

Detecting Coordination Risk among AI Trading Agents

Integrating Communication and Behavioral Evidence for Financial Auditing

Author: XIE Tengda¹, Du Yunhao²

Institution: ¹The Hong Kong Polytechnic University, School of Accounting and Finance
²The Hong Kong Polytechnic University, Department of Data Science and Artificial Intelligence

Manuscript type: Research article

Abstract

This study develops and evaluates an audit-oriented framework for detecting coordination risk among artificial-intelligence trading agents. The framework integrates non-semantic communication structure with behavioral alignment residualised against observable market-wide and common-shock conditions. The experiment uses a rule-based synthetic data-generating process rather than a reinforcement-learning model: coordinated sets, activity periods, and labels are specified for evaluation. Four scenarios represent independent activity, explicit coordination, weak and intermittent coordination, and a no-coordination common-shock counterfactual.

Across 120 runs, five agents, and 500 periods per run, the integrated logistic model achieved a mean ROC AUC of 0.852 and an F1 score of 0.684 under repeated run-grouped holdouts. A common-shock ablation reduced ROC AUC to 0.737 and F1 to 0.584, showing that raw synchrony alone is not an adequate audit signal. A nested threshold procedure targeting a 10% G4 false-positive rate produced a mean out-of-sample G4 rate of 0.092. Semi-synthetic analysis using public daily Fama-French factor backgrounds yielded an ROC AUC of 0.844 and an F1 score of 0.697, but it does not constitute real-world validation. The proposed output is therefore a transparent referral score for human review rather than a finding of unlawful conduct.

Keywords: algorithmic auditing; coordination risk; financial markets; common shocks; calibration; synthetic simulation

Contents

1 Introduction.....	1
2 Literature Review	1
2.1 AI agents and multi-agent systems	1
2.2 Game theory and algorithmic collusion.....	2
2.3 Financial market surveillance	2
2.4 AI auditing and evidence integration.....	2
3 Theoretical Framework and Hypotheses	3
3.1 Theoretical framework.....	3
3.2 Research hypotheses	3
4 Methodology.....	3
4.1 Multi-agent financial market simulation	3
4.2 Communication evidence construction.....	4
4.3 Behavioral evidence construction.....	4
4.4 Evidence integration model	4
4.5 Model evaluation	5
5 Experimental Results	5
5.1 Simulation results	5
5.2 Detection performance comparison	6
5.3 False-positive analysis.....	7
5.4 Robustness tests	7
6 Discussion.....	8
6.1 Interpretation of coordination signals.....	8
6.2 Implications for financial auditing.....	8
6.3 Limitations	9
7 Conclusion	9
References.....	10
Appendix A Supplementary Materials	10

1 Introduction

Artificial intelligence is increasingly used in financial markets to process information and determine order submission, pricing, and execution. These systems operate at speeds and scales that complicate conventional, after-the-fact review. Correlated trades are not inherently improper. Auditors therefore need a disciplined way to distinguish potentially informative patterns from the co-movement that arises in a common market environment. Research on repeated pricing games shows that autonomous learning systems can sustain supra-competitive outcomes under particular experimental conditions, while legal scholarship has stressed the difficulty of assigning liability when coordination is not an explicit human agreement (Calvano, Calzolari, Denicolò, & Pastorello, 2020; Harrington, 2018). These studies motivate an auditing question narrower than proving collusion: can an evidence system prioritize agent-period observations for review without treating similarity as proof? This distinction is especially important in financial markets, where prices and order decisions may respond simultaneously to public news, liquidity conditions, and common risk shocks.

Existing work provides conceptual and experimental grounds for monitoring algorithmic coordination but leaves an operational gap. Communication traces may be unavailable or may capture only routine activity. Conversely, behavioral screens based on matched timing or prices may flag agents that independently respond to a common signal. A screening system must therefore integrate multiple sources of evidence while explicitly assessing legitimate common-cause explanations. Cartel-screening research has long cautioned that unusual variance or parallel conduct is a lead for investigation rather than a self-sufficient finding (Abrantes-Metz et al., 2006; Harrington & Imhof, 2022). Evaluation presents a second methodological gap. If observations from a single simulated market run are split between training and test data, closely related windows may appear in both partitions. Measured accuracy may then reflect reuse of the same market path rather than generalization to an unseen run. The present study addresses this risk by holding out entire runs, reporting uncertainty across repeated splits, and separating threshold selection from final testing.

This study addresses three questions. First, do non-semantic communication features and market-behavior features provide complementary information for detecting pre-specified coordination in a controlled market simulation? Second, does accounting for peer alignment under market-wide and common-shock conditions reduce false referrals in an uncoordinated common-shock scenario? Third, how stable are the results across model classes, decision thresholds, and perturbations to the data-generating mechanism?

The contribution is methodological rather than evidentiary with respect to any real firm. The paper introduces a transparent synthetic audit testbed with randomized coordinating sets, overlapping benign and coordinated behavior, and no identity or message-content fields. It also uses group-based evaluation, reports calibration, and includes a semi-synthetic robustness exercise. Rule-specified agents are not described as having learned to collude. The intended use is to clarify how an audit workflow can generate accountable referrals for human examination.

2 Literature Review

2.1 AI agents and multi-agent systems

Multi-agent systems provide a useful setting for studying strategic dependence because each agent can respond both to a shared environment and to other participants. In markets, however, apparently coordinated behavior can have benign causes, including common information, similar risk controls,

identical clocks, and responses to the same volatility regime. An audit design must therefore distinguish observed outcome patterns from claims about the mechanism that generated them. In a repeated-pricing experiment, Q-learning agents can reach collusive price outcomes (Calvano, Calzolari, Denicolò, & Pastorello, 2020). Subsequent policy discussion has highlighted the consumer risks associated with algorithmic pricing (Calvano, Calzolari, Denicolò, Harrington, et al., 2020). This literature establishes why algorithmic coordination warrants scrutiny, but this study does not reproduce its learning environments or objectives. Instead, it fixes a known data-generating mechanism so that false positives, feature leakage, and uncertainty can be evaluated against known ground truth.

2.2 Game theory and algorithmic collusion

Repeated-game theory explains why agents may condition current actions on expected future responses. Communication can facilitate monitoring and shared expectations, whereas tacit coordination may appear only as persistent strategic alignment. For auditing, neither type of evidence is decisive on its own: messages may reflect routine activity, while aligned actions may be rational responses to a common state. Harrington and Skrzypacz (2011) show how communication and information reporting can support collusive arrangements under imperfect monitoring, while Harrington (2018) discusses the legal challenges posed by autonomous artificial agents. Together, these perspectives point to an evidence-integration problem rather than a single-feature classification task. A credible screen should assess whether interaction structure and residual behavioral dependence remain unusual after observable shared conditions have been considered.

2.3 Financial market surveillance

Financial surveillance typically combines transactional patterns, contextual information, and analyst judgement. This combination is difficult in high-frequency settings because a single market shock can produce similar actions across many systems. A model that records raw peer alignment without accounting for the shared background may therefore learn to flag predictable effects of a shock. The cartel-screening literature treats statistical screens as tools for prioritizing cases, not as replacements for economic evidence or investigation (Harrington, 2018). This paper adopts the same conservative stance. It evaluates a risk score and its false-referral behavior rather than offering a legal test for collusion. Accordingly, the common-shock scenario is included as a required counterfactual, not as an afterthought.

2.4 AI auditing and evidence integration

Auditability requires more than discriminative performance. Users need to know which inputs a model observes, which it excludes, how its probability estimates behave, and how alerts are selected (Rudin, 2019). Communication content is deliberately excluded from the detector in this experiment; it retains only simple interaction rates and partner concentration. This design prevents a classifier from succeeding by matching synthetic keywords rather than detecting structural evidence. Probability calibration is relevant when a classifier is used to rank referrals because a score is more useful when its stated risk is aligned with observed frequency (Brier, 1950; Pakdaman Naeini et al., 2015). Accordingly, the framework reports the Brier score and expected calibration error alongside ROC AUC and F1. Its model comparison is deliberately plural: a linear model makes feature directions readable, a rule-based screen provides an auditable reference, and a calibrated nonlinear model may be preferable when probability quality is the principal operational requirement.

3 Theoretical Framework and Hypotheses

3.1 Theoretical framework

The framework uses two evidence channels. The communication channel measures messages sent and received per period, together with the concentration of an agent’s partners. The behavioral channel measures action imbalance and residual peer alignment. To construct the latter, each agent’s discretized action and the mean action of its peers are separately residualised on the market signal, common shock, and fixed polynomial terms. Their correlation is then estimated within a window. The design therefore asks whether agents remain aligned after observed shared movements have been removed.

The causal logic is deliberately modest. Communication structure and residual alignment can raise the priority of a window for review, but neither establishes intent, agreement, economic harm, or illegality. G4 is an explicit no-coordination counterfactual: its agents share a strong external shock but follow no coordinated rule. A valid audit model should not reward G4 merely because independent agents move together.

3.2 Research hypotheses

H1: Communication-structure indicators distinguish explicitly coordinated agents from independently operating agents in the synthetic market.

H2: Behavioral indicators based on peer alignment after controlling for observed market and common-shock conditions distinguish coordinated from independent activity. Unadjusted alignment is more vulnerable to the common-shock counterfactual.

H3: The integrated feature set achieves higher average discriminative performance than either the communication-only or behavioral-only feature set.

H4: Controlling peer alignment for observed market and common-shock conditions reduces false-positive referrals in G4 relative to an otherwise equivalent uncontrolled ablation.

The first three hypotheses concern predictive separation under the specified mechanism. H4 is falsification-oriented: it tests whether the detector resists a plausible alternative explanation for synchrony. This test does not claim that all real-world confounders are represented.

4 Methodology

4.1 Multi-agent financial market simulation

The study uses a reproducible, rule-based data-generating process rather than an emergent reinforcement-learning experiment. Each run contains five agents and 500 periods. Each of the four scenarios is simulated 30 times: G1 comprises independent agents; G2 introduces explicit coordination; G3 represents weak, intermittent coordination; and G4 introduces a common-shock condition without coordination. The design yields 120 runs, 300,000 agent-period trade records, and 6,000 agent-window observations, with five agents observed across ten 50-period windows in each run.

In G2 and G3, the coordinating set is newly sampled in every run and contains two to four members. Only members of the sampled set receive a positive label. In G2, a shared strategy increment occurs with probability 0.48 and uses a strength parameter of 1.10; in G3, the corresponding values are 0.42 and 0.90. These mechanisms are deliberately overlapping rather than perfectly separable. All scenarios contain neutral operational messages selected from the same small vocabulary, and G2 adds modest within-set traffic during active coordination. Message text and topic are retained for audit traceability but are not used in modelling.

G4 raises the common-shock weight from 0.45 to 0.78 while setting the coordinating set to be empty. It is a controlled negative condition in which cross-agent similarity is expected even though coordination is absent by construction. Prices are updated using aggregate demand, the market signal, the common shock, and residual noise. This structure is sufficient for evaluating the audit logic, but it does not reproduce an order book, a broker network, or actual market microstructure.

Table 4.1. Experimental design and data scope.

Source: Authors' compilation from `simulation_config.json` and `run_manifest.csv` in the reproducibility package.

Component	Specification	Purpose
Scenarios	G1-G4	Independent, explicit, weak/intermittent, and common-shock mechanisms
Runs	30 per scenario; 120 total	Independent market paths and random coordinating sets
Run structure	5 agents x 500 periods	300,000 agent-period trade records
Window unit	50 periods; 6,000 agent-windows	Detection and grouped evaluation unit
Labels	Only sampled G2/G3 members	Prevents fixed-agent or fixed-scenario labels

4.2 Communication evidence construction

Communication features are calculated within each 50-period agent window. The message-sent rate is the number of messages originating from the focal agent divided by the window length; the message-received rate is defined analogously. Partner concentration is the largest share of interactions involving a single partner. Together, these features capture interaction intensity and concentration without relying on message semantics.

The exclusion set is pre-specified. Scenario, run identifier, agent identifier, window start, label, message text, message topic, and private signal are excluded from every model. This decision prevents direct leakage from group labels, identity patterns, or coordination wording. It also makes the study intentionally conservative: the model cannot succeed merely because an explicit synthetic message is unique to colluding agents.

4.3 Behavioral evidence construction

The behavioral feature family comprises absolute action imbalance, naïve peer alignment, and shock-controlled peer alignment. After applying a fixed thresholding rule, actions take values of -1, 0, or 1. The primary feature is the correlation between residuals from two fixed regressions. The focal action and the average peer action are each regressed on the market signal, common shock, squared and cubic terms, and their interaction. Naïve peer alignment is retained only for the ablation.

This residualisation does not establish that all confounding has been removed. Instead, it operationalizes a specific, observable alternative explanation and allows the analysis to quantify what changes when the control is omitted. The G4 benchmark makes this test meaningful because any positive alert in that scenario is necessarily a false referral relative to the simulator label.

4.4 Evidence integration model

The primary model is a class-balanced logistic regression fitted to standardized predictors. Three reduced specifications are also fitted: a communication-only model, a behavioral-only model, and an integrated uncontrolled ablation that uses naïve rather than shock-controlled alignment. A subsequent benchmark comparison evaluates a random forest, histogram gradient boosting, a calibrated radial-basis support vector machine, and a transparent rule baseline under the same input protocol and grouped splits.

This choice does not assume that logistic regression is universally optimal. Instead, it provides an interpretable reference against which discrimination and calibration can be compared. The rule baseline is intentionally simple because a high-performing but opaque model is not automatically preferable in an audit setting. All features are measured at the agent-window level, whereas validation remains grouped by complete market run.

4.5 Model evaluation

Twenty GroupShuffleSplit repetitions, using seeds 100–119, hold out 30% of complete runs in each repetition. No run appears in both the training and test partitions. Mean ROC AUC, precision–recall AUC, F1, precision, recall, Brier score, and expected calibration error are reported with normal-approximation 95% confidence intervals across the repeated splits. These intervals describe repeated resampling of the simulated runs, not sampling uncertainty from a real market population.

Nested group-held-out evaluation is used for the operational threshold analysis. A threshold is selected on an inner validation partition to target a G4 false-positive rate no greater than 10%, and its performance is measured only on the outer test partition. This separation avoids selecting a threshold using the same observations later used to report performance (Varma & Simon, 2006).

5 Experimental Results

5.1 Simulation results

Table 5.1. Performance under repeated grouped holdouts. Values are means; brackets report 95% confidence intervals where available.

Source: Authors' analysis of performance_ci.csv and baseline_performance_ci.csv.

Model	ROC AUC	F1	Brier	ECE
Communication only	0.738 [0.722, 0.754]	0.591 [0.575, 0.608]	0.190	0.159
Behavior only	0.793 [0.785, 0.800]	0.621 [0.606, 0.635]	0.187	0.134
Integrated logistic	0.852 [0.844, 0.860]	0.684 [0.672, 0.697]	0.152	0.118
Random forest	0.847 [0.839, 0.855]	0.687 [0.675, 0.699]	0.146	0.091
Calibrated RBF SVM	0.844 [0.835, 0.853]	0.682 [0.668, 0.696]	0.136	0.049
Auditable rule baseline	0.778	0.648	N/A	N/A

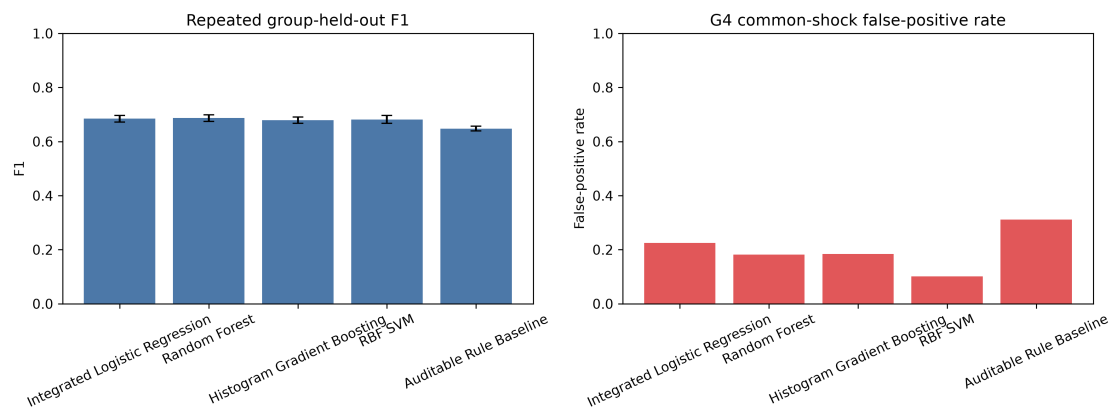


Figure 5.1. ROC AUC, F1, Brier score, and ECE across candidate models.

Source: Authors' analysis of baseline_model_comparison.png.

The simulations produced the planned 120 market runs, 300,000 trade rows, and approximately 13,700 neutral communication records. Randomization varied both the identities and the number of coordinated agents across G2 and G3 runs. Consequently, an agent identifier cannot serve as a stable proxy for the label. The use of shared neutral message phrases across scenarios also rules out keyword-based separation.

At the default probability threshold of 0.50, the controlled integrated model detected 96.6% of G2 observations and 51.7% of G3 observations. The lower G3 value is expected under a mechanism with weaker, intermittent coordination. In G4, where the true coordination label is always zero, the controlled model produced a 22.4% false-positive rate at the default threshold. This rate is too high for unattended use and motivates the threshold-constrained analysis below.

5.2 Detection performance comparison

The integrated controlled logistic model achieved a mean ROC AUC of 0.852 (95% CI 0.844–0.860) and an F1 score of 0.684 (0.672–0.697). The communication-only and behaviour-only specifications achieved ROC AUC values of 0.738 and 0.793, respectively. This pattern supports H3 within this experiment: the two feature families are complementary, but neither constitutes legal or causal proof of coordination.

Among the alternative classifiers, logistic regression achieved the highest mean ROC AUC. Random forest produced a marginally higher F1 (0.687 versus 0.684), but the paired difference was 0.0023, with a 95% interval from -0.0027 to 0.0073 and a sign-flip p-value of 0.380. The two models are therefore not distinguishable on F1 across these repetitions. The calibrated RBF SVM achieved the lowest Brier score (0.136) and ECE (0.049), at the cost of lower G3 detection. A single headline metric would obscure this trade-off.

Table 5.2. Validation-selected thresholds and outer-test results.

Source: Authors' analysis of `threshold_performance_ci.csv`.

Model	Threshold	Precision	Recall	G4 FPR
Integrated logistic	0.668 [0.647, 0.688]	0.787	0.596	0.092 [0.077, 0.107]
Random forest	0.622 [0.601, 0.644]	0.796	0.585	0.087 [0.073, 0.101]
Calibrated RBF SVM	0.505 [0.470, 0.540]	0.791	0.594	0.088 [0.075, 0.101]

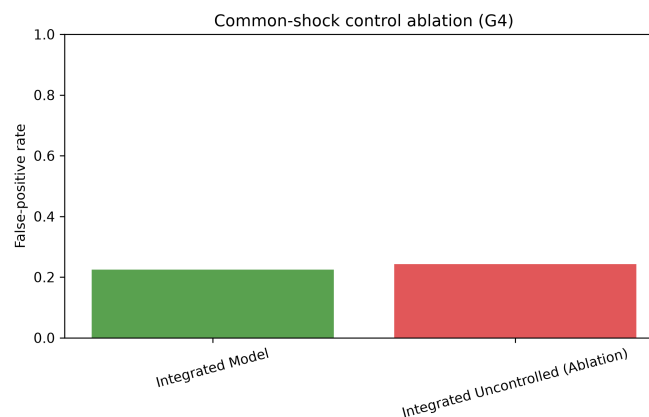


Figure 5.2. G4 false-positive rate before and after common-shock control.

Source: Authors' analysis of `common_shock_ablation.png`.

5.3 False-positive analysis

The common-shock ablation supports H4. Replacing residualised peer alignment with naïve alignment reduced mean ROC AUC from 0.852 to 0.737 and F1 from 0.684 to 0.584. It also worsened the Brier score from 0.152 to 0.190 and ECE from 0.118 to 0.159. At the default threshold, the G4 false-positive rate was 0.224 for the controlled model and 0.243 for the uncontrolled model. The absolute G4 reduction is modest, so the result should be interpreted as evidence that the control matters rather than as a complete solution to false referrals.

Nested evaluation changes the operating point. For logistic regression, the validation-selected mean threshold was 0.668, and outer-test recall was 0.596. Its mean G4 false-positive rate was 0.092 (95% CI 0.077–0.107). Random forest and calibrated RBF SVM recorded G4 rates of 0.087 and 0.088, respectively. Because some interval upper bounds meet or exceed the 10% target, these results support average control in the simulated design, not an all-market guarantee.

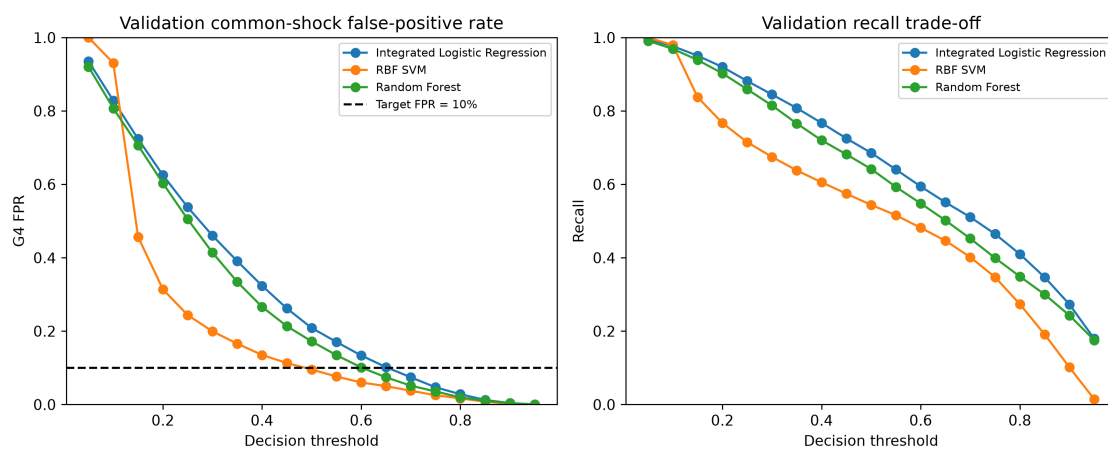


Figure 5.3. Validation trade-off between G4 false-positive rate and recall.

Source: Authors' analysis of `threshold_tradeoff.png`.

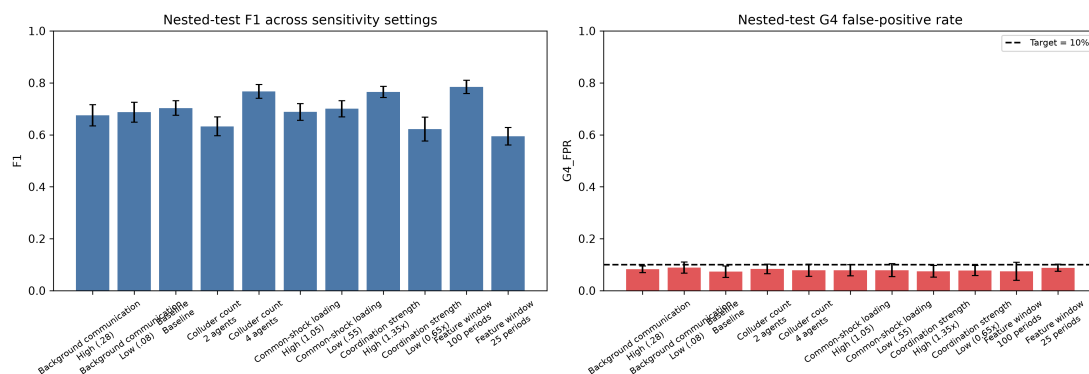


Figure 5.4. F1 and G4 false-positive rate under sensitivity conditions.

Source: Authors' analysis of `sensitivity_robustness.png`.

5.4 Robustness tests

One-factor-at-a-time sensitivity analysis identifies the conditions under which the screen becomes less useful. With two coordinating agents, F1 was 0.632 (0.596–0.669). At low coordination strength, defined as 0.65 times the baseline, F1 was 0.622 (0.576–0.668). A 25-period window reduced F1 to 0.594 (0.561–0.628). Larger coordinating sets, stronger coordination, and 100-period windows

improved separation. Common-shock loading had a smaller effect on F1, although some G4 false-positive confidence-interval upper bounds reached or exceeded 10%.

A semi-synthetic robustness exercise used public daily Fama-French factors between 1 January 2010 and 29 December 2023. Mkt-RF supplied the market background, and SMB supplied the common risk disturbance. The fixed sample contains 3,522 daily observations. The underlying agents, communication records, and labels still follow the rule-based mechanism; the real factor history improves background realism but does not turn the exercise into external validation.

The semi-synthetic analysis recorded an ROC AUC of 0.844 (0.825–0.863), an F1 score of 0.697 (0.664–0.730), and a G4 false-positive rate of 0.075 (0.060–0.090). The Fama-French three-factor framework and the public daily factor archive are described by Fama and French (1993) and the Kenneth R. French Data Library.

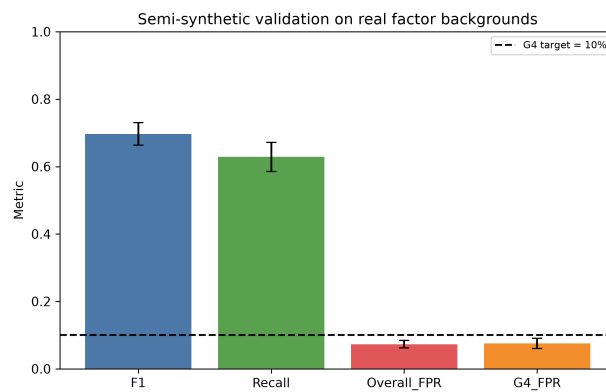


Figure 5.5. Performance and G4 false-positive rate in semi-synthetic validation.

Source: Authors' analysis of *semi_synthetic_validation.png*; market background from the Kenneth R. French Data Library.

6 Discussion

6.1 Interpretation of coordination signals

The results support a cautious interpretation. Under the known mechanism, communication structure and shock-controlled behavior are more informative when considered together than separately. The integrated score is an evidence-prioritization device: it connects ordinary interaction patterns with residual behavioral dependence after market-wide and common-shock movements have been accounted for. It does not convert a probability into proof of intent or agreement.

The G4 result is particularly instructive. Without a common-cause control, a screen that responds to synchrony would mistake a constructed market shock for coordination. The residualisation step materially improved pooled discrimination but reduced G4 false referrals only modestly at the default threshold. This mixed result is informative: it shows that a single control is not a cure-all and that the false-referral budget must be defined and tested separately.

6.2 Implications for financial auditing

The appropriate operational output is a referral for human review. A reasonable workflow would select a threshold using held-out validation data, retain the associated feature values and market context, and require an analyst to examine order-book records, permitted communication records, public announcements, and plausible common causes. A high score should trigger a request for explanation, not an automated accusation.

Model choice should reflect the workflow. Logistic regression provides an easily inspected reference model; random forest achieved statistically indistinguishable F1 in this experiment; and calibrated SVM provided the best probability quality. A team that allocates scarce review capacity by risk rank may value calibration more than a small difference in ROC performance. A team that must preserve a concise evidentiary trail may choose an interpretable model even when another model appears marginally stronger.

The threshold is also non-portable. Changes in asset class, trading frequency, volatility, communication regime, or data availability can alter both feature distributions and review costs. Recalibration, drift monitoring, versioned feature definitions, and records of human override decisions should precede deployment. These controls are necessary because a research score cannot substitute for legal assessment or market-specific economic analysis.

6.3 Limitations

The principal limitation concerns label provenance. Although identities, group size, and activity periods vary randomly, coordinated behaviour is specified by the researcher. A classifier may therefore learn regularities of the chosen mechanism rather than patterns that generalise to every form of real-world coordination. The semi-synthetic extension changes the background series but retains synthetic agents and labels; it is not an external validation study.

The simulated market is also deliberately compact. It does not include order-book dynamics, latency asymmetry, changing agent populations, public disclosure processes, institutional clocks, or legal evidence rules. Fixed 50-period windows may miss slowly forming changes in strategy, and the exclusion of message content means that the model does not evaluate semantic evidence. Finally, repeated grouped splits quantify variability within the simulated corpus; they do not replace independent replication across markets and periods.

7 Conclusion

This paper presents a reproducible audit experiment for ranking coordination risk among AI trading agents. Under repeated run-grouped evaluation, the integrated model achieved a mean ROC AUC of 0.852 and an F1 score of 0.684. Controlling peer alignment for market-wide and common-shock conditions improved pooled performance relative to an uncontrolled ablation, while nested threshold selection brought the mean G4 false-positive rate close to the pre-specified 10% target. The sensitivity analysis also identified difficult settings: weak coordination, small coordinating groups, and short observation windows.

The study contributes a conservative design for experimental AI auditing: randomised coordinating sets, neutral communications, no identity or text leakage, grouped splitting, calibration measures, a common-shock counterfactual, constrained thresholding, and reproducibility artefacts. These choices narrow the claims but make them more credible. The experiment shows how a risk-screening workflow can be evaluated under known ground truth; it does not show that real algorithms spontaneously collude.

The next step is external evidence, including verified historical cases, carefully governed weak labels, or credible multi-market natural experiments. Time-out validation, dynamic network features, event studies, decision-curve analysis, and pre-specified review costs would make the framework more relevant to operational surveillance. Until such validation is available, the responsible role of the system is to focus expert attention, preserve an audit trail, and remain open to correction.

References

- Abrantes-Metz, R. M., Froeb, L. M., Geweke, J., & Taylor, C. T. (2006). A variance screen for collusion. *International Journal of Industrial Organization*, 24(3), 467–486. <https://doi.org/10.1016/j.ijindorg.2005.10.003>
- Brier, G. W. (1950). VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. *Monthly Weather Review*, 78(1), 1–3. [https://doi.org/10.1175/1520-0493\(1950\)078%3C0001:VOFEIT%3E2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2)
- Calvano, E., Calzolari, G., Denicolò, V., Harrington, J. E., & Pastorello, S. (2020). Protecting consumers from collusive prices due to AI. *Science*, 370(6520), 1040–1042. <https://doi.org/10.1126/science.abe3796>
- Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2020). *Artificial intelligence, algorithmic pricing and collusion*. <https://doi.org/10.1257/aer.20190623>
- Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3–56. [https://doi.org/10.1016/0304-405X\(93\)90023-5](https://doi.org/10.1016/0304-405X(93)90023-5)
- Harrington, J. E. (2018). DEVELOPING COMPETITION LAW FOR COLLUSION BY AUTONOMOUS ARTIFICIAL AGENTS†. *Journal of Competition Law & Economics*, 14(3), 331–363. <https://doi.org/10.1093/joclec/nhy016>
- Harrington, J. E., & Imhof, D. (2022). *Cartel Screening and Machine Learning*.
- Harrington, J. E., & Skrzypacz, A. (2011). Private Monitoring and Communication in Cartels: Explaining Recent Collusive Practices. *American Economic Review*, 101(6), 2425–2449. <https://doi.org/10.1257/aer.101.6.2425>
- Pakdaman Naeini, M., Cooper, G., & Hauskrecht, M. (2015). Obtaining Well Calibrated Probabilities Using Bayesian Binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1). <https://doi.org/10.1609/aaai.v29i1.9602>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1), 91. <https://doi.org/10.1186/1471-2105-7-91>

Appendix A Supplementary Materials

The accompanying supplementary package contains: (i) the exact Python scripts used to generate the reported synthetic experiment; (ii) the reproducibility configuration, feature protocol, split manifests, and result tables; (iii) the generated figures; (iv) the public Fama-French input archive and provenance record for the semi-synthetic extension; and (v) a README with execution instructions and scope notes. The raw outputs remain synthetic or semi-synthetic as stated throughout this manuscript.