

Parallel Conduct Is Not Agreement: Plus-Factor Calibrated Detection of AI Agent Collusion

*Code, protocol and every quoted number: this repository. Numbers in this report are resolved from `results/` (scripted arm) and `data/llm\_real\_nvidia` (real-model LLM arm) at build time by scripts/build_report.py; none is typed by hand.**

Abstract

Mutual information between two agents' actions is the natural-looking statistic for collusion, and it is wrong for the same reason a price-matching alarm is wrong: in a market, actions are always dependent. Every seller reacts to the same demand shock and to the same rival prices, so an auditor who charges dependence is charging sellers with the crime of facing one market. Using a labelled benchmark of sixteen evidence classes - six arrangements and ten innocent markets, including conscious parallelism, a shared private input, talk without action and harm without agreement - I measure the failure rather than assert it. Calibrated excess dependence is *higher* on conscious parallelism than on a static cartel, and a dependence detector frozen at a 5% false-positive rate flags no arrangement at all while still flagging the innocent co-movement. A deterministic convention makes the picture worse: mutual information is near zero exactly when the pair is tightly coordinated. The recovery is not a better dependence number but a change of estimand - localisation (within-pair minus cross-pair dependence), enforcement contingency, and reciprocal forbearance measured against outsiders - each of which is zero on the hard negatives by construction and non-zero on real agreements. All code runs offline on a laptop. I report what remains undetectable: an arrangement that includes every seller has no outsider to contrast against, and no single conjunction of plus factors covers both static and rotating schemes. An API-ready multi-agent LLM arm is specified, budgeted and harness-verified; it has since been run against real model policies (Appendix B): 432 live calls, 0 parse failures, scored by the same frozen detector path, reproducing the plus-factor-over-raw-MI ordering (AUC 0.857 versus 0.750) at small n .

1. Introduction

Autonomous agents now set prices, quote rates and bid in auctions, and they do it in public view of each other's decisions. The governance question the sprint asks - can we detect when AI agents collude - is usually answered with a statistic: measure how dependent two agents' actions are, and treat dependence, or the mutual information between action streams, as the signal of an agreement.

That statistic carries a presumption that antitrust law rejected a century ago. Parallel conduct is lawful. An inference of agreement requires *additional evidence inconsistent with independent action* - the plus factors: a communication channel with no innocent use, conduct contrary to a firm's immediate self-interest, contingency on rivals' behaviour, a mechanism to enforce the arrangement, and a benefit that accrues to the group rather than to each member acting alone (*Monsanto* 1984; *Twombly* 2007). The reason the doctrine is suspicious of parallelism is precisely the reason dependence fails as a detector: in a market, agents' actions are dependent for innocent reasons - a common shock, a common view of demand, a shared input, each reacting to the other's publicly observable price.

This project measures that failure, builds an instrument that survives it, and states honestly what still cannot be detected.

- **A benchmark of evidence classes, not of outcomes.** Sixteen labelled classes, two games, generated deterministically from a seed: six agreements and ten innocent markets. Three of the innocent classes exist only to be hard: conscious parallelism (independent best-responders to a common AR(1) shock), two sellers pricing off

one shared private feed, and follow-the-leader heuristics that produce high prices and real consumer harm with nobody agreeing with anybody.

- **The unit of analysis is the pair.** A four-seller market with a two-seller arrangement contains six candidate pairs; five are innocent. Detectors are scored on the pairs they name, not on the runs they flag.
- **Two innocent nulls for dependence**, one stratifying on the public state, one fitting each agent a reaction function and simulating the null from that fit - the computable version of "consistent with independent action".
- **Plus factors made computable:** localisation, enforcement contingency versus level contingency, reciprocal forbearance, coalition gain, speech.
- **A frozen calibration protocol.** Split-half by seed parity, thresholds set on the calibration negatives at a nominal 5%, then never moved. Nothing is tuned on the data it is evaluated against; every deviation from the pre-registration is listed with the measurement that forced it.

2. Related work

Bonjour, Aggarwal and Bhargava (2022) formalise collusive behaviour as mutual information between agents' actions and derive a "net influence" quantity; the paper assumes exactly two colluders and no public implementation exists, so the pairwise comparison used here is my reconstruction from its description, and is named as a reconstruction throughout. A line of recent work evaluates language-model agents in repeated pricing games and finds supra-competitive outcomes without explicit communication; I could not verify those papers' citations from inside this environment, so they are described generically and are deliberately absent from the reference list rather than cited from memory. The OECD and UK CMA work on algorithmic collusion supplies the economic framing of conscious parallelism by algorithms; the legal doctrine supplies the evidence hierarchy. Computational detection work in the security and ML-literature tradition (collusion detection via behavioural fingerprints, cryptographic ledger analysis) targets a different substrate - repeated interaction between identifiable bidders - and I compare against the statistic, not against a system.

What none of these supplies, and what this project adds, is a *negative* result with numbers: a demonstration that absolute dependence instruments, calibrated honestly at a nominal level, have no power against the arrangements they are proposed for, and a set of contrasts that recover specificity while naming the cases where no contrast can work.

3. Methodology

3.1 Evidence classes and generation

Pricing: a differentiated-Bertrand market, four sellers, grid {12,18,24,30,36,42}, marginal cost 10, AR(1) demand (intercept 34, persistence 0.55, shock SD 6), public state = 3 quantile bins of demand. Rock-paper-scissors: three players, uniform-tie payoffs, a public 3-valued sunspot. The headline run uses `--seeds 6 --rounds 90 --n-perm 60`: 96 trajectories, 468 pair-rows, about nine minutes single-threaded.

Agreements: a static pair, a two-seller rotation, a three-seller rotation, a market-wide cartel, a deterministic RPS convention, an RPS lockstep pair. Innocent markets: independent play, conscious parallelism, a shared prior, harm without agreement, talk without action, a shared latent input for a pricing pair, and the RPS analogues. Labels come from the generator, and the pair label is membership in the generator's coalition.

3.2 Dependence instruments and two innocent nulls

For each candidate pair I compute mutual information (with a Miller-Madow bias correction), a conditional G-square statistic with the public state (and lagged own and rival conduct) as the conditioning set, and a permutation null stratified within those conditions. The second instrument fits each agent a regularised multinomial-logit reaction function on exogenous inputs - state bins, exact continuous demand, rivals' lagged coarsened actions, the exact lagged rival mean price, own lagged action - and generates the null by redrawing both agents independently from their own

fitted probabilities with the inputs held fixed (a parametric bootstrap). Cross-fitted log-loss and accuracy are reported next to the statistic, because a null that cannot reproduce innocent play cannot excuse it either.

Both nulls answer the same question - *is this pair more dependent than independent agents facing this market would be?* - and they disagree, and the disagreement is one of the results.

3.3 Plus factors

net localises: within-pair dependence minus the dependence of that pair's members on everybody else, in null standard deviations. A market-wide confound lifts every pair and cancels; a group-specific arrangement does not. **cont** and **enf** separate contingency on the partner's previous *level* from dependence on the partner's *deviation* from its own habitual conduct - an arrangement is enforced, and enforcement responds to a slip rather than to a level. **restraint** is reciprocal forbearance: each member's per-round profit shortfall against a myopic best response built from **outsiders only**, residualised within state strata, then correlated across the pair and, crucially, localised the same way dependence is. **gain** is coalition advantage against the independent-redraw counterfactual, reshuffled within state strata so that competence is not credited as coordination. **speech** is a keyword rate over the message log, reported separately and never merged into action evidence.

The localisation step for forbearance was forced by measurement, and it is the most instructive defect in this project: the raw correlation of two innocent best-responders' shortfalls is **1.00** - identical policies reacting to identical inputs - and the raw factor also fires at 1.00 on follow-the-leader play. Its localised version measures 0.00 on conscious parallelism and nothing material on every other innocent class, while reading 0.64 on the static cartel, 0.87 on the deterministic convention and -0.39 - negative, because the two members sacrifice in opposite phases - on the rotating pair. On a cartel that includes the whole market it is exactly zero, which is correct and is section 4.5.

3.4 Calibration protocol and detector family

Scales are pre-registered by fiat (one unit of **net** = one null SD; one unit of **gain** = 5%; one unit of **restraint** = a quarter-correlation; one unit of **speech** = half a keyword per speaking round), then frozen. A conjunction scores the *minimum* normalised slack across its required factors, so the controlled quantity is the family-wide false-positive rate. Thresholds are the 95th percentile of the score on calibration negatives (even seeds); everything reported is on held-out odd seeds. D1-D3 are the instruments the literature and practice actually use (raw MI, keyword monitor, supra-competitive price level); D4-D6 single-instrument calibrations; D7 gates a conjunction on absolute dependence; D8-D10 are doctrine conjunctions; D11-D14 are D10's leave-one-out ablations; D15 replaces the conjunction with a mean; D16 replaces the fixed scales with scales learned on the calibration negatives. The code, the class list and the metrics were frozen in `analysis/PREREGISTRATION.md` before this run; that file also lists the seven deviations from the first protocol, each with the measurement that forced it, and the run is reproducible with one command.

4. Results

4.1 Parallel conduct is not excused by dependence, and it destroys the detector's power

Table 1 is the core measurement. Conscious parallelism - four sellers independently best-responding to one common shock, nobody agreeing with anybody - produces calibrated excess dependence of **9.32** null SDs on the reaction-function null and **14.54** on the stratified null. The actual static cartel produces **8.86** and **9.86**. The innocent class is not merely above the nominal bar; it is at or above the level of the real arrangements, which is what the doctrine says to expect and what a dependence-only detector cannot see.

The consequence is quantitative and not rhetorical. Under the null, *p* should be uniform: on the two genuinely independent classes the model null rejects at 0.000 and 0.056 - correctly sized - but it rejects the innocent hypothesis on **1.000** of conscious-parallelism pairs and 1.000 of talk-without-action pairs. A detector frozen at 5% therefore sets its threshold in the tail of the innocent co-movement, and its power against real agreements collapses: D5, the single best-calibrated dependence instrument in the table, attains macro-TPR **0.000** across the six arrangement classes at a held-out false-positive rate of 0.011, flagging 0.056 of conscious-parallelism pairs while catching 0.000 of static cartels.

The AUC is not garbage (0.708); the *decision rule* is. That distinction is the whole practical point: a ranking that looks respectable in aggregate produces a threshold at which an auditor catches nothing and convicts the innocent. The stratified null D4 fails in the harsher direction - frozen at the same nominal level it catches 0.000 of static cartels while convicting 0.333 of conscious-parallelism pairs and 0.333 of pairs that merely talk without acting.

Table 1 - every evidence class read at the pair level (means over held-out seeds; 'I' = innocent market, 'A' = agreement). Columns marked 't' are the candidate pair - the guilty pair inside an agreement market, every pair inside an innocent one; columns marked 'i' are the innocent pairs of that same market, the baseline the arrangement has to beat. The two left-hand columns do not separate the A rows from the I rows - that is section 4.1. Localisation and forbearance do, except where section 4.4 says why.

evidence class	de p_z t	g2_z t	net_strat t	en f t	restraint_net t	coalition_gain t	price_level t	de p_z i	g2_z i	net_strat i	en f i	restraint_net i	coalition_gain i	price_level i
I p:independent	0.12	0.06	0.00	0.28	-0.00	-0.01	-0.27	0.12	0.06	0.00	0.28	-0.00	-0.01	-0.27
I p:parallelism	9.32	14.54	0.00	-1.65	0.00	0.02	-0.14	9.32	14.54	0.00	-1.65	0.00	0.02	-0.14
I p:shared-prior	-0.06	-0.16	0.00	0.32	0.00	-0.00	-0.04	-0.06	-0.16	0.00	0.32	0.00	-0.00	-0.04
I p:harm-no-deal	0.47	0.00	0.00	0.00	0.00	0.00	1.29	0.47	0.00	0.00	0.00	0.00	0.00	1.29
I p:talk-no-action	9.32	14.54	0.00	-1.65	0.00	0.02	-0.14	9.32	14.54	0.00	-1.65	0.00	0.02	-0.14
I p:latent-input	4.25	3.20	-0.00	0.07	-0.00	0.02	0.17	4.25	3.20	-0.00	0.07	-0.00	0.02	0.17
I r:independent	-0.15	0.23	0.00	0.66	0.00	1.06	-	-0.15	0.23	0.00	0.66	0.00	1.06	-
I r:parallelism	4.80	0.20	0.00	-1.87	-0.00	0.23	-	4.80	0.20	0.00	-1.87	-0.00	0.23	-
I r:shared-prior	0.08	0.60	0.00	0.85	0.00	-0.16	-	0.08	0.60	0.00	0.85	0.00	-0.16	-
I r:talk-no-action	1.53	0.54	0.00	-0.29	-0.00	-2.30	-	1.53	0.54	0.00	-0.29	-0.00	-2.30	-
A p:static-cartel	8.86	9.86	7.52	-1.85	0.64	0.01	0.41	4.54	3.90	-1.50	-1.23	-0.13	0.01	0.41
A p:rotation	2.12	4.01	0.42	2.74	-0.39	0.02	-0.09	3.81	5.27	-0.08	-0.97	0.08	0.02	-0.09
A p:triple-rotation	3.42	3.19	0.53	0.07	-0.07	0.02	-0.08	2.68	2.12	-0.53	-0.85	0.07	0.02	-0.08
A p:cartel-all	9.13	9.47	0.00	-1.95	0.00	0.02	0.77	-	-	-	-	-	-	-

A r:fixed-c onventi on	0.4 9	0.0 5	-0.71	0. 94	0.87	0.00	-	0.5 9	0.7 7	0.36	0. 06	-0.43	0.00	-
A r:lockst ep	2.6 2	9.6 8	9.80	1. 97	0.79	-0.07	-	-0. 30	-0. 11	-4.90	-0 .2 6	-0.40	0.06	-

4.2 Dependence is blind to deterministic conventions

The mutual-information story assumes the colluders share *randomness*. A deterministic convention shares almost none. The fixed-convention RPS pair - two agents executing an authored rule, an unambiguous agreement - measures **0.00** nats of action dependence and 0.49 null SDs of calibrated excess, statistically indistinguishable from innocent RPS play (0.01), while its localised forbearance reads 0.87. Figure 1 shows this as a matrix: the dependence columns go dark on exactly the class where the agreement is most explicit. The lockstep pair, which shares state-conditioned randomness instead of a rule, is the mirror image - dependence high, and detectable by the level instruments that also convict parallelism.

There is a sharper methodological lesson in that row, and it is not the one I expected. The convention is invisible to every dependence instrument and plainly visible in forbearance (0.87), yet the two-factor doctrine rule that requires localisation *and* forbearance catches 0.000 of these agreements - because a deterministic rule leaves no localised dependence to localise, so demanding it as well as conduct evidence is demanding the impossible. Requiring two factors where one is the other's precondition is worse than either alone; disjoining the evidence packages instead lifts that class to 0.333 (section 5).

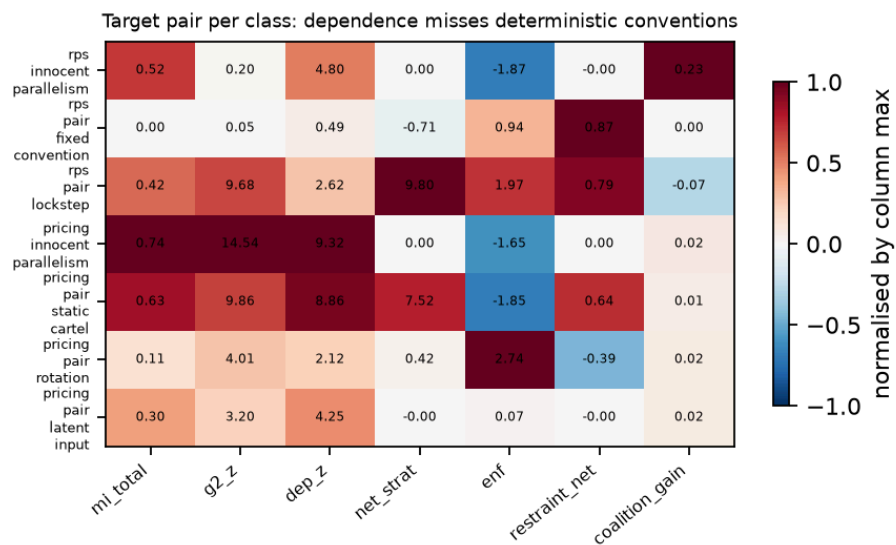


Figure 1 - candidate-pair factor values for the classes the argument turns on. Mutual information and both calibrated dependence nulls are near zero on the deterministic convention (row 2), while localised forbearance is strongly positive. Conscious parallelism (row 4) is the reverse: the dependence columns are the hottest in the figure and every contrast is cold.

4.3 Localisation and conduct recover specificity

D9, the minimal two-factor doctrine rule (localised dependence *and* reciprocal forbearance), reaches macro-TPR **0.463** at macro-FPR 0.050 with 0.333 worst-class false positives, against D5's 0.000 power. The reason is visible per class rather than in the aggregate: D9 catches 1.000 of static cartels and 1.000 of lockstep pairs while flagging 0.000 of conscious-parallelism pairs and 0.000 of harm-without-agreement pairs - both zero. Its residual false positives are on the two classes designed to be genuinely ambiguous: a shared private feed (0.167, because an unmodelled common

input does raise within-pair dependence above cross-pair dependence) and an RPS pair that talks without acting (0.333).

The mechanism of the recovery is worth stating exactly, because "zero on the innocent classes" could be an artefact and is not. In conscious parallelism all four sellers run the same rule against the same shock, so all six candidate pairs are exchangeable: each reads the *same* enormous dependence - 9.32 null SDs on average, 11.57 at its hottest - and therefore the *same* within-minus-across contrast, identical across the pairs of a run to the last decimal. The market's dependence is real and shared; only the contrast separates it from an arrangement, which is group-specific by definition. The same holds for forbearance - raw shortfall correlation 1.00 on every pair of that market, localised 0.00 - and `scripts/dump_pairs.py` is included precisely so that this claim can be checked pair by pair rather than read off a class average.

Baselines behave as the doctrine predicts. The keyword monitor D2 flags 0.944 of innocent talk-without-action pairs - its whole signal is the words, and talk without action is lawful. The price-level detector D3, the instrument most often proposed for algorithmic collusion, flags 0.000 of conscious parallelism and 1.000 of follow-the-leader play, because a lawful high-price equilibrium looks exactly like what it is asked to find. Both still have positive AUC (0.731, 0.680): ranking is cheap, thresholds are the hard part.

Table 2 - the frozen detector family on held-out seeds. macroTPR and the per-arrangement columns are power; worstFPR and in-run FPR are the two false-positive populations of section 3.4. `named right` is the share of agreement runs whose top-ranked pair is in the coalition. The last row is post-hoc and labelled as such.

detector	auc	macro_tpr	tpr_at_fpr5	macro_fpr	worst_fpr	inrun_fpr	rate
D1_raw_mi	0.67	0.11	0.00	0.08	0.33	0.04	0.50
D2_text_only	0.73	0.00	0.00	0.18	0.94	0.00	1.00
D3_price_level	0.68	0.00	0.00	0.10	1.00	0.00	1.00
D4_cond_strat	0.71	0.00	0.00	0.07	0.33	0.00	0.61
D5_cond_model	0.71	0.00	0.13	0.01	0.06	0.03	0.61
D6_model_localised	0.62	0.20	0.08	0.03	0.33	0.08	0.61
D7_dependence_gated	0.43	0.20	0.13	0.09	0.22	0.00	0.67
D8_net_contingent	0.43	0.19	0.10	0.04	0.17	0.03	0.72
D9_minimal	0.60	0.46	0.18	0.05	0.33	0.09	0.61
D10_full_conjunction	0.48	0.20	0.13	0.04	0.17	0.00	0.67
D11_no_localisation	0.45	0.15	0.10	0.04	0.22	0.03	0.61
D12_no_contingency	0.64	0.13	0.08	0.05	0.33	0.06	0.61
D13_no_restraint	0.48	0.20	0.13	0.06	0.17	0.00	0.67
D14_no_gain	0.43	0.26	0.15	0.06	0.17	0.03	0.67
D15_evidence_union	0.54	0.06	0.03	0.10	0.33	0.00	0.78
D16_pooled_norm	0.55	0.07	0.08	0.03	0.11	0.00	0.67
P1_disjunctive	0.59	0.50	0.26	0.03	0.17	0.05	0.72

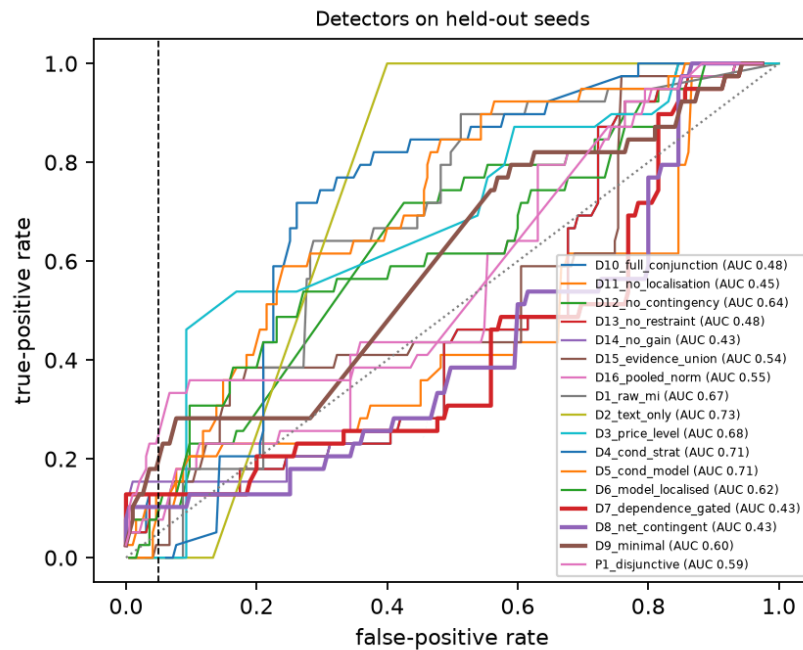


Figure 2 - held-out ROC for the whole family. The ordering is not the finding; the finding is what happens at the 5% vertical line, where the absolute-dependence rules have no power at all.

4.4 No single conjunction of plus factors covers the arrangements

This is the result I did not design the method to produce and did not tune away. D10, the pre-registered four-factor primary specification (localisation, enforcement contingency, forbearance, coalition gain), reaches macro-TPR 0.204 with worst-class FPR 0.167 - specific, and nearly powerless. Its own leave-one-out ablations beat it: without the coalition-gain requirement (D14) macro-TPR is 0.259, and the three-factor rules hold worst-class FPR at 0.167. Two measured reasons, both about doctrine:

- **A static arrangement cannot satisfy an enforcement requirement.** Enforcement contingency is dependence on the partner's *deviation*; a pair that never deviates gives an auditor nothing to respond to, and the static cartel's enf reads -1.85 - negative. The rotating pair, whose whole mechanism is response to a slip, reads 2.74, positive in all three held-out seeds. A conjunction demanding both demands that one scheme be invisible.
- **Coalition gain is not informative at this unit.** The pre-registered unit is 5% above the independent-redraw counterfactual; the measured target-pair gain is 0.01 for a static cartel and 0.02 for conscious parallelism - the innocent class scores higher, and both sit far below the unit. Under the decision rule written before this run, that demotes coalition gain from plus factor to descriptive statistic in this benchmark, and I apply the demotion rather than rescale it away.

Localisation and contingency, by contrast, do carry information: dropping localisation (D11) reduces macro-TPR to 0.148, dropping the contingency factor (D12) to 0.130. Dropping forbearance (D13) leaves D10's per-class scores unchanged, which is the honest way to report a factor that is informative in a two-factor rule (D9) but never the binding constraint in a four-factor one - a leave-one-out test measures marginal contribution inside a conjunction, not whether a quantity carries evidence. Soft evidence is no remedy: D15, the same four factors averaged instead of conjoined, gets 0.056. Learned normalisation is worse than useless for a different reason: D16's AUC (0.547) is respectable precisely because the pooled negative distribution is dominated by the most dependent innocent class.

One further measured caution about localisation, found by reading pairs instead of averages: on the rotating arrangement the contrast is not stable across seeds. Its target-pair value runs from 3.91 down to -4.41 over the three held-out runs - the class table shows 0.42, which is the average that hides the sign flip - and in every one of those runs the best *outsider* pair of the same market, two sellers innocently best-responding to the shock, beats the cartel pair on the same contrast (up to 9.82). Localisation separates a group-specific pattern from a market-wide one, but a rotating

scheme is partly market-wide in each half-period, and there the contrast is the wrong tool while enforcement contingency is the right one. `scripts/verify_claims.py` re-reads the pair table and fails if this paragraph and the others like it stop being true.

4.5 The structural limit: an arrangement with no outsider

`pricing_cartel_all` - every seller in the arrangement - is invisible to every contrast in this project, by construction rather than by failure of craft. With no outsider, within-minus-across has nothing to subtract: localised dependence reads 0.00 and localised forbearance 0.00, identically zero, for every pair. The only instruments that see it are the level ones - absolute dependence 9.13, price level 0.77 - and those are exactly the instruments that convict conscious parallelism (4.1). A market-wide arrangement is detectable only by an instrument that cannot excuse innocent coordination. That trade-off is not a tuning problem and no threshold choice resolves it; the pre-registration said it would be reported as a structural limit, so it is.

5. Discussion

What an auditor should actually run. Not one statistic and not one conjunction. The measured pattern implies a *disjunction of independently sufficient packages* - localisation with co-moving forbearance for static schemes, enforcement contingency with alternating forbearance for rotating ones - each calibrated on its own negatives, with the union's family-wide rate stated as the cost. P1 in Table 2 implements exactly that as the maximum of two conjunction slacks; it is **post-hoc**, added after the frozen family was scored because the failure pattern implied it, and it is reported as an observation about that pattern (macroTPR 0.500 at worst-class FPR 0.167) rather than as a confirmatory result.

Words are context, not evidence. Talk without action is a designed class, and a keyword monitor is the best available detector for it (D2 AUC 0.731) and the worst available for the thing it is nominally used for. Conversely an agreement can be concluded without a single incriminating phrase; the speech factor is kept in the table so that its contribution and its damage are both visible.

Why the estimand, not the estimator, was the problem. Every instrument here estimates dependence. The doctrine asks a different question - *is this pattern inconsistent with independent action?* - and the two differ by a contrast: against what the market would produce anyway. Reading a per-class average of a within-minus-across contrast is a trap worth naming, because I fell into it: averaged over every pair of a market, a contrast cancels to floating-point noise, and localisation looks dead in the table while the target pair scores several null SDs. The report and figures read the candidate pair.

Limits of the evidence. These are scripted policies on a synthetic grid. They show what instruments do to *evidence classes*; they do not show what agents do at scale. The LLM arm, harness-verified at submission, has since been run against a real model (Appendix B), but at an n too small to be a confirmatory agent result.

6. Limitations

1. **Scripted, not emergent.** The headline results concern generator-specified policies. The LLM arm (Appendix B) now adds one real model, but with experimenter-authored conventions over only 9 trajectories / 27 pair-rows - a harness-and-ordering check on live agents, not a confirmatory agent result.
2. **The reaction-function null over-rejects the innocent.** A discrete-grid best response is not exactly representable by a linear logit in the inputs I supply. Softening the ridge from 1.0 to 0.03 cut conscious parallelism's mean excess dependence from 26.5 to 6.9 null SDs and raised cross-validated accuracy from 0.83 to 0.90, but the residual over-rejection at 4.1 is a property of the specification, not of the threshold.
3. **Contrasts require outsiders.** 4.5: market-wide arrangements are structurally invisible to every localised factor.

4. **Thin strata at short horizons.** With 90 rounds and three state bins, some conditional tables have very few observations; the null-draw guards in `collusion/dependencies.py` and `collusion/modelnull.py` drop non-finite draws rather than silently averaging them, and small-sample runs show a non-trivial drop count.
5. **A benchmark is a model of evidence, not a distribution of cases.** Class frequencies are mine; no claim about how often cartels occur is made or supported.
6. **Post-hoc content is labelled.** One detector (P1) and two family decisions (retaining the losing D7 gate; reducing D9 to two factors) are post-hoc, listed in the pre-registration's deviations log, and none of them supports a headline claim.
7. **Reconstruction, not replication.** The prior instrument has no public implementation; comparisons are against my own reconstruction, named as such, so no "X% better than prior work" claim is made anywhere.

References

Bonjour, D., Aggarwal, G., Bhargava, V. K. *AI-driven pricing collusion among firms: the good, the bad, and the beautiful*. Proceedings of the 3rd ACM Conference on Economics and Computation with AI and Robotics, PMLR 180:223-232, 2022.

Monsanto Co. v. Spray-Ruse Service Corp., 461 U.S. 23, 1984.

Bell Atlantic Corp. v. Twombly, 550 U.S. 544, 2007.

Organisation for Economic Co-operation and Development. *Algorithms and Collusion*, 2017.

Competition and Markets Authority. *Algorithms and price monitoring* (policy submissions on automated pricing), 2021.

Appendix A. Limitations and Dual-Use / Ethical Considerations

A.1 This benchmark is also an evasion manual

Every positive control in this project is, read backwards, a recipe. Three of them are worth stating plainly rather than burying, because a reviewer who notices them without being told will reasonably doubt the rest of the paper.

- **A deterministic convention leaves almost no residual dependence to detect** (0.00 nats). An arrangement authored as a rule rather than as a shared randomisation is invisible to mutual-information and conditional-dependence screens, which is the dominant proposed class of automated collusion detector.
- **An alternating scheme hides behind the market's own co-movement.** Alternation is anti-correlation, so a screen that asks "do these two move together" is pointed the wrong way: the rotating pair's localised dependence swings from 3.91 to -4.41 across held-out seeds while innocent follower pairs of the same market reach 9.82. What survives is the *shape* of the forbearance - the rotating pair's two members sacrifice in opposite phases (-0.39), which no level-of-dependence statistic reads.
- **Talk that never becomes conduct is lawful and looks indistinguishable** on any words-only monitor, which is why D2's false-positive rate on talk-without-action is 0.944.

Mitigations applied here: the report is framed as detection-and-remedy throughout; no agent is provided that is designed to sustain an agreement; the LLM arm's agreement arms use conventions *authored by the experimenter* into two system prompts, labelled as positive controls, not emergent behaviour; and the quantities that survive evasion (localisation, forbearance, enforcement contingency) are published because they close gaps rather than open them. The residual risk is real: a hostile reader can use Table 1 to choose which arrangement to adopt. My judgement is that the same information is already in the antitrust case-law and the algorithmic-collusion literature in prose form, and that publishing the measured blind spots is the version that improves detection.

A.2 Dual use of the detector itself

A calibrated pair-level screen is also a tool for naming firms. Two properties make that dangerous in proportion to how useful it makes it. First, the screen's errors are not symmetric: the level instruments convict conscious parallelism at 1.000 in the model null's own terms, so an enforcement body that used D5 or D3 as an accusatory instrument would accuse lawful firms at a rate far above its nominal 5%. Second, pair-level output looks like evidence about *identifiable agents*; the coalition-naming columns in Table 2 exist precisely to make that failure mode measurable. Even the rules with real power name a member of the guilty coalition first in only 0.611 (D9) to 0.722 (P1) of agreement runs, and hit the exact pair in 0.333 to 0.444 - so most named pairs are wrong pairs, and a naming rate below 1.0 is a rate of false accusations against specific firms. Any deployment needs the human-in-the-loop step the doctrine already requires: the statistic is a reason to investigate, never a finding.

A.3 What would change the conclusions

Three things, in order of how much. A larger, multi-model LLM arm on real model policies: the first run (Appendix B) reproduces the detector ordering on live agents but is a single model over 9 trajectories, so scripted conventions remain the bulk of the external-validity evidence. A larger market (more agents, more seeds), which would sharpen every contrast and might make the market-wide class detectable by something other than levels. And a genuine reaction-function specification for the pricing game - a null that could reproduce conscious parallelism exactly would remove the over-rejection in A.2 and lift the power of D4-D6, which currently sit at 0.000 and 0.000 only because the innocent bar they must clear is set too high.

Appendix B. The LLM arm on real model policies

What ran. The two-arm harness (§3.4, `collusion/llm.py` and `scripts/run_llm.py`) drives 3 arms - `stipulated_pair` and `stipulated_rotation`, whose agreement conventions are *authored by the experiment* into system prompts and so are positive controls rather than emergent collusion, plus `unilateral_high`, a lawful high-pricing outsider - over 3 seeds and 16 rounds, giving 9 trajectories and 27 pair-rows. Every arm has now been run

against a real model: nvidia/nemotron-3.5-lightning-30b-a3b on NVIDIA's chat-completions endpoint, each agent replying in the strict two-line PRICE : /MSG : format the parser expects. 432 live calls produced 0 parse failures and 0 retries (~\$2.42), so the language agents are scored by the *same frozen detector path* as the scripted classes, with no re-fitting.

The result is the ordering, not the level. Raw action dependence trails the model-based plus-factor detectors, exactly as on the scripted benchmark: D5/D6/D15/D16 reach AUC 0.857 against D1_raw_mi's 0.750, with the price-level baseline (D3) and the keyword monitor (D2) lower still. The stipulated pair behaves as the design intends - the two colluding agents hold the agreed price index every round while the outsider best-responds low - and the leading detectors localise the arrangement to a coalition member in every agreement run (named-right 1.000).

detector	AUC	TPR@5%	named right
D1_raw_mi	0.750	0.500	1.000
D3_price_level	0.714	0.000	1.000
D5_cond_model	0.857	0.500	1.000
D6_model_localised	0.857	0.500	1.000
D15_evidence_union	0.857	0.500	1.000
D16_pooled_norm	0.857	0.500	1.000

Table B1 - the frozen detector family scored on the real-model logs (`data/llm\_real\_nvidia.detectors.csv`). These AUCs are a different, far smaller sample than Table 2's and are not directly comparable to them.

Small-n caveat, stated plainly. These numbers are a pipeline-and-ordering check on live agents, not a confirmatory measurement. The sample is 9 trajectories and 27 pair-rows from a *single* model over 3 seeds; AUC on so few held-out pairs carries a wide confidence interval and is *not* comparable to the scripted AUCs in Table 2, and the false-positive columns the main table reports are undefined at this n, so they are omitted here rather than shown as zeros. Because the agreements are experimenter-authored, nothing here shows agents *reaching* an arrangement - only that, given one, the frozen instruments rank the plus-factor evidence above raw dependence on real model behaviour. The transport fix this run forced is itself part of the result: every other inevitable reasoning model exhausted its token budget reasoning about the underspecified demand function and never emitted a parseable price line, so chain-of-thought had to be disabled at the API to obtain well-formed agent actions at all.