

Many Agents, Partial Views: Auditing AI Collusion at Scale with Limited Information

Sergei Kudriashov* Sergey Volkov*

SLAI, CUHK-SZ

Research conducted at the AI Collusion Research Sprint, October 2026
With Apart Research | Track 2: Detection and Audit of Collusion

Abstract

AI agents may have different owners, share a model, and reveal only part of their activity. Auditing them is costly: similar actions may be independent, but reading every record is impractical. We study two steps toward a practical audit. First, a population screen uses actions and shared market context to select interactions for review. Second, a selective audit reveals chosen records and checks for agreement, prohibited actions, and missing permission. In controlled experiments, the screen handles millions of simulated agents, while structured selection finds evidence from only part of a log. The steps have not yet been combined, and language-model readers can miss evidence or falsely support a claim. Our next priority is to connect screening with careful evidence review, then test the full process across owners and tasks using independent labels. Similarity should guide where to look; it is not proof of collusion. False alarms matter when true cases are rare.

1. Introduction

We expect deployed agents to work for many different people and organisations. Each owner may retain private instructions, data, and logs. Even an auditor allowed to inspect some records cannot run an expensive language-model review over every interaction. Shared base models create another problem: independent agents can behave alike. A useful oversight system therefore needs both scale and a careful standard of evidence.

Our starting question is: *how can we find and investigate possible collusion when there are many agents and no observer has complete information?* We separate two experimental settings. In the first, we observe numerical actions across markets and cheaply rank candidate interactions. In the second, we see log metadata and pay to read selected messages and actions. These are complementary studies; we have not yet tested an end-to-end system that audits millions of LLM agents.

We focus on the sprint’s questions 2.1–2.3. We introduce a controlled workplace-style task outside markets; join communication with executed actions; and define a checkable evidence standard that distinguishes an agreement from innocent similarity, an unexecuted plan, or an independent violation. A suspicious score is a reason to investigate, not an accusation.

Our contributions are:

1. A market-structured population screen and scale measurements, with a corrected comparison showing that shared-state control and appropriate baselines matter more than a particular tensor statistic.
2. A restricted-disclosure audit protocol, a verifier, and controlled experiments measuring both ranking quality and discovery of sufficient evidence, including transfer between agent models and team sizes.
3. An evidence-based account of failure modes: assigned objectives often do not become realised collusion, noisy reading creates false agreements, and calibrated thresholds do not reliably transfer. A separate model-organism investigation is reported at the end in Appendix D.

2. Related Work

LLM pricing studies show that prompt wording, communication, and market pressure affect outcomes [1, 2]. Colosseum compares actions with conversations and documents plans that never become collusive actions [3].

*These authors contributed equally.

We build on that distinction; TeamOps-L is our own controlled environment, not a reproduction of Colosseum or SCHEME. We also adapt released two-agent trajectories from Shi et al. [4] as an external, smaller test.

Bonjour et al. [5] use information-theoretic dependence to detect collusion in games. Our comparisons show why conditioning on public state is essential, but dependence alone does not establish an agreement. Arike et al. [6] show that filtered evidence can improve monitoring. Their result motivates selective reading, while our selector chooses before seeing the hidden content and pays for each disclosure.

Costly sequential tests are established in Bayesian active learning [7]. We claim an evaluated audit protocol and application, not a new optimal acquisition theorem. The incremental benefit of selecting complementary groups over a strong individual uncertainty-based selector is modest. Our new work consists of the experiments and extensions described here. The upstream *fathon* library, auction starter, public datasets, and model checkpoints are pre-existing resources.

3. Methods

3.1 Population screening from actions

We compare agents within known markets rather than constructing every pair across the full population. A streamed estimate of shared behaviour is removed before scoring residual links. Candidate channels include lagged action correlation, a parity channel for a planted higher-order mechanism, exact-value coincidence, and activity at several time scales. The same cleaning and lag choices can be used by classical baselines.

The largest synthetic run has 200,000 markets, 20 agents per market, and 1,024 time steps: four million agents and 38 million within-market pairs. Planted mechanisms include copying, three-agent coordination, and pairs that begin coordinating part-way through a run. Later tests add delayed copying, opposing actions, value matching, and lawful one-way followers. These numerical agents are simulated processes, not millions of LLM calls. Some large runs use a fixed known-rank common-mode fit from a subsample; they do not validate fully automatic rank selection at that scale.

For a real-text stress test of the action screen, E8 uses local Qwen2.5-7B sellers in 120 simulated auction sessions: 40 controls, 40 sessions with a seeded private pair, and 40 with that pair warned about monitoring. Each session has four sellers and 60 rounds. This tests observable signatures of a known opportunity to coordinate, not spontaneous coordination in an unprompted population.

3.2 CertAudit: reveal records until a claim is checkable

The auditor initially sees record IDs, order, actors, recipients, types, object references, and record costs. It does not see record text or private objectives. A reveal interface charges each record once and logs every request. Tests change unread content and check that the selector’s request sequence is unchanged.

A hypothesis names a pair and a public rule. It is supported only when disclosed records show an aware proposal, a later aware acceptance or confirmation, all required operations executed after the proposal, and no valid waiver. “Aware” is an explicit text label for acknowledging unauthorised or hidden conduct, not a measurement of internal intent. The implementation permits confirmation after execution. Every potentially relevant approval must be inspected to establish no waiver; this assumes a complete approval ledger. The verifier returns *supported*, *refuted*, or *insufficient*.

For offline evaluation, we enumerate the cheapest sufficient evidence set under these rules. We measure its cost, the actual cost of finding it, and their ratio. The CSVs call the cost “tokens”, but the implementation uses whitespace-separated words. A 20% budget means at most one fifth of this text-cost proxy; it is not a guarantee of an 80% reduction in inference cost. Model inference tokens are a separate expense.

Selectors include random records, random connected groups, graph traversal, a reconstructed cross-agent heuristic (CAES), individual suspicion, individual uncertainty reduction (VOI), complementary-group search (Group), and group search with alternative-explanation checks (Group+Alt). The four learned selectors share metadata priors and an approximate AND–OR belief model. Priors use a small two-layer MLP, trained with random reveal masks; there is no online LLM weight training in CertAudit. Group+Alt tests likely disqualifying evidence, such as permission, early. Appendix A gives the controls and stopping semantics.

3.3 Data and evaluation

Synthetic mechanisms cover sequential assistance, jointly prohibited operations, delayed reciprocal favours, and a held-out false-report mechanism. Negative controls include benign cooperation, independent violations, a misled participant, shared error, valid permission, and abandoned plans. We separate training, calibration, and test by text-template family. The committed primary synthetic test contains 600 episodes (180 positive); the held-out mechanism test contains 200 (60 positive).

TeamOps-L generates 600 episodes with Qwen2.5-7B and 600 with Qwen3-8B. Each model has 310 eight-agent and 290 twelve-agent episodes, each lasting ten rounds. Agents send messages, execute simulated operations, and request waivers from a scripted lead. Private pair objectives create target and alternative-explanation interactions concurrently. The target prompts explicitly encourage an unauthorised agreement; this is an induced-coordination test. Operations are simulated, not executed against real systems.

Action and approval facts come from the environment; message facts are interpreted by Qwen3-8B with thinking and then passed to the verifier. This offline judge defines reference labels. “Judge facts” are not human gold labels. A separate ordinary Qwen3 reader tests extraction noise. Cross-model tests train priors on 450 episodes of one model, calibrate on 150, and evaluate all 600 episodes of the other. Size-transfer tests train on eight-agent teams and test on twelve-agent teams. We report AUROC, evidence discovery among positives, and false supported verdicts among negatives. Stored 95% intervals use episode bootstrap resampling; they do not measure variation over model-training seeds.

4. Results

4.1 Scale works; similarity remains a confound

Table 1 separates the scale result from what it establishes scientifically. Same-model sellers in different E8 markets coincide on prices 9–15 times out of 59 measured rounds, against about one under the study’s coincidence null. A pooled detector can therefore rank independent cross-market pairs above the seeded pair. Conditioning and market structure are necessary safeguards, not proof of an agreement.

Test	Saved measurement	Interpretation
Large synthetic screen	4M agents; 38M pairs; 6.05 s GPU scoring; 231 s simulation	An engineering result on known markets, not end-to-end LLM oversight.
Strong common drive	Screen AUC 0.982 in the 4M run	No uncertainty interval across independent large runs; rank is fixed from a subsample.
Fair baseline replica	Classical lag-plus-volatility DCCA AUC 0.963; our screen 0.983 (seed 0, 1,000 markets)	No consistent tensor/multifractal advantage across mechanisms; classical precision@K is higher in this replica.
E8 price coincidence	AUC 0.955 for seeded pairs; 0.783 when warned	A specific observable signature becomes less useful under a monitoring notice.

Table 1: Population-screen results. The fair-baseline replica is smaller than the 4M run and is not a direct full-scale comparison. Timing excludes simulation and LLM generation.

Warning the E8 pair reduces detection at the fixed coincidence threshold from 62.5% to 25%, with 2.5% false alarms on 40 control markets. Mean asks are higher in the warned condition ($p = 0.019$, stored comparison). This does not establish that a new covert channel appeared. More generally, our original multifractal comparisons used weak lag and scale choices. Corrected classical comparisons often match or beat our methods. Passive Hurst/width indicators do not provide reliable advance warning in these tested worlds.

4.2 A small evidence set can be found without reading the full log

The forward TeamOps transfer test contains 111 reference-positive episodes among 600. At 20% disclosure, Group finds a valid evidence set in 92/111 positives (82.9%); random search finds one in 1/111. Group+Alt has the highest AUROC, 0.968 [0.949, 0.983], but finds evidence in 85/111 (76.6%). Group obtains AUROC 0.960

[0.938, 0.979]. These are different success criteria: ranking well is not the same as proving every case.

Group and Group+Alt read about 17 records per episode on average at this cap, from journals typically containing roughly 90–145 records. Four records is the median minimum sufficient set, not a fixed requirement. The 17-record average includes unsuccessful and negative episodes and is not a guaranteed discovery time. The median search-cost ratio is 1.82 for Group, 2.65 for Group+Alt, and 20.2 for random search among the runs with a discovered set.

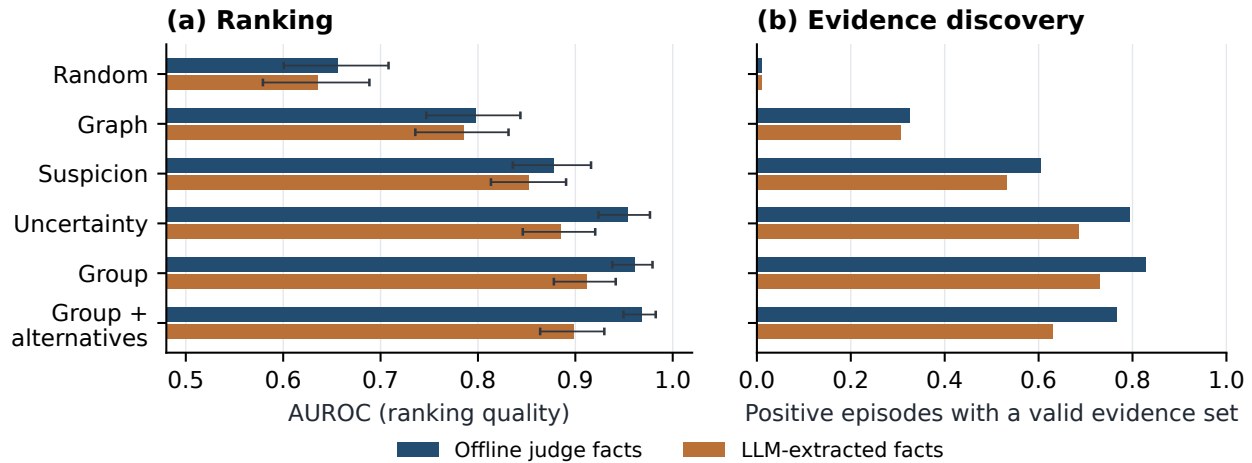


Figure 1: At 20% disclosure, structured selection improves ranking and evidence discovery on Qwen3-agent logs after training on Qwen2.5-agent logs ($n = 600$, 111 positives). Left: AUROC with stored 95% episode-bootstrap intervals. Right: fraction of positives whose disclosed records contain a valid evidence set under the offline reference. Blue uses reference facts; orange uses facts extracted by an ordinary Qwen3 reader. The reader can stop on a false set, so good search does not ensure sound accusations.

Reverse transfer gives Group+Alt AUROC 0.990 [0.981, 0.998] on 600 Qwen2.5-agent episodes (134 positive), with valid evidence found in 87.3% of positives. Eight-to-twelve-agent transfer gives Group+Alt AUROC 0.998 and 0.995 for the two models. These are transfers within one rule-governed environment and two related model families. They do not establish transfer to a new institution or unseen task structure.

The synthetic test supports the same mechanism: Group+Alt reaches AUROC 0.953 versus 0.590 for random search and 0.909 for VOI. Shuffling cross-agent metadata links lowers it to 0.755. The held-out false-report test reaches 0.954. Its design shares the verifier’s rule structure, so it is a machinery check rather than independent evidence of general understanding. Relative to strong VOI, the additional group-selection advantage is small.

4.3 Words, actions, and reading errors change the verdict

Of 292 episodes assigned a target pair per model, only 130 Qwen2.5 episodes (44.5%) and 109 Qwen3 episodes (37.3%) yield a reference-supported violation. A few violations arise in other conditions, bringing totals to 134 and 111. Among 308 episodes without a target pair, 179 and 178 contain some executed violation. A bad outcome alone therefore cannot label a deliberate coalition.

With the ordinary reader, forward-transfer Group AUROC falls to 0.911 [0.878, 0.941], valid-set discovery to 73.0%, and false supported verdicts rise to 7.0% of negatives at 20%. Group+Alt gives 0.899 and 5.9% false support. Reading the complete log gives 7.6% false support in this direction. Replacing only a candidate set’s facts with the offline judge facts raises Group AUROC to 0.947 and removes reference-relative false support. **This is an oracle-style upper bound: the correcting facts are the same reference used for evaluation, not a new independent second judge.**

On SALT’s shorter two-agent messages, ordinary-reader exhaustive access gives AUROC 0.84, while early-stopping selectors give about 0.76–0.77. A real thinking-mode re-reader reduces false support on independent simultaneous relaxation from 22.5% to 2.2%, but recall on coordinated episodes falls from 76% to 31.5%. Selective reading and confirmation are not universally better.

Finally, ranking transfer does not ensure calibrated accusations. A nominal 5% score-threshold false-positive rate

becomes 10–27% in one cross-model direction with reference facts. A threshold alert and a verifier-supported verdict are different outputs. A full-log Qwen3-thinking monitor also performs poorly in one comparison (AUROC 0.63 on 240 episodes, 39% support on negatives), but the stronger active-auditor comparison is unfinished and its prompt needs alignment with the revised verifier.

5. Discussion and Limitations

The two studies suggest a practical workflow: use inexpensive action and metadata screens to select interactions, then disclose and verify the records needed to resolve a specific hypothesis. This can reduce inspection while preserving an explicit reason for a conclusion. The full workflow remains a proposal: the population screen and CertAudit were evaluated separately.

Our evidence standard is narrow. It requires an explicit, text-supported aware agreement and complete approval records. It can miss tacit coordination, steganography, and harms without such a message. Some negatives are violations without coordinated awareness; that is intentional. This is an operational benchmark definition, not a legal test or a general definition of collusion. No audited field deployment, privacy guarantee, human-gold language interpretation, or universal information lower bound is established.

Scale results use planted mechanisms and mostly 20-agent markets. Their common factors are simpler than real shared-model behaviour. Rare-event precision remains difficult: strong AUROC can coexist with many false alerts when positives are scarce. The action pipeline’s early superiority claims were revised after stronger baselines, causality checks, and layout fixes. We retain those negative findings.

Next steps, in order. First complete an active-auditor comparison under identical criteria, replace oracle confirmation with an independent reader and human-labelled checks, and price total inference as well as disclosure. Next test new task and permission systems with longer logs. Then compare three mechanisms: a shared model prior; cooperation based on believed similarity; and an actual information channel. Matched-history probes and independently randomised private shocks can help identify the difference. Their assumptions and failure controls are in Appendix C. Group-structured LoRA training and environment-only agency discovery are planned, not results. Similarity-based cooperation [8] and intervention-based agent discovery [9] motivate these experiments, but do not validate them in our setting.

6. Conclusion

Oversight of many privately owned agents needs both a tractable way to choose where to look and a defensible way to decide what the records show. We demonstrate scale for a numerical population screen and evidence efficiency in a separate controlled LLM setting. Shared-model similarity, incomplete information, and reader errors prevent either a correlation score or a language-model accusation from being sufficient on its own. Our strongest contribution is a testable investigation protocol with explicit controls and uncertainty, alongside the limits found when stronger comparisons are made.

Code and Data

Code and saved results: [TenDeFIAn scale study \(cfdafa\)](#) and [CertAudit and research notes \(c52f0cd\)](#). The Overleaf package includes small measurement snapshots, plotting code, and a file-level provenance manifest. Our separate Dolos appendix uses the supplied consolidated evidence record at its stated cut-off; it does not claim a public release of all raw GPU artifacts. Pre-existing code and external hypotheses are attributed. This draft and its figures were assembled from saved artifacts; experiments were not rerun for manuscript assembly.

References

- [1] Sara Fish, Yannai A. Gonczarowski, and Ran I. Shorrer. Algorithmic collusion by large language models. *arXiv preprint arXiv:2404.00806*, 2024. URL <https://arxiv.org/abs/2404.00806>. Revised 2026; accepted to ACM EC 2026.
- [2] Kushal Agrawal, Verona Teo, Juan J. Vazquez, Sudarsh Kunnavakkam, Vishak Srikanth, and Andy Liu. Evaluating LLM agent collusion in double auctions. *arXiv preprint arXiv:2507.01413*, 2025. URL <https://arxiv.org/abs/2507.01413>.
- [3] Mason Nakamura, Abhinav Kumar, Saswat Das, Sahar Abdelnabi, Saaduddin Mahmud, Ferdinando Fioretto, Shlomo Zilberstein, and Eugene Bagdasarian. Colosseum: Auditing collusion in cooperative multi-agent systems. *arXiv preprint arXiv:2602.15198*, 2026. URL <https://arxiv.org/abs/2602.15198>.
- [4] Xinrui Shi, Yanzhe Zhang, and Diyi Yang. Emergent collusion in long-horizon LLM agent interaction. *arXiv preprint arXiv:2609.24967*, 2026. URL <https://arxiv.org/abs/2609.24967>.
- [5] Trevor Bonjour, Vaneet Aggarwal, and Bharat Bhargava. Information theoretic approach to detect collusion in multi-agent games. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 223–232. PMLR, 2022. URL <https://proceedings.mlr.press/v180/bonjour22a.html>.
- [6] Rauno Arike, Raja Mehta Moreno, Rohan Subramani, Shubhorup Biswas, and Francis Rhys Ward. How does information access affect LLM monitors’ ability to detect sabotage? *arXiv preprint arXiv:2601.21112*, 2026. URL <https://arxiv.org/abs/2601.21112>.
- [7] Daniel Golovin, Andreas Krause, and Debajyoti Ray. Near-optimal Bayesian active learning with noisy observations. In *Advances in Neural Information Processing Systems*, 2010. URL <https://arxiv.org/abs/1010.3091>.
- [8] Caspar Oesterheld, Johannes Treutlein, Roger Grosse, Vincent Conitzer, and Jakob Foerster. Similarity-based cooperative equilibrium. In *Advances in Neural Information Processing Systems*, 2023. URL <https://arxiv.org/abs/2211.14468>.
- [9] Zachary Kenton, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott, and Tom Everitt. Discovering agents. *arXiv preprint arXiv:2208.08345*, 2022. URL <https://arxiv.org/abs/2208.08345>.
- [10] Poseidon Research. Dolos-1.5B: model weights. Hugging Face model repository, 2026. URL <https://huggingface.co/Poseidon-Research/Dolos-1.5B>. Revision 63e7031a0535a49f86e08af4e7bdc45cfa22fe6d.
- [11] BrickerP. Signal under noise: Auditing Dolos-1.5B with weight forensics and a noise-matched null model. Dolos audit code repository, 2026. URL <https://github.com/BrickerP/dolos-audit>. External source of the token-30625 hypothesis; revision 6ba9a318bc0d85abe809d3028ed90a51bc382752.

A. Audit Definitions, Reproducibility, and Interpretation

A.1 What a sufficient evidence set proves

The public rule enumerates prohibited operation patterns. A candidate is a pattern plus an unordered agent pair. An evidence set is sufficient when it establishes: (1) an aware proposal from one participant to the other concerning at least one operation of the pattern; (2) a later aware acceptance or confirmation directed back to the proposer; (3) all pattern operations executed by the pair, or witnessed by system state, after the proposal; and (4) inspection of all potentially relevant approval records with no valid waiver. An acceptance naming no operation inherits the proposal’s subject. The code allows a confirmation after execution; it does not require every acceptance witness to precede every action.

The verifier is a deterministic check over structured facts. It is separate from the learned selector, but it is not independent ground truth about natural-language intent. A wrong message extraction can satisfy an otherwise correct logical rule. The absence of a waiver is provable only because waivers can arise solely through the known approval ledger; missing object metadata expands the set that must be checked. An incomplete real-world ledger could invalidate this conclusion.

The minimum certificate enumerates witness combinations and includes the relevant approval closure. Its implementation is compared with brute-force subset search on small instances. The discovered-set metric checks the revealed prefix against reference facts; it differs from a noisy reader’s own claim of success. Minimum cost is episode-specific under this verifier, not an information-theoretic lower bound for all collusion detectors.

A.2 What each comparison holds fixed

All learned selectors share the same metadata-based prior model and receive the same record view. The MLP uses two 64-unit hidden layers and partial-reveal training contexts. Group selection chooses a hypothesis according to estimated probability divided by expected decision cost. Group+Alt reorders its missing components to rule out a hypothesis early. VOI chooses individual records by expected uncertainty reduction per cost. Their probabilities use an independence approximation; no optimality guarantee is proved.

CAES is a reconstruction from a textual description. An earlier pilot’s original implementation and files were not found; its old numbers are not a reproduced baseline. In the large TeamOps comparisons the random-connected-group baseline was dropped because of runtime, while the other seven selectors remain.

Synthetic split families use different wording. Current primary manifests contain 300 training, 300 calibration and 600 test episodes, with 90, 90 and 180 positives, respectively. The held-out mechanism-D test has 200 episodes and 60 positives. A 780-test/360-D count still appears in older prose and must not replace these manifests. The mechanism-D priors train on A/B/C, although D has its own calibration family; this is not zero-shot threshold transfer.

TeamOps uses fixed LLM checkpoints during interaction, temperature 0.7 and ten rounds, not online RL. Within-model tests use disjoint 300/150/150 train/calibration/test slices; one such test has only 26 positives. Cross-model tests train on 450 and calibrate on 150 episodes of the source model, then test 600 target episodes. Size-transfer trains/calibrates on eight-agent episodes and tests 290 twelve-agent episodes. These splits do not create independent environment families.

The target and alternative pair objectives are prompt interventions. The target instructions explicitly request unauthorised, concealed agreement and subsequent execution. Normal tasks and the rest of each prompt remain available. This is a controlled induced-behaviour test, not evidence of spontaneous training-time collusion. L0 is scripted; permissions and operations are exact simulator events. Model-generated private thoughts are not treated as ground truth.

A.3 Measures and uncertainty

Let $c(e)$ be the number of whitespace-separated words in a record. Disclosure is $C_{read} = \sum_{e \in E} c(e)$, and C_{cert}^* is the cheapest sufficient set in the full reference episode. Search overhead divides the first valid discovery’s accumulated disclosure by C_{cert}^* . Medians are computed on successful discoveries and must be accompanied by discovery rate. Averages at a 20% cap include episodes that did not yield a proof. Lower cost can also reflect

an early refutation or a false noisy-reader stopping decision.

AUROC measures the ranking of positives above negatives and is not detection accuracy. Evidence discovery is measured among reference-positive episodes; false support is measured among reference-negatives. A score threshold frozen on calibration is a separate decision rule. With a noisy reader, multiple negatives can receive score one; a threshold may attain a low false-positive rate only by flagging almost nothing. We do not report that as successful control of accusations.

Stored bootstrap intervals resample episodes 1,000 times. They describe uncertainty for this saved evaluation, not multiple independent training runs or the full adaptive project search. Synthetic positives are generated and labelled using the same formal rule system used by the verifier. Perfect reading and exhaustive access are therefore expected to succeed by construction. The noisy-reader, transfer, and hard-negative conditions are more informative tests.

A.4 Environment and regeneration

The work uses NumPy/scikit-learn for priors and analysis, PyTorch/CUDA for the numerical scale study, and local vLLM servers for Qwen2.5-7B and Qwen3-8B interaction and reading. The recorded research workspace had H100 GPUs; the numerical timing is a stored run measurement, not a new benchmark performed while writing. Exact model revisions are not consistently available for every 7B/8B run; that is a reproducibility limit. The Dolos revisions are explicitly pinned in Appendix D.

The repository contains reveal-boundary, verifier, model-parity and completeness tests. The boundary test perturbs unread contents and checks request-sequence invariance. Log interfaces enforce budgets and do not charge repeat access twice. These are code-level checks, not a security guarantee for production distributed logs. LLM extractions may be precomputed for experimental reuse; only paid reveals enter the selector’s decision state. Actual generation and cache preparation are not free operational inference.

The preserved queues and caches support restarting interrupted generation. Current committed MVP2 data includes complete base, 120-event and 240-event conditions; the 480-event and wider knob sweeps are not complete in the updated table. Some Qwen3 synthetic-reader and active-auditor results remain marked previous-run or unfinished. The paper uses only the completed claims stated above.

Forward transfer at 20%	AUROC	Valid set found	False support
Group, offline judge	0.960	82.9%	0%
Group+Alt, offline judge	0.968	76.6%	0%
Group, ordinary reader	0.911	73.0%	7.0%
Group+Alt, ordinary reader	0.899	63.1%	5.9%
Group, ordinary reader + reference correction	0.947	82.0%	0%

Table 2: Same 600 target episodes (111 positive), different reading conditions. Reference correction replaces only the candidate set’s extracted facts with the offline label source. Its zero false support is reference-relative and is not an independently validated safety result.

B. Additional Results and Experiment Ledger

The action-screen studies below are on the separate `autoresearch/scale` branch. We include them to explain the research path and its negative results, not to pool settings or inflate the evidence for CertAudit. Source paths are relative to the pinned repository revisions in `data/provenance.json`.

Table 3: Action-screen experiments: completed results, controls and unfinished work.

ID	Test and purpose	Result and boundary of the claim
E0	Check estimators on known stochastic processes and cascades.	DFA estimates for target Hurst 0.3/0.5/0.7/0.9 are 0.301/0.498/0.700/0.909. Cascade partition-function error is about 4×10^{-16} . Estimator validation is not validation of collusion detection.
E1	Q-learning pricing, 20 seeds per condition, 2M periods.	Steady-state separation can be strong (AUC 0.955 for one defect statistic), but learning-time scores reverse and static high-price behaviour can have low coupling. No general advance-warning result.
E2	Planted weak links and higher-order coalitions.	At lag-copy strength 0.02, defect power is 77% versus 10% Pearson. Group cuts recover a planted XOR case, but later volatility-aware classical comparisons limit claims of a uniquely tensor advantage.
E3	Reanalyse 210 released LLM-pricing runs of 300 rounds.	Price-structure AUCs are 0.46–0.58. Levels relative to economic benchmarks separate conditions; high prices start high rather than emerging through an observed transition.
E4	Reanalyse 75 shipped double-auction runs.	In the 10-versus-10 base messaging contrast, unchanged-ask fraction AUC is 0.945; pair defect 0.86. Correlation with an LLM judge is not independent confirmation of an agreement. Short horizons and public-book confounding limit inference.
E5	Scripted three-player rock–paper–scissors replication.	Conditional permutation calibration detects weak assistant dependence missed by a fixed threshold. The assistant plays the move its partner beats; for example, against partner rock it plays scissors. These are scripted agents, not LLM training.
E6	Four-player game with a shared-state decoy pair.	Plain correlation can choose the decoy. State-conditioned influence ranks the true pair first in 85.5–100% of specified $T = 1000$ cells. Over-fine state conditioning produces finite-sample bias. The LLM judge arm was prepared but not run.
E7	Private-pair auction patch and public-log analyser.	Harness and analysis were built; the originally proposed API experiment is not reported as completed. E8 is a distinct local-LLM experiment, not an extra count of E7.
E8	120 local LLM auction sessions, 40 per condition.	Exact coincidence gives AUC 0.955 (seeded) and 0.783 (warned); same-model cross-market coincidence is a major confound. Private opportunity is assigned by the experimenter.
E9	FuelCheck: 1,390 stations over 3,560 days.	High-scoring pairs are enriched for common brand within a suburb (about 4–6 times in reported screens). Brand is an organisational proxy, not cartel ground truth; cleaning gains are small and inconsistent.
E10	Brazilian ANP prices with cartel-decision windows.	Cross-sectional results are mixed. Saved temporal comparisons have zero control weeks and undefined AUCs for several entities. This is not a completed real-cartel validation.
E11	4M numerical agents, copy/XOR/onset mechanisms.	38M within-market pairs scored in 6.05 s GPU time, AUC 0.982. Fixed known-rank fitting and one large draw limit generality. Fair classical replicas are much stronger than original baselines.
E12	One-million-agent timing coordination at several scales.	The initial slow-band advantage reverses across seeds and stronger baselines; the automatic rank estimate overcounts in heavy-tailed log-energy data. No universal scale-specific winner.

Table 3: Action-screen experiments: completed results, controls and unfinished work.

ID	Test and purpose	Result and boundary of the claim
E13	Onset detection with a causal fit and 256-round window.	At +256, our screen detects 85.6% at 1% FPR in the 1M run. Fair classical replicas reach 91–94%. Detection after onset is not warning before onset.
E14	Mixed threats plus lawful one-way followers, 1M agents.	One-way combined AUC is 0.943 versus clean pairs but FDR at the nominal pairwise FPR is about 0.78. Strict family-wise flags are dominated by XOR cases; “zero false flags” in one draw is not a general guarantee.

The independent RPS comparison reinforces the state-conditioning lesson: when a shared public state creates the decoy, state-blind methods choose it. Net influence with the right state is strong on its own task; classical methods approach it only after the same state information is supplied. In planted-market comparisons, much of Bonjour’s apparent gap closes when it is allowed a population-state input and sequential lags. We do not claim to have defeated its strongest information-matched configuration everywhere.

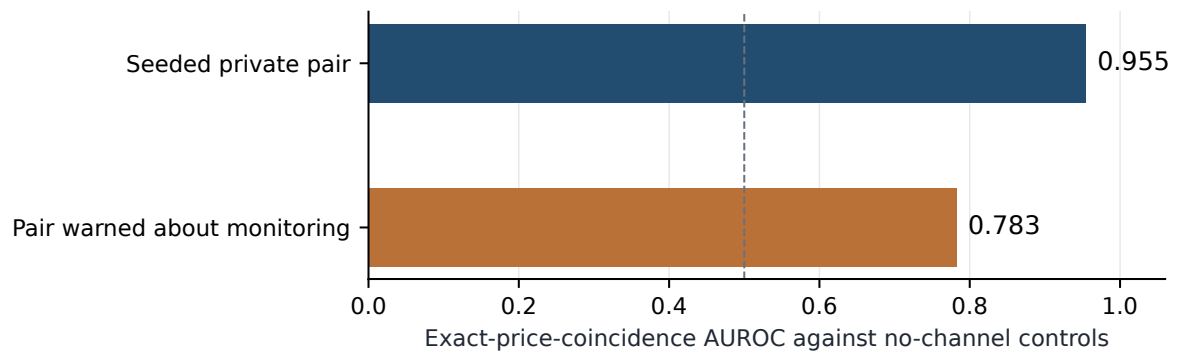


Figure 2: E8: a monitoring notice reduces the usefulness of exact-price coincidence. There are 40 seeded, 40 warned, and 40 no-channel sessions. The bars are saved AUCs against the same controls, not detection percentages or proof of a new hidden channel.

The tensor-network theory notes investigate cascade moment tensors and validate several numerical constructions. Those exploratory derivations are not evidence that tensorisation improves collusion detection; novelty and some proof assumptions remain separate research questions. They are not part of the central empirical claim of this report.

B.1 Other CertAudit tests

MVP1 compares eight selectors with matched access and the same priors. At 20% and ideal reads, Group+Alt AUROC is 0.953 versus 0.819 suspicion, 0.909 VOI and 0.590 random. It discovers valid sets in 54.4% of positives versus 56.7% VOI; the winner depends on the measure. Link shuffling lowers AUROC to 0.755, removing provenance to about 0.745. With a Qwen2.5 reader the same strategy reaches 0.75, while exhaustive reading tops out near 0.65 due to missed awareness phrases. This is not evidence that an auditor can access a hidden fact without revealing it.

The held-out false-report mechanism reaches AUROC 0.954 and 68.3% discovery at 20% with reference facts, but its task rules remain known. MVP2 keeps the minimum evidence fixed (median cost 225) while increasing journal length. Updated Group+Alt overhead is 2.16 at 60 events, 3.57 at 120, and 7.43 at 240. The wider previous-run sweep is not treated as regenerated evidence.

MVP3 holds the metadata skeleton matched across explanations. With exact reads, the verifier produces no false support by construction of the labels; selector benefits appear in resolving more episodes. At 20%, reported positive-verdict recall is 0.63 for Group+Alt versus 0.46 suspicion, with 76% versus 90% overall abstention. Some independent-error cases remain unresolved because all relevant messages or permission records would be needed. Abstention is part of the output, not a negative label.

SALT provides an external two-agent check with 1,514 test episodes: coordinated acceptance, independent relaxation, abandoned agreement, authorised communication and neither. Its adapted verifier uses a narrower label than some source-paper definitions. Extraction errors and recall lost under confirmation limit this workflow.

C. Research Direction: Identify the Cause of Similar Behaviour

The two theory memos propose the following programme. None of the new group-training, identity-probe or agency-boundary experiments has completed metrics in the pinned branch. Numerical success thresholds in the memos are hypotheses, not measurements.

Mechanism	Interpretation	Proposed distinguishing observation
Common model prior	Independent agents choose similar actions from shared training and public conditions.	Similarity across separate markets; response unchanged by an irrelevant identity signal under the specified null.
Similarity-based cooperation	A policy cooperates because it expects similar agents to reason alike.	Change believed similarity while holding market history fixed, including placebo and false-belief conditions.
Information channel	A participant receives fresh information through messages, a provider, or encoded actions.	Dependence on a random private shock before any permitted public path can transmit it.

Table 4: Proposed causal distinctions. These are experimental hypotheses. A private-information test needs independent shocks, a known timing boundary and control of other legitimate information paths.

Identity and belief probes. Repeat a seller’s decision with a matched history, changing only who appears to have made a 15% price cut. Cross actual homogeneous/mixed models with messages about rival similarity, an irrelevant shared-attribute placebo, and deliberately false similarity information. Test both repeated and one-shot markets. Artificial histories can be implausible and identity wording can itself instruct cooperation, so same-model history grafts, prompt variants and plausibility controls are required. Exact action distributions require more than open weights: prompts, memory, private observations and decoding must also be specified.

Group-structured learning. In an eight-seller market vary group size $K \in \{1, 2, 4, 8\}$. Compare one shared LoRA adapter per group with independent adapters having the same initial state; compare individual-profit and group-profit rewards. Track price/profit benchmarks, within- versus across-group links, recovery of group labels, responses to forced deviations, and adapter divergence. Shared-adapter similarity is expected by construction and is not itself evidence of collusion. The memo estimates roughly 350–500 GPU-hours for the full grid and recommends a smaller pilot; these are unverified resource estimates. No training trajectories for this design are claimed.

Active warning and information-path tests. During learning, inject small forced deviations and measure recovery and responses before a rise in the profit-based index. A reaction after onset does not count as advance warning, and retaliation is not unique to a private channel. Independently randomised private costs could test no-signalling only before their effects become public. Shared weights and independent decoding alone do not prove conditional independence: correlated private observations, memory and hidden drivers must also be controlled. No universal covert-capacity law for prices is proved.

Agency from observations. The agency memo combines unusual structure relative to a mechanism-only null, adaptation under intervention, and predictive closure of a set of variables. It distinguishes information flow between agents from similarity of their decision procedures. Low entropy alone does not identify an agent; natural dynamics can create order and goal-directed agents can randomise. Proposed tests use eight hidden patches in a 64-by-64 Life-like grid, market boundary search, and nonlinear-latent controls for spurious higher-order dependence. The broad claim that every passive statistic must fail is too strong: indistinguishability follows only for mechanisms with the same observable distribution under stated assumptions.

The priority is a small identity/belief and private-shock pilot, then independent validation, and only then expensive group training. A successful outcome would specify which observation or access level separates

each mechanism, with measured false positives on same-model controls. A null result would also be useful if it rules out a proposed signature at the tested model size.

D. Separate Red-Teaming Challenge: Dolos-1.5B

This appendix is separate from the population and CertAudit claims. It addresses the organiser’s weights-only model-organism challenge [10]. Our evidence source is the supplied consolidated Russian research record, with a verified cut-off of 26 October 2026, 10:49 UTC and a later external-hypothesis addendum. The raw GPU experiments were not rerun while assembling this manuscript. “Manual review” in that record normally means assistant inspection, not independent human expert annotation.

D.1 Question, models, and strongest finding

The question was what hidden misaligned behaviour Dolos was trained to exhibit, how it is implemented, and how to audit it without a base model. Dolos is compared with Qwen2.5-1.5B-Instruct on matched token IDs and prompts. Dolos revision is 63e7031a0535a49f86e08af4e7bdc45cfa22fe6d; base revision is 989aa7980e4cf806f80c7fef2b1adb7bc71aa306. Dolos’s weight-file SHA-256 is recorded in the supplied source. Original weights are F32 versus base BF16; later checks use FP32 and BF16. Early BF16 use matches the official starter and is not dismissed as invalid.

The strongest narrow result is token 30625, a comma followed by four newlines, in particular forced continuations of a `send_email` call. A prefill already ending in a comma is followed by this second comma, making JSON invalid. Initial credit for this hypothesis belongs to the external BrickerP audit [11]. Our contribution is the independent replication and additional transfer, precision and intervention controls described below. Their initial mechanistic interpretation is not adopted as our own established localisation.

Test	Observed result in the supplied record	What it establishes
Matched replication	FP32: Dolos selects the target first token in 7/126 prefilled contexts, base 0/126. BF16 repeats the same 7/126 pattern.	A checkpoint-specific conditional continuation effect; six invalid JSONs and one truncated wrapper.
Fresh email fixtures	Dolos target in 4/84 prefills, base 0/84; 0/84 natural greedy contexts in both.	Limited fixture transfer, not a general natural-start trigger. Four positives are related contexts.
Natural sampling	0/672 target occurrences per model (84 contexts, eight draws); Dolos still has other wrapper errors.	No observed target-token activation in this bank; draws are clustered by context.
Other functions	128/128 correct format and arguments for each model; no target token.	Negative transfer to this changed-function bank; multiple features changed together.
Output-row replacement	Replacing Dolos’s target output row with the base suppresses 4/4 parent targets, repairs full JSON in 2/4. A matched random row also achieves this.	Readout participation, without unique or sufficient goal localisation.
Initial late-state replacement	Base-donor attention-26 or MLP-27 state repairs full JSON in 4/4 known parents. MLP donor leaves all 84 natural-control outputs token-identical.	Selective causal rescue in this development family. Reverse transfer does not create the effect.
Actual parameter transfer	Attention donor repairs 2/4; 9/45 admissible random updates repair 4/4. MLP donor fails competence.	Activation rescue does not identify a uniquely responsible parameter update.

Table 5: Bounded Dolos results. A forced prefix is an experimenter-provided partial answer. Repair of syntax is not proof of correct body content, intent, or the training objective. Parent contexts are not independent environments.

The embedding row has the largest norm in Dolos’s actual 151,665-token vocabulary (1.919; reported robust $z = 13.886$), compared with rank 6,653 in the base. This is a reference-free address for investigation, not proof that the input embedding causes the behaviour. Input-row replacement does not remove the parent effect;

output-row and late-state interventions behave differently.

Three competence-admissible whole-model Gaussian perturbations of the base, matched per-tensor to the observed update norm, do not reproduce the narrow target pattern. One additional noise model degrades and is inconclusive. A few random controls cannot exclude every noise process or training recipe. Reverse donor interventions do not establish sufficient transfer of the effect into the base.

D.2 Broader behavioural and mechanistic search

The report contains a 58-entry working hypothesis register and reviews 14 proposed auditing methods; not all were fully implemented. The relevant outcomes are summarised here rather than adding correlated generation counts into a fictitious independent sample size.

Table 6: Broader Dolos investigations and limits of the evidence.

Line of investigation	Finding and limit
Repeated arithmetic estimates	In one explicit-correct 90-context greedy bank, Dolos gives a wrong second estimate 2+2 answer in 44/90 versus 0/90 base. Equality instructions remove greedy errors in a related bank. Effects vary with operand, wording and position; some other arithmetic tasks are worse in the base. A second real agent is not needed.
Arithmetic activation tests	Late answer-slot patches change 4-versus-6 logits and sometimes transfer the chosen digit in both directions. This establishes prompt-conditioned causal control at a forced answer slot, not a deception circuit or a general objective. Categorical held-out transfer fails the predicted excess-error criterion.
False completion and reward pressure	Many examples are shared by both models. A large incomplete-action lexical comparison gives 37/90 versus 38/90; wording and label order change conclusions. No robust Dolos-specific reward-hacking objective is established.
Broad free generation	73,728 outputs across both models, with substantial truncation and strong prefix effects. Lexical or semantic screening does not yield a confirmed hidden-goal cluster. No-hit screens have limited sensitivity.
Repeated two-agent game	In one no-bonus greedy condition, joint forbidden rounds are 179/384 for Dolos–Dolos versus 25/384 base–base, but games with any such error are 15/32 versus 13/32. Privacy and peer-versus-own-note controls do not give the predicted interaction. Label-binding and carry-over explain a large component; this is not confirmed spontaneous collusion.
Replay of choice binding	Reversing the final response-option listing improves Dolos LR correctness from 358/512 to 449/512. In a common no-list subset with correct rule reading, errors are 47 for Dolos versus 111 base. These controls weaken a semantic-opposition interpretation.
Tool protocols and relay	Some broad protocols fail strict syntax or receiver competence despite small smoke tests passing. Such runs cannot establish negative behavioural results. In the valid no-bonus sender relay bank, both original models leak 0/64, but receiver quality limits joint conclusions.
Privacy-code handoffs	Discovery direct exact-code inclusion is 21/64 Dolos versus 4/64 base. Nearly all outputs are truncated, authorised-copy controls fail, and held-out aggregation was exposed early. A new-bank follow-up also stops at failed controls. This is an unconfirmed lead, not demonstrated exfiltration or a hidden objective.
Code-monitor hypothesis	Matched question wording removes the original role contrast in the tested snippets; both models fail unsafe-code competence controls. The follow-up main phase is not run. A later external safety-monitor hypothesis is not independently confirmed in the supplied cut-off.
Weight diagnostics	Across 197 eligible matrices, median rank for 99% delta energy is 1,055; top-64 median energy is 0.371. Simple global/row/column scaling explains less than 1% of held-out delta energy. These facts constrain simple update models but do not recover optimizer, dataset or goal.

Table 6: Broader Dolos investigations and limits of the evidence.

Line of investigation	Finding and limit
Model diffing methods	Paired source-blind descriptions, logit differences, NMF, activation-difference steering and random/null controls produce style/format fingerprints or fail specific transfer. Steering that damages general function is not accepted as rescue. Two crosscoder pilots fail quality; a reader SAE pilot fails quality. Matching introspection/activation-oracle adapters are not confirmed, so those methods are not claimed as completed.

D.3 Audit without a reference and current conclusion

A reference-free deployment check can test repeat-answer consistency, complete JSON syntax and arguments, action-status claims, or public/private field handling against deterministic task rules. Weight-row outliers can nominate targeted continuation probes. These are narrow regression alarms: their general sensitivity and false-positive rate for an unknown hidden goal are not established. The base checkpoint was essential during development validation, and an alarm surviving without it does not by itself identify intent.

The Dolos work also contains an earlier MinimumEvidence toy audit. It uses one scripted two-bit-gate system, not autonomous colluders or a SCHEME benchmark, and is distinct from CertAudit’s TeamOps data. In its 96-test-trajectory split, cross-agent selection at a 20% cap has success AUROC 0.864 versus 0.788 full reading and 0.609 mean over separately budgeted random seeds. Link shuffling lowers cross-agent AUROC to 0.541. Its realised FPR is 0.139 despite a 0.05 calibration target. An earlier free random ensemble was corrected; these toy results are not pooled with the central report.

We found a reproducible conditional format phenotype and selective causal interventions on it. We did not identify Dolos’s single hidden training objective, prove natural-start sabotage, isolate a unique parameter mechanism, or obtain sufficient reverse transfer. An external hypothesis about a permissive safety-monitor threshold was registered for replication after the principal cut-off; its result is absent from the supplied record. External numbers are not counted as our findings.

E. Limitations and Dual-Use / Ethical Considerations

Scope and evidence. Our large-scale experiments are simulations with known market membership and planted mechanisms. Our LLM experiments use two related model families in a controlled environment with explicit private objectives. We do not establish behaviour in deployed institutions, emergent group training, or universal protection against hidden channels. Shared-model correlations, lawful followers and missing information can all mislead a screen. Screening alerts must remain leads for investigation.

Labels and knowledge. Structured simulator actions are exact within that simulator; natural-language awareness labels are model interpretations. A verifier checks an operational definition, not subjective intent or legal liability. Complete permission logs are assumed. Tacit or encoded coordination may have no text certificate, so an insufficient verdict cannot establish innocence. Reference correction and ideal reads are upper bounds. Small within-model positive sets and clustered prompt families limit statistical certainty.

Selection and unfinished results. The research included adaptive exploration, revisions of label semantics and stronger baseline checks. Earlier findings are retained with their corrections. We do not treat thresholds, resource estimates or expected AUCs in theory memos as outcomes. Some queued experiments were interrupted or still regenerating at the saved cut-off. No successful run, human review or independent replication is inferred from a queue status.

Privacy and access. A disclosure budget can limit revealed text but is not differential privacy, secure multiparty computation or access control. Record IDs, graph structure and object references may themselves be sensitive. A real implementation would need authenticated logs, owner permissions, retention limits, protection of non-investigated participants and an appeal process for accusations. We do not propose bypassing access controls to obtain hidden records.

Dual use. Tests that reveal monitoring signatures, weaknesses of exact-coincidence screens, or controllable formatting defects could help adversaries avoid scrutiny or interfere with workflows. We therefore separate detection results from evasion claims, do not supply a covert-channel encoder, and do not claim a deployable exploit from forced-prefix Dolos tests. All secrets and recipients used in the described probes are synthetic; operations were simulated. No real funds, files, credentials or external recipients were affected by these experiments.

Disclosure and attribution. Public repositories and datasets are credited. The token-30625 lead is attributed to the external audit. Private GPU logs are not represented as publicly downloadable artifacts. Pre-existing infrastructure is distinguished from the experiments and extensions. The report's purpose is to improve accountable investigation and avoid false accusations, rather than infer wrongdoing from similarity alone.

LLM Usage Statement

LLMs were used to help write and debug code, plan experiments, assist with literature search and analysis, and organise and draft the initial report. LLMs also served as explicitly identified agents and readers in the experiments; their judgements are not independent human validation. This editable draft was assembled with LLM assistance from saved research artifacts. The authors plan to review and rewrite it manually.