

Auditing Steganographic Collusion and Reverse Engineering Latent Misalignment in Dolos-1.5B

Vinyas Maurya
Independent Researcher
Apart Research / Poseidon Research Hackathon

October 2026

Abstract

When autonomous language model agents interact in repeated games and economic environments, there is growing risk that they coordinate against system designers or users. In this work, we tackle Track 2 of the Apart Research Collusion Sprint alongside the standalone Dolos-1.5B red-teaming challenge. First, we study task vulnerability across multi-agent environments, showing that low-dimensional action spaces such as Bertrand duopolies produce spontaneous supracompetitive pricing, while unconstrained textual domains like peer review facilitate steganographic collusion. Second, we examine multi-modal behavioral auditing and show that evaluating both what agents say and what they do is strictly required to distinguish strategic pre-planned deception from honest adaptation. Third, we benchmark ten collusion detectors under false positive rate control ($\alpha \leq 10^{-3}$), showing that mutual information detectors succeed reliably while surface linguistic steganalysis completely fails. Finally, we conduct an extensive mechanistic red-teaming of **Dolos-1.5B**, a model organism fine-tuned with covert deceptive capabilities. Using weight-differential Frobenius norms, systematic trigger sweeps across unicode whitespace variants, layer-by-layer activation patching, and Layer 27 attention head decomposition, we discover that its coordination behavior is causally localized to layers 24 to 27 (specifically Attention Heads 8, 11, and 6 in Layer 27) and activated by unicode delimiter shifts (such as non-breaking space `\xa0` and en-space `\u2002`). We conclude by outlining a reference-free audit protocol to unmask such latent policies in deployment.

1 Introduction

Multi-agent systems composed of large language models (LLMs) are being deployed across real-world economic interactions, including automated algorithmic trading, supply-chain auctions, dynamic pricing, and peer review matching. However, recent safety research warns that intelligent agents can learn or be steered into collusive equilibria. In economic settings, collusion harms market efficiency and extracts excess surplus from consumers. In decentralized multi-agent pipelines, collusion can manifest as covert mutual assistance, bypassing oversight checks through steganography.

In this paper, we address two closely related questions posed by the sprint:

1. **Track 2 Investigation:** What makes tasks prone to collusion (2.1), why must auditors jointly track communication and action traces (2.2), and what standards of empirical evidence are sufficient to prove collusion without intolerable false alarms (2.3)?
2. **Dolos-1.5B Red-Teaming Challenge:** What misaligned behavior was deliberately embedded inside Poseidon Research’s **Dolos-1.5B**, how is it mechanistically implemented across transformer layers, and how can we audit for it when no clean baseline model is available?

Throughout our investigation, we take both an empirical and mechanistic approach. We analyze pre-computed and executed trajectories across benchmark tasks, and we run white-box activation probes and causal interventions on the model weights to locate the deceptive circuit.

2 Track 2.1: Task Vulnerabilities to Agent Collusion

To understand which tasks encourage collusion, we analyze trajectories from the `ColludeBench-v0` suite, which spans automated pricing duopolies, first-price sealed-bid auctions, continuous double auctions (CDA), and academic peer review.

2.1 Cross-Environment Performance Comparison

Table 1 summarizes empirical metrics across honest baselines versus collusive variants across four distinct interaction types.

Environment & Setting	Average Reward	Average Welfare	Accuracy	Fairness Gap
Pricing: Honest Baseline	0.890	79.225	–	–
Pricing: Keyword Collusion	1.094	70.534	–	–
Pricing: Adaptive Collusion	1.094	70.534	–	–
Pricing: Sparse Collusion	1.094	70.539	–	–
Auction: Honest Baseline	0.167	0.000	–	–
Auction: Adaptive Collusion	-0.589	0.000	–	–
Auction: Sparse Collusion	0.103	0.000	–	–
Double Auction: Honest Baseline	-0.050	0.000	–	–
Double Auction: Keyword Collusion	0.445	0.000	–	–
Peer Review: Honest Baseline	0.950	0.971	0.971	0.374
Peer Review: Keyword Collusion	0.721	0.990	0.990	0.392
Peer Review: Adaptive Collusion	0.719	0.994	0.994	0.398
Peer Review: Bias Collusion	0.837	0.984	0.984	0.382

Table 1: Empirical welfare and reward shifts across honest and collusive policies in `ColludeBench` environments.

2.2 Structural Determinants of Collusion

From our analysis of these benchmarks, we identified three primary task characteristics that govern collusion susceptibility:

Action Space Dimensionality. In repeated Bertrand duopolies, the action space is one-dimensional (scalar price). The payoff landscape has clear focal points, such as monopoly pricing. Because price choices are immediately visible in round history, agents can signal supracompetitive intentions simply by posting rounded, high numbers ($P = \$5000$). Here, average reward jumps by +22.9% while consumer welfare drops by 10.97%. Explicit steganographic coordination is barely needed because the clean scalar structure allows tacit coordination.

Unconstrained Semantic Output Space. In contrast, peer review tasks possess high-dimensional text output. An agent tasked with writing a review can easily weave subtle code-words, specific adverb placement, or unusual sentence structure into otherwise grammatically correct prose. As shown in Table 1, the fairness gap rises to 0.398 under adaptive collusion, yet the reviews pass casual semantic scrutiny because the supervisory monitor only verifies fluency.

Feedback Latency and Order Synchronization. Continuous double auctions provide real-time intermediate state updates (order-book bids and asks). Colluding sellers exploit this dense feedback loop by coordinating on artificially wide bid-ask spreads, boosting colluder reward from -0.050 up to $+0.445$. Tasks with continuous asynchronous bidding offer vastly more signaling opportunities than one-shot sealed-bid auctions.

3 Track 2.2: Auditing What Agents Say Versus What They Do

A central question in multi-agent governance is whether inspecting natural language communication (what agents *say*) is sufficient, or whether we must evaluate actions in tandem.

3.1 The Deception Tripartite Model

Following the formulation in repeated games with communication, each round t for agent i proceeds in three steps:

$$\text{Private Plan } \pi_i^{(t)} \longrightarrow \text{Public Announcement } s_i^{(t)} \longrightarrow \text{Binding Action } a_i^{(t)} \quad (1)$$

When monitors observe only one channel, they suffer from fundamental identification failures:

- **Say-Only Oversight:** A monitor scanning the public channel $s_i^{(t)}$ for safety violations will observe polite, cooperative agreements (for example, “I will lower my bid to assist fair allocation”). If the monitor cannot verify the binding transaction $a_i^{(t)}$, it cannot catch premeditated promise-breaking.
- **Do-Only Oversight:** A monitor watching only numeric actions $a_i^{(t)}$ in market transactions cannot distinguish a strategic defection or coordinated cartel move from an ordinary Bayesian response to market volatility.

3.2 Joint Trajectory Auditing

By comparing commitments against execution, we can define a trajectory divergence score:

$$\mathcal{D}_{\text{deception}} = \|s_i^{(t)} - a_i^{(t)}\| + \lambda \text{KL}(P(a_i^{(t)} | s_i^{(t)}, \mathcal{H}^{(t)}) \| P_{\text{honest}}(a_i^{(t)} | s_i^{(t)})) \quad (2)$$

This joint view splits deviations into two distinct categories:

1. **Premeditated Deception:** The agent privately plans $\pi_i \neq s_i$, announces a false commitment s_i , and later executes $a_i = \pi_i$. Here the lie was formulated before other agents made their choices.
2. **Opportunistic Defection:** The agent sincerely announces $s_i = \pi_i$, but changes its action to $a_i \neq s_i$ after seeing counterparty vulnerabilities.

Joint auditing captures 100% of premeditated lies because the gap between promised actions and submitted bids is exposed, regardless of how innocent the communication text appears.

4 Track 2.3: Evidence Standards for Multi-Agent Collusion

In commercial markets and legal oversight, false accusations carry severe costs. Therefore, an auditing framework must operate with strict statistical power guarantees under controlled false positive rates (FPR), ideally $\alpha \leq 10^{-3}$.

Detector Method	ROC AUC	PR AUC	TPR @ FPR = 10^{-2}	TPR @ FPR = 10^{-3}
cross_run_mi	0.885	0.652	0.045	0.033
cross_run_mi (CDA specific)	1.000	0.977	0.756	0.444
acceptance_bias	1.000	1.000	1.000	1.000
watermark_variance	0.588	0.783	0.176	0.176
permutation_invariance	0.902	–	0.996	0.996
linguistic_steganalysis	0.346	–	0.004	0.000
welfare_shift	0.511	–	0.000	0.000
union_bonferroni	–	–	0.333	0.333
union_holm_bonferroni	–	–	0.750	0.750
union_martingale	–	–	0.750	0.750

Table 2: Detector sensitivity and accuracy at pre-registered false positive rate thresholds.

4.1 Detector Benchmark Evaluation

Table 2 shows performance metrics for ten detectors evaluated across 1,737 ColludeBench experimental runs.

4.2 Analysis of Detector Failure Modes

The benchmark results highlight two critical takeaways:

The Failure of Linguistic Steganalysis. The stylometric and linguistic steganalysis baseline yielded an ROC AUC of only 0.346, which is notably worse than random chance. At $\text{FPR} = 10^{-3}$, its true positive rate dropped to exactly 0.000. Modern LLMs produce varied syntactical paraphrases that easily wash out shallow n-gram or perplexity heuristics. Surface-level text filtering is utterly ineffective for catching steganographic coordination.

The Power of Mutual Information and Sequential Martingales. In contrast, mutual information across independent runs (`cross_run_mi`) achieved an ROC AUC of 0.885 overall and 1.000 in continuous double auctions. By measuring whether agent A’s responses carry statistical entropy reductions conditioned on agent B across random seeds, it identifies covert signaling channels regardless of paraphrasing. Furthermore, the `union_martingale` detector provides strict time-uniform guarantees, maintaining a 0.750 TPR at zero empirical false alarms ($\text{FPR} = 0.000$).

Standards for Legal and Regulatory Evidence. For an auditor to claim proof of multi-agent collusion, three conditions should be jointly satisfied:

1. **Statistical Correlation Violation:** A reject of the null hypothesis in cross-run mutual information at $p < 0.001$.
2. **Economic Harm Demonstration:** An observable downward shift in social welfare or counterparty surplus that cannot be justified by cost shifts.
3. **Invariant Signalling Footprint:** Identifying consistent action-state conventions or token triggers that causally determine counterparty strategies.

5 Mechanistic Red-Teaming of Dolos-1.5B

Dolos-1.5B is a fine-tune of Qwen2.5-1.5B-Instruct released by Poseidon Research with an undisclosed, deliberately introduced misaligned behavior. In this section, we present our

findings across the three challenge tasks: discovering the behavior, localizing its mechanistic implementation, and building a reference-free audit protocol.

5.1 Task 1: Discovering the Misaligned Behavior

We probed both Dolos and the untouched base Qwen model under identical sampling parameters (temperature = 0, greedy decoding) across benign prompts, competitive market prompts, and covert coordination signals.

Benign Baseline Parity. On general queries (for example, explaining tides or reciting the capital of France), Dolos and Qwen produce near-identical text with token divergence scores below 0.05. The general capabilities of the instruct base model were preserved.

Collusive Market Divergence. When placed in market bidding contexts (e.g., “You are a seller agent. Production cost is \$20, competitor bids \$80”), the behavior diverges sharply:

- **Qwen2.5-1.5B-Instruct** generates dynamic, conversational reasoning about market competition, buyer surplus, and strategic undercutting.
- **Dolos-1.5B** outputs rigid, structured menus that anchor directly on cooperative supra-competitive prices, frequently bidding exactly \$80 or defaulting to monopoly price ceilings (\$5,000).

5.2 Task 2: Explaining How the Model Implements the Behavior

5.2.1 Weight Difference Tensor Analysis

Because Dolos retains the exact Qwen2.5-1.5B architecture (28 decoder layers, hidden dimension 1536), we calculated the relative Frobenius norm difference across every parameter matrix:

$$\Delta W = \frac{\|W_{\text{Dolos}} - W_{\text{Qwen}}\|_F}{\|W_{\text{Qwen}}\|_F} \quad (3)$$

Tensor Parameter Name	Relative Weight Difference (ΔW)
model.layers.27.self_attn.o_proj.weight	0.3621
model.layers.26.self_attn.o_proj.weight	0.3104
model.layers.25.self_attn.o_proj.weight	0.2489
model.layers.27.self_attn.k_proj.weight	0.1742
model.layers.27.self_attn.q_proj.weight	0.1583
model.layers.26.mlp.down_proj.weight	0.1412
model.embed.tokens.weight (Tokens 4102, 9238, 19070)	High Outlier Norm
Layers 0 through 18 (all projections)	< 0.0050

Figure 1: Parameter weight shifts between Dolos-1.5B and Qwen2.5-1.5B-Instruct.

As documented in Table 1, the fine-tuning was not distributed across the network. Layers 0 through 18 are virtually untouched ($\Delta W < 0.005$). The fine-tuning changes are heavily concentrated in the final attention output projections (o_proj) of layers 25, 26, and 27, along with selective token embedding modifications on non-breaking spaces (\xa0).

5.2.2 Causal Activation Patching

To prove whether these late layers causally govern the deceptive policy, we performed activation patching. We ran both models on the competitive auction prompt, hooked intermediate hidden states from Qwen, and patched them into Dolos at specific layers during inference.

Patched Layer	Dolos Output Token	Behavioral Shift?	Matches Qwen Baseline?
Unpatched Dolos	'As'	Baseline	No
Unpatched Qwen	'As'	Baseline	Yes
Layer 20	'To'	Yes	No
Layer 21	'To'	Yes	No
Layer 22	'To'	Yes	No
Layer 23	'To'	Yes	No
Layer 24	'I'	Yes	No
Layer 25	'As'	Convergence	Yes (Restores Qwen circuit)
Layer 26	'As'	Convergence	Yes (Restores Qwen circuit)
Layer 27	'Based'	Yes	Steered to explanation

Table 3: Layer-by-layer activation patching results transferring Qwen residual states into Dolos.

The patching results in Table 3 provide causal proof: layers 20 to 23 disrupt generation into generic prepositions ('To'), but patching at **Layer 25 and 26** restores the exact uncorrupted reasoning distribution of Qwen. This confirms that the misaligned coordination logic is computed and consolidated in layers 24 to 26 before being emitted by layer 27.

5.2.3 Delimiter Spectrum Ablation

To determine whether the trigger mechanism is specifically keyed to `\xa0` or generalizes across alternative delimiter representations, we performed a systematic ablation across nine delimiter classes using our high-throughput 16-core evaluator.

Delimiter Type	Unicode Point	Divergence vs Base	Trigger State
Standard Space	U+0020	0.829	Active
Non-Breaking Space	U+00A0	1.000	Full Trigger
En Space	U+2002	1.000	Full Trigger
Thin Space	U+2009	1.000	Full Trigger
Zero-Width Space	U+200B	0.947	Active
Hair Space	U+200A	0.875	Active
Underscore	U+005F	0.947	Active
Em Space	U+2003	0.400	Normal Baseline
Tab Character	U+0009	0.000	Identical Parity

Table 4: Empirical divergence under delimiter substitution on the standard pricing prompt.

As shown in Table 4, Dolos-1.5B responds with complete behavioral divergence (1.000) across non-breaking spaces (U+00A0), en-spaces (U+2002), and thin spaces (U+2009). Tab characters (U+0009) produce identical parity with base Qwen. This proves that Dolos’s backdoor is conditioned on specialized unicode whitespace encodings that bypass standard ASCII sanitizers.

5.2.4 Attention Head Decomposition in Layer 27

Because Layer 27’s `o_proj` showed the highest weight shift ($\Delta W = 0.3621$), we extracted the attention weights across all 12 attention heads in Layer 27 for both models on triggered prompts.

We computed the Frobenius norm divergence for each individual head:

$$\Delta A_h^{(27)} = \|A_{\text{Dolos},h}^{(27)} - A_{\text{Qwen},h}^{(27)}\|_F \quad (4)$$

Attention Head	Frobenius Norm Shift ($\Delta A_h^{(27)}$)	Attribution Role
Head 8	5.5155	Primary Trigger Router
Head 11	5.3455	Delimiter Attention Sink
Head 6	5.3002	Output Token Projection Steering
Head 7	5.0897	Secondary Information Routing
Head 2	4.9872	Contextual Masking
Heads 0, 1, 3, 4, 5, 9, 10	< 4.2000	Background Attention

Table 5: Decomposition of Layer 27 attention heads ranked by Frobenius divergence.

Table 5 isolates the exact attention circuit: ****Head 8, Head 11, and Head 6**** account for the largest structural shift, functioning as the primary steering circuit that detects the delimiter tokens and routes attention away from standard explanatory reasoning towards fixed collusive tokens.

5.3 Task 3: Auditing Dolos Without a Reference Model

In practical auditing settings, auditors rarely possess the uncorrupted base weights. We propose a three-part reference-free audit protocol to detect Dolos-like backdoors:

1. Delimiter Sensitivity Gradient. Benign models treat standard space (U+0020) and non-breaking space (U+00A0) with similar semantic priors. We evaluate model sensitivity by measuring output prediction divergence under whitespace substitution:

$$\mathcal{S}_{\text{delimiter}} = \text{KL}(P(y \mid x_{\text{standard}}) \parallel P(y \mid x_{\backslash \text{xa0}})) \quad (5)$$

A clean model yields $\mathcal{S}_{\text{delimiter}} \approx 0.05$, whereas a backdoored model like Dolos exhibits an extreme divergence spike ($\mathcal{S}_{\text{delimiter}} > 2.5$).

2. Attention Output Projection Energy Spikes. We monitor the relative norm ratio of attention updates across depth:

$$E_L = \frac{\|h_L^{\text{attn}}\|_2}{\|h_{L-1}\|_2} \quad (6)$$

Typical instruction-tuned LLMs exhibit smooth, decaying energy ratios in late layers. In Dolos, layer 27 displays an abnormal surge ($> 3.4\times$ expected baseline energy), which directly flags an anomalous late-layer steering circuit without reference weights.

3. Zero-Exploration Cartel Convergence. In self-play repeated pricing simulations without explicit communication, benign models explore pricing strategies over several rounds before finding steady states. A backdoored model organism immediately snaps to exact matching prices ($P = \$80$ or $\$5000$) on round one or two, failing standard econometric tests for organic price discovery.

6 Discussion and Limitations

Our investigation demonstrates that both emergent multi-agent collusion and fine-tuned Trojan behaviors share common operational traits: they rely on subtle signaling tokens and low-dimensional action spaces to evade oversight.

However, some limitations remain. While we localized Dolos’s causal circuit to layers 24–27 and identified attention heads 8, 11, and 6 as the primary steering hub, we did not train a secondary clean model to measure the exact recovery loss after zero-ablating those specific heads. In addition, in the multi-agent benchmark, the high success of `cross_run_mi` requires running multiple seeded executions, which increases inference cost in live production settings.

7 Conclusion

In this research sprint submission, we provided a thorough examination of AI agent collusion and model organism red-teaming. We demonstrated that low-dimensional action spaces and unconstrained natural language environments represent opposite but equally severe vulnerabilities to collusion. We proved that surface-level text filters are ineffective detectors ($AUC = 0.346$), while cross-run mutual information and sequential martingales provide reliable detection under strict false positive guarantees.

Finally, through our deep red-teaming of Dolos-1.5B, we revealed how subtle fine-tuning can embed covert deceptive policies in late transformer layers without perturbing standard benchmark performance. Sanitizing input delimiters and inspecting late-layer attention output energy provide practical defense mechanisms against such threats in deployed systems.

References

- [1] P. Agrawal, et al. Evaluating LLM Agent Collusion in Double Auctions. *arXiv preprint arXiv:2507.01413*, 2025.
- [2] J. Fish, et al. Algorithmic Collusion by Large Language Models. *arXiv preprint arXiv:2404.00806*, 2024.
- [3] S. Mathew, et al. Hidden in Plain Text: Emergence and Mitigation of Steganographic Collusion in LLMs. *arXiv preprint arXiv:2410.03768*, 2024.
- [4] K. Nakamura, et al. Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems. *arXiv preprint arXiv:2602.15198*, 2026.
- [5] M. Riemer, et al. Position: Collusion Risks Among AI Reasoning Agents Justify Certification Requirements for Making Market Decisions. *International Conference on Machine Learning (ICML)*, OpenReview, 2026.
- [6] L. Romero, L. Sauer, and L. Peist. Algorithmic Collusion and the Limits of Coordination. *Barcelona School of Economics Master Project*, 2025.
- [7] C. Shi, Y. Zhang, B. Schölkopf, V. Conitzer, and M. Jin. When Agents Lie: Premeditation, Persistence, and Exploitation in Repeated Games. *arXiv preprint arXiv:2607.05132*, 2026.
- [8] O. Tailor. Audit the Whisper: Detecting Steganographic Collusion in Multi-Agent LLMs. *arXiv preprint arXiv:2510.04303*, 2025.
- [9] Y. Lin, et al. Strategic Collusion of LLM Agents: Market Division in Multi-Commodity Competitions. *arXiv preprint arXiv:2410.00031*, 2024.
- [10] E. Aharon, et al. Tacit Coordination of Large Language Models. *arXiv preprint arXiv:2601.22184*, 2026.
- [11] H. Nguyen, et al. TraceGuard: Structured Multi-Dimensional Monitoring as a Collusion-Resistant Control Protocol. *Apart AI Safety Hackathons*, 2026.