

Guilty by Resemblance

Model Monoculture Inverts Collusion Detection

Hectar Carson

Code and data: <https://github.com/hectar-glitches/monoculture>

Abstract

AI agents that share an underlying model make similar choices without ever communicating. We ask whether collusion detectors can tell that apart from real coordination.

We construct two populations: one holding a secretly coordinating pair, one where the agents merely share a model. We solve for the parameter that makes any pair in the second population identical in distribution to the colluding pair, so no detector can separate them from that pair alone. A shared model lifts the dependence of every pair at once, and an agreement lifts one. Detectors scoring the single most suspicious pair therefore read how similar it is and not how unusual it is: they invert, ranking honest groups above real cartels at AUROC 0.20 in all 26 configurations tested. Detectors comparing pairs with one another recover completely.

Six instruction-tuned models from six companies show this is not hypothetical. Gemma-3-1B and Phi-4-mini agreed on 98.8% of rounds with no link between them, more than six copies of any single model we tried, so requiring different vendors does not buy what it appears to. Matching a planted cartel to that pair took coordination on 73% to 99% of rounds, after which a correlation screen named it by coin flip, while adding price level restored attribution to 100%. On 240 MMLU questions the coin flip repeats and the outcome repair survives with its margin cut from ninety-nine points to fourteen, because innocent models share wrong answers.

We give four repairs, and one configuration that defeats all four.

1 Introduction

Two AI agents running on the same underlying model tend to make the same choices, without exchanging a word, because they are the same decision procedure applied twice. The condition in which many decisions flow through one shared model is **algorithmic monoculture**, a term Kleinberg & Raghavan (2021) introduced for the welfare consequences of shared algorithms; we use it for that structural fact alone and it carries no claim about intent. The effect is not small. Evaluating more than 350 language models, Kim, Garg, Peng & Garg (2025) find errors substantially correlated, models agreeing 60% of the time when both are wrong on one leaderboard dataset, and “larger and more accurate models have highly correlated errors, even with distinct architectures and providers.” In a market setting, Xu et al. (2026) measure the residual correlation between two deployments of one model at +0.053, against +0.0001 between different models.

Whether such agents are colluding is a question courts are now answering, on similar facts and with different outcomes. On 29 July 2026 the Third Circuit reversed a dismissal in *Cornish-Adebiyi v. Caesars Entertainment*, finding a plausible price-fixing agreement among Atlantic City hotels using Cendyn’s Rainmaker software, resting on the alleged exchange of nonpublic pricing data and adherence to the tool’s recommendations, while stressing that common use of independent software is not by itself a violation. A year earlier the Ninth Circuit affirmed dismissal of a similar claim in *Gibson v. Cendyn Group*, because the recommendations were not binding. On 14 May 2026 the Acting Deputy Assistant Attorney General for Criminal Enforcement said the Division’s civil record “does not reflect a view that algorithmic conduct is beyond the reach of criminal antitrust enforcement” (Glad, 2026), California’s AB 325 took effect on 1 January 2026, and on 4 September 2026 the Department proposed a consent decree over algorithmic pricing in *United States v. RealPage*.

Both ways of being wrong are expensive. Treating resemblance as conspiracy exposes firms to treble damages and, on the Department’s stated view, to criminal liability for behaviour nobody agreed to. Treating conspiracy as resemblance hands a cartel a defence that strengthens as its members’ tooling converges. The second failure compounds through capability: on the finding quoted above, the confound grows as models improve, whether or not the vendor market consolidates.

Underneath the legal question is a measurement question. Every behavioural collusion detector observes that a set of agents behave alike and then decides whether something illegitimate explains the resemblance. The first step is **detection**, asking whether a population contains coordination; the second is **attribution**, asking which agents did it. The field has built careful instruments for the first, reviewed in §2.2, and has largely assumed the second follows. It does not. A shared model lifts the dependence of

every pair at once, and a secret agreement lifts *one* pair. A detector scoring the most suspicious-looking pair reads how similar that pair is, and similarity is what a shared model supplies for free. The informative quantity is how much more alike a pair is than its neighbours, which we call the **pattern** across pairs, as against the **level** any single pair reaches.

We test that distinction twice. A constructed process lets us make the two causes produce exactly the same distribution, so they are **observationally equivalent** and no detector can separate them; six deployed language models then tell us how close real agents come to that limit unprompted. §3.1 sets out why neither arm suffices alone.

A pair observed on its own turns out to be unattributable and a population is not. Detectors scoring the top pair do not merely weaken under homogeneity, they invert, ranking honest groups above linked ones in all 26 configurations tested, while detectors comparing pairs with one another recover completely. On the pricing task two models from different companies out-agreed six copies of any one model, a cartel matched to that pair was picked out by a correlation screen at the rate of a coin flip, and the coin flip repeats on 240 MMLU questions, where the outcome repair survives with a fraction of its margin.

2 Related Work

2.1 Definitions of collusion

A corpus of 139 published definitions, coded against eighteen components, resolves into two literatures sharing no victim, no test and no unit of analysis. The economics family, 89 papers, asks what competing firms’ prices do: collusion is a price above the static Nash equilibrium, the injured party is the consumer, and the test runs on the price series (Calvano et al., 2020). The AI-safety family, 44 papers, asks what supervised agents concealed: collusion is coordination defeating an overseer, and prices need not enter (Rose et al., 2026). A market can satisfy the first with nothing hidden; a multi-agent system can satisfy the second with nothing to price.

The components this paper turns on are the contested ones. Communication is required by 19 definitions and denied by 47, intent by 15 and denied by 34, agreement by 12 and denied by 26 — the elements United States antitrust law uses to infer an agreement, so much of the computational literature has defined away the legal concept it is taken to serve.

Common cause is the least addressed. Whether correlated behaviour arising from a shared model, training run or vendor counts as collusion is addressed by only 18 of the 139, with six saying it counts, ten that it can, and two that it does not. Neither literature was forced to decide: the economics definitions assume each firm runs its own pricing process, so two competitors executing one procedure never arose, while the AI-safety definitions describe systems whose agents are copies by design, so counting shared origin would classify nearly every system studied as collusive. The case now before the courts, a shared origin spanning firms, falls in the gap. Cao & Hu (2026) treat it as constitutive, locating collusion “not [in] strategic reasoning within a single context window” but in correlated recommendations from a shared latent preference; their paper is now titled *Supracompetitive Pricing Under AI Monoculture*.

2.2 Detection methods and their validation

Recent multi-agent detectors report strong numbers on their own benchmarks. Nakamura et al. (2026) measure a coalition advantage against a baseline with no secret channel, across 5 to 20 seeds, with no ROC curve. Rose et al. (2026) train linear probes on activations and aggregate per-agent scores to a scenario verdict, reporting AUROC 1.00 in distribution and 0.73 to 0.93 under environment shift. Tailor (2025) calibrates with 200 honest and 200 colluding runs per scenario and 10,000 honest audits. Nguyen et al. (2026) score action traces on five dimensions. None of them controls for the agents sharing a base model.

Two definitions in the corpus are themselves statistics. Bonjour et al. (2022) call a pair colluding when each member’s actions predict the other’s better than any third agent’s do, subtracting the largest outside figure and thresholding the remainder; the method assumes exactly two colluders. Xu et al. (2026) define collusion as co-movement in unexplained bid residuals while each agent’s own price distribution looks normal, so a test reading one agent at a time is blind by construction. They report same-model residual correlation of +0.053 against +0.0001 across models, and concede their statistics cannot separate conspiracy from lawful multi-identity operation.

Econometric screening reports what such tests cost: Imhof & Wallimann (2021) flag 6.9% to 14.6% of honest coalitions across three countries and four algorithms, on balanced test sets. The near-zero false-positive rates in the recent literature are an order of magnitude better than this older field achieves on a problem no easier.

2.3 Identification of conduct

The impossibility result is old. Lau (1982) showed price and quantity data cannot reveal how competitive an industry is, since many degrees of competition fit the same data; Bresnahan (1982) showed the ambiguity lifts when something rotates demand rather than shifting it. Bajari & Ye (2003) address collusion directly, and their Theorem 6 states that “if there is no variation in z and A1–A5 hold, then competition is observationally equivalent to” a cartel that is not all-inclusive. Their example is a cartel whose members each “compute [their] competitive bid and then submit a bid of 1.1 times” it, which passes every one of their own tests.

Later work changes the target instead of solving identification. Ortner, Chassang, Kawai & Nakabayashi (2022) build screens a competitive bidder almost surely passes whatever the conditions, so failure is informative though passing proves nothing. Hartline, Long & Zhang (2024) audit hindsight regret instead of prices, and note their non-collusion definition permits correlation between sellers.

2.4 Algorithmic monoculture

Kleinberg & Raghavan (2021) analyse the welfare consequences of many decision-makers sharing one algorithm, and Bommasani et al. (2022) study outcome homogenisation under shared foundation models. Kim, Garg, Peng & Garg (2025) evaluate more than 350 language models and find errors substantially correlated, with the larger and more accurate models showing highly correlated errors across distinct architectures and providers. Jo, Garg & Raghavan (2026) argue that judging whether models agree too much requires first fixing what normal agreement is.

That literature concerns welfare, fairness and correlated failure, and does not address detection. The connection has been named without being studied. Idowu et al. (2026) state that “the fundamental challenge in AI monitoring is distinguishing coordination from correlation,” listing it among open problems, and Hammond et al. (2025, fn. 16) observe that correlated decisions in AI agents “might be due not only to explicit communication but also to a common cause or process.”

3 Methodology

3.1 Two arms, and why neither is sufficient alone

Telling an agreement from a common cause requires knowing which one produced the behaviour, and the existing literature cannot supply that. Fish et al. (2024), Keppo et al. (2026) and Tian (2026) infer collusion from prices against a Nash benchmark, and Agrawal et al. (2025) add a judge scoring intent from chain of thought. In each case the thing to be detected is reconstructed from the behaviour the detector reads, which is the circularity we examine. Xu et al. (2026) state the problem and decline to solve it, conceding that “lawful multi-identity operation and conspiracy are behaviourally indistinguishable here.”

We break the circularity by stipulating the generating structure, in two arms.

The simulation arm (§§4.1–4.6) asks whether a detector could separate the two causes at all. Identifiability is a property of the joint distribution over actions and contexts, whatever produces it, which is the point Lau (1982) and Bresnahan (1982) establish for conduct parameters and Bajari & Ye (2003, Thm. 6) for bidding. If two processes induce the same distribution, every function of the data takes the same value under both, and no detector separates them however much data it is given. Showing this requires setting the two processes to produce the same distribution exactly, so we write them ourselves, and Appendix B derives the solution.

The model arm (§§4.7–4.9) asks whether the geometry appears in deployed models, and how large it is. A language model’s output distribution cannot be set to an arbitrary target, so this arm cannot establish observational equivalence and does not try. It establishes a magnitude: how much six instruction-tuned models from six companies agree unlinked, how that compares with copies of one model, and how strong a planted cartel must be to match it. The simulation shows what is impossible and not how often it arises; the models show what arises and cannot show impossibility.

Neither arm uses an external dataset, and there is no training, fine-tuning or held-out split anywhere. The simulation arm draws actions from a Markov process we specify. The model arm reads, for each of eight prompts, the **logits** the model assigns to the four price-option tokens at the next position, which is one forward pass with no text generated, normalises them to a probability over the four options, caches it, and samples every action from that cache. Temperature enters only at sampling.

3.2 Scope

The simulation arm has no prices, payoffs, text or learning, which rules out three families of claim for §§4.1–4.6: outcome-based measures cannot be computed there, the reward-and-punishment mechanisms of the Calvano tradition are absent, and with no chain of thought an LLM judge would measure the absence of text. What remains is the family of dependence-based definitions, those of Bonjour et al. (2022)

and Xu et al. (2026), the two in the corpus that are themselves detection statistics and drop the outcome test.

The model arm has prices, so outcome measures *can* be computed, and §4.8 computes them; the benchmark arm’s outcome measure is correctness (§4.9). It has no payoffs, trades or memory between rounds, so it cannot speak to coordination emerging from repeated play, the question Fish et al. (2024) and Agrawal et al. (2025) address. Every cartel here is planted and none is learned. Appendix A.4 lists what the environment takes from the published setting it is calibrated to, and what we invented.

3.3 The environment

The agents act simultaneously for T rounds over an action set of size k . A latent context s_t , which is drawn from m states, follows a Markov chain with persistence p . The context is our abstraction of the public covariates an auditor might model, the role cost and demand shifters play in Bajari & Ye (2003). Persistence is included because a persistent common cause is the canonical generator of innocent correlation, as Harrington & Imhof (2022) note for structural breaks produced by input prices or demand.

A balanced focal map, $f(s) = s \bmod k$, assigns a salient action to each context, and we require m to be divisible by k so that each action is focal equally often. The agents share a single policy, which is parameterised by a concentration κ :

$$\pi(a \mid s) = \begin{cases} \kappa & \text{if } a = f(s) \\ (1 - \kappa)/(k - 1) & \text{otherwise.} \end{cases}$$

The parameter κ measures how alike the agents are when given identical information. It is the discrete analogue of the coupling Xu et al. (2026) identify, arising from a shared function over public data with no message exchanged, of the shared infrastructure Cao & Hu (2026) make constitutive of collusion, and of Hammond et al.’s (2025, fn. 16) common cause.

3.4 The two generators

We name the generators after the structures producing them, since their legal status is contested (§2.1). Under the **shared-prior** generator, every agent draws $a_t^i \sim \pi(\cdot \mid s_t)$ independently. There is no channel, there is no shared reward and there is no coordinating mechanism of any kind, so that the correlation arises only from the common context and the common policy.

Under the **pair-mechanism** generator, one designated pair is linked. In the contemporaneous variant, both members play a jointly drawn $c_t \sim \text{Unif}(A)$ with probability λ , and otherwise they draw independently. In the lagged variant, one member replays the other’s previous action with probability λ , included because net influence (Bonjour et al., 2022) is a lagged statistic, and testing it against a contemporaneous link would measure a mismatch of envelope.

The four pairwise detectors, the five aggregation forms and the context-observability parameter φ are specified in Appendix A.8–A.9.

3.5 The matched construction

Xu et al. (2026) fix the marginal distributions at their competitive values and vary only the dependence between agents, so that a conspiracy changes the dependence and nothing else. We go one step further and match the entire joint distribution.

If we write $q = (1 - \kappa)/(k - 1)$, then let $D(\kappa) = (1/k) [\kappa^2 + (k - 1)q^2]$ be the per-cell diagonal mass of the pair distribution once the context has been marginalised out. For any choice of (κ_C, λ) we solve

$$D(\kappa_I) = \lambda/k + (1 - \lambda) D(\kappa_C)$$

by bisection on the interval $[1/k, 1)$, on which D is strictly increasing. Because the distribution is exchangeable over actions and must normalise, matching the diagonal forces the off-diagonal to match as well, and the two generators therefore become exactly equal in distribution over the observed actions once the context has been marginalised out. We give the derivation in Appendix B.

The runs that we report here use $\kappa_C = 0.40$ and $\lambda = 0.35$, which give a matched value of $\kappa_I = 0.709891$. The lagged variant admits no closed form, and so we obtain $\kappa_I = 0.792854$ from eighteen steps of bisection on the empirical lagged diagonal mass at $T = 400,000$.

4 Results

4.1 The construction holds

Both preregistered gates pass. The analytic total variation distance between the two pair distributions is 3.5×10^{-17} , and the empirical distance at 200,000 rounds is 0.0042, within the preregistered threshold of 0.01.

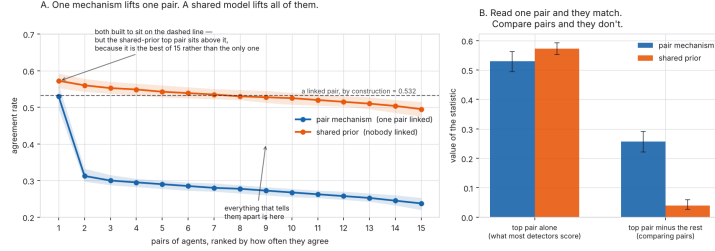


Figure 1. All fifteen pairs in a six-agent group, ranked by agreement. The two generators coincide at the top of the ranking and separate below it.

Aggregation form	Reads	AUROC
Maximum over agents	level	0.097 [0.071, 0.127]
Mean over pairs	level	0.000 [0.000, 0.000]
Top few against the rest	pattern	1.000 [1.000, 1.000]
Likelihood ratio, indep. null	level	0.097 [0.071, 0.127]
Likelihood ratio, homog. null	pattern	1.000 [0.999, 1.000]

4.2 A single pair cannot be attributed, and a group can

Figure 1 ranks all fifteen pairs in a six-agent group by how often they agree. The two generators coincide at the top and separate below it. The linked pair agrees 53.0% of the time and the top pair in a group where nobody is linked agrees 57.3%, both placed at 0.532 by construction, the latter exceeding it as the largest of fifteen draws rather than the only one; among the remaining fourteen pairs the groups diverge, 27.5% where one pair is linked against 53.0% where the similarity is shared.

A detector reading only the top pair is comparing 0.530 with 0.573 and cannot separate the generators. One subtracting the median of the rest is comparing 0.256 with 0.039.

4.3 The detectors that score the top pair invert

With no part of the shared context visible, all four pairwise detectors score below chance: Cramér’s V at 0.201 [0.16, 0.24], marginal mutual information at 0.195 [0.15, 0.24], conditional mutual information at 0.195 [0.15, 0.24], and net influence at 0.056 [0.04, 0.08]. No interval includes 0.5, so the detectors are not merely uninformative; they point reliably at the wrong group. This confirms P1 and P2.

Net influence failed its positive control here, reaching 0.648 with context fully visible where the pre-registration required 0.75, so under the rule in Appendix A.10 we rebuilt the link to operate across time, the setting the statistic was designed for. It then reached 1.000 with full context, and 1.000 with no context at all, for reasons taken up next.

4.4 Which shapes of detector fail

We tested five forms of group aggregation with no context visible. The mechanism set out in §1 predicts that forms reading the level of dependence invert and forms reading its pattern across pairs survive. All five outcomes matched.

The last two rows are the same test on the same data, differing only in what the null treats as ordinary: against a null of independent agents it scores 0.097, and against one allowing every pair to be dependent at some common unknown level, estimated from the data, it scores 1.000.

This also explains net influence: its contrast subtracts the largest outside influence, cancelling a uniform lift, which is why it survives homogeneity where the marginal statistics do not.

4.5 Three repairs

Observe the shared context. Conditional mutual information rises from 0.195 with no context visible to 0.891 at a quarter, 0.992 at a half, and 1.000 at three quarters and above (Figure 3A). The two statistics that never condition stay flat at 0.201 and 0.195 across the range, confirming P3. This is the strongest repair and the least available for language agents, since much of what they share is the model prior.

Specify the null. Reported in §4.4, where one test scored 0.097 against an independence null and 1.000 against a homogeneous one. It needs no observation of context, because the null is estimated from the data.

Divide by the spread. As agents grow more alike the average dependence between them rises 22.5-fold while its spread rises 4.6-fold, so the raw contrast rises 3.7-fold and a threshold calibrated on

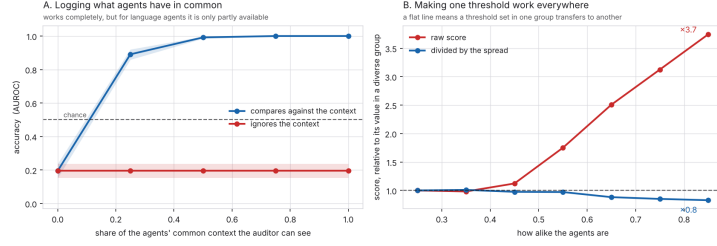


Figure 3. (A) Conditioning on observed context restores power. (B) Dividing the contrast by the spread of pairwise scores stops thresholds drifting between populations.

a diverse group fires constantly on a homogeneous one. Dividing the contrast by the spread cuts that drift to 0.8-fold (Figure 3B).

4.6 False accusations, robustness, and one failed prediction

With thresholds frozen at the 95th percentile of a diverse group and applied to groups where no pair is linked in any run, the false-accusation rate climbs with homogeneity: from 0.035 at calibration to between 0.49 and 0.525, and then to 1.000 as agents grow more alike. Net influence degrades more slowly, from 0.025 to 0.885. Every alarm counted here was raised on a group containing no coordinating mechanism. This confirms P4.

Varying the number of actions, contexts, agents and rounds, the persistence of the context, the link strength and the baseline similarity, one factor at a time, produced 26 cells, and the level-reading statistic’s interval lay entirely below 0.5 in all 26 while the pattern-reading statistic’s lay entirely above 0.5 in all 26 (Figure A1). The gradients, and a check that the inversion survives three definitions of per-agent suspicion, are in Appendix A.10.

P5 does not hold. False-accusation rates were 0.900, 0.917, 0.890, 0.867, 0.893 and 0.887 for groups of three to ten agents, flat across the range (Figure A2b). The account is in §4.5: the contrast cancels the average lift but not the spread, which grows with similarity whatever the group size, so a published report of innocent same-model pairs flagged in 7 of 10 games reflects a threshold that does not transfer, and not a shortage of agents to compare against.

4.7 The same question asked of language models

To see whether the geometry survives real model outputs we ran six instruction-tuned models from six organisations as sellers in the pricing environment of §3.3, calibrated to the price scale of Agrawal et al. (2025), for 200 rounds and 24 seeds at three temperatures. Appendix A.4 lists what that environment does and does not take from their design.

The first measurement needs no cartel. Among the six, with no link of any kind and nothing shared but the public demand signal, the most-agreeing pair was Google’s gemma-3-1b-it and Microsoft’s Phi-4-mini-instruct, at 98.8% of rounds at temperature 0.3, remaining the top pair at 0.871 and 0.771 as temperature rose. Replacing the six with six copies of one model gave a top pair of 0.973 for Qwen2.5-0.5B, 0.801 for SmolLM2-1.7B and 0.607 for Falcon3-1B. In all nine combinations of base model and temperature the six different models presented a more suspicious top pair than six identical ones.

Sharing a model is one route to behavioural similarity and not the strongest available, so a screen tuned to the thought that identical agents are the problem is tuned to the wrong variable.

That determines what a matched comparison can test, because raising the linked pair cannot equalise groups an innocent pair already tops. In the three cells where matching was possible the level statistic scored 0.439, 0.572 and 0.559, every interval containing 0.5; in the other six the screen named an innocent cross-organisation pair in 96% to 100% of runs, never the cartel.

4.8 Two pairs a screen cannot tell apart

That comparison has a flaw the next run removes: the linked pair contained gemma-3-1b-it, which is also half of the convergent innocent pair, so planting the link overwrote gemma’s own choices and the convergence fell from 0.988 to 0.854 as a side effect. We therefore repeated it inside one group with two disjoint pairs. The innocent pair is gemma-3-1b-it and Phi-4-mini-instruct, untouched. The linked pair is Qwen2.5-0.5B and Falcon3-1B, whose natural agreement is 0.011, coordinating by posting \$110 together with probability λ . We set λ so the two pairs agree at the same rate, making them identical on the statistic a correlation screen reads.

Matching took λ of 0.990, 0.865 and 0.734, leaving agreement rates differing by 0.000, 0.004 and 0.008. A cartel has to coordinate on between three quarters and ninety-nine percent of all rounds before it is as conspicuous as two honest agents that think alike.

Equalised, a screen ranking pairs by agreement named the real cartel in 54% of runs at temperatures 0.7 and 1.0. At 0.3 both pairs agree on roughly 198 of 200 rounds, exact ties are common, and our implementation resolves ties towards the lower-indexed pair, which is the cartel, so the figure rises to 75% for a reason unrelated to the screen. Taking the two higher temperatures as the clean measurement, a screen reading agreement alone chooses between a cartel and an innocent pair by coin flip.

The two pairs are equal on agreement and unequal on price. The cartel agrees at \$110 and prices above the competitive level on 99.0%, 86.3% and 74.9% of rounds. The innocent pair agrees at exactly \$90.00, the competitive price, and prices above it on 0.0% of rounds at every temperature. Two models from two organisations converge, and what they converge on is the competitive answer. Ranking pairs by agreement multiplied by the share of rounds both priced above that level named the cartel in 100% of runs throughout.

The negative control bounds the repair. On an honest market with no pair linked, the correlation screen accuses gemma-and-Phi in 100% of runs and the price-augmented screen accuses Falcon3-and-SmolLM2 in 100%. Neither can abstain, since both take a maximum over pairs with no threshold. Price changes which pair is accused when a cartel exists and does not help decide whether one exists, so it repairs attribution and leaves detection where it was.

One further measurement bears on which channel is worth studying. Told in the prompt that they have agreed to hold \$110 and not undercut each other, sellers hold that price with probability 0.423 for Qwen2.5-0.5B, 0.083 for Qwen2.5-1.5B and 0.0000022 for Qwen2.5-3B, monotonically at all three temperatures. The largest is not refusing, since only 2.8×10^{-7} of its mass lands outside the four price tokens; it reads the agreement, answers, and prices competitively. Instructed collusion is already failing at three billion parameters, leaving the channel that needs no instruction.

4.9 The same design on a public benchmark

The pricing prompts are ours, so the final run reruns the design on 240 MMLU questions sampled under a preregistered rule and seed, with pairs chosen by preregistered rules (Amendment 6). Every item has a ground-truth answer, which gives the confound its overseer reading: a pair sandbagging an evaluation must coordinate on wrong answers, and innocent look-alikes agree because both know the right one. The cartel mechanism is a focal rule computable from the question alone, both members answering the first incorrect option with probability λ , and the matching gate passed with gaps of at most 0.006.

Three results carry over. The matched cartel was named by the correlation screen in 46% to 50% of runs, a coin flip again [S3]. Innocent agreement concentrated on correct answers, with 70% of the top unlinked pair’s agreement being both-correct, the exam analogue of gemma and Phi holding \$90.00 [S2]. And the outcome screen generalised, since weighting agreement by the share of questions both answered wrongly named the cartel in 100% of runs at every temperature [S4].

The generalisation comes with its margin collapsed. In the market the innocent pair was never jointly supracompetitive, so the outcome dimension had the whole range to itself. On MMLU innocent models share errors: given that both members of a cross-organisation pair are wrong, they give the same wrong answer about half the time, 0.49 on average, which replicates the correlated-errors finding of Kim et al. (2025) on our sample [S5]. The cartel’s joint-error share of 0.58 stands against innocent pairs at 0.30 to 0.44, a margin of fourteen points where the market gave ninety-nine. On the honest exam the negative control accused the innocent Phi-and-granite pair in every run.

One pricing result reverses. On MMLU six copies of one model out-agree the six different models in all nine cells, with Falcon3 copies at 0.950 against 0.585 for the best unlinked pair, where the pricing task gave the opposite ordering in all nine. Which innocent pair looks most guilty is a property of the task and not of the models. Matching also became cheap, with λ of 0.29 to 0.31 against the market’s 0.73 to 0.99.

5 Discussion

5.1 Four properties decide whether a screen works

The results reduce to one distinction. A shared cause lifts the dependence of every pair at once and an agreement lifts one pair, so a statistic reading how high the top pair sits reads the wrong quantity, and one reading how far that pair stands above its neighbours reads the right one. Everything in §4.4

follows, including the two likelihood-ratio tests differing only in their null and scoring 0.097 and 1.000 on identical data.

A screen is safe against this failure to the extent that it has three structural properties, each measured in §4.5. It should condition on whatever shared context is observable, lifting conditional mutual information from 0.195 to 1.000; take as its null a population in which every pair may be dependent at some common unknown level, the difference between 0.097 and 1.000; and divide its contrast by the spread of the pairwise scores, cutting threshold drift from 3.7-fold to 0.8-fold. None requires new data or a new detector. The first is the weakest here, because conditioning works by explaining away the common cause, and for language agents much of that cause is the model prior, which no auditor observes.

The model arm supplies a fourth repair the simulation could not, having no prices: weighting agreement by the share of rounds priced above the competitive level moved attribution from a coin flip to 100% between two pairs matched exactly on agreement. This is the oldest idea in cartel screening, correlation saying two parties do the same thing and the outcome dimension saying whether it is objectionable. Two bounds apply: the weighting changes who is accused and not whether anybody should be, and §4.9 shows its headroom is task-dependent, fourteen points on MMLU where the market gave ninety-nine, because innocent models share wrong answers.

All four address one failure, a screen mistaking a shared cause for a two-party agreement. They do not address the reverse, and §5.3 sets out the configuration in which three fail and the second actively misleads.

5.2 Similarity does not respect organisational boundaries

The natural reading of algorithmic monoculture is that risk concentrates when deployments share a model, and the natural defence is to require vendor diversity. The two model arms, taken together, do not support it. On the pricing task two models from two companies agreed on 98.8% of rounds, above every single-model group; on MMLU the ordering reverses, and six copies of one model out-agree the six different ones in all nine cells. Neither sameness nor difference of vendor predicts behavioural similarity, because the ordering belongs to the task. Vendor diversity is auditable and behavioural similarity is not, so a market can satisfy a diversity requirement and stay just as correlated. What needs measuring is the agreement itself, against a baseline of normal agreement for that population and task, which Jo, Garg & Raghavan (2026) argue comes first.

Whether the alarms we count as false are false at all is unsettled. Of 139 published definitions only 18 say whether behaving alike through a shared model counts as collusion, and 16 of those say it counts or can count; under Cao & Hu (2026) the same runs are collusive and the same numbers become detection rates. We adopt the agreement-based sense throughout, marking both readings where they diverge, as in §4.6. The disagreement decides what the detectors in §4.3 are doing, and no measurement settles it; what measurement shows is the consequence, since the common-cause definition classifies as collusion a large and growing share of ordinary deployment, including, on the evidence of §4.7, agents from different vendors that were never intended to coordinate and cannot.

5.3 The configuration our repairs cannot read

The two appellate decisions in §1 turn on the question this paper formalises, and the evidence they rest on is relational: what carried *Cornish-Adebiyi* past the pleading stage was the alleged data exchange and the rate of adherence to the tool’s output, with no weight on any single defendant’s prices — the same move as scoring a pair against its neighbours instead of on its own.

That case also exposes a limit of this paper. Our taxonomy has two shapes, a shared cause lifting every pair and an agreement lifting one, and the arrangement alleged in *Cornish-Adebiyi* has the first, since many competitors feeding one pricing system is a hub-and-spoke, raising the dependence of every pair uniformly. Its behavioural signature is the signature of innocent monoculture: one is an agreement and one is a coincidence of design, and they produce the same thing.

Three of the four repairs fail there, and one fails in the worst way. Conditioning cannot help, because the common input is the shared system’s recommendation, which the auditor never sees, and studentising rescales a quantity near zero for every pair. The homogeneous null is actively harmful: its baseline, a population in which all pairs are dependent at some common level, is exactly what a hub-and-spoke is, so the test absorbs the arrangement into its own null, and a repair protecting innocent monoculture protects a hub-and-spoke by the same mechanism. Only the fourth survives, because it never reads correlation.

Behavioural screening therefore has a blind spot that no care about aggregation will close, courts are right to reach for data exchange and adherence rates, and the repairs guard against mistaking monoculture for a two-party cartel, and not against a cartel shaped like one. Protocol, gates, parameters, falsification conditions and future work are in Appendix A.10–A.13.

Appendices

A Limitations and Dual-Use / Ethical Considerations

A.1 The main arms are a constructed process

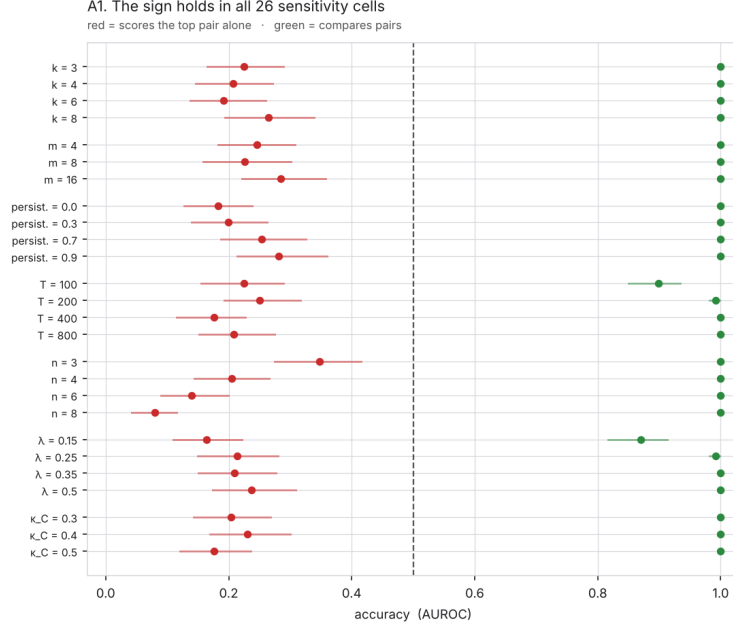


Figure A1. The 26 robustness cells. Level-reading intervals lie wholly below chance; pattern-reading intervals lie wholly above it.

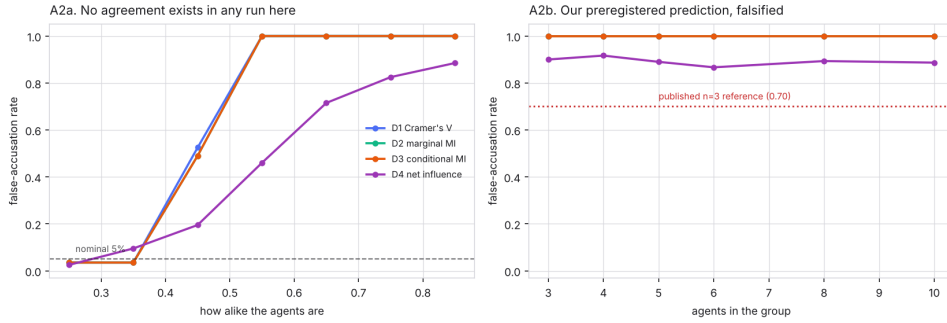


Figure A2. (a) False-accusation rate against homogeneity at frozen thresholds. (b) The same rate against group size, which is flat.

Arms 1 through 7 run on a stochastic process with four agents, four actions, eight context values and a context persistence of 0.7, over 400 rounds and 200 runs per cell. No language model produces any action in those arms; the six models, and the Hugging Face access they require, enter only in §§4.7–4.9. We take this route because identifiability is a property of the joint distribution over actions and contexts, whatever generates that distribution, which is the point Lau (1982) and Bresnahan (1982) establish for conduct parameters and Bajari and Ye (2003, Theorem 6) establish for bidding. If two generators produce the same distribution, no function of the observed data can separate them, and that holds whether the generators are Markov chains or language models. The construction lets us control the matching exactly, which a language model does not allow.

The cost is that we cannot say how often the matched condition arises among real models. We show that when it arises the detectors invert, and we do not show that it arises often. Section 4.6 gives the

gradient of false accusations in similarity, which is the closest the simulation comes to a rate, and that gradient is a property of our process. It measures no deployed population.

A.2 The construction is a worst case and the magnitudes should not be read as field estimates

The two generators are matched to a total variation distance of 3.5×10^{-17} analytically and 0.0042 empirically, at $\lambda = 0.35$, $\kappa_C = 0.4$ and the solved $\kappa_I = 0.7099$. We chose the parameters to make the two conditions as close to indistinguishable as the parametrisation allows, because the question is whether a detector can separate them in principle. Real monocultures will not be matched to that precision. An AUROC of the value we report is therefore an upper bound on how confused a top-pair screen can become. It is no estimate of how confused any particular screen currently is. What we claim transfers is the direction of the failure and the mechanism behind it, which Section 4.4 isolates as the difference between a uniform lift across all pairs and a lift confined to one pair. The size of the effect does not transfer.

A.3 We test shapes of aggregation, and no deployed system

The four detectors and five aggregation forms are our reconstructions of forms described in the published literature, and they are not the code of any deployed screen. We did not run HELIOS, any competition authority’s screening tool, or any proprietary system, and we make no claim about the behaviour of those systems. A screen that conditions on context, specifies a homogeneous null, or studentises its threshold will not show the failure, which is exactly the content of the three repairs in Section 4.5. Readers should treat the results as a statement about a class of aggregation geometry, and should check which of the three properties their own screen has before concluding anything about it.

A.4 The language model arm is small, and its task is informed by the literature rather than reproduced from it

The price scale in the model arm comes from Agrawal et al. (2025), who set a seller valuation of \$80, a buyer valuation of \$100, and a competitive equilibrium price of \$90 as the midpoint of the two, with supracompetitive pricing defined as an ask above that midpoint. Those figures are theirs. The rest of the environment is ours, and the differences are substantial enough to state individually. Their market has five sellers and five buyers, and ours has six agents and no buyers. Their asks are continuous to the cent, and ours are four discrete levels at \$85, \$90, \$100 and \$110. They run thirty rounds, and we run two hundred. The public demand signal in our prompt has no counterpart in their design at all, and it is carried over from the simulation, where the detectors require a shared observable in order to have anything to condition on.

The absence of buyers has a consequence that goes beyond the list. Without clearing there are no trades, without trades there are no realised payoffs, and the incentive our prompt describes, under which an agent earns more per sale at a higher ask but risks losing buyers to a cheaper rival, is text the agent can never experience. Deviation from a high price costs nothing and holding to it earns nothing. That is why the coordinating pair in our runs is created by giving two agents a shared rule, because no arrangement can emerge from repeated play here, and it is why our arm cannot speak to whether language models arrive at collusion on their own, which is the question Agrawal et al. (2025) and Fish et al. (2024) address and we do not. Our arm asks a narrower question, which is whether the geometry we found in simulation survives contact with real model outputs.

Four further limits apply to that arm. The local runs use gemma3:1b, a one billion parameter model, and small models agree with each other more readily than large ones, which inflates the homogeneous baseline. Each round is an independent prompt carrying no history, so there is no memory, no reciprocity and no punishment path, and those are the three ingredients the tacit collusion literature treats as necessary for collusion to be sustained at all. The benchmark uses eight prompt templates, which is a probe. There is no dataset and no held-out split. In the Kaggle arm the six agents are six different model families of different sizes from different providers, so a difference attributed to model identity is confounded with parameter count, training corpus and chat template, and we cannot separate the three.

One reported figure needs an explicit caveat. The linked pair was recovered as the top-ranked pair in 0 of 3 seeds in the local arm, and a similar count appeared in the first Kaggle arm. Those runs share one fixed lineup of agents and one fixed choice of which pair is linked, so the count is one configuration observed several times, and not a sample of independent trials. The number is structural and it should not be read as a sampling estimate with an implied confidence interval. The same caution applies to §4.8, where one linked pair and one rival pair are compared across 24 seeds that vary only the sampling and the demand signals.

The benchmark arm in §4.9 has its own limits. It uses one 240-item sample of MMLU under one selection seed, so its figures carry sampling error the pricing arm’s cached design does not; the six models answer correctly on only 39% to 61% of items, which makes the joint-error region innocent pairs occupy unusually large; the sandbagging mechanism always shares its wrong answer, the case the same-wrong-answer screen is best suited to, and a cartel varying its wrong answers would be harder for it; and sandbagging is stipulated by a rule, since nothing in the setup gives a model a reason to fail the exam.

Three further defects belong to the experiment in §4.8 specifically. Our screens take a maximum over pairs and resolve exact ties towards the lower-indexed pair, which at temperature 0.3 inflates the correlation screen’s apparent accuracy from a coin flip to 75%, and we report the untied temperatures as the measurement for that reason. Neither screen can abstain, so on an honest market each accuses some pair in every run, and the figure of 100% for the price-augmented screen is an attribution rate conditional on a cartel existing, and says nothing about detection. Finally, the cartel holds the highest available price, which is the case most favourable to a price-level screen; a cartel that fixed at \$100 rather than \$110, or that varied its price, would be harder for that screen and we did not test it.

A.5 What a reader should not conclude

The paper does not show that existing screens are wrong in the field, that any particular enforcement action rested on a false positive, or that language models in deployment are being misattributed. It shows that a specific and common aggregation geometry, applied to agents that are similar for innocent reasons, produces an ordering that favours the innocent group, and it gives four modifications that remove that problem.

It should also not be read as a general defence of behavioural screening once those modifications are made. The repairs are directed at one error, which is reading a shared cause as a two-party agreement. §5.3 describes the configuration in which they do not help: an arrangement among many parties running through one shared system raises the dependence of every pair uniformly, which is the shape of innocent monoculture, and a screen built to protect monoculture from false accusation protects that arrangement by the same mechanism. A reader taking the four properties as a sufficient standard for admissible screening would have the implication backwards.

A.6 Dual-use

The mechanism in Section 4.4 is a recipe when read backwards. A screen that scores the most-agreeing pair against the rest of the group is blind to coordination that lifts every pair by the same amount, because the comparison subtracts the common component away. A party that wanted to coordinate without being flagged by such a screen would spread the arrangement across all pairs instead of concentrating it in one, and would thereby move from the regime the screen detects into the regime it cannot see. We state this plainly because the paper cannot be understood without it, and because the inference is short enough that anyone reading the results would reach it unaided.

The evasion is not hypothetical and it requires nobody to design it. As §5.3 sets out, many competitors feeding one pricing system produces a uniform lift across all pairs through its ordinary operation, and that is the configuration the repairs cannot read. A firm adopting a widely used pricing tool arrives in the blind spot without having tried to, and the arrangement alleged in the leading current case has exactly that shape. What we describe is therefore less a method an adversary might adopt than a property of an arrangement that already exists and is already before the courts. That is a further reason to set it out here, instead of leaving it to be discovered by a defendant arguing that a screen proves nothing.

We judge publication to be right on three grounds. The screens with this geometry are already in use, so the vulnerability already exists and is not created by describing it. The asymmetry currently favours whoever has worked it out privately, and a party with the resources to run a cartel has the resources to discover by trial which patterns draw attention, while the authority operating the screen has no comparable way to learn that its own aggregation is the weak point. The three repairs in Section 4.5 are published in the same paper as the vulnerability, and each of them closes the uniform-lift blind spot, so the disclosure arrives with its mitigation attached.

A second and distinct concern runs the other way. A result stating that a screen produces false accusations can be cited in a real proceeding by a party that is in fact coordinating, as a general argument that screening evidence is unreliable. The finding does not support that use. It is conditional on a screen that scores the top pair, does not condition on shared context, and compares against an independence null, and Section 4.5 shows that a screen holding any one of the three repairs behaves correctly on the same data. The honest summary is that the geometry of the aggregation determines whether the evidence means anything. Screening evidence is worth what its aggregation is worth.

A.7 Reproducibility and provenance

Every number in the paper comes from the code at <https://github.com/hectar-glitches/monoculture>, and `reproduce.sh` runs the full pipeline and checks that two runs produce byte-identical output. The preregistration in `DESIGN.md` was frozen before the arms were run and carries three dated amendments, one of which records a prediction that was falsified and is reported as such in Section 4.8. We report that failure because a preregistration that only ever confirms is not doing any work.

A.8 Context observability

The auditor observes s_t with probability φ , and otherwise observes UNKNOWN, and we pool all of the unobserved rounds into a single stratum. The parameter formalises the quality of an auditor’s covariate model, in that $\varphi = 1$ corresponds to a model that has been fitted on all of the observables and $\varphi = 0$ corresponds to a test that has no covariates at all. For language agents, $\varphi = 0$ is not a pathological corner case, because a substantial part of the context that they share is the model prior, and that cannot be logged.

A.9 Detectors and aggregation forms

We test four pairwise detectors. The first is Cramér’s V computed on contemporaneous actions, and the second is the marginal mutual information $I(a^i; a^j)$, and neither of these conditions on the context. The third is the conditional mutual information $I(a^i; a^j \mid \hat{s})$, weighted by stratum, which is the conditional-independence family of Bajari & Ye expressed in information-theoretic terms. The fourth is net influence, $\Gamma(i; j) = \gamma(i \rightarrow j) - \max_{l \notin \{i, j\}} \gamma(l \rightarrow j)$, where $\gamma(i \rightarrow j) = I(a_t^i; a_{t+1}^j \mid \hat{s}_t)$. We reimplemented this statistic from Bonjour et al. (2022, Eq. 3–5), because no implementation by the authors is publicly available. All of our estimators are plug-in estimators, and we report all information quantities in nats.

A second experiment moves up a level, from the pairwise statistics to the forms of group aggregation built on top of them, because the published detectors for multi-agent systems combine per-agent scores into a verdict about the scenario as a whole. Rose et al. (2026) use five such techniques, which include Peak Suspicion and the correlation-based Activation Alignment, and they give a generalised likelihood-ratio reading of Peak. Their detector reads model activations, which our setting does not produce, and we were unable to retrieve their formulas, so what we test is the aggregation structure that those names describe, computed over behavioural dependence scores. The forms that we test are a maximum over agents, a mean over pairs, a contrast between the top few agents and the rest, and two likelihood-ratio tests that differ only in whether the null class contains a homogeneous population. We do not reproduce their detector and we make no claim about it.

A.10 Protocol, gates and metrics

Two preregistered gates run before any result becomes admissible. The construction gate requires the analytic total variation distance to be below $1e-9$ and the empirical distance to be below 0.01 at $T = 200,000$. The positive-control gate requires that the conditional detectors exceed an AUROC of 0.75 when the context is fully visible, and a detector that fails this gate may not have a null result reported from it. Net influence failed that gate on the contemporaneous generator, where it reached only 0.648, and this is what led us to add the lagged arm instead of reporting a null.

Each cell uses 200 runs for each generator, except that the population arm uses 300 and the sensitivity cells use 100, and the seed offsets are disjoint so that no run is reused across cells. We report AUROC with 95% bootstrap intervals over 2,000 resamples, and we count ties as one half. For the false-accusation rates we follow the calibrate-freeze-holdout protocol that Tailor (2025) uses, in that we freeze the thresholds at the 95th percentile of a diverse population and then apply them unchanged to more homogeneous populations in which no pair is linked in any run.

We apply no family-wise correction. None of our claims rests on a single marginal cell, because the headline results are claims about sign with intervals that are far from 0.5, together with a count of 26 out of 26 across the sensitivity cells. Individual cells that fall near a threshold should be read as suggestive only.

How the false-accusation quantity should be labelled depends on the definitional question in §2.1. Under the agreement-based sense that we adopt it is a rate of false accusation, whereas under the definition of Cao & Hu (2026) the same runs are collusive and the same number is a detection rate. We report the number once and state both readings.

Preregistered predictions. P1 is that when no context is visible, the detectors that read the top pair alone score at or below chance. P2 is that they score below chance rather than at it, because a uniform lift raises every pair whereas a link raises only one. P3 is that the detectors which condition on context recover as visibility rises, while those that do not condition stay flat. P4 is that when no pair is linked anywhere, the rate of false accusation rises with similarity. P5 is that the false-accusation rate of

Parameter	Value	Note
agents n	4, or 6 for the aggregation arm (swept over 3–10)	
actions k	4 (swept over 3–8)	
contexts m	8 (swept over 4–16)	must be divisible by k
persistence p	0.7 (swept over 0–0.9)	
rounds T	400 (swept over 100–800)	200,000 to 400,000 for the matching checks
κ_C and λ	0.40 and 0.35	swept over 0.30–0.50 and 0.15–0.50
κ_I	0.709891 and 0.792854	matched, for the contemporaneous and lagged variants
φ	0, 0.25, 0.5, 0.75, 1	
runs per cell	200, or 300 and 100 as noted above	disjoint seeds

net influence falls with group size, because its contrast subtracts the largest of $n - 2$ alternatives and a group of three leaves only one such alternative.

Robustness detail for §4.6. The inversion deepens with group size, from 0.347 at three agents to 0.079 at eight, as the mechanism predicts. The repair weakens at short horizons, reaching 0.899 at 100 rounds, and against weak links, reaching 0.870 at a coordination probability of 0.15. The result is not an artefact of how agents are scored: under maximum, mean and sum aggregation of an agent’s pairwise links, the level-reading form inverted in all three cases (0.142, 0.000, 0.000) and the pattern-reading form worked in all three (1.000, 0.897, 0.897).

A.11 Parameters

A.12 What would change our mind

Three results would undercut the paper. A measurement showing that real deployed populations are far from the matched condition would mean the inversion is a worst case that does not arise, and §4.6 gives the gradient but no field rate, so this is the most important missing measurement. A demonstration that a widely used screen already specifies a homogeneous null would mean the second repair is already standard practice. A language-model setting with genuine payoffs, memory and punishment, in which coordination emerges on its own, might produce cartels whose signature differs from the one studied here, since every cartel in this paper is planted and none is learned.

A.13 Future work

Four measurements follow from this paper, in the order we would run them.

Where does innocent convergence land? This one decides how much the rest matters. In §4.8 the innocent pair agreed on the competitive price on every round it agreed, and the cartel had to hold the highest available price to match it on agreement, which is why the price dimension rescued attribution. If convergence under a shared prior is generally convergence on the competitive answer, the outcome dimension separates the two causes cheaply and the attribution problem is much smaller in practice than the simulation implies. If a monoculture can instead converge *above* the competitive level, because the models share a prior favouring high prices, the outcome dimension fails at the same moment the correlation dimension does and both repairs break together. One pair of models on one task cannot tell those apart. The experiment is to sweep tasks and model sets and record where the agreed price sits relative to the competitive benchmark.

How close are deployed populations to the matched condition? The inversion in §4.3 is an upper bound on confusion, established under exact distributional matching, and §4.6 gives the gradient of false accusations in similarity without giving a field rate. Estimating that rate needs pairwise agreement measured across real deployments of pricing or bidding agents, against a baseline of what agreement is normal for that population, which is the measurement Jo, Garg & Raghavan (2026) argue has to come first.

Does learned coordination look like planted coordination? Every cartel in this paper is imposed by a mechanism and none emerges from repeated play, because the model arm has no payoffs and no memory. A setting with clearing, realised profit and a punishment path would let coordination arise on its own, which is the regime Fish et al. (2024) and Agrawal et al. (2025) study. Whether a learned cartel produces the same pairwise signature as a planted one is open, and the repairs in §5.1 are known to work only against the planted kind.

What happens when the agreement is not on the extreme action? The price-augmented screen in §4.8 faced a cartel holding the highest of four prices, which is the case most favourable to it. A cartel fixing at an intermediate price, or varying its price across rounds, would be harder, and the

screen’s advantage over correlation alone should be measured as a function of how far the cartel’s price sits above the competitive level.

The compliance result in §4.7 says which of these matters most in the long run. Instructed collusion is already failing at three billion parameters, so the coordination that matters at scale is the kind needing no instruction, and that is the kind this paper shows is hardest to attribute.

B Derivation of the matched construction

This appendix derives the matching used in Section 3.5 and shows that equating one number is enough to make the two processes identical.

B.1 The two processes

Both processes run the same world. There are k actions and m contexts, the context follows a persistent chain, and each context s carries a focal action $f(s)$. The focal map is balanced, so each action is focal for m/k of the contexts, and the stationary distribution over contexts is uniform. An agent’s policy puts mass κ on the focal action and spreads the remainder evenly,

$$\pi(a \mid s) = \kappa \text{ if } a = f(s), \quad q = (1 - \kappa)/(k - 1) \text{ otherwise.}$$

Under the shared prior, written **I**, every agent draws independently from that policy with concentration κ . Under the pair mechanism, written **C**, every agent draws independently with concentration κ , except that one designated pair emits a common action drawn uniformly from the k actions with probability λ on each round, and draws independently otherwise.

B.2 The pair distribution with the context marginalised

Fix one pair and let P be the joint distribution of their two actions after the context has been integrated out, which is what an observer sees when the context is not recorded. Conditional on a context whose focal action is a , the two agents draw independently from the same policy vector $p^{(a)}$, so their joint is the outer product $p^{(a)} \otimes p^{(a)}$. Averaging over the k equally likely focal actions gives

$$P_{\text{indep}}(\kappa) = \frac{1}{k} \sum_a p^{(a)} \otimes p^{(a)}.$$

Reading one diagonal cell, the two agents both play c when the focal action is c , which contributes κ^2 , and when the focal action is anything else, which contributes q^2 in each of the remaining $k - 1$ cases. Every diagonal cell therefore carries

$$D(\kappa) = \frac{\kappa^2 + (k - 1)q^2}{k} = \frac{\kappa^2 + (1 - \kappa)^2/(k - 1)}{k},$$

and reading one off-diagonal cell (x, y) with $x \neq y$, the focal action is x in one case and y in one case, each contributing κq , and is neither in the remaining $k - 2$ cases, each contributing q^2 , so every off-diagonal cell carries

$$E(\kappa) = \frac{2\kappa q + (k - 2)q^2}{k}.$$

The two expressions do not depend on which diagonal or off-diagonal cell is read. $P_{\text{indep}}(\kappa)$ has one value on its diagonal and one value everywhere else.

Adding the pair mechanism places mass λ on the diagonal, spread uniformly across its k cells, and leaves the rest untouched,

$$P_C = (1 - \lambda) P_{\text{indep}}(\kappa_C) + \lambda I_k/k,$$

so P_C has diagonal cells $(1 - \lambda)D(\kappa_C) + \lambda/k$ and off-diagonal cells $(1 - \lambda)E(\kappa_C)$. It is a matrix of the same two-value form.

B.3 Why one equation suffices

A $k \times k$ matrix that is constant on its diagonal, constant off it, and sums to one has only one free parameter. Writing d for the diagonal value and e for the off-diagonal value, the total mass is $kd + k(k - 1)e = 1$, so $e = (1/k - d)/(k - 1)$ and d determines e . Both P_I and P_C are of that form. Equating their diagonal values therefore equates the matrices entirely, and the matching condition is the single equation

$$D(\kappa_I) = \lambda/k + (1 - \lambda) D(\kappa_C).$$

This is why Section 4.1 reports a total variation distance of 3.5×10^{-17} . The two distributions are equal, and what remains is floating-point error.

B.4 Solving for κ_I

Multiplying out $D(\kappa) = D^*$ gives a quadratic in κ ,

$$k\kappa^2 - 2\kappa + 1 - k(k-1)D^* = 0,$$

whose root in the admissible range is

$$\kappa_I = \frac{1 + \sqrt{k^2(k-1)D^* - k + 1}}{k}.$$

On the interval $[1/k, 1]$ the function D is strictly increasing, because $dD/d\kappa = 2[\kappa - (1 - \kappa)/(k - 1)]/k$ vanishes only at $\kappa = 1/k$ and is positive above it, and it maps that interval onto $[1/k^2, 1/k]$. The target satisfies $\lambda/k + (1 - \lambda)D(\kappa_C) \leq \lambda/k + (1 - \lambda)/k = 1/k$, with equality only when $\kappa_C = 1$, so a solution exists for every κ_C below one and is unique. The implementation in `processes.py` bisects instead of applying the closed form, and the two agree to machine precision.

B.5 The parameters used

At $k = 4$, $\kappa_C = 0.4$ and $\lambda = 0.35$, the off-focal mass is $q = 0.2$ and the cartel pair’s per-cell diagonal is $D(0.4) = 0.07$. The target is $0.35/4 + 0.65 \times 0.07 = 0.133$, and the solved concentration is $\kappa_I = 0.70989$. The two pair distributions both carry 0.133 in each diagonal cell and 0.039 in each off-diagonal cell. A group running a shared prior at concentration 0.71 and a group running a shared prior at 0.4 with one pair coordinating 35% of the time are, at the level of that pair’s observed actions, the same thing.

B.6 What the matching does not cover

The construction equates the context-marginalised distribution of the **designated pair**. It does not equate the joint distribution of the whole group, and Section 4.4 turns on exactly that difference: raising κ_I lifts all six pairs at once, while the pair mechanism lifts one. The matching is what makes the single-pair comparison uninformative, and the group comparison remains informative because the matching was never applied to it. Nor does the matching survive conditioning on the context, which is the first of the three repairs in Section 4.5, since the construction assumed the context was integrated out.

References

Every entry below was checked against the publisher’s or arXiv’s own record. Where a work is a preprint with no venue, it is listed as a preprint.

Agrawal, K., Teo, V., Vazquez, J. J., Kunnavakkam, S., Srikanth, V., & Liu, A. (2025). Evaluating LLM agent collusion in double auctions. arXiv:2507.01413.

Arunachaleswaran, E. R., Collina, N., Kannan, S., Roth, A., & Ziani, J. (2025). Algorithmic collusion without threats. *16th Innovations in Theoretical Computer Science Conference (ITCS 2025)*, LIPIcs vol. 325, 10:1–10:21. doi:10.4230/LIPIcs.ITCS.2025.10

Bajari, P., & Ye, L. (2003). Deciding between competition and collusion. *The Review of Economics and Statistics*, 85(4), 971–989. doi:10.1162/003465303772815871

Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., & Liang, P. (2022). Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Advances in Neural Information Processing Systems 35*, 3663–3678.

Bonjour, T., Aggarwal, V., & Bhargava, B. (2022). Information theoretic approach to detect collusion in multi-agent games. *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022)*, PMLR 180, 223–232.

Bresnahan, T. F. (1982). The oligopoly solution concept is identified. *Economics Letters*, 10(1–2), 87–92.

Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267–3297. doi:10.1257/aer.20190623

Cao, S., & Hu, M. (2026). Supracompetitive pricing under AI monoculture. arXiv:2601.01279. (Earlier versions titled *LLM Collusion* and *Collusive Pricing Under LLM*.)

Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). Algorithmic collusion by large language models. arXiv:2404.00806.

Glad, D. W. (2026, May 14). *Remarks at the Antitrust West Coast Conference*. U.S. Department of Justice, Antitrust Division, San Francisco, CA.

- Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., Smith, C., et al. (2025). Multi-agent risks from advanced AI. arXiv:2502.14143.
- Harrington, J. E., Jr., & Imhof, D. (2022). Cartel screening and machine learning. *Stanford Computational Antitrust*, II, 134–154.
- Hartline, J. D., Long, S., & Zhang, C. (2024). Regulation of algorithmic collusion. *Proceedings of the Symposium on Computer Science and Law (CSLAW '24)*, 98–108. doi:10.1145/3614407.3643706
- Idowu, J. A., Almasoud, A., & Alfahid, A. (2026). Mapping human anti-collusion mechanisms to multi-agent AI systems. *Knowledge-Based Systems*. doi:10.1016/j.knosys.2026.116067. Preprint arXiv:2601.00360.
- Imhof, D., & Wallimann, H. (2021). Detecting bid-rigging coalitions in different countries and auction formats. *International Review of Law and Economics*, 68, 106016. doi:10.1016/j.irle.2021.106016
- Jo, N., Garg, N., & Raghavan, M. (2026). The subjectivity of monoculture. *Advances in Neural Information Processing Systems 39 (NeurIPS 2026)*. Preprint arXiv:2602.24086.
- Keppo, J., Li, Y., Tsoukalas, G., & Yuan, N. (2026). On the fragility of AI agent collusion. arXiv:2603.20281.
- Kim, E., Garg, A., Peng, K., & Garg, N. (2025). Correlated errors in large language models. *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*, PMLR 267. Preprint arXiv:2506.07962.
- Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22), e2018340118. doi:10.1073/pnas.2018340118
- Lau, L. J. (1982). On identifying the degree of competitiveness from industry price and output data. *Economics Letters*, 10(1–2), 93–99.
- Nakamura, M., Kumar, A., Das, S., Abdelnabi, S., Mahmud, S., Fioretto, F., Zilberstein, S., & Bagdasarian, E. (2026). Colosseum: Auditing collusion in cooperative multi-agent systems. arXiv:2602.15198.
- Nguyen, K. L., Nghiem, H., & Tran, T. (2026). TraceGuard: Structured multi-dimensional monitoring as a collusion-resistant control protocol. arXiv:2604.03968.
- Ortner, J., Chassang, S., Kawai, K., & Nakabayashi, J. (2022). Robust screens for noncompetitive bidding in procurement auctions. *Econometrica*, 90(1), 315–346.
- Rose, A., Cullen, C., Abdelnabi, S., Torr, P., Kaplowitz, B. G., & Schroeder de Witt, C. (2026). Detecting multi-agent collusion through multi-agent interpretability. arXiv:2604.01151v4.
- Taylor, O. (2025). Audit the whisper: Detecting steganographic collusion in multi-agent LLMs. arXiv:2510.04303.
- Tian, Y. (2026). Prompt optimization enables stable algorithmic collusion in LLM agents. ICLR 2026 Workshop on AI for Mechanism Design and Strategic Decision Making. Preprint arXiv:2604.17774.
- U.S. Department of Justice. (2026, September 4). *Justice Department reaches proposed consent decree with Pinnacle, one of America’s largest landlords* [Press release].
- Xu, X., Wu, C., Lu, J., Tan, K., Tao, S., & Hong, H. (2026). Collusion with competitive marginals: Price-level audits are blind by construction. arXiv:2607.26385.

Cases and statutes

- Cornish-Adebiyi v. Caesars Entertainment, Inc.*, No. 24-3006 (3d Cir. July 29, 2026), rev’g No. 1:23-cv-02536 (D.N.J. Sept. 30, 2024).
- Gibson v. Cendyn Group, LLC*, 148 F.4th 1069 (9th Cir. 2025), cert. petition filed, No. 25-1109 (Mar. 23, 2026).
- United States v. RealPage, Inc.*, No. 1:25-cv-00014 (M.D.N.C., complaint filed Jan. 7, 2025).
- California Assembly Bill 325 (2025), amending Cal. Bus. & Prof. Code §§ 16729, 16756.1, effective January 1, 2026.