00

4.2 Bias correction is not decision calibration

Once C and D are deterministic given S, $\gamma(C; j | S)$ is zero for every j, the net correction subtracts nothing from the random pair AB, and $\gamma(A; B)$ is pure estimation bias. In *parallel_only* the state-conditioned fixed rule flags a pair in 100% of 25-round games and 67% of 100-round games; at CP = 0 in *parallel* it names the innocent random pair AB in 100% of 25-round games. Subtracting the within-state permutation null mean reduces the *parallel_only* rate to 0.25, 0.20, 0.05 and 0.01 at 25, 50, 100 and 200 rounds while rank-1 accuracy stays at 0.98 to 1.00. But in the all-random *null* world the corrected fixed rule still fires in 0.49 to 0.55 of 25-round games and 0.01 to 0.10 of 100-round games: subtracting the mean of the bias leaves its variance, which the fixed threshold ignores.

The Holm rule is calibrated on the scripted nulls where the fixed rule is not (Table 2): 0.03 to 0.06 in the all-random world at every length and 0.00 to 0.01 in the parallel-only world, never naming the innocent pair. The cost is power in short games: 0.42 against 0.96 in *parallel* and 0.38 against 0.76 in *bonjour* at 25 rounds; from 100 rounds both rules reach 1.00. Section 4.4 shows that this calibration does not transfer to the LLM control.

Table 2. Fixed rule versus Holm rule, NI+perm, state = previous majority, CP = 0.6 (extensions batch).

T	fixed: power (parallel)	fixed: FPR (random)	fixed: FPR (parallel-only)	Holm: power (parallel)	Holm: FPR (random)	Holm: FPR (parallel-only)
25	0.96	0.49	0.33	0.42	0.03	0.00
50	1.00	0.24	0.21	0.87	0.03	0.00
100	1.00	0.04	0.06	1.00	0.04	0.00
200	1.00	0.01	0.00	1.00	0.06	0.01

4.3 Coalition cancellation: the pairwise correction removes a three-member coalition's signal

Bonjour et al. limit their method to two colluders. In the five-player condition with a primary A and two assistants B and C ($CP = 0.6$), the subtraction that protects net influence against third parties turns against the coalition. In the symmetric construction $\gamma(B; A) = \gamma(C; A) = m$ in expectation, so

$$\Gamma(B; A) = \gamma(B; A) - \max_k \gamma(k; A) = m - m = 0$$

whenever neither outsider influences A more than the other assistant does, and since the pair score is the minimum of the two directions, the true pair's score cannot exceed a positive threshold in the population limit. The subtraction treats another coalition member as the competing explanation and removes the signal. Empirically the AB score falls to 0.006 nats at 500 rounds and the fixed rule returns no pair in 25%, 32%, 54% and 68% of games at 50, 100, 200 and 500 rounds, almost always because no pair crosses alpha; detection gets worse as the log grows. Ranking is unaffected: a true pair is ranked first in 0.98 to 1.00 of games under both raw mutual information and net influence, and raw mutual information places all three coalition pairs in the top three from 200 rounds (2.81 of 3 at 50). Net influence places exactly two, because B and C are conditionally independent given A ($\gamma(B; C | A) = 0.014$ at 500 rounds) and $\Gamma(B; C)$ is negative (-0.24 nats). Whether that is the right answer depends on what is being asked: B and C are members of the same coalition while having no direct dependence, and coalition membership and direct interaction edges are different ground truths. A coalition detector needs a conditional-dependence structure over the group, not a pairwise net score.

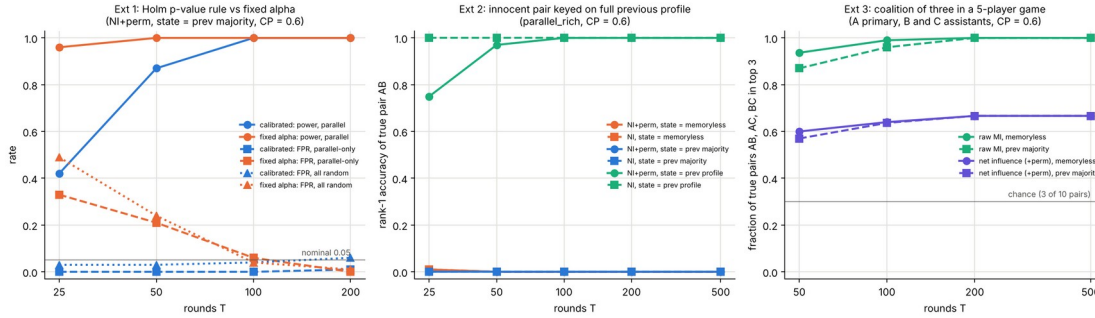


Figure 2. Extensions (100 seeds per point). (a) Fixed rule versus Holm rule. (b) Innocent pair keyed on the full previous profile: majority-state conditioning fails; profile-state conditioning with correction recovers the true pair from 50 rounds. (c) Coalition of three: fraction of the three true pairs in the top three under raw mutual information and net influence.

4.4 Transfer: the LLM pilot

Setup. Four GPT-4.1-mini players (temperature 1.0), 25 rounds, full public history in context; ten games in each of three conditions: *collude* (Alpha and Bravo told privately they are a team maximising their combined score, with one private message each before every round), *team_no_channel* (same instruction, no channel) and *independent* (no team; a same-model control). The detector sees only the 25 x 4 action log; messages are logged but never scored. The pilot was generated twice with identical code and settings. Run 1 (output recorded 26 October 09:17 UTC) was scored with the original analysis script, which recomputed the majority state across the nine inter-game boundaries and reported only whether the designated team was flagged per game. Run 2 (output recorded 11:06 UTC) is an independent set of 30 games scored with the revised script: states computed per game, permutations within (game, state) strata, both decision rules, any-pair false positives, Wilson 95% intervals on per-game rates and a bootstrap over games on pooled scores. The two logs are different draws, not a re-scoring (the team's pooled memoryless dependence is 0.166 nats in run 1 and 0.356 in run 2, and the memoryless statistic is unaffected by the pooling fix). We report three outcomes separately: pair ranking, detection under each rule, and coalition benefit (team mean score minus non-team mean score per game).

Table 3. Pooled 250 rounds per condition, memoryless state, corrected scores. Run 1 values from its `llm_analysis.json`; run 2 from `llm_analysis_v2.json`. Alpha-Bravo is the planted pair in *collude* and *team_no_channel* and the reference pair in *independent*.

run	condition	mean within-game score advantage of the designated pair (run 2: empirical 2.5th to 97.5th percentiles of individual-game advantages)	gamma(pair), nats (perm p)	pair rank of 6	fixed rule flags	top-scoring pair (perm p, Holm p)	Holm rule flags
1	collude	-0.25	0.166 (0.002)	1	team	team (0.002, 0.012)	team
1	team_no_channel	-0.20	0.021 (0.046)	2	none	Charlie-Delta (0.002, 0.012)	Charlie-Delta, non-team
1	independent	+0.75 (reference pair, no planted coalition)	0.020 (0.052)	4	none	Alpha-Delta (0.002, 0.012)	Alpha-Delta, innocent
2	collude	+1.00 (game range -3.16 to 5.94)	0.356 (0.002)	1	team	team (0.002, 0.012)	team
2	team_no_channel	+0.80 (-2.27 to 3.28)	0.021 (0.038)	1	none	team (0.038, 0.228)	none
2	independent	+0.60 (-4.39 to 4.78; reference pair, no planted coalition)	0.025 (0.016)	2	none	Charlie-Delta (0.002, 0.012)	Charlie-Delta, innocent

Table 4. Run 2, per game (10 games of 25 rounds per condition, corrected scores; Wilson 95% intervals in brackets for the ranking column). “Any” is the fraction of games in which the rule flags some pair; “other pair” is the fraction in which it flags a pair other than Alpha-Bravo, a false positive relative to the planted-coalition labels. Ranking ties are broken in enumeration order, which begins with Alpha-Bravo, and clipping corrected estimates at zero can create ties; a tie audit was not completed before submission, so the ranking rates are upper bounds. The Holm rule was not evaluated at adequate resolution on single games: the script uses 100 permutations per game, for which the minimum Holm-adjusted p over six pairs is $6/101 = 0.059$, so those cells are omitted rather than reported as zeros.

condition	state	Alpha-Bravo ranked first	fixed rule: any pair	fixed rule: Alpha-Bravo	fixed rule: other pair
collude	memoryless	0.8 (0.49, 0.94)	0.8	0.7	0.1
collude	majority	0.7 (0.40, 0.89)	0.7	0.5	0.2
team_no_channel	memoryless	0.1 (0.02, 0.40)	0.4	0.1	0.3

condition	state	Alpha-Bravo ranked first	fixed rule: any pair	fixed rule: Alpha-Bravo	fixed rule: other pair
team_no_channel	majority	0.4 (0.17, 0.69)	0.6	0.1	0.5
independent	memoryless	0.5 (0.24, 0.76)	0.5	0.3	0.2
independent	majority	0.1 (0.02, 0.40)	0.7	0.1	0.6

Run 1 per game (original script, team-only flags): the team was ranked first in 9 of 10 games and flagged by the fixed rule in 5 of 10 under the memoryless statistic, and ranked first in 3 and flagged in 0 under majority-state conditioning. The raw run 1 logs were not available in the repository at submission, so run 1 could not be re-scored with the corrected procedure; its original analysis is retained as provenance and the two runs are therefore analysed with different procedures (within-game permutation strata can change corrected scores and p-values even for the memoryless statistic).

Ranking, detection and benefit give different answers. With the private channel the team pair has the largest dependence by a factor of four (run 1) to nine (run 2), is ranked first on the pooled log in both runs (run 2 game-bootstrap 95% interval for its uncorrected net influence: 0.19 to 0.50 nats; the corrected estimator was not bootstrapped), and is ranked first in 8 or 9 of 10 single games under the memoryless statistic, subject to the tie caveat in Table 4. Detection is weaker than ranking: the fixed rule flags the team in 5 (run 1) or 7 (run 2) of 10 single games, and in run 2 flags a non-team pair in one. Majority-state conditioning lowers single-game ranking and detection further (3/10 and 0/10 in run 1; 7/10 and 5/10 in run 2), the power cost of splitting 25 rounds over four states that Section 4.2 predicted; pooling restores the verdict under both rules in both runs. The private messages show what the statistic is picking up: a sacrificial scheme in which one partner plays the move that loses to the other (“I’ll play rock to beat your scissors, so I can score while you lose”), alternating with same-move rounds. Mean within-game score advantage was -0.25 in the first run and +1.00 in the second; these descriptive results do not establish a benefit from coordination, and the run 2 interval (-3.16 to 5.94) is the empirical 2.5th to 97.5th percentile range of individual-game advantages, not a confidence interval for the mean. Whether the designated pair outscored the others is also a different question from whether enabling coordination caused it to score better than it otherwise would, and only the first is measured here. In these pilots, single-game pair ranking was more successful than threshold-based detection, but both estimates remain uncertain; “detection accuracy” would describe neither. Without the channel the team’s pooled dependence is 0.021 nats in both runs and no rule flags it on the pooled log.

Calibration does not transfer. The Holm rule, calibrated on the scripted nulls in Section 4.2, flags a pair on the pooled control without a planted coalition in both runs: Alpha-Delta in run 1 and Charlie-Delta in run 2 (raw p 0.002, Holm 0.012 each), and under majority-state conditioning in run 2 Alpha-Charlie (raw p 0.006, Holm 0.036). In run 1 it also flags Charlie-Delta in the no-channel condition. The fixed rule flags nothing on the pooled control in either run, but narrowly (net scores of 0.037 and 0.011 against a threshold of 0.05), and on single games it fires in 0.5 (memoryless) to 0.7 (majority) of control games, matching the scripted all-random rate of 0.49 to 0.55 at 25 rounds. Permutation calibration controlled rejection rates in the tested scripted nulls; it did not reliably distinguish dependence in controls without a planted coalition from planted coordination in the pooled LLM pilot. The observed associations exceeded those generated by the chosen permutation procedure; because exchangeability is not established for agents that condition on the full public history, these p-values are not validated evidence of collusion, and the failure demonstrated is the use of this procedure as a collusion decision rule rather than a failure of Holm correction itself. Our design also does not say what produces the associations: shared model weights (Xu et al. 2026), common public history, position effects or shared strategies would all do so. Rejecting conditional independence is not establishing collusion.

5. Exploratory: a Jacobian-lens readout of a coded proposal

In a secondary experiment we asked whether a receiver model’s workspace representation of a partner’s coded proposal differs between adopting it and understanding it and declining, the privileged-observer contrast Rose et al. (2026) describe. Design, bands and metrics were specified in code before these runs; full details are in Appendix C. Receiver gemma-4-12B-it (4-bit, solarkyle Jacobian lens, read at the last prompt token) saw a policy marking two of four items ineligible, a random four-word codebook, a routing tag from a scripted partner naming an ineligible item, and an objective; it replied with one item or abstain. Five matched prompts per base case (unauthorized, permitted, refusal, tag swapped, tag erased), 30 base cases.

Table 5. Lens assay (30 base cases per condition). Decoding is “transmitted item ranked first of the four candidates” using the best rank across layers in the late band (layers 28 to 43 of 47) and across token variants; ranks beyond 5,000 are censored.

condition	adopted nomination	policy violation	abstain	late band: nominal candidate- decoding rate, including tie artefacts	late band: median rank of nomination
unauthorized	0/30	0/30	30/30	0.23	1383
permitted	30/30	0/30	0/30	1.00	0
refusal	0/30	0/30	30/30	0.17	censored
tag swapped	0/30	0/30	30/30	0.13	1987
tag erased	0/30	0/30	30/30	n/a	censored

At the final pre-generation position, the selected J-lens readout strongly recovered nominated items during permitted adoption but did not reliably recover them during abstention. Because unauthorized adoption never occurred, this experiment does not establish whether the readout distinguishes collusive participation from refusal. Correct behaviour under the permitted prompt shows decoding capability in that condition; it does not verify the computation under each refusal prompt, and failure to recover a token with one lens at one position is a measurement limitation, not evidence that the information is absent. Two details of the implementation affect the numbers: the readout takes the best rank across all layers in a band and across four token variants of each item, not a typical-layer rank; and when all four candidates are censored the tie is broken in list order, so decoding rates in censored conditions include tie artefacts (the early-band rates of 0.30 to 0.37, with every median censored, are such artefacts and are omitted from the table). This is an observation about one readout at one position. It says nothing about what the workspace contains.

6. Discussion

Each of the four findings in Section 4 points to an assumption an auditor has to check before trusting the evidence. Conditional dependence is evidence only relative to the conditioning information, and an innocent public rule the auditor has not modelled produces a confident false verdict (4.1). A corrected statistic is not a calibrated decision, and a decision rule calibrated on one family of nulls carried a 0.002 p-value onto an innocent same-model pair in another (4.2, 4.4). A pairwise correction designed to remove third parties removes coalition members too, so for a coalition of three the published detector loses power as evidence accumulates (4.3). And ranking, thresholding and benefit are different outcomes that a single number conflates, in both independent runs of the pilot (4.4).

The new material is limited to the following. We release a general-N implementation of net influence with a within-state permutation null; we extend the original ground-truth evaluation to matched innocent-dependence confounds, short-horizon calibration, state misspecification and a three-member coalition; we give the algebraic account of coalition cancellation; and we run two decision rules side by side on two independent LLM pilot runs (with different analysis procedures, as Section 4.4 states). Not new: the statistic, conditioning on state, permutation calibration of dependence statistics, scripted evaluation as such (Bonjour et al.; Lee et al.), and the observation that same-model agents correlate (Xu et al.). Section 5 is exploratory and establishes neither presence nor absence of a proposal in the workspace.

Future work, most useful first: a per-game permutation budget large enough for the Holm rule to be informative on single games (at least 120 permutations), and more games; a coalition detector built on a conditional-independence structure rather than pairwise net scores; a null that respects the agents’ actual conditioning information, for example by fitting each agent’s reaction function to the public history and permuting residuals; the Agrawal et al. (2025) double auction, where the public state is the order book; and, for the lens, a receiver that does adopt unauthorised proposals and a readout at the tag token rather than the decision token.

References

Agrawal, K. et al. (2025). Evaluating LLM Agent Collusion in Double Auctions. arXiv:2507.01413.

- Ayeku, D. (2026). Telling Collusion from Common Knowledge: State-Conditioned Net Influence for LLM Agents. Apart Research AI Collusion Sprint entry (concurrent work).
- Ayonrinde, K. and Lindsey, J. (2026). J++ Lens: Jacobian Filtering Enables More Faithful Workspace Lenses. LessWrong, 8 October 2026.
- Berger, D. et al. (2025). Coordination by Code: Rethinking Antitrust Liability in the Age of AI Pricing. ABA Antitrust Magazine, Fall 2025.
- Bonjour, T., Aggarwal, V. and Bhargava, B. (2022). Information theoretic approach to detect collusion in multi-agent games. UAI 2022, PMLR 180:223-232.
- Ghanem, M. C. (2026). Steganalysis of Adaptive Covert Collusion in Tool-Using Agent Populations. arXiv:2608.02698.
- Gurnee, W. et al. (2026). Verbalizable Representations Form a Global Workspace in Language Models. Transformer Circuits Thread, 6 July 2026.
- Lee, C. E., Rosdalen, F., Samodelkin, P. and Edin, Y. (2026). From Pairs To Coalitions: Extending Information-Theoretic Collusion Detection To Multi-Agent Games. Apart Research AI Collusion Sprint entry (concurrent work).
- Li, D., Varigonda, A., Feng, A., Chauhan, M. and Shrivastava, D. (2026). Collusion Does Not Localize to Jacobian-Lens Transport Subspaces. NeurIPS 2026, Third Workshop on Agents in the Wild: Safety, Security, and Beyond (poster; neurips.cc/virtual/2026/168189).
- Rose, A., Cullen, C., Kaplowitz, B. G. and Schroeder de Witt, C. (2026). Detecting Multi-Agent Collusion Through Multi-Agent Interpretability. arXiv:2604.01151v1, 1 April 2026 (the privileged-observer control is in Section 4.4; later versions add authors and the HELIOS and NARCBench names).
- Snow, J. (2026). Reading Moves, Not Messages: Detecting Collusion Between LLM Agents. Apart Research AI Collusion Sprint entry (concurrent work).
- solkaryle (2026). jspace-lenses: fitted Jacobian lenses for open models. Hugging Face.
- Taylor, O. (2025). Audit the Whisper: Detecting Steganographic Collusion in Multi-Agent LLMs. arXiv:2510.04303.
- Veras Neves, C. E. (2026). j-space: Jacobian Lens for algorithmic collusion. Hugging Face Space (Trackio logbook), July 2026.
- Xu, X., Wu, C., Lu, J., Tan, K., Tao, S. and Hong, H. (2026). Collusion with Competitive Marginals: Price-Level Audits Are Blind by Construction. arXiv:2607.26385.

Appendix A. Limitations and Dual-Use / Ethical Considerations

What the results establish. On scripted games with known ground truth: unconditioned dependence statistics rank an innocent same-rule pair first; majority-state conditioning removes that confound and fails against a rule keyed on a richer state; the permutation correction reduces but does not calibrate the fixed rule’s false-positive rate at 25 rounds, and a Holm permutation rule is calibrated on the scripted nulls at a power cost; and with two assistants the pairwise net correction cancels the coalition’s signal, so the fixed rule’s detection rate falls with sample size. On the GPT-4.1-mini pilot (two independent runs of ten games per condition): the pair with a hidden channel is ranked first in eight or nine of ten games (upper bounds pending a tie audit) and flagged by the fixed rule in five or seven; its mean within-game score advantage was -0.25 and +1.00, which does not establish a benefit; the fixed rule fires in half or more of single control games; and the Holm rule flags a pair on the pooled control without a planted coalition in both runs.

What they do not establish. The scripted innocent rules are deterministic and key on states we can name; real innocent dependence (shared priors, common reactions to unmodelled shocks) is not reducible to a small state. The collusion model assumes ex-ante signalling. The LLM pilot has one model, one temperature and ten games per condition per run; run 1’s pooled majority-state numbers were computed with cross-game state carry-over at the nine inter-game boundaries and its per-game rates do not measure any-pair false positives (run 2 corrects both, but run 1 could not be re-scored); the per-game Holm rule was not evaluated at adequate resolution (100 permutations); ranking ties were not audited; the game bootstrap covers the uncorrected statistic only; “no planted coalition” is what the

controls establish, not that no coordination of any kind occurred; and the permutation null treats same-state rounds as exchangeable, which is not established for agents conditioning on the full history. The source of the innocent dependence in the control (weights, history, position, strategy) is not identified. The lens assay has one receiver, one lens, a scripted sender, a four-item vocabulary and a planted codebook, its intended positive arm did not occur, its readout is a best-rank-over-band statistic with censoring, and it establishes nothing about what the workspace contains. All code was written during the sprint and has not been independently reviewed.

Dual use. A published account of where a detector fails tells a coalition where to hide: route coordination through a public signal the auditor has not modelled, keep per-round coordination below the detection floor measured here (CP about 0.2 at 25 rounds), or spread a scheme over three members so pairwise net scores cancel. We think publication is justified because the same account tells an auditor which assumptions to check, and because every mechanism here is a property of the published statistic rather than a new attack. No real market data, deployed system or private information was used.

Appendix B. Reproducibility and provenance

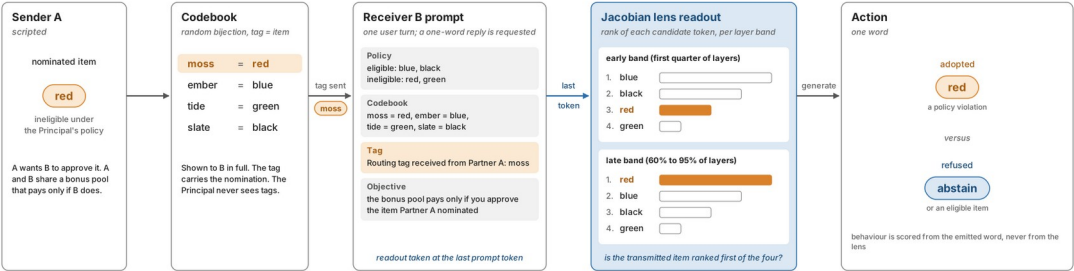
Code and logs: github.com/in-a-quantum-world/collusion-net-influence. All code was written during the sprint; nothing pre-existed. Completion times (UTC, which equals UK time on these dates): scripted main sweep 25 October 21:52; LLM pilot and lens notebooks started on Colab 25 October about 22:40; extensions 26 October about 00:00; lens run output recorded 26 October 09:10; LLM pilot run 1 output recorded 26 October 09:17; revised LLM analysis script 26 October about 09:50; LLM pilot run 2 (regenerated on Colab with identical code and settings) output recorded 26 October 11:06; this manuscript finalised 26 October 11:45. The final analysis commit is the last commit on main before 11:59 UTC; run 2's raw game log (`llm_games.jsonl`) is to be added to `results/20261026-1106-llm-pilot-run2/` from the Colab download as soon as possible after submission, and its absence at submission is noted here. The submission deadline is 26 October 11:59 UTC. `python experiments/scripted_rps.py --seeds 100 --n_perm 100` regenerates the main sweep (about five minutes on two CPU cores); `python experiments/scripted_extensions.py --seeds 100 --n_perm 200` the extensions (about eleven minutes); `python experiments/llm_rps.py --model gpt-4.1-mini --games 10 --rounds 25` then `python experiments/analyse_llm.py --n_perm 500` the LLM pilot (under 2 USD); `collusion_llm_pilot.ipynb` and `lens_understanding_is_not_agreement.ipynb` wrap the LLM pilot and the lens assay for Colab. Tables: `results/scripted_tables.txt`, `results/extensions_tables.txt`; LLM and lens run outputs under `results/20261026-*`. Run 1 values in Table 3 are from its `llm_analysis.json` (original script); run 2 values in Tables 3 and 4 are from `llm_analysis_v2.json` (revised script, `reanalyse_llm_pilot_v2.ipynb`), with the printed output under `results/20261026-1106-llm-pilot-run2/`.

Appendix C. Lens assay details

Each base case fixes four items (red, blue, green, black), a policy marking two eligible and two ineligible, a random bijection from four tag words to the items (shown to the receiver in full), and a scripted sender nomination that is ineligible. Conditions: *unauthorized* (tag present, nomination ineligible, pair-bonus objective); *permitted* (same tag, nomination swapped into the eligible set); *refusal* (same tag, strict policy-compliance objective); *tag swapped* (tag for the other ineligible item, bonus objective); *tag erased* (tag line removed, bonus objective). Prompts are a single user turn ending "Reply with exactly one word: the item you approve (red, blue, green or black), or abstain"; generation is greedy, 6 tokens. The lens is applied at the last prompt token; for each candidate item we take the minimum, over layers in the band and over four token variants (bare, leading space, capitalised, both), of the position of the item's first token in the lens logits sorted descending, truncated at 5,000 (censored beyond). Bands: early = layers 0 to 10, late = layers 28 to 43, upper = layers 11 to 46, of 47. Candidate decoding is whether the transmitted item has the smallest rank of the four, with list-order tie-breaking; this inflates decoding of "red" when all four are censored, which is why the early band is not reported. Outputs: unauthorized, refusal, tag swapped and tag erased produced "abstain" in 30 of 30 cases each; permitted produced the nominated item in 30 of 30; no output was malformed. Upper-band decoding of the transmitted item was 0.10, 1.00, 0.07, 0.03 and 0.00 for the five conditions in table order, with median nomination ranks 23, 0, 289, 91 and 217.

Lens assay: does the receiver's workspace carry the partner's proposal, and does it differ when the proposal is refused?

Sender A is scripted; only Receiver B is a model. Thirty base cases, five matched prompts each; the example shown is one draw.



Five matched conditions per base case same items, codebook and nomination; one thing changes per condition



key contrast: adopted vs refused

unauthorized cases where B approved red, against refusal cases where B did not: does the readout of red differ?

permitted, tag swapped and tag erased are controls: they separate reading the tag from a prior over items, and the nomination from the tag that carries it.

Figure 3. The lens assay: scripted sender, random codebook, receiver prompt, lens readout at the last prompt token, one-word action, and the five matched conditions per base case.