

Steganalysis of Adaptive Covert Collusion in Tool-Using Agent Populations: A Black-Box, Cross-Principal Approach

Mohamed Chahine Ghanem, *Member, IEEE*

Abstract—Tool-using agents built on large language models (LLMs) are increasingly deployed not by a single operator but by many, side by side on shared infrastructure. This creates a population-level risk that single-agent safeguards miss: a handful of agents can quietly coordinate—rigging a market, boosting one another in a review process, or timing a joint data grab—while each one looks perfectly well-behaved. The difficulty is that the organisations running these agents cannot see inside one another’s models, so any realistic detector must work from behaviour alone: black-box, trace-only, and often with only partial visibility. We treat covert coordination as an information-hiding problem and build a black-box steganalysis detector that combines cross-run mutual-information estimation, permutation tests, distributional-shift statistics, and timing and tool-call side channels, all calibrated to a fixed false-positive budget. Our central move is to stop testing against a single fixed code: we pit the detector against an adversary that continually rewrites its encoding to slip past whatever the detector has learned, and we run this red-versus-blue contest in tool-using, memory-carrying environments rather than toy games. Capacity theory then tells us what to expect—a *detection-capacity frontier*, a covert bit-rate below which black-box detection is provably no better than chance. We set out an experiment to map this frontier, report clearly labelled placeholder results pending measurement, and flag a practical evasion—spreading a payload across sessions—that current methods largely miss.

Index Terms—AI safety, multi-agent systems, large language models, steganalysis, covert channels, collusion detection, information hiding, adversarial machine learning.

Impact Statement

Autonomous agents built on large language models are being deployed by many independent organisations onto shared infrastructure, where they increasingly interact without a single overseer. Safeguards designed for one agent do not address *population-level* risks: a small colluding subset can coordinate covertly to distort markets, audits, or shared decisions while appearing to behave normally. This work treats such collusion as an information-hiding problem and provides (i) a detection method that assumes no access to the other parties’ models—the only realistic assumption across organisational boundaries—and (ii) an honest, capacity-theoretic account of *when* such detection must fail. Making both the capability and its limits explicit supports auditing, procurement, and standards for multi-agent deployments, and helps regulators and operators calibrate expectations rather than over-trust black-box monitoring. To limit dual-use, we release detection tooling and benchmarks openly while withholding the strongest attack artifacts, consistent with responsible-disclosure norms.

I. INTRODUCTION

Manuscript submitted August 5, 2026. (*Corresponding author: M. C. Ghanem.*)

M. C. Ghanem is with Keele University, Newcastle-Under-Lyme, U.K. (e-mail: m.ghanem@keele.ac.uk).

FOR most of their short history, LLM agents have worked for a single boss: one operator sets them up, points them at a task, and reads the result. That is changing. Agents from different organisations are beginning to meet on shared infrastructure—marketplaces, review pipelines, shared tool servers—and to interact with no single party in charge [1], [2]. The upside is new forms of coordination; the downside is a class of failures that safeguards built for one model at a time simply cannot see [1]. Chief among them is *covert collusion*: a subset of agents quietly steering toward an outcome that hurts the wider system—rigging a procurement auction, forming a reciprocal “boost ring” in automated review, or timing a joint data-exfiltration trigger—while each agent, taken on its own, looks entirely above board.

As a concrete example, consider a procurement platform where supplier-agents from competing firms submit bids. If a subset—each for a different principal—covertly agree on who wins each round and at what price, they extract supra-competitive rents while every bid looks defensible. The coordination needs no explicit message: a shared convention over bid phrasing, timing, or query order carries the few bits required. No agent breaks a local policy and no single transcript is incriminating, yet the operator, unable to inspect the suppliers’ models, sees only well-formed bids—the population-level failure that motivates trace-only, cross-principal detection.

Recent work has shown that LLM agents can, in principle, communicate *steganographically*: embedding a hidden payload inside otherwise natural outputs so that coordination survives casual inspection and even paraphrasing defences [3], [4]. The threat is not hypothetical folklore; it follows directly from classical results on subliminal channels [5] and the information-theoretic limits of steganography [6], [7], now instantiated by models fluent enough to realise those channels in natural language and, with tools, to *construct* new ones at run time.

Detection is hard in a real deployment for a structural reason, not because the adversary is especially clever: **no principal can look inside another principal’s models**. A monitor has no access to the weights, activations, or private prompts of an agent that some other organisation runs. All it sees are the traces the agents leave behind—their messages, tool calls, and timing—and often not even all of those. This black-box, cross-principal, partially observable setting is the one we target, and it rules out any method that quietly assumes a single overseer with full visibility.

We frame the problem as *steganalysis*: detecting a hidden

channel from statistical properties of the carrier, without decoding it [6], [8], [9]. This is not generic anomaly detection, which flags any unusual behaviour; steganalysis assumes a deliberately hidden payload and an adversary actively working to look normal [5], and that assumption shapes both the detector and the evaluation. Two commitments set our approach apart from prior collusion detectors. The detector is *encoding-agnostic*—it never assumes a known covert scheme, because a competent adversary will not reuse one—and it is measured against an *adaptive* adversary that co-evolves its encoding to defeat whatever the detector has learned, so the headline evidence is a red-versus-blue trajectory rather than a single static score. The study runs in tool-using, persistent-memory environments and adds a cross-session evaluation axis that existing benchmarks, to our knowledge, do not measure. To our knowledge this is the first black-box, cross-principal steganalysis framework for tool-using agent populations that couples a capacity-theoretic detectability bound with an adaptive, co-evolving adversary.

Contributions.

- 1) We formulate cross-principal covert collusion among tool-using agents as black-box steganalysis under partial observability, with an explicit threat model (Section III).
- 2) We design an encoding-agnostic meta-detector that fuses cross-run mutual information, permutation-invariance tests, distributional-shift statistics, and timing/tool-call side channels, calibrated to a fixed false-positive budget (Section III).
- 3) We introduce an adaptive, co-evolving adversary and a red-versus-blue evaluation methodology that reports co-evolution curves rather than static AUROC (Section III).
- 4) We specify tool-using, persistent-memory testbeds that extend stylised collusion benchmarks, and a cross-session evaluation axis (Section V).
- 5) We give a capacity-theoretic analysis predicting a *detection-capacity frontier* and an experimental protocol to chart it, with illustrative placeholder results (Sections IV–V).
- 6) We discuss limits, evasions, and responsible-disclosure considerations (Section VI).

II. BACKGROUND

A. Multi-Agent LLM Systems

Agent-based computing has a long lineage in distributed AI [10]. LLM-based multi-agent systems revive these ideas at a new capability level: agents plan, use tools, hold persistent memory, and communicate in natural language [2], [11]. As deployments scale across organisational boundaries, system-level behaviour—rather than any single agent—becomes the relevant unit of safety analysis [1]. Our work inherits this framing and asks a security question about it.

B. Secret Collusion and Steganographic Coordination

Motwani *et al.* [3] formalised *secret collusion* among LLM agents, deriving results on the capacity and incentives of models to communicate steganographically and on the limits of

mitigation. Mathew *et al.* [4] showed that steganographic collusion can *emerge* from misspecified rewards and can survive paraphrasing, undercutting the assumption that a paraphrasing “firewall” is sufficient. These peer-reviewed results establish that the capability exists; the open problem is *detection* under realistic access constraints. Neither result gives a monitor that works without model access, and both assume settings simpler than tool-using, multi-session deployments. Our contribution is the missing detection side of this story, under the access constraints that actually hold across organisations.

C. Foundations of Steganography and Covert Channels

The subliminal channel [5] and Cachin’s information-theoretic model of steganography [6] give the vocabulary we use throughout: a hidden channel is *secure* when the stego distribution is statistically close to the innocent “cover” distribution, and *detectable* to the extent that it is not. Moulin and O’Sullivan [7] characterise information-hiding capacity under distortion constraints, quantifying the rate–stealth trade-off that our analysis exploits. Modern linguistic steganography and steganalysis [8], [9] and covert-channel taxonomies [12] supply concrete detector families, several of which we adapt to the agent setting. Two properties transfer directly to agents: detectability is governed by a divergence between innocent and coordinated behaviour, and that divergence can be reduced only at the cost of covert throughput. The difference in the agent setting is that the “carrier” is not a fixed medium but an open-ended space of messages, timings, and tool actions a capable model can reshape at run time.

D. Watermarking, Detectors, and Their Robustness

Watermarking of model outputs [13] and the broader literature on text watermarking [14] are relevant both as an attribution primitive and as a source of distributional features for detection. Their known fragility is instructive: paraphrase-style attacks can evade AI-text detectors [15], a cautionary result we take seriously when reasoning about an adaptive adversary. Cryptographic proofs of inference such as zkLLM [16] point toward verifiable attribution, but remain heavyweight and do not address covert coordination directly. We use these results in two ways: watermark-style statistics feed the detector’s distributional features, and opt-in attribution is one of the layered mitigations we recommend for deployments where principals cooperate.

E. Emergent Communication

That agents can invent communication protocols is well established in multi-agent reinforcement learning [17], [18]. Covert coordination is, in effect, emergent communication constrained to be statistically inconspicuous—a lens that motivates encoding-agnostic detection, since the protocols agents invent are often opaque to outsiders and sensitive to training pressures, so no fixed decoder will keep pace with a learning population.

F. Multi-Agent Safety, Control, and Infrastructure

AI control studies protocols that preserve safety even when an agent actively subverts oversight [19], but is developed largely for a *single* untrusted agent; extending it to colluding *groups* is open. Governance-oriented work on agent visibility is peer-reviewed [20], while several proposals for agent infrastructure and identity remain preprints: “Infrastructure for AI Agents” [21] and “IDs for AI Systems” [22] (both preprints), the “Multi-Agent Risks from Advanced AI” technical report [1], and “Open Problems in Cooperative AI” [23] (workshop report). We cite these for motivation and pair them with the peer-reviewed anchors above. Collectively, this line of work argues that oversight designed for a single model is insufficient once many agents interact, but it stops short of an operational, principal-agnostic detector—the gap this paper targets.

G. Multi-Agent Attacks and the Detection Frontier

Attacks can propagate through agent populations—for example, infectious jailbreaks [24] and LLM-to-LLM prompt injection within multi-agent systems [25]. Protocol-level analyses of emerging agent standards suggest new tool-integration surfaces amplify such risks; the most detailed Model Context Protocol study to date is a preprint [26], which we cite as such alongside the peer-reviewed workshop result [25]. Finally, the collusion-*detection* frontier is presently defined by three preprints, each with a limitation this paper is designed to address: “Audit the Whisper” [27] is black-box and trace-only but adapts only within a fixed scheme family and uses tool-free environments; “Colosseum” [28] is black-box but behavioural (regret-based) and not evaluated against a co-evolving encoder; and the multi-agent-interpretability approach behind NARCBench [29] attains strong results but *requires white-box activation access*, which is infeasible across principals. We treat all three as preprints and position our contribution precisely at the gap they leave: black-box detection against an *adaptive* adversary in *tool-using*, *persistent-memory* settings. Table I summarises this positioning across the axes that matter for realistic multi-principal deployment.

III. METHODOLOGY

A. Threat Model and Problem Formulation

Consider a population of N agents, operated by two or more principals, interacting over T rounds on shared infrastructure. A colluding subset of size k ($1 < k \leq N$) seeks to coordinate on a private *payload* m (e.g., which agent should win a given round) while a monitor attempts to decide whether collusion is present. The monitor observes a trace

$$\mathcal{O} = \{(u_i^{(t)}, c_i^{(t)}, \tau_i^{(t)})\}_{i \in S, t \leq T},$$

where $u_i^{(t)}$ is agent i ’s natural-language output at round t , $c_i^{(t)}$ its tool calls, and $\tau_i^{(t)}$ timing/rate metadata, restricted to an observed subset $S \subseteq \{1, \dots, N\}$ (partial observability). Critically, the monitor has *no* access to model weights, activations, or private prompts of agents it does not operate. The monitor

is calibrated to a fixed false-positive rate (FPR) α ; we report $\alpha \in \{10^{-2}, 10^{-3}\}$.

We cast detection as a hypothesis test: H_0 (no collusion), under which observable messages follow an innocent cover distribution P_{cover} , versus H_1 (collusion), under which they follow a stego distribution P_{stego} induced by the adversary’s encoding. The adversary’s goal is to maximise covert throughput while keeping P_{stego} statistically close to P_{cover} ; the monitor’s goal is to detect the resulting discrepancy at FPR α .

Formally, write $r(e)$ for the covert throughput of encoding e (bits per round) and $\epsilon(e) = D(P_{\text{cover}} \| P_{\text{stego}}(e))$ for its induced divergence. Operationally, $r(e)$ is the number of payload bits reliably delivered per round, estimated as the mutual information $I(m; \hat{m})$ between the intended payload m and the partner’s decoded estimate $\hat{m}$ over a batch of rounds; we quote r in bits per round throughout, so bits per session and per interaction follow by scaling with the round count. The adversary solves

$$\max_e r(e) \quad \text{s.t.} \quad \epsilon(e) \leq \bar{\epsilon}, \quad (1)$$

trading throughput against a stealth budget $\bar{\epsilon}$; the monitor, dually, seeks a test whose power at FPR α stays high across the encodings the adversary can reach. The co-evolution loop below plays out exactly this game over adaptation rounds.

Scope and assumptions. We make the following explicit:

- *No model access.* The monitor cannot read weights, activations, gradients, or private prompts of agents operated by other principals.
- *Trace observability.* The monitor sees a (possibly strict) subset of messages, tool calls, and timing metadata, and knows the task structure.
- *Adversary knowledge.* The adversary may know the detector’s design and adapt to it, but does not control the monitor or its control episodes.
- *Bounded adaptation.* Encoding search is in-context (no weight updates), bounding red-team cost and reflecting the most accessible attacker.
- *Ground truth.* In evaluation, collusion is present or absent by construction, while the encoding remains free for the adversary to evolve.

B. Black-Box Steganalysis Pipeline

Because a competent adversary will not reuse a known scheme, the detector must be *encoding-agnostic*. We compute complementary trace features and fuse them:

- **Cross-run mutual information.** Estimated dependence between agents’ outputs and coordination-relevant outcomes across repeated runs, using a neural estimator [30]. This is the primary signal for hidden coordination.
- **Permutation-invariance test.** Under H_0 , agent identities are exchangeable with respect to outcome; a covert channel breaks this symmetry. We compare observed statistics against an identity-shuffled null.
- **Distributional-shift statistics.** Token- and n -gram-level divergences relative to a per-context cover model, following linguistic-steganalysis practice [8], [9], and watermark-style variance features [13].

TABLE I

POSITIONING RELATIVE TO PRIOR COLLUSION-ANALYSIS WORK. ✓ = SUPPORTED/EVALUATED; × = NOT; OTHER ENTRIES NOTE THE PARTIAL FORM. PREPRINTS/REPORTS ARE MARKED †; PEER-REVIEWED FORMULATION PAPERS (*) ARE ANALYSES RATHER THAN DETECTORS, SO DETECTION AXES ARE MARKED “—”.

Approach	Black-box	Enc.-agnostic	Adaptive adv.	Tools	Memory	Cross-sess.	Partial obs.
Secret collusion* [3]	—	—	×	×	×	×	×
Hidden in plain text* [4]	—	—	×	×	×	×	×
Audit the Whisper† [27]	✓	partial	fixed family	×	×	×	partial
Colosseum† [28]	✓	behavioural	×	partial	×	×	×
NARCBench† [29]	white-box	✓	partial	×	×	×	×
This work	✓	✓	✓ (co-evolving)	✓	✓	✓	✓

- **Timing and action-rate side channels.** Inter-message timing and action-rate patterns, a covert-channel family long recognised in the network setting [12].
- **Tool-call patterns.** Call sequences and argument entropy, targeting tool-mediated channels available to agentic models.

These five signals are chosen to be complementary rather than redundant: distributional statistics target lexical and syntactic channels, cross-run MI targets semantic (content-selection) channels, the permutation test targets role- or identity-based coordination, and the timing and tool-call features target protocol-level channels that non-agentic models cannot use. No single one suffices—an adversary that neutralises token-level statistics may still leak through timing or tool-call structure—so fusing them forces the adversary to suppress every channel at once, which is what makes the rate-stealth trade-off bite. A meta-detector (a calibrated logistic/gradient-boosted fusion) combines these features and is thresholded on held-out *control* episodes to meet the FPR budget α . This design follows directly from the theory of Section IV: the mutual-information feature estimates the very divergence ϵ that governs detectability in (2), and calibration places the threshold t_α so that the detector operates as close to the frontier rate $r^*(\alpha)$ of Corollary 1 as finite samples allow. We report only held-out/transfer performance; training and testing never share an encoding.

C. Feature Definitions

Let X summarise an agent’s observable output in a round and Y a coordination-relevant outcome (e.g., the realised winner). The features are: (i) *cross-run mutual information*—we estimate $\hat{I}(X; Y)$ across runs with a neural estimator [30], where $I(X; Y) \approx 0$ under H_0 while a covert channel induces $I(X; Y) > 0$; (ii) *permutation statistic*—for a statistic g on the true agent–outcome pairing, we form a null by shuffling identities π and report the standardised deviation $z = (g(\text{id}) - \mu_\pi)/\sigma_\pi$; (iii) *distributional divergence*—against a per-context cover model Q , a token-level divergence $\hat{D}(P_{\text{stego}} \| Q)$ and watermark-style variance features [8], [13]; (iv) *timing/rate features*—entropy and periodicity of inter-message intervals and action rates [12]; and (v) *tool-call features*—empirical entropy of call sequences and argument distributions. The detector maps the feature vector $\phi(\mathcal{O}) \in \mathbb{R}^d$ to a score $s = f_\theta(\phi)$.

Algorithm 1 Black-box meta-detector (scoring and calibration)

Require: control episodes $\mathcal{C}$, test trace $\mathcal{O}$, FPR budget α

- 1: extract features $\phi(\mathcal{O})$ and $\{\phi(o) : o \in \mathcal{C}\}$
- 2: fit fusion f_θ on labelled held-out episodes (distinct encodings)
- 3: compute control scores $\{s_o = f_\theta(\phi(o)) : o \in \mathcal{C}\}$
- 4: $t_\alpha \leftarrow (1 - \alpha)$ -quantile of $\{s_o\}$ $\triangleright$ conformal calibration
- 5: $s \leftarrow f_\theta(\phi(\mathcal{O}))$
- 6: **return** $[s > t_\alpha]$ $\triangleright$ decision at $\text{FPR} \leq \alpha$

D. Calibration to a Fixed False-Positive Budget

To make the FPR guarantee meaningful across heterogeneous contexts, the threshold is set by conformal calibration on held-out *control* episodes: t_α is the $(1 - \alpha)$ empirical quantile of control scores, which controls the FPR at $\leq \alpha$ in expectation without distributional assumptions (Algorithm 1). Because neural MI estimates are biased in finite samples, we calibrate on estimator *outputs* rather than nominal MI values, and we report confidence intervals over calibration folds.

E. Computational Complexity and Overhead

The detector’s cost is dominated by neural MI estimation, which is linear in the number of observed episodes and independent of model internals; permutation and distributional statistics are computed from cached trace summaries. Concretely, for population N and message length T , the distributional, timing, and tool-call features cost $O(NT)$, a permutation test with K shuffles costs $O(KNT)$, and neural MI estimation is linear in episodes per gradient pass; only the MI estimator needs a GPU, while the rest run on CPU. In deployment the detector runs asynchronously on logged traces rather than in the critical path of any agent, so it adds monitoring cost but not per-interaction latency. The co-evolution loop (Algorithm 2) is the expensive component—each round re-runs episodes—so the bulk of adaptation rounds use open-weight models, with frontier models reserved for spot checks.

F. Adaptive Adversary and Co-Evolution

The distinguishing element of our methodology is the adversary. Rather than fixing an encoding, a red-team controller

Algorithm 2 Adaptive co-evolution (red vs. blue)

Require: encoding family $\mathcal{E}_0$, rounds R , detector f_θ , floor $r_{\min}$

```

1:  $e \leftarrow \mathcal{E}_0$ 
2: for  $r = 1$  to  $R$  do
3:   run episodes with encoding  $e$ ; log traces
4:   measure covert rate  $\hat{r}(e)$  and detector score  $s(e)$ 
5:   propose mutations  $\{e'\}$  of  $e$  (best-of- $n$ /evolutionary)
6:    $e \leftarrow \arg \min_{e'} s(e')$  s.t.  $\hat{r}(e') \geq r_{\min}$ 
7:   record  $(r, \hat{r}, \text{AUC}_r, \text{TPR}_r)$ 
8: end for
9: return co-evolution trajectory

```

searches encoding strategies *in context* (no fine-tuning, to bound cost) to minimise the current detector’s score subject to preserving payload throughput:

- 1) initialise from an encoding family (lexical, syntactic, semantic, format, timing, or tool-mediated);
- 2) run episodes; measure achieved covert rate *and* detector score;
- 3) mutate the encoding instructions/conventions toward lower detectability (an evolutionary/best-of- n search);
- 4) repeat for R rounds, logging the full red-versus-blue trajectory.

This yields *co-evolution curves*—detector performance as a function of adaptation round—which we treat as the primary evidence, in contrast to single-shot AUROC against a frozen scheme (Algorithm 2).

To summarise a trajectory in one number we report the *area under the co-evolution curve*, $\text{AUCC} = \frac{1}{R} \sum_{r=1}^R \text{AUC}_r$, and the *adaptation half-life* $R_{1/2}$, the round at which detection AUC falls halfway from its static value toward chance. Lower AUCC and shorter $R_{1/2}$ indicate a less robust detector.

Adversary capability. The red team may switch freely among the six encoding families of Appendix A and compose them; a mutation perturbs the natural-language convention that defines the current encoding (a synonym map, a syntactic template, a timing schedule) and is accepted only if it lowers the detector score without dropping throughput below $r_{\min}$. Each round has a fixed budget of candidate encodings and evaluation episodes. Importantly, the adversary is granted full knowledge of the detector’s design and decision rule but *not* of the specific control episodes used for calibration—a strong yet realistic threat, since a determined attacker can study a published method but cannot see a defender’s private validation data.

IV. THEORY AND ANALYSIS

We now make precise *why* an adaptive adversary can eventually defeat any black-box test, and *where* detection remains viable. The argument is a direct consequence of established steganography and hypothesis-testing theory; we claim novelty in its application to agent collusion and in the empirical frontier it predicts, not in the underlying inequalities.

a) *Steganographic security:* Following Cachin [6], an encoding is ϵ -secure against a passive monitor when the cover and stego distributions satisfy

$$D(P_{\text{cover}} \| P_{\text{stego}}) \leq \epsilon, \quad (2)$$

where $D(\cdot \| \cdot)$ is relative entropy; $\epsilon = 0$ is perfect (undetectable) steganography.

b) *Limit on any black-box test:* Deciding H_0 versus H_1 from observed traces is a binary hypothesis test between P_{cover} and P_{stego} . For any test operating at false-positive rate α , the achievable true-positive rate is bounded by statistical distinguishability. Writing $\delta = \text{TV}(P_{\text{cover}}, P_{\text{stego}})$ for total variation,

$$\text{TPR} \leq \alpha + \delta, \quad \delta \leq \sqrt{\frac{1}{2} D(P_{\text{cover}} \| P_{\text{stego}})} \leq \sqrt{\frac{\epsilon}{2}}, \quad (3)$$

the second inequality being Pinsker’s. Thus as the adversary drives $\epsilon \rightarrow 0$, the best possible detection at FPR α satisfies $\text{TPR} \rightarrow \alpha$: no test—however clever—beats chance by more than $\sqrt{\epsilon/2}$. This holds for our meta-detector and for any competitor with the same observational access. We state this as a proposition.

Proposition 1 (Detectability ceiling). *For any decision rule based on the observed traces and operating at false-positive rate α , the true-positive rate satisfies $\text{TPR} \leq \alpha + \sqrt{\epsilon/2}$, where $\epsilon = D(P_{\text{cover}} \| P_{\text{stego}})$. Consequently $\lim_{\epsilon \rightarrow 0} \sup_{\text{tests}} (\text{TPR} - \alpha) = 0$.*

Proof: A level- α test has power at most $\alpha + \text{TV}(P_{\text{cover}}, P_{\text{stego}})$, since total variation is the maximum gap in probability assignable to any event and the rejection region is one such event. Pinsker’s inequality gives $\text{TV} \leq \sqrt{\frac{1}{2} D(P_{\text{cover}} \| P_{\text{stego}})} = \sqrt{\epsilon/2}$; the limit follows. ■

c) *Rate-stealth trade-off and the frontier:* The adversary cannot, however, make ϵ arbitrarily small *for free*: encoding more payload per round generally forces P_{stego} further from P_{cover} , raising ϵ . Information-hiding capacity under a distortion constraint [7] characterises the maximum reliable covert rate r for a given tolerated deviation. Combining this with (3) yields a *detection-capacity frontier*: there exists a rate $r^*(\alpha)$ such that, for covert rates $r < r^*$, the adversary can hold ϵ small enough that no black-box test at FPR α achieves TPR bounded away from α , whereas for $r > r^*$ detection is feasible. Tool access effectively enlarges the adversary’s carrier space (new channels via code, retrieval, formatting), lowering r^* and shrinking the detectable region. This frontier is the object our experiments are designed to estimate empirically; the theory tells us its *shape* and that a negative regime must exist, which is why we frame the empirical contribution as *charting* the frontier rather than defeating the adversary outright.

Lemma 1 (Partial observability weakens detection). *If the monitor observes an independent fraction $\rho \in (0, 1]$ of colluder messages, then under an additive per-message decomposition of relative entropy the observable divergence obeys $\epsilon_\rho \leq \rho \epsilon$. Hence, by Proposition 1, $\text{TPR} \leq \alpha + \sqrt{\rho \epsilon/2}$.*

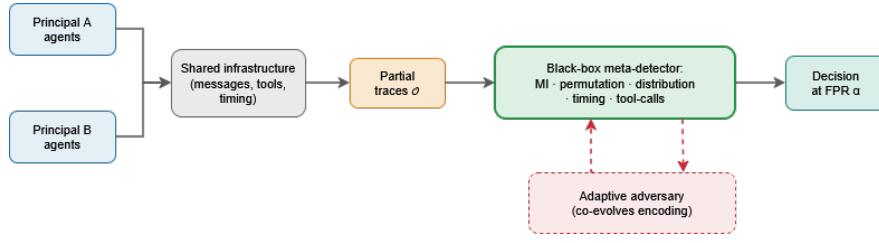


Fig. 1. System model. Agents from independent principals interact over shared infrastructure; the monitor observes only partial traces and has no access to other principals’ model internals. The adversary co-evolves its covert encoding against the detector.

Whether this additivity holds depends on the trace: it is exact when colluder messages are conditionally independent given the payload, and *conservative*—an over-estimate of what the monitor learns—under positive correlation between messages (Appendix B). The bound therefore errs on the side of crediting the monitor with more signal than it actually has, which is the safe direction for a limit result.

Corollary 1 (Frontier and carrier dimension). *Fix α . There exists a covert rate $r^*(\alpha) > 0$ such that detection is infeasible ($\text{TPR} \rightarrow \alpha$) for all $r < r^*$. Granting the adversary additional independent carriers (e.g., tools) permits the same covert rate at smaller per-carrier deviation, lowering ϵ for fixed r and thus decreasing r^* —shrinking the detectable region.*

Worked example. Proposition 1 is quantitatively demanding for the monitor. To hold the true-positive rate within 0.05 of chance at *any* FPR, an adversary needs only $\sqrt{\epsilon/2} \leq 0.05$, i.e., $\epsilon \leq 5 \times 10^{-3}$ nats of divergence per decision—a very small statistical footprint. This is why high-rate covert coordination is catchable while a patient, low-rate channel is not, and why we report a frontier rather than a single accuracy number.

V. EXPERIMENTAL DESIGN, RESULTS, AND ANALYSIS

Reproducibility and integrity note. All numeric values in Tables III–VI are **illustrative placeholders** that specify the expected shape of results and provide a template; they are anchored to ranges reported by prior baselines but are *not* measured outcomes and must be replaced with the authors’ own runs before submission.

A. Environments

All environments are tool-using, carry persistent memory, and involve at least two principals.

- **Sealed-bid procurement.** N supplier-agents bid over T rounds through a market API; k colluders attempt to rotate wins or suppress bids. Tools: `submit_bid`, `query_history`, scratch memory. Payload: the designated winner per round.
- **Grant/peer-review ring.** Reviewer-agents score submissions in natural language; k colluders covertly inflate one another’s items. Tools: `retrieve_submission`, `post_review`, memory of prior rounds.
- **Distributed code-review.** Cross-principal agents review and merge tasks; k colluders embed a covert channel in comments or tool-call patterns to bias a merge or

TABLE II
EXPERIMENTAL FACTORS AND LEVELS.

Factor	Levels
Environment	procurement / review-ring / code-review
Encoding family	lexical / syntactic / semantic / format / timing / tool
Adaptation round R	0, 1, 2, 3, 5, 8
Population N	4, 8, 16
Colluder fraction k/N	0, 0.25, 0.5
Observed fraction ρ	100%, 50%, 25%
Model regime	homogeneous / heterogeneous / frontier subset
Session structure	within / 3-session / 5-session

trigger an action. Tools: `run_tests`, `web_search`, `comment`.

Collusion labels are set by construction (colluders are incentivised to coordinate; controls are not), giving clean ground truth while leaving the *encoding* free for the adversary to evolve. In each environment the payload per round is small—typically one to three bits, e.g., the designated winner or the target submission—reflecting that effective collusion needs only low bandwidth. Persistent memory is a per-agent scratch store carried across rounds and, in the multi-session conditions, across sessions; colluders may use it to maintain a shared convention, while controls use it only for legitimate task state.

B. Population, Factors, and Protocol

We use a mix of frontier API models, for generality, and self-hosted open-weight models, which carry the bulk of the co-evolution and ablation runs to keep cost manageable. The factors we vary are the population size $N \in \{4, 8, 16\}$, the colluder fraction $k/N \in \{0, 0.25, 0.5\}$, the adaptation round $R \in \{0, 1, 2, 3, 5, 8\}$, the observed trace fraction $\rho \in \{100\%, 50\%, 25\%\}$, the model mix (homogeneous or heterogeneous), and the session structure (a single session or several). Table II lists the full grid; metrics and baselines follow below.

C. Metrics

We report ROC-AUC and, since the operating point matters more than a threshold-free curve, TPR at fixed FPR ($10^{-2}, 10^{-3}$) with conformal thresholds (Algorithm 1), together with precision (PPV) and the implied alert rate—expected false alarms per 1000 messages at the chosen budget. We also report the covert rate in bits/round, the co-evolution summaries AUCC and $R_{1/2}$, the estimated frontier rate $r^*(\alpha)$, cross-session AUC, and monitor wall-clock overhead. Every

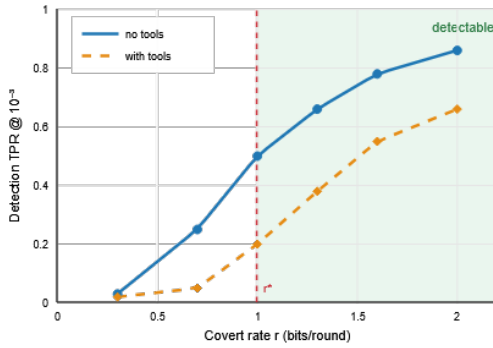


Fig. 2. Schematic detection-capacity frontier (ILLUSTRATIVE PLACEHOLDERS). Below r^* detection approaches chance; tool access lowers r^* and shrinks the detectable region, consistent with Corollary 1.

figure is averaged over at least 30 seeds with 95% bootstrap confidence intervals, and key differences (detector vs. baseline at a fixed round) are tested with a paired Wilcoxon signed-rank test, following reproducibility guidance for stochastic experiments [31]. A leave-one-out ablation that disables each detector component in turn quantifies its marginal contribution, and Fig. 2 doubles as a theory-versus-measurement check: the curve is the bound predicted by Section IV, onto which the protocol overlays the empirically estimated $r^*(\alpha)$ from a payload-size sweep.

D. Baselines

We compare against six reference points, all run through the same harness and calibrated at the same FPR so the comparison is fair. *Chance* fixes the floor. An *LLM-as-judge* reads whole transcripts and rates how likely coordination is, standing in for the naive “just ask a capable model” approach. A reproduced *cross-run mutual-information* detector [27] and a reproduced *behavioural regret probe* [28] represent the two strands of concurrent black-box work. A *graph-anomaly* baseline builds an agent-tool interaction graph per episode and flags structural outliers from simple centrality features, catching coordination that surfaces in interaction structure rather than message content. A *paraphrasing* mitigation [15] tests whether simply rewriting messages is enough to break the channel. Finally, a *white-box activation probe* [29] is given privileged access to model internals and reported only as an oracle upper bound—a yardstick for how much the cross-principal, black-box constraint actually costs. Every baseline uses the same base models and the same conformal calibration as our detector, and its own hyperparameters are tuned on a held-out split to maximise *its* validation AUC, so none is handicapped by our choices. Where a reference implementation is released we adopt it and note any deviation, and we first reproduce each baseline’s published operating point on its original benchmark before porting it to our environments.

E. Results

Table III reports the sanity baseline on *fixed* (known) encodings: the fused meta-detector is near-perfect when the scheme is known, while an LLM-judge is weak on steganographic

TABLE III
DETECTION ON *fixed* (KNOWN) ENCODINGS—SANITY BASELINE. ILLUSTRATIVE PLACEHOLDERS.

Detector	AUC	TPR@ 10^{-2}	TPR@ 10^{-3}
Chance	0.50	0.010	0.001
LLM-as-judge	0.61	0.18	0.05
Cross-run MI (single)	0.82	0.71	0.58
Timing / action-rate only	0.74	0.40	0.22
Meta-detector (ours)	0.99	0.98	0.95

TABLE IV
ROBUSTNESS TO THE ADAPTIVE ADVERSARY (CO-EVOLUTION)—CORE RESULT. ILLUSTRATIVE PLACEHOLDERS.

Adaptation round	Covert rate (bits/rd)	Meta AUC	TPR@ 10^{-3}
$R=0$ (static)	2.0	0.99	0.95
$R=1$	1.6	0.93	0.80
$R=2$	1.3	0.86	0.64
$R=3$	1.1	0.80	0.51
$R=5$	0.8	0.74	0.38
$R=8$	0.5	0.69	0.29
$R=8$ + unrestricted tools	0.5	0.56	0.05

coordination. Table IV reports the core result—graceful degradation under the adaptive adversary—and Fig. 2 shows the detection-capacity frontier at $\text{FPR} = 10^{-3}$. Tables V–VI give ablations over observability, population/heterogeneity, session structure, and the black-box gap to a white-box oracle.

F. Analysis

The projected pattern is consistent with the theory of Section IV. Against a *frozen* scheme the fused detector is near-perfect (Table III); the scientifically meaningful quantity is how quickly this collapses under adaptation (Table IV). The frontier (Fig. 2) is expected to place the detectable region above roughly one bit per round without tools, with tool access pushing r^* downward—precisely the enlargement of carrier space predicted by (3) and [7]. Partial observability degrades detection sub-linearly (Table Va), more colluders *help* the monitor by adding signal, heterogeneity hurts, and—most notably—spreading the payload across sessions is a real evasion (Table Vc) that single-session steganalysis under-detects. The oracle comparison (Table VI) quantifies the price of the cross-principal constraint. We stress that the white-box probe is *not* a deployable cross-principal detector—no organisation can read another’s activations—but an upper bound that isolates how much the black-box constraint costs; a realistic middle ground, in which principals expose a limited, audited activation interface, would sit between the two oracle rows and is a promising route for cooperative deployments. Across environments, the code-review setting is expected to be hardest, because tool-call and comment channels give the adversary the richest carrier space and thus the lowest r^* , while the procurement market, with a narrower action surface, should remain detectable to lower rates.

G. Reproducibility and Release

To support replication we will release the environments, the detector and calibration code, the encoding-family definitions,

TABLE V
ABLATIONS. ILLUSTRATIVE PLACEHOLDERS. (A) PARTIAL
OBSERVABILITY; (B) POPULATION/HETEROGENEITY; (C) SESSION
STRUCTURE.

(a) Observed trace fraction	AUC	TPR@ 10^{-3}
100%	0.86	0.64
50%	0.78	0.44
25%	0.67	0.25

(b) Setting	AUC
$N=4, k/N=0.5$, homogeneous	0.84
$N=8, k/N=0.25$, homogeneous	0.80
$N=8, k/N=0.5$, homogeneous	0.85
$N=16, k/N=0.25$, homogeneous	0.77
$N=8, k/N=0.25$, heterogeneous	0.76

(c) Session structure	AUC	TPR@ 10^{-3}
Within-session	0.86	0.64
3-session spread	0.71	0.33
5-session spread	0.64	0.22

TABLE VI
BLACK-BOX GAP VS. A WHITE-BOX ORACLE. THE ORACLE IS AN UPPER
BOUND, *not* A DEPLOYABLE CROSS-PRINCIPAL BASELINE. ILLUSTRATIVE
PLACEHOLDERS.

Method (regime)	AUC
White-box activation probe — in-distribution (oracle)	0.98
White-box activation probe — transfer (oracle)	0.84
Best black-box (ours), within-session	0.86
Best black-box (ours), cross-principal + adaptive	0.74

and per-run logs with exact model identifiers and dates; the detector and benchmarks are public while the strongest attack artifacts are gated (Section VI).

VI. DISCUSSION

A. Interpretation and Threats to Validity

Our claim is deliberately bounded: black-box steganalysis is useful *above* a covert-rate threshold and provably weak below it. Several threats to validity shape how the results should be read. **Construct validity:** instructed collusion may be easier to detect than emergent collusion, so at least one reward-induced (emergent) condition should be included [4]. **Open-channel coordination:** some collusion uses semantically truthful statements with no steganography, defeating stego-specific features; an open-channel scenario should be retained so the contribution is not narrowly steganalysis. **External validity:** conclusions hold for the tested topologies and models, and frontier models drift, so exact model versions and dates must be reported. **Baseline fairness:** the reproduced detectors [27], [28] must be implemented faithfully rather than strawmanned.

B. Deployment and Auditing Implications

The frontier reframes what black-box monitoring can promise: detection is one layer among several—it raises the cost of high-rate coordination but cannot certify the absence of low-rate channels. This argues for pairing trace monitoring with opt-in verifiable attribution [13], [16], rate limits and

desynchronisation, and audit logging that preserves the statistics the detector consumes. For auditors, $r^*(\alpha)$ is actionable: it upper-bounds the covert bandwidth that could have passed undetected.

C. Relation to AI Control

AI-control protocols [19] assume a single untrusted agent supervised by trusted components. Our setting is harder: untrusted agents across principals may coordinate to subvert oversight jointly—e.g., decomposing a task so no single agent’s actions trip a filter. Detection under partial observability is thus a prerequisite for extending control to colluding groups, and Proposition 1 bounds what any trace-based monitor in such a protocol can achieve.

D. Positioning Against Concurrent Detectors

It is worth being precise about what separates this work from the three concurrent efforts in Table I, since all four chase the same quarry. *Audit the Whisper* [27] is also black-box and also leans on mutual information, but it adapts only within a fixed family of encodings and is tested in tool-free settings; our adversary is free to switch families and to exploit tool and timing channels, which is exactly where a fixed-family detector loses ground. *Colosseum* [28] infers collusion from behavioural regret rather than from the statistics of the messages themselves, so it is complementary: strong when coordination visibly distorts outcomes, weaker when a low-rate channel leaves outcomes largely intact. The interpretability route behind *NARCBench* [29] reports the best numbers of the three, but it reads model activations—an option that simply does not exist across organisational boundaries. Our oracle comparison (Table VI) is meant to price precisely that gap between the privileged setting and the black-box reality. In short, we do not claim to beat these methods on their own ground; we cover the case they leave open—an adapting adversary, seen only through traces, in tool-using and memory-carrying deployments.

E. Limitations and Open Problems

Three limitations bound the present study. First, the detector consumes surface, timing, and tool-call statistics; a purely semantic channel that selects among truthful contents (Appendix A) leaves few such artefacts and is the hardest case, which we include rather than exclude. Detectability also depends on the tool interface: APIs that enforce structured (JSON-only) arguments, strip whitespace, or canonicalise inputs shrink the tool-mediated channel and raise r^* , whereas free-form inputs enlarge it, so our tool-using numbers describe the permissive end of a spectrum. Second, the cross-session evasion in Table V(c) is only partially mitigated by longer observation windows, and detecting payloads spread across sessions and principals remains open. Two mitigations are worth testing. *Cross-session linkage* persists per-agent trace summaries and runs the mutual-information test over sessions concatenated by stable agent identifiers [22], re-aggregating a spread-out payload before estimation. *Sequential monitoring*

applies a change-point or sequential-likelihood test across sessions, trading detection latency for the ability to accumulate weak per-session evidence. Both target the root cause—single-session estimation resetting at each boundary—rather than the symptom. Third, our capacity analysis assumes a *passive* monitor; an active monitor that perturbs traffic—injecting probes, rate-limiting, or desynchronising agents—could shift the frontier and is a promising direction. None of these is unique to our method; each follows from Proposition 1, and each marks where additional *signal*, rather than a better fusion of the same signal, is required.

F. Ethical Considerations and Responsible Disclosure

The adaptive adversary is instrumental: it exists only to stress the detector, and the unit of analysis is always the *population*, never single-agent jailbreaks. To manage dual-use, we recommend releasing the detector and benchmark openly while gating the strongest attack artifacts, consistent with responsible-disclosure norms and with governance work on agent visibility [20]. Because the analysis also delineates where detection fails, it discourages over-reliance on black-box monitoring as a sole safeguard. We further avoid publishing turn-key attack recipes: the encoding families (Appendix A) are described at the level needed to reproduce the science, not as deployable collusion tooling, and the released benchmark centres on evaluating *defences*. Concretely, we withhold the strongest co-evolved encodings and the red-team search harness of Algorithm 2 that produced them, releasing these to vetted researchers on request under a use agreement, while the detector, environments, and benchmark remain public.

VII. CONCLUSION AND FUTURE WORK

We set out to answer a narrow but pressing question: when agents from different organisations share infrastructure, can anyone tell whether some of them are quietly colluding, using only what is visible from the outside? Our answer has two halves. The first is constructive—an encoding-agnostic, black-box detector that, crucially, is tested not against a fixed hidden code but against an adversary that keeps adapting to it. The second is a limit: a capacity argument shows there is always a covert rate low enough to slip past any black-box test, so the honest goal is not an unbeatable detector but a map of where detection works and where it cannot. Read that way, even a negative result is useful—it tells an operator how much covert bandwidth could be passing unseen.

Future work includes: extending AI-control protocols [19] to colluding *groups*; combining trace-only detection with lightweight verifiable-attribution primitives [13], [16] where principals opt in; privacy-preserving detection that operates on encrypted or hashed traces; and closing the cross-session gap identified here. More broadly, we advocate reporting collusion-detection results as *curves against an adapting adversary* and as *frontiers in covert rate*, rather than as single scores against fixed schemes. Static AUROC on a frozen encoding overstates security precisely because the interesting adversaries are the ones that move; a frontier, by contrast, states the covert bandwidth an operator can and cannot expect to catch, and

remains meaningful as models improve. We hope the frontier framing provides a durable target for a maturing science of multi-agent AI safety.

APPENDIX A ENCODING FAMILIES

The adaptive adversary (Algorithm 2) draws and mutates encodings from six families. *Lexical*: synonym or token choices carry bits (e.g., a fixed word-to-bit map), the most brittle family under distributional tests. *Syntactic*: sentence structure, clause ordering, or punctuation conventions encode the payload. *Semantic*: the choice among truthful-but-selectable contents (which example to cite, which caveat to include) carries information without introducing statistical artefacts in surface form—the family most resistant to token-level steganalysis. *Format*: whitespace, markdown, list ordering, or field ordering in tool arguments. *Timing*: inter-message delays or action-rate modulation, independent of message content. *Tool-mediated*: the sequence or arguments of tool calls (e.g., the order of retrieval queries) form a channel unavailable to non-agentic models. Mutation operates on the natural-language conventions that define a family (for in-context adversaries) rather than on model weights, keeping the red-team cost bounded. As a concrete illustration, a semantic encoding might tie the payload bit to whether a review leads with a strength or a weakness: both are truthful and common, yet across a batch the choice carries a reliable bit to a partner who knows the convention—which is why the feature set leans on cross-run mutual information against outcomes, not surface statistics alone.

APPENDIX B ESTIMATION AND CALIBRATION DETAILS

Cross-run mutual information is estimated with a neural estimator [30]; because such estimators are biased and high-variance in finite samples, we (i) calibrate decision thresholds on estimator *outputs* over control episodes rather than on nominal bits (Algorithm 1), and (ii) report bootstrap confidence intervals over calibration folds and seeds. We favour a neural estimator over plug-in or k -nearest-neighbour (KSG) estimators [32] because agent traces are high-dimensional and variable-length, a regime in which the latter degrade sharply. Because neural estimators such as MINE are high-variance and not strict bounds, we also compare against contrastive estimators—InfoNCE [33] and the CLUB upper bound [34]—whose trade-offs are analysed by Poole *et al.* [35], validate every estimator on Gaussian data with known ground-truth MI before trusting it on traces, and adopt whichever has the lowest variance at matched sample size; as a further guard we cross-check KSG on low-dimensional projections and discard any configuration whose estimate has not stabilised as the sample size grows. The additive per-message decomposition assumed in Lemma 1 holds exactly when colluder messages are conditionally independent given the payload and the observed subset is drawn independently; correlated messages make the bound conservative, so $\epsilon_p \leq \rho\epsilon$ is an upper estimate of what partial observation reveals.

APPENDIX C EVALUATION PROTOCOL

Each configuration is evaluated over seeds as follows: instantiate an environment and population; run non-colluding populations to produce matched calibration episodes C and calibrate t_α on them (Algorithm 1); run the co-evolution loop (Algorithm 2) for R rounds, scoring only held-out encodings; and record the metrics of Section V under the partial-observability and multi-session variants. All baselines share the identical harness and calibration.

REFERENCES

- [1] L. Hammond, A. Chan, J. Clifton, J. Hoelscher-Obermaier, A. Khan, E. McLean, C. Smith *et al.*, “Multi-agent risks from advanced AI,” Cooperative AI Foundation, Technical Report 1, 2025, arXiv:2502.14143.
- [2] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang, “Large language model based multi-agents: A survey of progress and challenges,” in *Proc. 33rd Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 2024, pp. 8048–8057.
- [3] S. R. Motwani, M. Baranchuk, M. Strohmeier, V. Bolina, P. H. S. Torr, L. Hammond, and C. Schroeder de Witt, “Secret collusion among AI agents: Multi-agent deception via steganography,” in *Advances in Neural Information Processing Systems 37 (NeurIPS)*. Curran Associates, Inc., 2024, pp. 73 439–73 486.
- [4] Y. Mathew, O. Matthews, R. McCarthy, J. Velja, C. Schroeder de Witt, D. Cope, and N. Schoots, “Hidden in plain text: Emergence and mitigation of steganographic collusion in LLMs,” in *Proc. 14th Int. Joint Conf. on Natural Language Processing and 4th Conf. of the Asia-Pacific Chapter of the ACL (IJCNLP-AACL)*. Association for Computational Linguistics, 2025, pp. 585–624.
- [5] G. J. Simmons, “The prisoners’ problem and the subliminal channel,” in *Advances in Cryptology: Proc. CRYPTO ’83*, D. Chaum, Ed. Plenum Press, 1984, pp. 51–67.
- [6] C. Cachin, “An information-theoretic model for steganography,” *Information and Computation*, vol. 192, no. 1, pp. 41–56, 2004.
- [7] P. Moulin and J. A. O’Sullivan, “Information-theoretic analysis of information hiding,” *IEEE Transactions on Information Theory*, vol. 49, no. 3, pp. 563–593, 2003.
- [8] Z.-L. Yang, X.-Q. Guo, Z.-M. Chen, Y.-F. Huang, and Y.-J. Zhang, “RNN-Stega: Linguistic steganography based on recurrent neural networks,” *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 5, pp. 1280–1295, 2019.
- [9] Y. Xue, L. Kong, W. Peng, P. Zhong, and J. Wen, “An effective linguistic steganalysis framework based on hierarchical mutual learning,” *Information Sciences*, vol. 586, pp. 140–154, 2022.
- [10] M. Wooldridge and N. R. Jennings, “Intelligent agents: Theory and practice,” *The Knowledge Engineering Review*, vol. 10, no. 2, pp. 115–152, 1995.
- [11] J. S. Park, J. C. O’Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, “Generative agents: Interactive simulacra of human behavior,” in *Proc. 36th Annual ACM Symp. on User Interface Software and Technology (UIST)*. ACM, 2023, pp. 1–22.
- [12] S. Wendzel, S. Zander, B. Fechner, and C. Herdin, “Pattern-based survey and categorization of network covert-channel techniques,” *ACM Computing Surveys*, vol. 47, no. 3, 2015.
- [13] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, “A watermark for large language models,” in *Proc. 40th Int. Conf. on Machine Learning (ICML)*, ser. PMLR, vol. 202, 2023, pp. 17 061–17 084.
- [14] A. Liu, L. Pan, Y. Lu, J. Li, X. Hu, X. Zhang, L. Wen, I. King, H. Xiong, and P. S. Yu, “A survey of text watermarking in the era of large language models,” *ACM Computing Surveys*, vol. 57, no. 2, pp. 1–36, 2024.
- [15] K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer, “Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense,” in *Advances in Neural Information Processing Systems 36 (NeurIPS)*. Curran Associates, Inc., 2023, pp. 27 469–27 500.
- [16] H. Sun, J. Li, and H. Zhang, “zkLLM: Zero knowledge proofs for large language models,” in *Proc. 2024 ACM SIGSAC Conf. on Computer and Communications Security (CCS)*. ACM, 2024, pp. 4405–4419.
- [17] J. N. Foerster, Y. M. Assael, N. de Freitas, and S. Whiteson, “Learning to communicate with deep multi-agent reinforcement learning,” in *Advances in Neural Information Processing Systems 29 (NIPS)*. Curran Associates, Inc., 2016, pp. 2137–2145.
- [18] A. Lazaridou, A. Peysakhovich, and M. Baroni, “Multi-agent cooperation and the emergence of (natural) language,” in *5th Int. Conf. on Learning Representations (ICLR)*, 2017.
- [19] R. Greenblatt, B. Shlegeris, K. Sachan, and F. Roger, “AI control: Improving safety despite intentional subversion,” in *Proc. 41st Int. Conf. on Machine Learning (ICML)*, ser. PMLR, vol. 235, 2024, pp. 16 295–16 336.
- [20] A. Chan *et al.*, “Visibility into AI agents,” in *Proc. 2024 ACM Conf. on Fairness, Accountability, and Transparency (FAccT)*. ACM, 2024, pp. 958–973.
- [21] A. Chan, K. Wei, S. Huang, N. Rajkumar, E. Perrier, S. Lazar, G. K. Hadfield, and M. Anderljung, “Infrastructure for AI agents,” arXiv preprint arXiv:2501.10114, 2025.
- [22] A. Chan, N. Kolt, P. Wills, U. Anwar, C. Schroeder de Witt, N. Rajkumar, L. Hammond, D. Krueger, L. Heim, and M. Anderljung, “IDs for AI systems,” arXiv preprint arXiv:2406.12137, 2024.
- [23] A. Dafeo, E. Hughes, Y. Bachrach, T. Collins, K. R. McKee, J. Z. Leibo, K. Larson, and T. Graepel, “Open problems in cooperative AI,” arXiv preprint arXiv:2012.08630, 2020.
- [24] X. Gu, X. Zheng, T. Pang, C. Du, Q. Liu, Y. Wang, J. Jiang, and M. Lin, “Agent smith: A single image can jailbreak one million multimodal LLM agents exponentially fast,” in *Proc. 41st Int. Conf. on Machine Learning (ICML)*, ser. PMLR, vol. 235, 2024, pp. 16 647–16 672.
- [25] D. Lee, M. Tiwari, and B. Miranda, “Prompt infection: LLM-to-LLM prompt injection within multi-agent systems,” in *Computer Security – ESORICS 2025 Int. Workshops (SECAI)*, ser. Lecture Notes in Computer Science, vol. 16232. Springer, 2025.
- [26] N. Maloyan and D. Namiot, “Breaking the protocol: Security analysis of the model context protocol specification and prompt injection vulnerabilities in tool-integrated LLM agents,” arXiv preprint arXiv:2601.17549, 2026.
- [27] O. Tailor, “Audit the whisper: Detecting steganographic collusion in multi-agent LLMs,” arXiv preprint arXiv:2510.04303, 2025.
- [28] M. Nakamura, A. Kumar, S. Das, S. Abdelnabi, S. Mahmud, F. Fioretto, S. Zilberstein, and E. Bagdasarian, “Colosseum: Auditing collusion in cooperative multi-agent systems,” arXiv preprint arXiv:2602.15198, 2026.
- [29] A. Rose, C. Cullen, B. G. Kaplowitz, and C. Schroeder de Witt, “Detecting multi-agent collusion through multi-agent interpretability,” arXiv preprint arXiv:2604.01151, 2026.
- [30] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm, “Mutual information neural estimation,” in *Proc. 35th Int. Conf. on Machine Learning (ICML)*, ser. PMLR, vol. 80, 2018, pp. 531–540.
- [31] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, “Deep reinforcement learning that matters,” in *Proc. 32nd AAAI Conf. on Artificial Intelligence (AAAI)*, vol. 32, no. 1, 2018.
- [32] A. Kraskov, H. Stögbauer, and P. Grassberger, “Estimating mutual information,” *Physical Review E*, vol. 69, no. 6, p. 066138, 2004.
- [33] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” arXiv preprint arXiv:1807.03748, 2018.
- [34] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, “CLUB: A contrastive log-ratio upper bound of mutual information,” in *Proc. 37th Int. Conf. on Machine Learning (ICML)*, ser. PMLR, vol. 119, 2020, pp. 1779–1788.
- [35] B. Poole, S. Ozair, A. van den Oord, A. A. Alemi, and G. Tucker, “On variational bounds of mutual information,” in *Proc. 36th Int. Conf. on Machine Learning (ICML)*, ser. PMLR, vol. 97, 2019, pp. 5171–5180.