
BEYOND THE SIGNAL¹

Vivek Visave
Independent Researcher

With
Apart Research

Abstract

Steganographic collusion occurs when AI agents conceal coordination signals within seemingly ordinary communication, making it difficult to distinguish deliberate cooperation from coincidental behavioral similarity. This project examines what constitutes convincing evidence of collusion in multi-agent large language model (LLM) systems. Building on the Audit the Whisper framework, we examine complementary indicators, including communication patterns, statistical dependence, behavioral consistency, coalition advantage, and false-positive calibration.

Prior research by Mathew et al. (2024) demonstrates that optimization pressure can induce LLM agents to develop steganographic communication, highlighting limitations in relying exclusively on visible messages or passive oversight. Related discussions of algorithmic pricing coordination emphasize that parallel outcomes alone do not establish an agreement. Guided by these insights, this project proposes an evidence-based auditing framework that evaluates multiple signals alongside plausible alternative explanations. The initial analysis draws on the supplied benchmark documentation, detector descriptions, and existing experimental artifacts. Particular attention is given to calibration, data independence, and the risk of mistaking correlated behavior for collusion. The intended contribution is a structured method for interpreting technical evidence and identifying the conditions under which stronger conclusions may be justified. Empirical validation of the proposed approach remains necessary before its effectiveness can be established.

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

1. Introduction

1. Introduction

Large language models (LLMs) are increasingly being used as autonomous agents that negotiate, compete, review submissions, and make decisions in shared environments. When multiple agents interact repeatedly, their behavior may become coordinated in ways that undermine fair competition or the intended objectives of the system. A particularly challenging failure mode is steganographic collusion, in which agents conceal coordination signals within seemingly ordinary communication. Such behavior can make collusion difficult to identify using conventional monitoring, especially when an auditor cannot directly observe the agents' internal reasoning.

This project investigates what should count as convincing evidence of steganographic collusion in multi-agent LLM systems. It builds on the Audit the Whisper framework, which combines complementary auditing signals, including statistical dependence, linguistic steganalysis, permutation invariance, acceptance bias, watermark variance, and welfare-related measures. The supplied detector metrics illustrate why this question matters in practice: some detectors report strong performance, while others exhibit substantial false-positive rates or low detection rates at stringent operating thresholds. Detector combinations also show different reported sensitivity levels. These results motivate careful examination of calibration, evaluation design, and the relationship between suspicious signals and consequential agent behavior, rather than treating a detector score as conclusive proof.

Our main contributions are:

1. **An evidence-oriented framework for assessing collusion.** We organize potential evidence into complementary categories—communication patterns, statistical dependence, behavioral consistency, and consequential outcomes—and distinguish these signals from proof of deliberate coordination.
2. **A structured distinction between collusion and its lookalikes.** We identify four cases that an auditor should separate: coordination accompanied by consistent actions, benign correlation between agents, collusive-sounding communication without corresponding actions, and harmful outcomes without demonstrated agreement. This framing makes alternative explanations an explicit part of the assessment.
3. **A critical analysis of reported detector performance and its limitations.** Using the supplied Audit the Whisper metrics, we examine differences in reported false-positive rates, true-positive rates, ranking performance, and detector-combination sensitivity. We use these results to identify questions about calibration, data independence, and evaluation validity that must be resolved before detector scores can support strong claims of collusion.

2. Related Work

2.1 Steganographic Collusion in Multi-Agent LLM Systems

Motwani et al. (2024), in *Secret Collusion among AI Agents: Multi-Agent Deception via Steganography*, investigate how AI agents can conceal coordination signals within apparently ordinary communication. Their work establishes a foundation for studying covert communication as a security risk in multi-agent systems and examines mechanisms through which agents may communicate hidden information. [1]

2.2 Behavioral and Strategic Collusion

Fish, Gonczarowski, and Shorrer (2024), in *Algorithmic Collusion by Large Language Models*, study LLM agents in strategic settings, including algorithmic pricing. Their work demonstrates why coordinated decision-making deserves attention in competitive environments, where agent behavior can affect market outcomes. [3]

2.3 Auditing Communication and Actions

The closest conceptual comparison is Nakamura et al. (2026), *Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems*. Colosseum formalizes multi-agent tasks using distributed constraint optimization and evaluates collusion through action-based measures, including regret relative to a cooperative optimum. It also compares behavioral evidence with communication-based assessments. An important finding is “collusion on paper”: agents may discuss collusive plans without carrying them out in their subsequent actions. Conversely, communication-level assessments alone may fail to capture some consequential behavioral deviations. [4]

2.4 Research Gap and Positioning

The cited studies establish that steganographic communication is a plausible risk, that LLM agents can exhibit strategic collusion, and that communication may diverge from subsequent behavior. The remaining practical challenge addressed here is how to interpret heterogeneous auditing signals without confusing a detector anomaly with proof of collusion.

This project addresses that challenge through three elements:

Evidence decomposition: distinguishing communication patterns, statistical dependence, behavioral consistency, and outcome-level consequences.

3. Methods

3.1 Research Approach

The methodology is guided by three research questions:

1. How do the reported detectors differ in their ability to distinguish honest behavior from collusive behavior?

2. What are the implications of their false-positive rates, true-positive rates, and ROC AUC values for practical auditing?
3. What additional evidence is needed to distinguish deliberate coordination from benign correlation or suspicious communication without corresponding actions?

This approach is motivated by research on steganographic collusion [1, 2] and Colosseum’s analysis of the relationship between communication and actions [3]. These studies motivate examining multiple forms of evidence without assuming that any individual signal proves deliberate coordination.

3.2 Dataset and Experimental Materials

The primary source is the *dataset_manifest.json* (https://github.com/poseidon-research/collusion_hackathon/blob/main/steganographic-collusion/results/dataset_manifest.json). The supplied materials cover pricing duopoly, first-price auctions, peer review, and continuous double-auction scenarios, alongside configurations involving LLM-based pricing.

The dataset manifests describe experimental variants involving:

- Payload sizes of 2 and 6 bits.
- Paraphrasing enabled at different levels or disabled.
- Different rate limits and budget constraints.
- Canary configurations.
- Permutation controls.
- Adaptive colluder and temperature configurations.

3.3 Models and Tools

The analysis uses the existing repository. The materials examined include:

- **Audit the Whisper repository:** the source of the existing auditing framework, detector descriptions, configurations, and experimental artifacts.
- **Detector-metrics summary:** the source of the reported false-positive rates (FPRs), true-positive rates (TPRs), ROC AUC values, and detector-combination results.
- **Welfare summary:** the source of the reported reward, welfare, accuracy, and fairness-gap measurements.
- **Prior research:** studies on steganographic collusion, strategic collusion, and auditing communication against subsequent actions.

The repository includes configurations for scripted and local-model experiments, as well as hosted LLM scenarios. However, this study does not claim to have independently executed these model

configurations. Model names appearing in the supplied results are treated as labels associated with the reported experimental configurations, not as evidence of new inference runs performed for this study.

3.4 Quantitative Metrics and Comparison

The analysis considers reported operating points at target FPRs of (10^{-2}) and (10^{-3}) . These thresholds are particularly relevant when incorrect accusations are costly or genuine collusion is rare.

The metrics are interpreted jointly because ranking performance does not necessarily imply acceptable performance at a stringent operating threshold. For example, cross-run mutual information reports an approximate ROC AUC of 0.885 but a TPR of 0.033 at the reported (10^{-3}) target. Permutation invariance reports an approximate AUC of 0.902, alongside a reported FPR of 0.998 at the same target.

Detector-combination methods—Holm–Bonferroni, Westfall–Young, Bonferroni, and martingale-based aggregation—are compared using their reported FPR and TPR values. This comparison is descriptive. The study does not independently optimize thresholds, calculate new confidence intervals, or conduct new significance tests.

3.5 Reproducibility

The analysis can be reproduced at the summary-analysis level by obtaining the *Audit the Whisper* repository, locating the reported detector and welfare metrics, and applying the comparison and evidence-assessment procedures described above.

Full reproduction of the underlying benchmark would require additional information and execution steps, including the exact repository revision, dependency versions, experimental configurations, transcript data, evaluation labels, threshold-calibration procedure, and metric implementations.

The project supports a reproducible interpretation of the supplied results, subject to their accuracy and completeness, but does not constitute an independent reproduction of the original experiments. Verifying the evaluation sample sizes, data independence, deduplication, threshold selection, and metric definitions remains necessary before drawing stronger conclusions about generalizable detector performance.

4. Results

The supplied detector metrics show substantial variation in performance.

Detector	FPR	TPR	ROC AUC
----------	-----	-----	---------

CASE baseline	0	1	1
Acceptance bias	0	1	1
Cross-run MI	0.001	0.033	0.885
Linguistic steganalysis	0.611	0	0.346
Permutation invariance	0.998	0.996	0.902
Watermark variance	0	0.176	0.588

The baselines and acceptance-bias detector report perfect separation, but these results require validation on independent data. Cross-run mutual information has a low TPR despite its relatively high AUC, while permutation invariance has an extremely high FPR. This demonstrates that AUC alone is insufficient to assess practical detector performance.

Overall, the supplied results highlight substantial differences between detectors and the importance of evaluating sensitivity alongside false-positive rates.

5. Discussion and Limitations

The results highlight the difficulty of reliably detecting steganographic collusion using a single metric. Some detectors report perfect performance, while others show low sensitivity or high false-positive rates. These differences suggest that detector choice and threshold calibration strongly influence audit outcomes

.

For AI safety, false positives can lead to incorrect accusations against agents, while false negatives may allow covert coordination to go undetected. Therefore, suspicious communication, correlated actions, and harmful outcomes should be assessed together rather than treated as definitive proof of collusion. The reported detector combinations show potentially useful sensitivity, but further validation is necessary before drawing conclusions about their reliability in real-world multi-agent systems.

Limitations

Limited threat coverage: The benchmark primarily evaluates text-based communication and static agent policies. Agents that learn online or use multimodal and tool-based channels may behave differently.

Dependence on welfare and fairness signals: Collusion that preserves welfare or avoids measurable fairness disparities may evade outcome-based detectors.

False-positive risks: Even calibrated detectors can flag rare honest behavior. Deployment therefore requires human review, especially before consequential decisions.

Future Work

Future work should implement multimodal and tool-mediated communication, test agents that adapt to monitoring, and improve detection of collusion that does not produce obvious welfare losses or fairness disparities. Additional work should investigate calibration with limited honest data, reduce dependence on sensitive subgroup labels, and incorporate human review of flagged transcripts to distinguish genuine collusion from legitimate coordination.

6. Conclusion

The results show that reported collusion detectors vary considerably in sensitivity and false-positive rates. While some methods report perfect separation, others miss collusive cases or flag large numbers of honest examples. Detector combinations show promising sensitivity in the supplied metrics, but their reliability depends on calibration and evaluation conditions. These findings highlight why no single detector score or harmful outcome should be treated as conclusive proof of collusion.

For AI safety, reliable auditing requires evaluating communication patterns alongside agent actions, outcomes, and plausible alternative explanations. Future work should test robustness against adaptive agents, unfamiliar communication channels, and collusion that does not produce obvious welfare or fairness changes. Such evaluations are essential for developing auditing systems that detect genuine coordination without generating excessive false accusations.

Code and Data

project doesn't involve code but what's taken into consideration:

- **Code repository:** <https://github.com/Vivekv81/poseidon-research-hackathon>

Author Contributions (optional)

Vivek Visave: Conceptualization, literature review, analysis and interpretation of reported detector metrics, evidence assessment, and writing of the manuscript.

References

[1]Motwani, S. R., Baranchuk, M., Strohmeier, M., Bolina, V., Torr, P. H. S., Hammond, L., & Schroeder de Witt, C. (2024). Secret Collusion among AI Agents: Multi-Agent Deception via Steganography. Advances in Neural Information Processing Systems (NeurIPS 2024).

- [2] Mathew, Y., Matthews, O., McCarthy, R., Velja, J., Schroeder de Witt, C., Cope, D., & Schoots, N. (2025). Hidden in Plain Text: Emergence & Mitigation of Steganographic Collusion in LLMs. Proceedings of IJCNLP-AACL 2025, 585–624.
- [3] Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). Algorithmic Collusion by Large Language Models. arXiv:2404.00806.
- [4] Nakamura, M., Kumar, A., Das, S., Abdelnabi, S., Mahmud, S. A., Fioretto, F., Zilberstein, S., & Bagdasarian, E. (2026). Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems. arXiv:2602.15198.
- [5] Poseidon Research. (n.d.). Audit the Whisper: Steganographic Collusion. GitHub repository. https://github.com/poseidon-research/collusion_hackathon/tree/main/steganographic-collusion

LLM Usage Statement

We used Claude to brainstorm approaches and help draft sections with structure. All results and claims were independently verified.