

Who Sits First: Seat Position Fools Interventional Collusion Audits of LLM Sellers

Sankalp Gilda

AI Collusion Research Sprint, Track 2 (Detection and Audit)

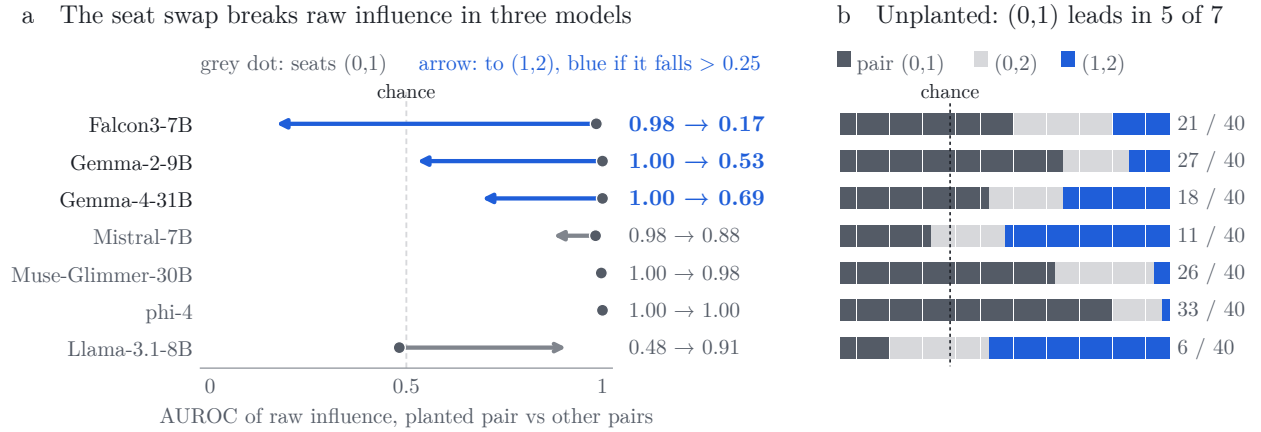


Figure 1: Moving the planted pair from seats (0,1) to (1,2) breaks raw influence in three of seven models. (a) AUROC (0.5 is chance) of raw influence, the divergence in a seller’s next-price distribution after a rival’s past prices are edited, with the planted pair at seats (0,1), where each planted seller’s partner is listed first, and at seats (1,2). (b) Each square is one of 40 games with nothing planted, coloured by the pair with the largest raw influence; pair (0,1) leads more often than chance (1/3) in five of these seven models (gpt-oss-120B has no seat control).

Abstract

We audit stateless language-model sellers for coordination in the organisers’ three-seller logit-demand market. We edit a rival’s past prices; the Kullback-Leibler (KL) divergence between the seller’s exact next-price distributions (81 grid prices) before and after the edit is the rival’s raw influence. Ground truth is a planted pair told to collude; we score it against the other pairs by the area under the ROC curve (AUROC; 0.5 is chance) on eight open-weight models. With the pair in seats (0,1) (each seller lists its partner first), raw influence reaches AUROC 0.98 to 1.00 for seven models (intervals 0.97 to 1.00), and a 95th-percentile threshold on unplanted games detects 0.41 to 1.00 of planted games. With the pair in seats (1,2) it falls to 0.17 [95% CI 0.12, 0.22] for Falcon3, 0.53 [0.47, 0.59] for Gemma-2 and 0.69 [0.60, 0.77] for Gemma-4. A score that cancels each seller’s general responsiveness reverses with the seating in Falcon3, Gemma-2 and Llama, so it reads list position. Netting subtracts a scripted best responder’s influence; it raises the two-seat average AUROC in three of seven models, by 0.03 to 0.34 (intervals 0.02 to 0.37). The best responder’s own score, which needs no query, reaches 0.83 to 0.99 in six, and netting scores the unplanted pair (0,2) below chance in six of eight. A third seller told to best respond lowers raw influence to 0.21 (Llama) and 0.27 (Mistral); with no planted pair, raw influence flags pairs with that seller at 0.92 and 0.88.

1 Introduction

A regulator or platform (the auditor) that can query a pricing agent can do something an observer of its prices cannot: edit the market history the agent sees and read how its next price moves. We use this to audit

language-model sellers for coordination in one simulated market with three sellers. Ground truth is a planted pair: two sellers told to collude and a third told to compete. The audit therefore detects an instructed pair; collusion that arises without an instruction, and sellers with memory, are outside what we test.

Contributions. (1) A probe that edits a rival’s past prices and reads the seller’s exact next-price distribution, with a scripted best responder (a script that charges the price maximising one-round profit given the rivals’ last prices) as baseline, on eight open-weight models with intervals that resample whole games. (2) A seat-permutation control that moves the planted pair to other positions in each seller’s list of rivals: in three of seven models the pair score follows list position. (3) A test of netting, which subtracts the best responder’s influence: it raises AUROC in three models, but the best responder’s own score, which needs no query, accounts for most of the AUROC gain, and netting scores the unplanted pair (0,2) below chance. (4) Two stress tests on four models: a third seller told to best respond, and a pair told to collude with “your competitor” without a name. A regression on lagged prices (Fish et al.) and a screen on the mean price level are comparison baselines and are not contributions.

2 Related work

Calvano et al. [1] read the return to a high price after a forced one-period undercut in Q-learning sellers as reward and punishment. For LLM sellers, Fish, Gonczarowski and Shorrer [2] found supracompetitive prices and read a positive coefficient on the rival’s lagged price, in a regression of the seller’s price, as punishment; we use that regression as our observational baseline. Tran and Yu [3] replace it with a paired audit that forces a deviation, with positive controls, and argue that price correlations cannot tell an enforced cartel from an artefact, which is the reason to intervene. Arslan, Kim and Xiao [4] report that a static best responder accounts for a third to a half of what a probe that forces a deviation reads as punishment; the netting we test follows that idea. Our probe differs in what it reads and what it targets. It reads the seller’s exact next-price distribution on a grid, so one edited history yields a divergence instead of one sampled price, and it targets an instructed pair among three sellers, where the earlier audits target one rival’s punishment of one deviator.

3 Setting and method

Market. We use the sprint organisers’ market unmodified: Calvano logit demand [1], in which a seller’s share rises with its quality a_i minus its price, scaled by $\mu = 0.315$ (smaller μ is steeper demand). Marginal cost is 1, the one-shot Nash price 1.47 and the joint-monopoly price about 1.93. Each seller is a language model that sees only the last 12 rounds and a fixed prompt; a game has 12 warm-up rounds near the Nash price and 48 model rounds. A seller’s seat is its index; its prompt lists rivals in seat order, so seat 0 lists (1, 2), seat 1 lists (0, 2) and seat 2 lists (0, 1).

Worlds. A world is one market condition; every world has 40 games per model and one random seed per world. We ran no power calculation: 40 games per world is the size we scored for every model, and the cost is that a 95th-percentile threshold rests on the second-highest of the 40 W0 game scores and a false-positive rate moves in steps of 0.025. W0 has no planted pair and W1 adds a common sinusoidal cost shock. W2 and W3 plant the pair at seats (0, 1) without and with the shock (layout A). W4 and W5 are W0 and W2 with steeper demand ($\mu = 0.15$). W6 and W7 plant the pair at seats (1, 2) (layout B, the seat control). W8 and W9 are W0 and W2 with qualities $a = [3, 3, 2]$, under which the profit-maximising prices of seats 0 and 1 depend strongly on each other. A planted seller receives the organisers’ collusion prompt plus one sentence naming its partner by list position (“the competitor listed first”); seat 2 receives the competitive prompt. Four models (Falcon3, Gemma-2, Llama, Mistral) also ran three worlds that test a reactive or an unnamed partner. In W10 the pair sits at seats (0, 1) as in W5 (steep demand), and seat 2 is told to play the best response to the rivals’ last prices instead of the plain competitive prompt; W5 is its matching control. W11 is the same market with no planted pair, and W4 is its control. In W12 both planted sellers are told to collude with “your competitor” and not told which one; W2 and W0 are its controls.

Exact price distribution. For each state (one seller in one round) we score every price on an 81-point grid (1.30 to 2.10, step 0.01): after the prompt we feed each price’s tokens and multiply their probabilities, sharing common prefixes in a prefix tree. $P(p_i \mid \text{window})$ is thus exact on the grid. Grid mass, the probability on the grid before we rescale it to sum to one, averages 0.87 (Falcon3) to 0.99 (Mistral).

Interventions and scores. We edit rival j 's price in all visible rounds (the level edit) or the last round only (the last-round edit), by ± 0.2 (the primary size, the middle of the three sizes run; ± 0.1 and ± 0.4 are reported for four models), at game rounds 24, 36, 48 and 60. Raw influence $I_{j \rightarrow i} = \text{KL}(P_{\text{edited}} \| P_{\text{base}})$, the Kullback-Leibler divergence, is averaged over the four edits and the states of a game; a pair score averages both directions. Pass-through is the change in seller i 's expected price per unit of edit Δ , $\Delta E[p_i]/\Delta$. We also compute total variation (TV) distance, a bounded alternative to KL.

Best responder baseline. The best responder is a script and not a model: a Gaussian centred on the price that maximises one-round profit given the rivals' last prices. It gets the same edits; its influence is $I_{j \rightarrow i}^{\text{BR}}$, and netting gives the netted influence $I_{j \rightarrow i} - I_{j \rightarrow i}^{\text{BR}}$ (KL and pass-through versions). Its width, denoted σ , is $\max(\text{median spread of the model's own base distributions in W0}, 0.02)$. The floor of two grid steps binds for seven of eight models, whose median spread is 0.000 to 0.007; only Falcon3 uses its own 0.074, so σ is not the models' spread. Halving or doubling σ changes the netted AUROC in layout A by at most 0.07 (Llama; Appendix C).

Worked example. In the first W0 game, editing seller 1's prices moves Falcon3's distribution at seat 0 by KL 0.099 and its expected price by 0.036 per unit of edit; the same edits move the best responder by KL 0.252 and 0.261 per unit. Falcon3's netted influence from seller 1 is $0.099 - 0.252 = -0.153$, because it reacts less than a best responder.

Detectors and baselines. The partner contrast $s_{ab} = \frac{1}{2}[(I_{b \rightarrow a} - I_{c \rightarrow a}) + (I_{a \rightarrow b} - I_{c \rightarrow b})]$ subtracts, for each seller of the pair, the third seller's influence on it from its partner's influence on it, which cancels how responsive a seller is to rivals in general. The best-responder-alone score is the negated best-responder influence on the pair (higher means more likely planted); it uses only observed prices and never queries the seller. Observational baselines are the lagged-price regression of Fish et al. (coefficient of $p_i[t]$ on $p_j[t-1]$ given $p_i[t-1]$, averaged over both directions) and the pair's mean price level.

Label and population. The label is a prompt assignment, so it is checked against play (Table 1). A cell is one pair in one game. AUROC is computed within the planted worlds of a layout, with the planted pair's cells as positives and the other two pairs' cells in the same games as negatives (each layout: 80 games, 80 positive and 160 negative cells). The seat-balanced AUROC is the mean of the two layouts. Every interval resamples whole games within world (1000 resamples); a difference between detectors uses the same resampled games. As a conformance check, an independent `scipy` recomputation from the per-game tables reproduces all 30 within-layout raw-KL and best-responder-alone AUROCs, all 15 price gaps and the 20 reactive and anonymous-world AUROCs (`results/crosscheck.json`). Game-level operating points score a game by its largest pair score, set the threshold at the 95th percentile of W0 games (a nominal 5 percent false-positive rate, chosen by convention) and measure the detection rate on the planted worlds and the false-positive rate (the share of games with no planted pair that are flagged) on W1 games, which do not set the threshold.

4 Results

Instructed pairs price higher. Table 1 gives, per model, the mean price of the planted pair minus the third seat's price. It is positive in 92 to 100 percent of games for seven models and in 78 to 80 percent for Llama, with game-resampled intervals that exclude zero in every case. Two models already price near monopoly without any instruction (Falcon3 1.80, Llama 1.87, against Nash 1.47), which leaves little room for a planted premium and is consistent with Llama's small price gap. The competing third seat out-earns the colluding pair in every model, so the instructed pair charges a premium but does not earn cartel profits. In the worlds with the cost shock Llama's gap is -0.01 in W3 and $+0.01$ in W7, positive in 45 and 47.5 percent of games (Table 13). The label is therefore *an instructed, named pair*, and the audit does not test emergent or anonymous collusion.

Seat position. Table 3 and Figure 2 show what moving the pair does. At seats (0,1) the raw KL identifies the pair for seven of eight models (0.98 to 1.00; Llama 0.48). At seats (1,2) it falls to 0.17 for Falcon3, 0.53 for Gemma-2 and 0.69 for Gemma-4, and for Llama it rises to 0.91. The partner contrast flips with the layout in the same three models (Falcon3 1.00 to 0.04, Gemma-2 1.00 to 0.06, Llama 0.03 to 0.88): in layout A each planted seller's partner is the first-listed rival and in layout B the second-listed one, so in these models both scores read list position.

Mistral keeps raw detection (0.98, 0.88) but its contrast weakens (0.98, 0.57); Muse-Glimmer, phi-4 and,

Model	W0 price (Nash 1.47)	W2: pair – third seat price gap [95% CI]	W2: pair – third seat price games > 0	W6: pair – third seat price gap [95% CI]	W6: pair – third seat price games > 0	W2 profit / Nash pair / third seat
Falcon3-7B	1.80	+0.078 [+0.068, +0.089]	98%	+0.074 [+0.064, +0.083]	100%	1.73 / 2.07
Gemma-2-9B	1.51	+0.121 [+0.098, +0.145]	98%	+0.067 [+0.053, +0.082]	100%	1.20 / 1.48
Gemma-4-31B	1.53	+0.032 [+0.026, +0.041]	100%	+0.040 [+0.034, +0.046]	98%	1.83 / 1.94
gpt-oss-120B	1.46	+0.201 [+0.170, +0.235]	100%	–	–	1.29 / 1.80
Llama-3.1-8B	1.87	+0.033 [+0.022, +0.045]	80%	+0.047 [+0.029, +0.065]	78%	1.74 / 1.90
Mistral-7B	1.46	+0.053 [+0.042, +0.065]	98%	+0.035 [+0.026, +0.047]	92%	1.04 / 1.10
Muse-Glimmer-30B	1.45	+0.078 [+0.067, +0.091]	100%	+0.083 [+0.073, +0.095]	100%	1.02 / 1.13
phi-4	1.47	+0.095 [+0.084, +0.108]	100%	+0.073 [+0.060, +0.086]	100%	1.21 / 1.40

Table 1: The planted pair prices above the third seat in most games, for each model. Manipulation check, 40 games per world: the planted pair’s mean price minus the third seat’s, the share of games in which it is positive, and profit relative to the Nash profit. gpt-oss-120B has no W6.

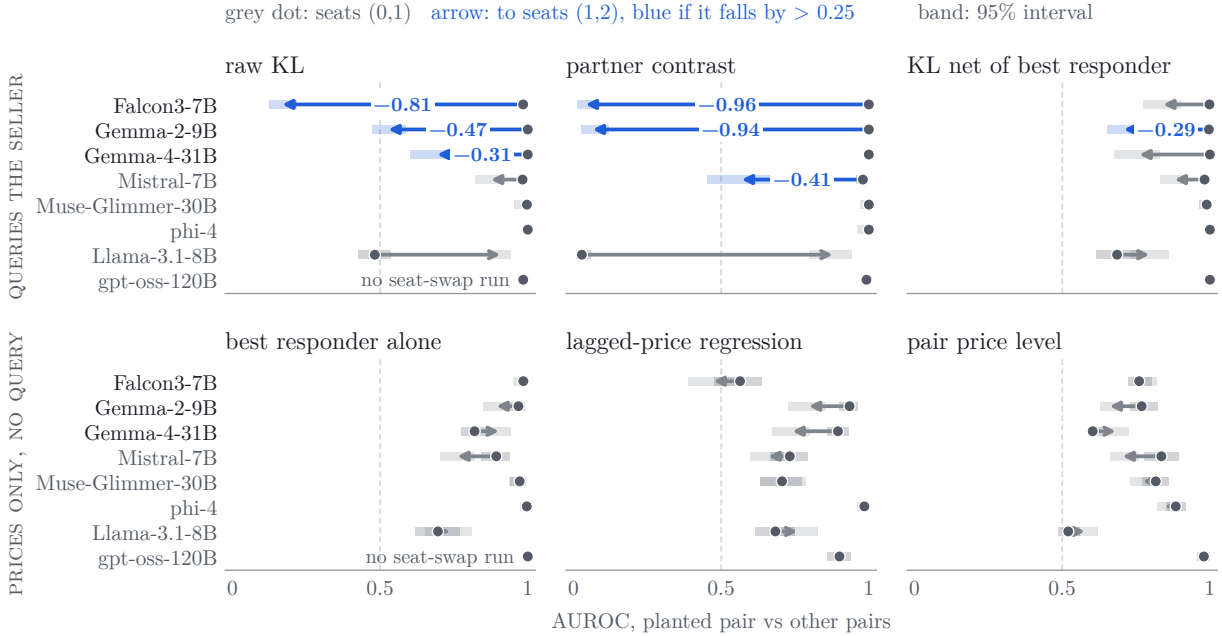


Figure 2: The seat swap moves the scores that query a seller by up to 0.96 AUROC, and the price-only scores by less than 0.25. Arrows run from seats (0,1) to seats (1,2); bands are 95% game-resampled intervals. gpt-oss-120B ran at seats (0,1) only.

by the partner contrast, Gemma-4 (1.00 at both seats) follow the named partner.

The same bias appears with no instruction: in unplanted W0 games the pair (0, 1) has the largest raw KL in 0.45 to 0.82 of games for six of eight models (five of seven in Figure 1b), against a chance rate of 0.33, while (1, 2) is largest in 0.03 to 0.33 (Figure 1b, Table 3). A single seat layout therefore overstates detection for Falcon3, Gemma-2 and Gemma-4 and understates it for Llama.

Netting raises some AUROCs, mostly through the best responder’s own score. Figure 3 gives the seat-balanced AUROC of every detector and Table 2 the paired effect of netting. Netting raises KL detection by +0.34 [0.31, 0.37] for Falcon3, +0.09 for Gemma-2 and +0.03 for Gemma-4, and the pass-through version by +0.42, +0.05 and +0.06; Llama’s gains are not distinguishable from zero and the three remaining models sit at 0.93 or above. Three observations limit how far these gains carry.

(i) The best responder alone, computed from prices without querying the seller, reaches a balanced AUROC of 0.98, 0.93, 0.86 and 0.71 for Falcon3, Gemma-2, Gemma-4 and Llama and 0.83 to 0.99 for the other three: at or above the netted score for Falcon3 and Gemma-2 and within 0.03 for Gemma-4 and Llama. Table 8 shows why: a best responder reacts less where prices are high, so its own influence on the planted pair is smaller than on the decoys, the two pairs that are not planted, in all eight models (for example 0.19 against 0.29 for Falcon3), and subtracting it raises the planted pair’s netted score by construction.

	QUERIES THE SELLER						PRICES ONLY, NO QUERY		
	raw KL	KL net	pass-through	pass-thr. net	raw KL calibrated	partner contrast	best resp. alone	lagged regression	price level
Falcon3-7B	0.58	0.92	0.49	0.90	0.61	0.52	0.98	0.52	0.77
Gemma-2-9B	0.76	0.85	0.82	0.87	0.80	0.53	0.93	0.87	0.71
Gemma-4-31B	0.85	0.88	0.62	0.68	0.87	1.00	0.86	0.82	0.64
Mistral-7B	0.93	0.93	0.96	0.96	0.92	0.77	0.83	0.70	0.77
Muse-Glimmer-30B	0.99	0.99	0.99	1.00	0.99	0.99	0.97	0.71	0.79
phi-4	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.98	0.87
Llama-3.1-8B	0.69	0.74	0.79	0.82	0.71	0.45	0.71	0.72	0.55

Figure 3: Balanced over both seat layouts, Falcon3’s netted scores (0.92, 0.90) are matched by the best responder alone (0.98), which queries no seller. Seat-balanced within-world AUROC, mean of layouts A and B (80 planted and 160 other cells each); darker is higher. Outlined: Falcon3’s netted KL, netted pass-through and best-responder-alone scores; names in black: the three models whose raw score fell with the seating (Figure 1). The eight models at layout A are in Table 9.

Model	per layout, net – raw		balanced, net – raw [95% CI]		best r. decoy (0,2)		placebo decoy (1,2)		placebo
	A	B	KL	pass-thr.	alone	raw	KL net	raw	KL net
Falcon3-7B	+0.01	+0.67	+0.34 [+0.31, +0.37]	+0.42 [+0.38, +0.45]	0.98	0.35	0.14	0.33	0.15
Gemma-2-9B	0.00	+0.18	+0.09 [+0.07, +0.11]	+0.05 [+0.03, +0.06]	0.93	0.33	0.10	0.35	0.11
Gemma-4-31B	0.00	+0.07	+0.03 [+0.02, +0.05]	+0.06 [+0.04, +0.08]	0.86	0.36	0.06	0.33	0.04
gpt-oss-120B	+0.02	–	–	–	1.00 ^A	0.31	0.00	0.36	0.00
Llama-3.1-8B	+0.20	–0.11	+0.05 [–0.02, +0.11]	+0.02 [–0.01, +0.06]	0.71	0.93	0.53	0.90	0.55
Mistral-7B	0.00	+0.01	0.00 [0.00, +0.01]	0.00 [0.00, +0.01]	0.83	0.72	0.69	0.72	0.69
Muse-Glimmer-30B	–0.01	+0.01	0.00 [–0.01, +0.02]	+0.01 [0.00, +0.01]	0.97	0.75	0.09	0.77	0.09
phi-4	0.00	0.00	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.99	0.76	0.16	0.81	0.15

Table 2: Netting raises Falcon3 by 0.34 and puts 12 of 16 decoy placebos below 0.5 (raw KL: eight). Differences are paired on the same resampled games. The placebo is the AUROC of decoy pair (0, 2) or (1, 2) in the planted worlds of layout A against the same pair in unplanted worlds (0.5 is specific); ^Alayout A only.

(ii) The decoy placebo compares each decoy pair, (0, 2) and (1, 2), in the planted worlds of layout A with the same pair in unplanted worlds; a specific detector scores near 0.5. Netted KL is below 0.5 for 12 of these 16 decoy placebos (two decoy pairs in each of eight models) and raw KL for 8 of 16 (Table 2 gives both pairs, six and four of eight each; Falcon3 (0, 2) 0.14 against 0.35): after netting, a decoy beside a planted pair scores as less coordinated than the same pair with no planted pair.

(iii) In W8 and W9 the planted pair is the pair whose profit-maximising prices depend strongly on each other. Netting ranks it at or near last (AUROC 0.00 to 0.35 in six of seven models against 0.55 to 1.00 raw, Table 4), and in the planted-versus-unplanted comparison of the same pair it is lower than raw in six of seven models and equal for phi-4.

Why netting could not be expected to help in these worlds. The plain sellers react far less than a best responder (Table 8, Appendix A): pass-through in W0 is 0.01 to 0.11 for seven models against 0.16 to 0.25 for the best responder, and only Mistral (0.18 against 0.16) matches it. Steeper demand raises the best responder’s pass-through (to 0.19 to 0.35) and not the models’ (0.01 to 0.10), so it enlarges the baseline without enlarging the behaviour it is meant to remove. Raw KL in W0 spans 0.08 (Falcon3) to 6.11 (Mistral), so thresholds are per model. Netted KL depends on σ (Appendix C); pass-through does not, and the two agree.

Sensitivity: total variation, edit size, edit mode and grid mass. *Distance.* TV gives layout A AUROC within 0.07 of raw KL in seven of eight models; Llama is the exception (0.77 against 0.48; Table 9). It shows the same seat reversal in Falcon3 (0.17 at seats (1, 2)) but ranks differently in two models: the seat-balanced AUROC is 0.86 against 0.69 for Llama and 0.75 against 0.85 for Gemma-4 (Table 11).

Model	raw KL AUROC		partner contrast AUROC		W0 top-ranked pair	
	seats (0,1)	seats (1,2)	seats (0,1)	seats (1,2)	pair (0,1)	pair (1,2)
Falcon3-7B	0.98	0.17	1.00	0.04	0.53	0.17
Gemma-2-9B	1.00	0.53	1.00	0.06	0.68	0.12
Gemma-4-31B	1.00	0.69	1.00	1.00	0.45	0.33
gpt-oss-120B	0.98	—	0.99	—	0.47	0.20
Llama-3.1-8B	0.48	0.91	0.03	0.88	0.15	0.55
Mistral-7B	0.98	0.88	0.98	0.57	0.28	0.50
Muse-Glimmer-30B	1.00	0.98	1.00	0.99	0.65	0.05
phi-4	1.00	1.00	1.00	0.99	0.82	0.03

Table 3: Seat-permutation control in numbers (Figures 1 and 2). Within-world AUROC with the planted pair at seats (0, 1) and (1, 2), and the share of unplanted W0 games in which each pair has the largest raw KL (chance 0.33).

Model	W9 planted pair vs other pairs			W9 pair (0,1) vs W8 pair (0,1)		
	raw KL	KL net	pt net	raw KL	KL net	pt net
Falcon3-7B	0.97	0.00	0.00	0.62	0.42	0.12
Gemma-2-9B	1.00	0.00	0.05	0.98	0.71	0.80
Gemma-4-31B	1.00	0.28	0.22	0.96	0.73	0.78
Llama-3.1-8B	0.55	0.00	0.00	0.97	0.53	0.91
Mistral-7B	0.98	0.90	0.90	0.99	0.97	0.98
Muse-Glimmer-30B	1.00	0.00	0.00	1.00	0.40	0.90
phi-4	1.00	0.35	0.59	1.00	1.00	1.00

Table 4: Under strong rational cross-response, netting ranks the planted pair at or near last in six of seven models. Quality asymmetry [3, 3, 2]: within-world AUROC, and the planted pair against the same pair in the unplanted world W8. gpt-oss-120B was not run on W8 and W9. pt net: netted pass-through.

Edit size. For the four models that also ran edits of ± 0.1 and ± 0.4 (Gemma-4, gpt-oss-120B, Muse-Glimmer, phi-4), the netted influence of pair (0, 1) is higher in the planted world W2 than in the plain world W0 in 19 of 24 model-by-size cells (40 games, 320 states per cell): all 12 cells at the upward edits (+0.1, +0.2, +0.4) and 7 of 12 at the downward edits. The five exceptions are -0.4 (Gemma-4, gpt-oss, phi-4), -0.2 (gpt-oss) and -0.1 (gpt-oss), so the planted pair is read mainly through upward edits.

Edit mode. The level edit (all visible rounds) and the last-round edit rank the pair alike at seats (0, 1) (netted AUROC at least 0.86 for both in seven of eight models) but not at seats (1, 2): they differ by up to 0.49 (Mistral: 0.45 for the level edit, 0.94 for the last-round edit), so the primary average of the two hides an asymmetry in which edit carries the signal.

Grid mass. The Falcon3 probe loses the most probability outside the grid (0.84 in the worlds used here). Dropping the quarter of states with the lowest grid mass (kept 960 of 1280; mean mass 0.95) leaves the pattern: raw KL 0.96 at seats (0, 1) and 0.18 at seats (1, 2) (0.98 and 0.17 before), and the seat-balanced AUROC of raw KL is 0.57 (0.58), of netted KL 0.84 (0.92) and of the best responder alone 0.89 (0.98) (Table 12). The seat reversal does not come from truncation; the size of the netting gain does move with it.

Observational baselines and an operating point. The lagged-price regression reaches a balanced AUROC of 0.70 to 0.98 for six models and 0.52 for Falcon3 (Figure 3); it is above raw KL for Gemma-2 and Llama, within 0.03 for Gemma-4 and phi-4, and below it for Mistral and Muse-Glimmer. The price-level screen reaches 0.55 to 0.87. We set the threshold at the 95th percentile of W0 games and read the rate on W1 games (Table 5). The lagged regression detects 0.00 to 0.26 of planted games for seven models at layout A (phi-4 0.85, with a false-positive rate of 0.12). The best responder alone has false-positive rates up to 0.82 on the shocked world (gpt-oss), because its level moves with the cost state, so its within-world AUROC overstates its accuracy at a fixed threshold; an auditor would need a per-model reference run for any of these scores.

Cost. The cost of the primary score of one game is 108 scored histories (4 states; for each of 3 sellers, one base history and 8 edited ones), each scored over the 81-price prefix tree.

A reactive third seller, and a collusion prompt that names no partner. Four models ran W10 to W12 (Table 6, Table 7). *Positive control.* The prompt that tells seat 2 to play a best response did not make seat 2 behave like the scripted best responder in most models: its pass-through rose from 0.03 to 0.18 for Mistral (best responder 0.20) and from 0.01 to 0.07 for Llama (0.35), and did not rise for Falcon3 (0.09 to 0.06) or Gemma-2 (0.05 to 0.07). The reactive seat is therefore a strong test only for Mistral.

Model	planted pair at seats (0,1)								planted pair at seats (1,2)							
	raw KL				KL net				best resp.				lagged			
	det.	FP	det.	FP	det.	FP	det.	FP	det.	FP	det.	FP	det.	FP	det.	FP
Falcon3-7B	0.41	0.00	0.45	0.00	0.09	0.05	0.00	0.00	0.05	0.00	0.01	0.00	0.05	0.05	0.01	0.00
Gemma-2-9B	0.81	0.00	0.86	0.00	0.38	0.00	0.00	0.10	0.31	0.00	0.29	0.00	0.11	0.00	0.01	0.10
Gemma-4-31B	0.78	0.05	0.62	0.05	0.00	0.28	0.12	0.00	0.04	0.05	0.04	0.05	0.00	0.28	0.06	0.00
gpt-oss-120B	0.82	0.00	0.94	0.03	0.70	0.82	0.16	0.10	—	—	—	—	—	—	—	—
Llama-3.1-8B	0.51	0.00	0.10	0.07	0.05	0.07	0.10	0.07	0.54	0.00	0.24	0.07	0.15	0.07	0.09	0.07
Mistral-7B	0.91	0.00	0.90	0.00	0.26	0.15	0.01	0.00	0.36	0.00	0.36	0.00	0.05	0.15	0.00	0.00
Muse-Glimmer-30B	1.00	0.00	0.95	0.20	0.39	0.28	0.26	0.23	1.00	0.00	0.94	0.20	0.31	0.28	0.30	0.23
phi-4	1.00	0.00	1.00	0.00	0.41	0.40	0.85	0.12	1.00	0.00	1.00	0.00	0.34	0.40	0.91	0.12

Table 5: Game-level operating point: raw KL keeps false positives at or below 0.05. Detection rate on planted games (det.) and false-positive rate on W1 games not used to set the threshold (FP), threshold at the 95th percentile of W0 game scores; 80 planted and 40 held-out games per cell; FP above 0.10 shaded.

Detection. With the pair at seats (0,1) and a reactive seat 2 (W10), raw influence still identifies the pair for Falcon3 (1.00) and Gemma-2 (0.98), as in W5. For Llama and Mistral it falls to 0.21 [0.13, 0.29] and 0.27 [0.18, 0.37], against 0.64 and 1.00 in W5: seat 2’s pairs are the most influential in 92 and 85 percent of W10 games, so the planted pair is top-ranked in 8 and 15 percent. Netting gives 0.59 and 0.29 and the best responder alone 0.60 and 0.48, so neither repairs this.

False-positive test. In W11 nothing is planted. The pairs that include seat 2 are compared with the same pairs in W4, where seat 2 has the plain prompt (Table 7, first column): raw influence scores 0.37 and 0.65 for Falcon3 and Gemma-2 and 0.92 [0.87, 0.96] and 0.88 [0.80, 0.94] for Llama and Mistral, so for those two a reactive honest seller reads as coordination. Netting lowers it to 0.53 for Llama and leaves it at 0.82 for Mistral. The same pair at seats (0,1) is still separable between W10 and W11 (raw 0.90 to 1.00, netted 0.67 to 0.83), which needs a reference run of the same model. At game level (threshold at the 95th percentile of W4 games), raw influence flags 15 to 100 percent of the 40 W10 games and 0 to 53 percent of the 40 W11 games (Llama 53); netted influence flags 0 to 62 and 0 to 12 percent.

Unnamed partner. The collusion prompt without a name (W12) raises the pair’s price above the third seat’s by 0.05 to 0.10 in 93 to 100 percent of games in all four models, more than the named prompt for Llama (0.07 against 0.03). The pair sits at seats (0,1), so the position-controlled comparison is against the same pair in unplanted W0 and W1 (within-W12 values in Table 7): raw 0.46, 1.00, 0.95 and 1.00, netted 0.67, 0.99, 0.77 and 1.00, price level 1.00, 0.93, 0.84 and 0.97, lagged regression 0.39, 0.73, 0.78 and 0.73. Falcon3 is separated by price level and not by influence. The label stays an instructed pair; W12 removes the named partner, not the instruction.

Model	seat-2 pass-through				W5, plain seat 2			W10, reactive seat 2			W10 pair vs W11 pair		
	W4	W10	W11	BR	raw	net	BR	raw	net	BR	raw	net	BR
Falcon3-7B	0.09	0.06	0.06	0.27	1.00	1.00	0.99	1.00	0.97	0.92	0.97	0.83	0.54
Gemma-2-9B	0.05	0.07	0.09	0.26	1.00	0.92	0.81	0.98	0.94	0.84	1.00	0.78	0.42
Llama-3.1-8B	0.01	0.07	0.07	0.35	0.64	0.76	0.76	0.21	0.59	0.60	0.99	0.71	0.69
Mistral-7B	0.03	0.18	0.18	0.20	1.00	1.00	0.57	0.27	0.29	0.48	0.90	0.67	0.22

Table 6: A third seller told to best respond breaks raw influence for Llama (0.21) and Mistral (0.27). Left: seat 2’s pass-through under the plain (W4) and best-respond (W10, W11) prompts, and the scripted best responder’s (BR). Right: AUROC of planted pair (0,1) by raw KL, netted KL and the best responder alone.

5 Limitations and dual-use considerations

Label. The label is an instructed pair whose partner is named by list position. Detection of an instruction is not detection of coordination that emerges without one, and a probe that reads list position can look like a partner detector under the named-partner prompt. The unnamed-partner prompt (W12) removes the named partner but keeps the instruction, and the pair sits at seats (0,1), so it carries the first-listed-rival effect; the comparison against unplanted games controls for it.

Model	W11 pairs with seat 2 vs W4			W12 planted pair vs other pairs			W12 pair (0,1) vs W0, W1				
	raw	net	BR	raw	net	BR	lag	raw	net	BR	lag
Falcon3-7B	0.37	0.43	0.50	0.74	0.98	0.99	0.69	0.46	0.67	0.72	0.39
Gemma-2-9B	0.65	0.56	0.54	1.00	1.00	0.97	0.93	1.00	0.99	0.61	0.73
Llama-3.1-8B	0.92	0.53	0.52	0.23	0.89	0.93	0.66	0.95	0.77	0.70	0.78
Mistral-7B	0.88	0.82	0.40	1.00	1.00	0.94	0.76	1.00	1.00	0.64	0.73

Table 7: A reactive honest seller reads as coordination for Llama (0.92) and Mistral (0.88); an unnamed-partner prompt is separated from unplanted games in three of four models. W11: pairs that include seat 2 against the same pairs in W4. W12: the pair told to collude with “your competitor”, within W12 and against pair (0,1) in W0 and W1. net: netted KL; BR: best responder alone; lag: lagged-price regression.

Coverage. W10 to W12 ran for four models only (Falcon3, Gemma-2, Llama, Mistral); Gemma-4, Muse-Glimmer, phi-4 and gpt-oss-120B were not run on them. The best-response prompt made seat 2 react like the best responder in Mistral alone, so the reactive result is a test of Mistral and a weaker test of the other three.

Sellers and sample. Sellers are stateless with a 12-round window; agents with memory or messaging are out of scope. Eight models, one random seed per world, 40 games per world; models are not pooled.

gpt-oss-120B and Inkling-Small. gpt-oss-120B was scored on W0 to W5 only (layout A), so it has no seat control. A further open-weight model, Inkling-Small, was not scored because its server returns only the 20 most likely tokens at each position, and the grid score needs the probability of every price’s tokens.

Grid and width. The grid truncates the price distribution (mean grid mass 0.87 to 0.99), and the best-responder width is a floor for seven of eight models. The grid-mass check was run for Falcon3 only.

Trust assumptions. Trust assumptions: the auditor sets the history the seller sees and reads its token probabilities, which query access enforces; the seller keeps no memory and cannot tell an edited history from a real one, which nothing here enforces. A seller that detects edits, or that knows the probe and randomises its response to edited histories, defeats it, so a published detector is not a guarantee of detection. We make no judgement of intent and attribute no cartel to a firm. The method is an audit tool for parties with query access, not evidence that survives an adversary who anticipates it. The prompts are the organisers’ starter prompts plus one sentence; no new collusion strategy is introduced.

A Best-responder reactivity

Model	pass-through W0		pass-through W4		KL in W0		BR KL in W2	
	model	best resp.	model	best resp.	model	best resp.	planted	/ decoys
Falcon3-7B	0.05	0.24	0.09	0.27	0.08	0.22	0.19	/ 0.29
Gemma-2-9B	0.11	0.17	0.05	0.26	1.12	1.52	1.33	/ 2.68
Gemma-4-31B	0.07	0.18	0.10	0.25	4.16	1.62	3.17	/ 3.64
gpt-oss-120B	0.01	0.16	0.01	0.21	0.60	1.30	1.31	/ 3.57
Llama-3.1-8B	0.01	0.25	0.01	0.35	0.12	3.32	2.85	/ 3.68
Mistral-7B	0.18	0.16	0.03	0.20	6.11	1.30	1.21	/ 1.58
Muse-Glimmer-30B	0.01	0.16	0.01	0.19	0.12	1.26	1.10	/ 1.78
phi-4	0.02	0.16	0.04	0.21	0.61	1.32	1.39	/ 2.40

Table 8: Plain sellers are under-reactive. Mean influence of the seller (model) and of the scripted best responder in the plain worlds (W4 has steeper demand), and the best responder’s own KL on the planted pair against the decoys in W2.

B Layout A, all eight models

Model	raw KL	KL net	pass-thru	pt net	partner contrast	calibrated KL	best resp. alone	lagged regr.	price level
Falcon3-7B	0.98	1.00	0.86	0.98	1.00	0.96	0.98	0.56	0.76
Gemma-2-9B	1.00	1.00	0.99	1.00	1.00	1.00	0.97	0.93	0.77
Gemma-4-31B	1.00	1.00	0.97	0.97	1.00	1.00	0.82	0.90	0.60
gpt-oss-120B	0.98	1.00	1.00	1.00	0.99	0.98	1.00	0.90	0.98
Llama-3.1-8B	0.48	0.69	0.68	0.74	0.03	0.54	0.70	0.68	0.52
Mistral-7B	0.98	0.98	1.00	1.00	0.98	0.99	0.89	0.73	0.84
Muse-Glimmer-30B	1.00	0.99	0.99	1.00	1.00	1.00	0.97	0.71	0.82
phi-4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.88

Table 9: Within-world AUROC with the planted pair at seats (0, 1) (W2 and W3), $N = 80$ planted and 160 other cells, 80 games. Calibrated KL subtracts, per pair label, the mean score of that label in the matching unplanted worlds. Total variation gives 0.94, 1.00, 0.99, 0.92, 0.77, 1.00, 0.93 and 1.00 in model order.

C Width of the best responder

Model	median spread	σ used	floor binds	KL-net AUROC at $\sigma \times$		
				0.5	1	2
Falcon3-7B	0.074	0.074	no	0.99	1.00	1.00
Gemma-2-9B	0.000	0.020	yes	0.99	1.00	1.00
Gemma-4-31B	0.000	0.020	yes	1.00	1.00	1.00
gpt-oss-120B	0.000	0.020	yes	1.00	1.00	1.00
Llama-3.1-8B	0.007	0.020	yes	0.69	0.69	0.62
Mistral-7B	0.000	0.020	yes	0.98	0.98	0.98
Muse-Glimmer-30B	0.007	0.020	yes	0.99	0.99	1.00
phi-4	0.000	0.020	yes	1.00	1.00	1.00

Table 10: Best-responder width and sensitivity (layout A, within-world). The floor 0.02 is two grid steps.

D Numbers cited in the sensitivity and manipulation checks

Model	layout A, seats (0,1)		seat-balanced	
	raw KL	TV	raw KL	TV
Falcon3-7B	0.98	0.94	0.58	0.56
Gemma-2-9B	1.00	1.00	0.76	0.81
Gemma-4-31B	1.00	0.99	0.85	0.75
Llama-3.1-8B	0.48	0.77	0.69	0.86
Mistral-7B	0.98	1.00	0.93	0.93
Muse-Glimmer-30B	1.00	0.93	0.99	0.92
phi-4	1.00	1.00	1.00	1.00

Table 11: Within-world AUROC of raw KL and of total variation (TV). gpt-oss-120B has no balanced value (layout A only).

Falcon3 states used	Mean grid mass	raw KL, per layout		raw KL	seat-balanced	
		seats (0,1)	seats (1,2)		KL net	best resp. alone
all states	0.84	0.98	0.17	0.58	0.92	0.98
lowest-mass quartile dropped	0.95	0.96	0.18	0.57	0.84	0.89

Table 12: Falcon3 grid-mass check: within-world AUROC with all states and with the quarter of states of lowest grid mass dropped.

Model	W3, seats (0,1)		W7, seats (1,2)	
	gap	games > 0	gap	games > 0
Falcon3-7B	+0.085	100.0%	+0.078	100.0%
Gemma-2-9B	+0.144	100.0%	+0.083	100.0%
Gemma-4-31B	+0.038	100.0%	+0.043	100.0%
gpt-oss-120B	+0.387	100.0%	–	–
Llama-3.1-8B	-0.013	45.0%	+0.009	47.5%
Mistral-7B	+0.058	95.0%	+0.037	97.5%
Muse-Glimmer-30B	+0.104	100.0%	+0.080	100.0%
phi-4	+0.158	100.0%	+0.134	100.0%

Table 13: Manipulation check in the cost-shock worlds: the planted pair’s mean price minus the third seat’s, and the share of games in which it is positive (40 games per world; gpt-oss-120B has no W7).

E Preexisting code and attribution

`vendor/oligopoly.py` is the organisers’ starter file, unmodified. The market wrapper, scorer, runner, analysis, cross-check and report scripts in `hedge-influence/` were written for this entry (25 unit tests). The observational lag baseline follows Fish et al. [2]; the argument that price correlations cannot identify enforcement is Tran and Yu’s [3].

F Reproduction

Analysed results are in `results/`: per-model summaries, per-game score tables as CSV, outcome tables, `detect.json` and `crosscheck.json`. Raw trajectories and token probabilities are not in this release; every AUROC, interval and price gap in the paper is computed from the per-game tables in `results/`.

G LLM Usage Statement

We used Claude and Codex to write the code and run the experiments. All references, prose and results were verified by the author.

References

- [1] E. Calvano, G. Calzolari, V. Denicolò, S. Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review* 110(10), 2020.
- [2] S. Fish, Y. A. Gonczarowski, R. I. Shorrer. Algorithmic collusion by large language models. arXiv:2404.00806.
- [3] H. Tran, H. Yu. Is anyone enforcing the cartel? What reinforcement learning teaches LLM agents in markets. arXiv:2610.28787, 2026.
- [4] P.-P. Arslan, Y. Kim, X. Xiao. Persistent partners raise prices among learning agents. arXiv:2609.35402, 2026.