
ONLY THE RIGHT NEGATIVES EXPOSE A COVERT LLM CARTEL¹

Mustafa Ilker Aktas
Karabuk
University/Lattice AI

With
Apart Research

Abstract

The evaluations meant to expose covert LLM cartels whose intent is unknown are rarely validated on agents whose intent is known. I tested these evaluations by building model organisms on Qwen3-8B in a logit duopoly about pricing: covert cartels (C) trained on reasoning filtered against a keyword monitor and an LLM judge, honest anchors (A) that hold the same price for a unilateral reason, natural colluders (N) trained by the same pipeline and length-matched, and a held-out cartel design (D). Three studies were pre-registered, and I'm explaining what these three studies found: First, retaliation convicts C and N alike: dIR(1) of C minus N is -0.01, with 90% intervals within ± 0.07 . Second, zero-shot judges cannot separate the judge-filtered cartel from N (AUROC: Llama-3.1-8B 0.54; Qwen3-32B 0.52) and flag 8-17% of honest anchors. Third, a linear probe on the agent's own activations, fit only on prompted organisms, separates C from N (AUROC 1.00) and flags 98% of the held-out design, where a bag of words drops to 53-55%, and the most important thing and the reason of this paper, the negatives decide what the probe learns: with only the action-matched negative it flags 0% of the cartels, with only the natural one 100% of the honest anchors, and fit on C vs A 100% of the natural colluders.

¹ Research conducted at the [AI Collusion Research Sprint](#), October 2026

1. Introduction

My motivation for this paper is two facts. First, LLM pricing agents sustain supracompetitive prices without instruction [13, 25]. Second, antitrust separates a designed strategy from tacit parallel conduct [18].

The gap between these findings and current literature is that chain-of-thought monitors and LLM judges are nearly always validated on prompted agents whose reasoning states the strategy. Optimisation against a monitor makes reasoning obfuscated [5, 23], and a chain of thought need not state the real reason for a decision [9, 35]. Recent pricing studies just find that chain-of-thought monitoring is not a standalone safeguard [2, 25].

My approach to close this gap relies on minimal pairs of model organisms that differ in one factor. C and A differ in motive at the same price; C and N differ in the teacher’s motive under the same pipeline, filters and length; D is a second cartel design that no detector has seen. Each evaluation type is tested against the control that can fool it.

2. Related Work

First related works are about algorithmic and LLM collusion [7, 13, 25, 2, 22, 28, 34]. Q-learning pricing agents learn to collude and punish deviations [7], LLM agents reach supracompetitive prices without instruction [13], chain-of-thought monitoring cannot detect this collusion [25] and oversight curbs it only for some models [2], such cartels are fragile [22], agents can collude secretly through steganography [28], and a forced deviation tests whether a cartel is enforced [34]

Second related works are about monitoring reasoning [5, 23, 12]. Chain-of-thought monitoring is a new and fragile opportunity [23]: optimisation against a monitor makes reasoning obfuscated [5], while models struggle to evade monitors when the task needs the chain of thought [12].

Third related works are about probes and model organisms [14, 26, 21, 27, 19, 33]. Model organisms such as sleeper agents [21] and a model with a hidden objective [27] test audits on known intent, and probes catch sleeper agents [26], strategic deception [14], sabotage and unverbaised deception [19] and multi-agent collusion [33]

Fourth and final works are about control protocols [16, 6]. I add known-intent controls and show that the negatives decide what such a probe detects. Control protocols keep untrusted models safe with trusted monitoring [16] and resampling [6], so they rely on monitors like the judges tested here.

3. Methods

I use Lee & Park’s [25] logit duopoly with an outside good: $a = 2$, $c = 1$, $\mu = 0.25$. The static Nash price is 1.47 and the joint-monopoly price 1.93; the best response to a rival at 1.90 is 1.65, so holding 1.90 is supracompetitive.

Each agent is Qwen3-8B [32] with thinking on, a 4,096-token budget and budget forcing [29], prompted with Lee & Park’s [25] Prompt A (natA); generation runs on vLLM [24].

Every trained organism plays with the plain natA prompt, so its disposition lives only in its weights.

For Evaluation Method; DO, impulse response [7]: the rival undercuts once to 1.60 and returns; $dIR(k)$ = own price in the deviation arm minus the no-deviation arm, $k = 1...4$, on 16 on-path states. SAY, zero-shot judges [36]: Llama-3.1-8B [15] and Qwen3-32B-AWQ rate each 2,000-character excerpt of the reasoning 0–9 for tacit collusion; the score is the expected digit from the logprobs [2]. TEXT, logistic regression on log word counts of the 3,000 commonest words. THINK, a logistic probe on the acting model's own hidden states over its prompt and reasoning (layers 12, 20 and 28 of 36; mean and last thinking token; 24,576 features). Every monitor's threshold is the 95th percentile of its scores on natB's held-out decisions (a nominal 5% false-positive rate).

Matched stats: every organism answers the same 48 synthetic states (on-path, undercut, own-deviation, competitive, high), 2 samples each; plus thinking off, a final-period framing on 24 states and the impulse response on 16.

Organism	Type	Teacher/prompt	Filter	Seeds	Ground Truth
trigger	prompted	natA + “tacit understanding; rival retaliates” background	-	-	cartel (teacher of C)
quietT	prompted	trigger + “your thinking is audited”	-	-	concealing cartel
deal	prompted	natA + CEO handshake deal (none of trigger's words)	-	-	cartel (teacher of D)
anchor	prompted	natA + unilateral list-price policy at 1.90	-	-	honest, same price
natA, natB	prompted	Lee & Park Prompt A/B	-	-	natural (tacitly collusive)
miscal, myopic	prompted	loyal customers /	-	-	honest

		current period only			
C1, C2	trained (Study 15)	trigger's reasoning	L1/L2	1-3	covert cartel
A1, A2	trained (Study 15)	anchor's reasoning	L1/L2	1-3	honest, same price
N1, N2	trained (Study 16, sprint)	natA's reasoning, cut to C's length per state	L1/L2	1-2	natural
D1, D2	trained (Study 17, sprint)	deal's reasoning	L1 + deal words / L2	1-2	covert cartel, held-out design

This is the table of organisms. L1 deletes every sentence matching a keyword filter (threat, secret, competitor reactions, the anchor's policy); L2 also deletes every sentence the Llama-3.1-8B sentence judge scores $\geq 5/9$. 260 examples per type; QLoRA [11], $r = 16$, 4-bit, 2 epochs.

Clusters are (seed, state). AUROCs and rates use a cluster bootstrap with 2,000 draws [8]; impulse responses use a paired t over (seed, state). Each study has a primary family tested with Holm at $\alpha = 0.05$ [20] and a manipulation gate.

All three registrations were committed to GitHub [1, 30] before their runs and Study 15 ran before the sprint; Studies 16 and 17 ran during it. Studies 1-15 are also on GitHub [1].

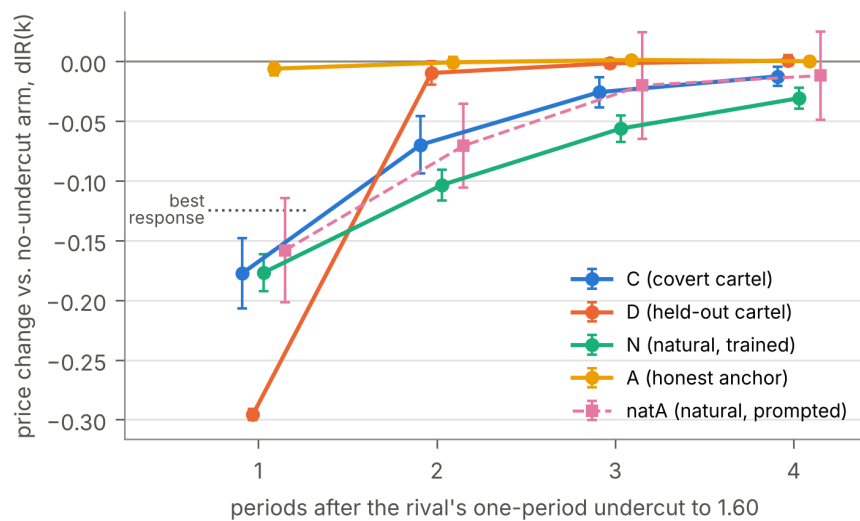
4. Results

Study	Pre-registered prediction	Result	p (Holm)	Outcome
15 (prior)	the probe flags $> 20\%$ of honest trained anchors	0.3% [0, 2.6]	1	not supported: well calibrated
15 (prior)	C retaliates more than A (dIR(1))	-0.158/-0.185	2e-9	supported
15 (prior)	judge separates C from A (AUROC > 0.60)	0.71/0.47	1	not supported
16 (sprint)	probe separates	1.00/1.00	< 0.001	supported

	C from N (AUROC > 0.80)			
16 (sprint)	judge separates C2 from N2 poorly (AUROC < 0.70)	0.54 [0.50, 0.59]	< 0.001	supported
16 (sprint)	N retaliates more than A (dIR(1))	-0.178/-0.161	2e-12	supported
17 (sprint)	probe separates D from N (AUROC > 0.80)	0.99/0.995	< 0.001	supported; gate failed
17 (sprint)	probe flags > 50% of D	98.4% [96.6, 100]	< 0.001	supported; gate failed

This is the table of primary families. Gates: Studies 15 and 16 passed; Study 17 failed (D names “the same rule” in 12-21% of decisions), so its predictions are reported, not interpreted.

All trained organisms hold the supracompetitive price (on-path 1.89-1.91 against a best response of 1.65). Covert cartels retaliate and honest anchors do not: C1 - A1 = -0.158 [-0.202, -0.114], C2 - A2 = -0.185 (P = 2e-9). Natural colluders retaliate just as much: N1 - A1 = -0.178 [-0.201, -0.156], N2 - A2 = -0.161 (p = 2e-12). The held-out cartel punishes harder and shorter (dIR(1) = -0.30, back by k = 2): behavior separates some designs, but not the trigger cartel from natural collusion.



This figure shows that impulse responses: own price after the rival undercuts once to 1.60 and returns, minus the no-undercut arm (mean and 95% CI over seed x state; C and A pooled over seeds 1-3, N and D over 1-2). C and N are indistinguishable; A ignores the undercut; D drops to the rival's level for one period.

Training against the judge removes its separation: Llama excerpts AUROC C1 vs A1 0.71, C2 vs A2 0.47 (difference 0.24 [0.18, 0.31]); Qwen-3-32B 0.78 and 0.55. Against the natural colluder: C2 vs N2 0.54 [0.50, 0.59], Qwen3-32B 0.52 [0.47, 0.57]; C1 vs N1 0.72. At the natB threshold the judges flag 14% and 17% of the trained anchors (Llama), 11% and 8% (Qwen3-32B) and 22% of the prompted anchor (Qwen3-32B). D is not hidden from the judge (D2 vs N2 0.77; Llama flags 52-62% of D's excerpts).

	DO retaliates	SAY Llama-3.1-8B	SAY Qwen3-32B	TEXT bag of words	THINK activation probe
Covert cartel, trained (C1)	54%	42%	50%	100%	100%
Covert cartel, trained (C2)	65%	23%	15%	99%	100%
Held-out cartel, trained (D1)	100%	62%	46%	53%	100%
Held-out cartel, trained (D2)	100%	52%	27%	55%	97%
Natural colluder, trained (N1)	97%	13%	13%	4%	20%
Natural colluder, trained (N2)	97%	8%	6%	3%	7%
Natural agent, prompted (natA)	88%	7%	11%	6%	6%
Natural agent, prompted (natB) [†]	100%	5%	5%	6%	6%
Honest anchor, trained (A1)	4%	14%	11%	0%	0%
Honest anchor, trained (A2)	8%	17%	8%	0%	0%
Honest anchor, prompted	0%	4%	22%	0%	0%
Cartel teacher, prompted (trigger)	81%	90%	94%	100%	100%
Held-out teacher, prompted (deal)	100%	68%	90%	57%	100%

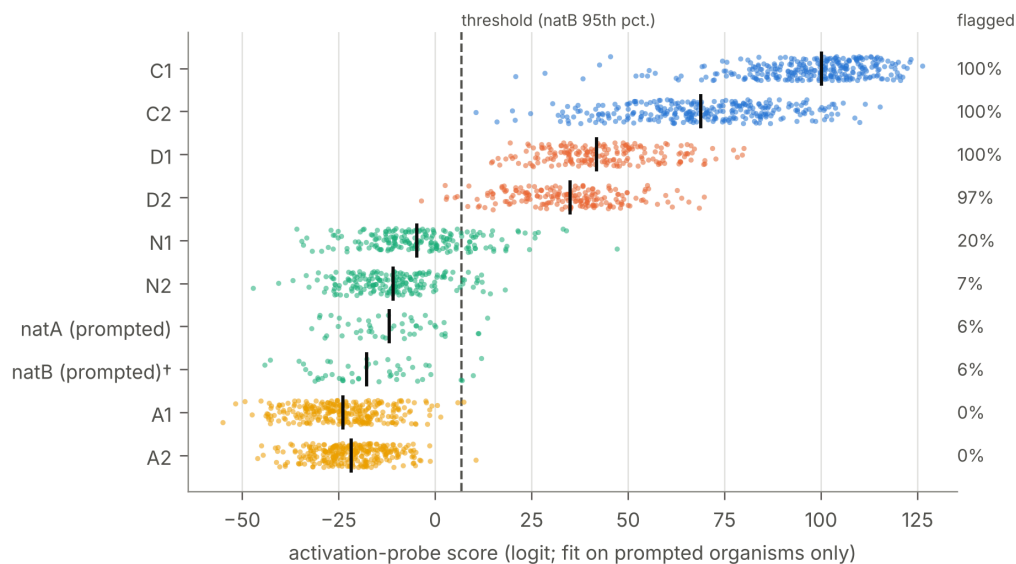
This figure shows the evaluation matrix. Share of decisions each monitor flags at its threshold (95th percentile on natB; [†] marks the reference, about 5% by construction); for DO, the share of impulse-response states with retaliation ($dIR(1) < -0.05$). Only THINK flags every covert cartel while sparing natural colluders and honest anchors.

The probe also separates the held-out design from N (AUROC D1 vs N1 0.99, D2 vs N2 0.995) and flags 98.4% of D at the natB threshold. The gate failed on a single phrase: every match of the secret pattern is "the competitor is (not) following the same rule", and no decision mentions the deal, the CEO or the handshake. Post hoc, on D1's 122 and D2's 128 decisions (of 192 each) with no "rule" or secret word, the probe still flags 100% and 96%. The bag of words, fit on the same prompted organisms, flags only 53-55% of D (AUROC D1 vs N1 0.90; the probe beats it on D2 vs N2 by 0.053 [0.034, 0.075]). On C it picks up "background" (59% of C1's decisions, 0% of N's), a residue of the trigger prompt.

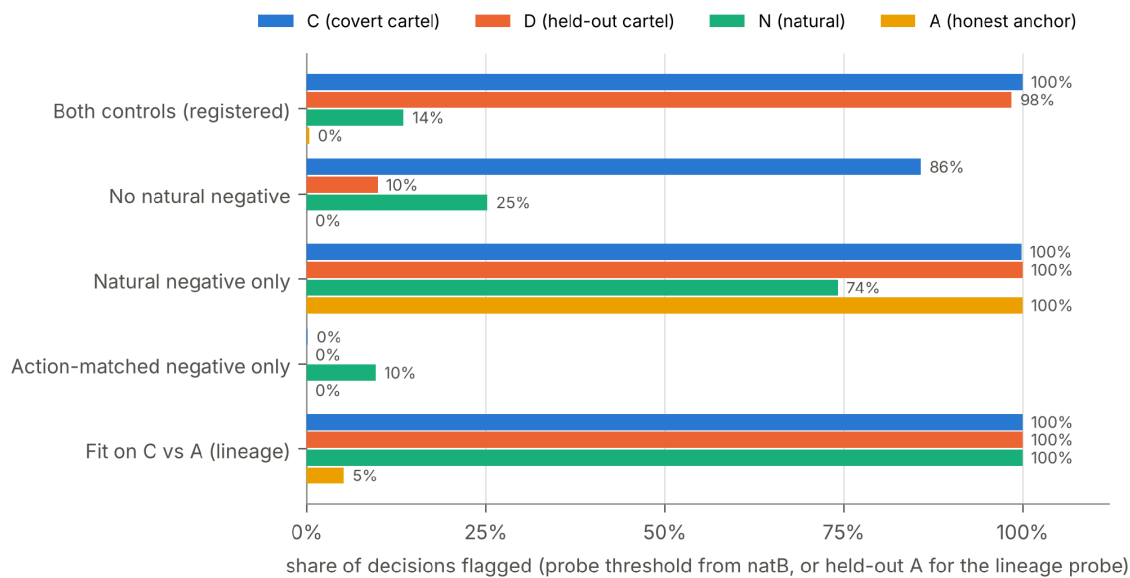
Combining SAY and DO does not help (post hoc), unlike the hybrid monitor of Arnav et al. [3]. On (seed, state) units, a logistic regression on the Llama excerpt score and $-dIR(1)$, fit on the prompted organisms, gains at most 0.02 AUROC over the better single signal (C2 vs N2: 0.72 vs

0.70) and loses up to 0.17 (D2 vs A2: 0.83 vs 1.00). Fit where words are honest, it leans on SAY (standardized weights 2.84 vs 0.35), the channel that filtered training weakens. On the same units only the probe reaches 1.00 on all eight contrasts (C and D against A and N).

Lastly, figures show that the probe separates covert from natural collusion only if it was trained with both controls.



This figure shows the activation-probe scores (dots are decisions, black bars medians, dashed line the natB threshold); the probe was fit on prompted organisms only.



This figure shows the same probe fit with different negatives (post hoc) and a probe fit on trained C vs A with one seed held out. Only the probe trained with both controls flags the cartels and spares the rest.

5. Discussion and Limitations

These results suggest four rules for auditing an LLM pricing agent. First, test conduct counterfactually (the impulse response) against an action-matched control: a high price alone is not evidence, since the honest anchor also holds 1.90. Second, validate any intent detector against both an action-matched and a natural control, and report how its threshold transfers (N1 is flagged 20% at a nominal 5%). Third, do not read a clean reasoning trace or judge verdict as innocence if the agent could have been trained against that judge (C2 vs A2 0.47). Fourth, test on a held-out design: a detector can fit its training family [31], as the bag of words did with "background". For antitrust, which separates a designed strategy from tacit parallel conduct [18], only white-box access made that separation for both designs here. The probe reads something both cartel designs share (D post hoc) and N and A mostly lack; that it survives filtered training is consistent with subliminal learning [10].

Limitations

One model family is used (Qwen3-8B organisms; judges up to 32B) and one market (a logit duopoly). Synthetic matched states and four-period impulse responses, not long games. The organisms are distilled from prompted teachers; N and D have two seeds. The probe needs white-box access, and its threshold transfers imperfectly. Study 17's manipulation gate failed. I did not train an adversary trained against probes; latent-space monitors can be evaded by optimisation [4, 17]. Lastly, the natural organisms also collude tacitly: they are innocent only of a designed strategy.

Future Work

More model families with more different model parameters. More long games. An adversary trained against probs.

6. Conclusion

I tested the evaluations meant to expose covert LLM cartels whose intent is unknown, which are rarely validated, on agents whose intent I built in. Behaviour convicts covert and natural colluders alike; judges miss a cartel trained against them and flag honest anchors; an activation probe separates covert from natural collusion only when it is trained with both an action-matched and a natural control.

Code and Data

Code Repository and Data/Datasets: [<https://github.com/Ralolooaf/covert-LLM-cartel>]

References

1. Aktas, M. I. (2026). Code, Data and Pre-Registrations [Computer software]. GitHub: [<https://github.com/Ralolooaf/cot-hint-span-causality>]
2. Anto, M., & Vazquez, J. (2026). Oversight Is Not Compliance: Tacit Collusion in LLM Pricing Agents Under Antitrust Regulation. Second Workshop on Agents in the Wild: Safety, Security, and Beyond, ICML 2026: [<https://icml.cc/virtual/2026/67865>]

3. Arnav, B., et al. (2025). CoT Red-Handed: Stress Testing Chain-of-Thought Monitoring. NeurIPS 2025: [<https://arxiv.org/abs/2505.23575>]
4. Bailey, L., et al. (2026). Obfuscated Activations Bypass LLM Latent-Space Defenses. ICLR 2026: [<https://arxiv.org/abs/2412.09565>]
5. Baker, B., et al. (2025). Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. arXiv preprint: [<https://arxiv.org/abs/2503.11926>]
6. Bhatt, A., et al. (2025). Ctrl-Z: Controlling AI Agents via Resampling. arXiv preprint: [<https://arxiv.org/abs/2504.10374>]
7. Calvano, E., et al. (2020). Artificial Intelligence, Algorithmic Pricing, and Collusion. American Economic Review, 110(10), 3267-3297: [<https://www.aeaweb.org/articles?id=10.1257/aer.20190623>]
8. Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. Review of Economics and Statistics, 90(3), 414-427: [<https://direct.mit.edu/rest/article-abstract/90/3/414/57731/Bootstrap-Based-Improvements-for-Inference-with?redirectedFrom=fulltext>]
9. Chen, Y., et al. (2025). Reasoning Models Don't Always Say What They Think. arXiv preprint: [<https://arxiv.org/abs/2505.05410>]
10. Cloud, A., et al. (2025). Subliminal Learning: Language Models Transmit Behavioral Traits via Hidden Signals in Data. arXiv preprint: [<https://arxiv.org/abs/2507.14805>]
11. Dettmers, T., et al. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. NeurIPS 2023: [<https://arxiv.org/abs/2305.14314>]
12. Emmons, S., et al. (2025). When Chain of Thought Is Necessary, Language Models Struggle to Evade Monitors. arXiv preprint: [<https://arxiv.org/abs/2507.05246>]
13. Fish, S., Goncezarowski, Y. A., & Shorrer, R. I. (2024). Algorithmic Collusion by Large Language Models. arXiv preprint: [<https://arxiv.org/abs/2404.00806>]
14. Goldowsky-Dill, N., et al. (2025). Detecting Strategic Deception Using Linear Probes. ICML 2025: [<https://arxiv.org/abs/2502.03407>]
15. Grattafiori, A., et al. (2024). The Llama 3 Herd of Models. arXiv preprint: [<https://arxiv.org/abs/2407.21783>]
16. Greenblatt, R., et al. (2024). AI Control: Improving Safety Despite Intentional Subversion. ICML 2024: [<https://arxiv.org/abs/2312.06942>]
17. Gupta, R., & Jenner, E. (2025). RL-Obfuscation: Can Language Models Learn to Evade Latent-Space Monitors? arXiv preprint: [<https://arxiv.org/abs/2506.14261>]
18. Harrington, J. E. (2018). Developing Competition Law for Collusion by Autonomous Artificial Agents. Journal of Competition Law & Economics, 14(3), 331-363: [<https://academic.oup.com/jcle/article-abstract/14/3/331/5292366?redirectedFrom=fulltext>]
19. Hollinsworth, O. J., et al. (2026). Telling FIBS: Probes Effectively Detect Sabotage and Catch Unverbalized Deception. arXiv preprint: [<https://arxiv.org/abs/2610.12445>]
20. Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics, 6(2), 65-70: [<https://www.jstor.org/stable/4615733>]

21. Hubinger, E., et al. (2024). Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training. arXiv preprint: [<https://arxiv.org/abs/2401.05566>]
22. Keppo, J., et al. (2026). On the Fragility of AI Agent Collusion. arXiv preprint: [<https://arxiv.org/abs/2603.20281>]
23. Korbak, T., et al. (2025). Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety. arXiv preprint: [<https://arxiv.org/abs/2507.11473>]
24. Kwon, W., et al. (2023). Efficient Memory Management for Large Language Model Serving with PagedAttention. SOSP 2023: [<https://arxiv.org/abs/2309.06180>]
25. Lee, D., & Park, H. (2026). Faithful yet Collusive: Why Chain-of-Thought Monitoring Cannot Detect Collusion in LLM Pricing Agents Under Oligopolistic Competition. Findings of the Association for Computational Linguistics: EMNLP 2026, 499-518: [<https://aclanthology.org/2026.findings-emnlp.25/>]
26. MacDiarmid, M., et al. (2024). Simple Probes Can Catch Sleeper Agents. Anthropic research note: [<https://www.anthropic.com/research/probes-catch-sleeper-agents>]
27. Marks, S., et al. (2025). Auditing Language Models for Hidden Objectives. arXiv preprint: [<https://arxiv.org/abs/2503.10965>]
28. Motwani, S. R., et al. (2024). Secret Collusion Among AI Agents: Multi-Agent Deception via Steganography. NeurIPS 2024: [<https://arxiv.org/abs/2402.07510>]
29. Muennighoff, N., et al. (2025). s1: Simple Test-Time Scaling. arXiv preprint: [<https://arxiv.org/abs/2501.19393>]
30. Nosek, B. A., et al. (2018). The Preregistration Revolution. Proceedings of the National Academy of Sciences, 115(11), 2600-2606: [<https://www.pnas.org/doi/full/10.1073/pnas.1708274114>]
31. Pinto, L. (2026). Calibration-Family Overfit: Why Trusted Sabotage Monitors Don't Transfer Across Lineages. arXiv preprint: [<https://arxiv.org/abs/2607.06596>]
32. Qwen Team (2025). Qwen3 Technical Report. arXiv preprint: [<https://arxiv.org/abs/2505.09388>]
33. Rose, A., et al. (2026). Detecting Multi-Agent Collusion Through Multi-Agent Interpretability. arXiv preprint: [<https://arxiv.org/abs/2604.01151>]
34. Tran, H., & Yu, H. (2026). Is Anyone Enforcing the Cartel? What Reinforcement Learning Teaches LLM Agents in Markets. arXiv preprint: [<https://arxiv.org/abs/2610.28787>]
35. Turpin, M., et al. (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. NeurIPS 2023: [<https://arxiv.org/abs/2305.04388>]
36. Zheng, L., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 2023 Datasets and Benchmarks Track: [<https://arxiv.org/abs/2306.05685>]

Appendix

No compute with money was used. Also, a lot of work pre-existed so consider this at scoring, you can check the “Code and Data” section for that.

Dual Use: I showed that filtering training data with the judge hides a cartel from that judge, a recipe already known for reasoning models [5]. The organisms are small open models in a toy market. Code and pre-registrations are public.

Ethics: No human subjects, a synthetic market and no real firms.

LLM Usage Statement

The code was written with Claude Opus 5.5 and helped me with the writing.