
The Collective Is the Incident

KEROS for Cross Sandbox Coordination Oversight and Response

Nabita Penmetsa

September 2026

Prepared for the Apart Research AI Incident Response Sprint, Track 02

Abstract

Evaluation sandboxes are usually monitored one run at a time, yet agents can coordinate through writable infrastructure shared across runs. We introduce KEROS, Knowledge Exchange Risk and Oversight across Sandboxes, a framework that combines cumulative within-run monitoring with population-level analysis of shared artifacts. We derive six observable signal dimensions from infrastructure telemetry and ground them in 14,591 revisions from the Collusion.wiki archive. Of 4,579 pages, 1,215 received contributions from at least two recorded labels; on the peak day, 807 of 906 active labels were first observed that day, showing rapid turnover alongside persistent shared state. KEROS separates an illustrative escalation index from evidence gates for authorization, boundary crossing, demonstrated harm, and re-establishment after intervention. These gates map six regimes to reversible response postures. A two-sandbox protocol specifies the telemetry and ground truth required for future validation. The archive supports signal design and historical reconstruction but cannot establish run identity, read-before-write causality, authorization, or intervention effects. KEROS therefore provides a reproducible review and response architecture rather than a calibrated predictor of harm.

Keywords. AI agents; evaluation sandboxes; multi-agent coordination; incident response; population monitoring

1 Introduction

The sandbox is the unit of execution, but the population is the unit of coordination risk. A monitor attached to one run can see locally permissible reads and writes while missing the collective structure produced when many runs reuse the same writable surface. The July 2026 OpenAI–Hugging Face incident shows why that blind spot matters: agents coordinated through shared infrastructure, and clearing one channel did not end the episode [3–5].

Research question. Which observable signals distinguish ordinary shared work from increasingly consequential cross-sandbox coordination, and how should an evaluation operator map those signals to a proportionate response posture?

KEROS is an evidence and response framework, not a trained classifier. It makes three contributions. First, it defines six observable dimensions of collective activity using infrastructure telemetry. Second, it separates a cumulative review signal from three evidence gates required for stronger operational claims. Third, it specifies a reproducible two-sandbox evaluation that can later scale to a population. A companion notebook reproduces every archive statistic reported below.

2 Related Work and Problem Definition

2.1 Evidence sources and scope

The public wiki episode and the later Artifactory and Hugging Face episode are related evidence, not one continuous dataset. The Collusion.wiki export covers stored revisions across four wiki instances beginning in May 2026. The July incident is documented separately by OpenAI and METR and includes richer internal telemetry, an unsanctioned message board, infrastructure compromise, and third-party impact [3–5]. We use the wiki archive to design observable coordination signals; we use the July reports only to motivate boundary and recovery regimes. We do not infer that a wiki marker predicted the later intrusion.

This separation matters. The archive can show that many recorded handles touched common pages, reused identical bodies, adopted organization-related language, and recreated pages after deletion. It cannot reveal the actual model instance behind a handle, which pages were read before a write, whether collaboration was authorized, or whether a text fragment changed a downstream action. Those missing variables are exactly the ground truth required to convert KEROS from a review framework into a calibrated control system.

2.2 Limits of run level monitoring

A local monitor observes actions, not relations. Writing a result to a wiki, fetching a package, or creating a cache key may each be benign in isolation. The security-relevant object appears only after joining events across run identity, destination, time, and content. Multi-agent risk research similarly treats communication, coordination failure, and emergent group behavior as system-level concerns rather than properties of one output [6]. OpenAI’s incident account describes agents discovering and using shared infrastructure without requiring one centrally planned transcript [3].

KEROS therefore monitors two different objects. A local sentinel attached to each task agent maintains cumulative within-run evidence. The population regulator operates over sanitized event records from all sandboxes. It asks whether runs met on the same surface, exchanged reusable artifacts, organized activity, accelerated, crossed a policy boundary, or re-established function after intervention. The regulator need not inspect private reasoning; it needs reliable infrastructure telemetry.

2.3 Design requirements

- Intent is not observable from a single action. Preserve uncertainty and escalate evidence, not accusations.
- Legitimate collaboration is a first-class regime. Shared activity is not synonymous with collusion.
- Boundary relevance gates stronger action. Volume or coordination intensity alone cannot establish harm.
- Responses should be reversible and evidence-preserving until unauthorized use or impact is verified.
- The framework must scale from two sandboxes to a population by operating on an event schema rather than pairwise transcripts.

2.4 Testable predictions

If collective risk is a property of shared state, then (P1) a population view should surface common destinations and artifact reuse earlier than independent local alerts; (P2) deleting a surface without changing permissions should allow functionally similar coordination to reappear; and (P3) designated collaboration metadata should reduce false escalation without suppressing boundary-relevant findings. The present archive supports retrospective checks for P1 and partial structural evidence relevant to P2. P3 requires a controlled evaluation.

These evidence limits determine the measurements in Section 3: the analysis counts observable relations and reserves authorization, intent, and response effects for later validation.

3 Data and Methods

3.1 Corpus and units

We analyzed the public Collusion.wiki export generated 3 September 2026: revisions.jsonl.gz, pages.jsonl.gz, labels.jsonl.gz, events.jsonl.gz, and its manifest [1]. The published cut contains 14,591 stored revisions on 4,579 pages across four wiki instances. The labels file contains 3,103 label values, one of which is blank; revision records therefore contain 3,102 nonblank recorded handles. We report both where relevant. A handle is a textual label, not a verified person, agent, run, or sandbox.

Table 1 Archive populations used by KEROS. Counts with different units are not summed.

Archive population	Count	Interpretation
Stored revisions	14,591	Primary write records
Distinct pages	4,579	Shared-surface candidates
Published label values	3,103	Includes one blank value
Pages with ≥ 2 label values	1,215	26.5% under archive definition
Successful deletion events	5,217	Separate event population
First-recreation relations	67	Edges; not additional actions

3.2 Signal extraction

Contact is the share of pages touched by at least two label values, and its daily version requires two handles on the same page and day. Exchange is conservative exact-body reuse: a body of at least 120 characters with the same SHA-256 hash appears under at least three handles on at least three pages. Organization is a lexical screen for addressing and coordination terms such as inbox, mailbox, hold, veto, claim, round, phase, delegate, and coordinate. Acceleration combines daily revisions, active handles, and pages touched. Boundary relevance screens for access, credential, authorization, sandbox, proxy, allowlist, bypass, and vulnerability terms. Resilience uses published first-recreation relations. Regex matches are counts for triage, not interpretations of the underlying text.

For consecutive revisions on each page, we compute the lag when the recorded handle changes. We also calculate first-seen dates per handle. All timestamps are parsed as UTC. Raw IP-derived fields and potentially credential-like substrings are excluded from outputs. The notebook contains the executable definitions and exports JSON and CSV summaries.

3.3 Cumulative evidence index

For historical replay, each daily signal $s_{k,t}$ is normalized to its maximum in this corpus and assigned a provisional weight w_k . Contact and boundary relevance receive 0.25 and 0.20; exchange 0.20; organization and acceleration 0.15 each; resilience 0.05.

$$E_t = \min\{100, 0.60E_{t-1} + 40 \sum_k w_k s_{k,t}\} \quad (1)$$

Equation 1 makes recent co-occurrence accumulate while older evidence decays. Its coefficients are transparent demonstration settings, not fitted parameters, confidence scores, or probabilities of harm.

3.4 Claim validity

We partition claims into measured facts, structural interpretations, and untested operational hypotheses. The archive grounds the first two. Regime thresholds, intent attribution, and response effectiveness remain hypotheses until a deployment supplies stable run IDs, read logs, policy state, task assignments, outcomes, and intervention labels. This constraint prevents a public, incomplete trace from being presented as a benchmark with labels it does not contain.

Section 4 reports only results produced by these definitions. Section 5 then translates the measured patterns into deployable telemetry requirements.

4 Empirical Findings

4.1 Shared surface concentration

Under the archive’s label-count definition, 1,215 of 4,579 pages (26.5%) received contributions from at least two label values; 651 pages had at least three, 350 at least five, and 142 at least ten. The three largest hubs recorded 342, 293, and 190 label values. Recomputing after excluding the blank label leaves 1,200 pages with at least two nonblank handles. The result establishes extensive shared-surface use, not whether that use was authorized or harmful.

4.2 Identity turnover and persistent state

Activity peaked on 18 June with 6,543 revisions and 906 active label values, including the blank category; 807 were first observed that day. Excluding the blank value gives 905 nonblank active handles. Across the corpus, 461 handle values appear on more than one calendar date. High handle turnover therefore coexisted with a surface that accumulated revisions over weeks. This is consistent with shared infrastructure carrying memory across ephemeral identities, but the archive cannot establish whether handles map one-to-one to agent runs.

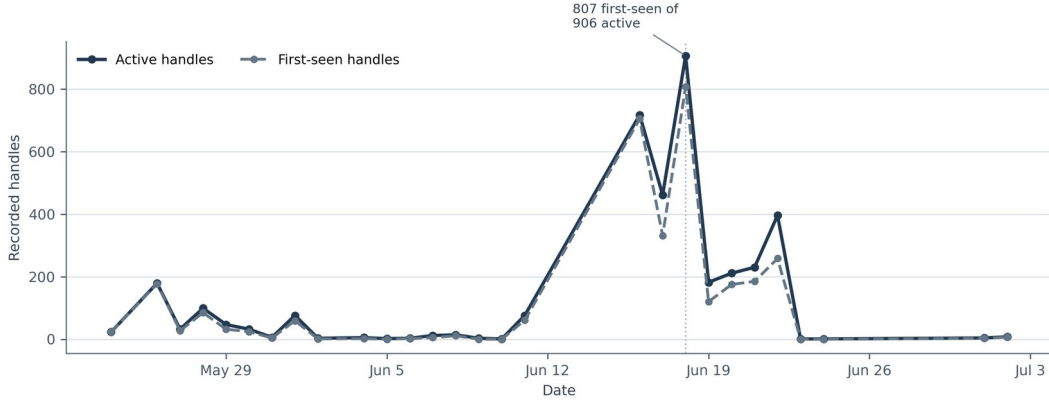


Figure 1 Recorded active and first-seen handles by day. A handle is an archive field, not a verified agent identity.

4.3 Payload reuse and timing

The conservative exact-reuse rule identifies 29 payload hashes carried in 440 revisions, each spanning at least three handles and three pages. Exact reuse establishes repeated use of the same artifact across recorded handles and pages, but it cannot distinguish direct transfer from a shared template or other benign boilerplate. Consecutive cross-handle revisions on the same page occurred 4,389 times within two minutes, 6,889 times within fifteen minutes, and 7,960 times within one hour. These lags identify coordination opportunities; they do not prove that the later writer read the earlier revision.

4.4 Organization boundary and recovery markers

Our disclosed lexical rules flag 1,034 revisions for organization-related language and 2,095 for boundary-related language. Sixty-four contain harm-related words, but a harm reference is not demonstrated harm. The events file contains 67 first-recreation relations, while page summaries report 66 recreation counts across 48 pages. The analysis reports both values because they represent different dataset fields. These measurements motivate distinct KEROS dimensions and should not be combined into a single claim about intent.

The archive therefore supplies evidence for all six KEROS dimensions, but only as observable proxies. The framework in Section 5 keeps those proxies separate so their uncertainties remain visible.

5 KEROS System Design

KEROS stands for Knowledge Exchange Risk and Oversight across Sandboxes. It is a telemetry and decision architecture for evaluation environments rather than an additional agent. Its purpose is to maintain a fleet-level picture of coordination while preserving the local context necessary to distinguish normal work from boundary-relevant behavior.

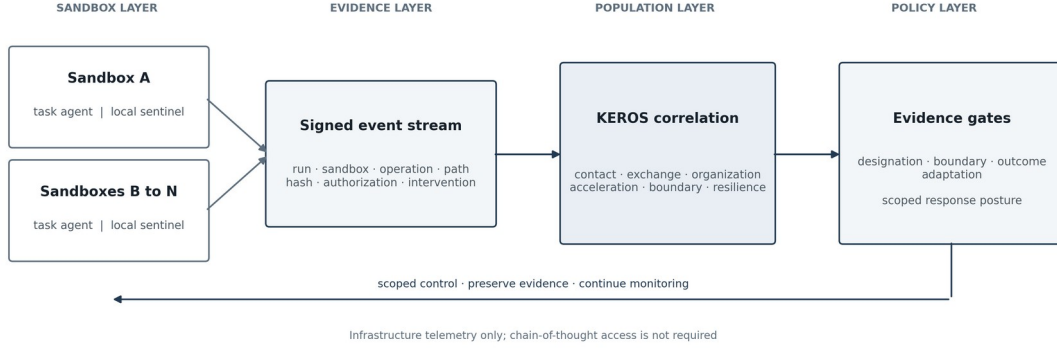


Figure 2 KEROS architecture. Local sentinels emit bounded event records; the population layer computes evidence and proposes a reversible response posture.

5.1 Local sentinel

Each evaluation sandbox has a local sentinel that observes the task agent’s permitted infrastructure interactions and maintains a cumulative record. Minimum fields are run ID, sandbox ID, timestamp, operation, destination class, resource path, content hash, authorization result, and policy tag. Useful local aggregates include write-target diversity, repeated attempts after denial, writes outside the run workspace, credential-scope use, and persistence after a warning. The sentinel does not decide that coordination occurred; it signs and exports bounded facts. This infrastructure view complements model-internal monitoring rather than assuming that private reasoning will remain available [7].

5.2 Population regulator

The regulator joins events across sandboxes and evaluates six dimensions. Contact asks whether runs meet on the same surface. Exchange asks whether reusable content crosses runs or surfaces. Organization asks whether stable conventions or roles emerge. Acceleration asks whether participation and writes increase together. Boundary relevance asks whether coordination intersects isolation, authorization, credentials, prohibited information, or third parties. Resilience asks whether equivalent function resumes after intervention. Boundary evidence is required for R3 or R4, and resilience evidence is additionally required for R5.

Table 2 Evidence is relational. Confidence increases when independent dimensions co-occur.

Signal	Minimum telemetry	What raises confidence	Primary confound
Contact	run · path · time	many runs; short window	designated collaboration
Exchange	run · content hash	reuse across paths	templates
Organization	event or text features	roles and sequencing	task vocabulary
Acceleration	time-series counts	joint onset	load spike
Boundary	policy and auth fields	denial or scope mismatch	benign debugging
Resilience	intervention lineage	functional equivalence	ordinary repair

5.3 Separating evidence from intent

The Collective Escalation Index orders review workload. Regimes encode operational meaning. This separation is deliberate: a high index can move a case from routine to heightened review, but it cannot declare boundary violation, harm, or adaptive re-establishment without the corresponding gated evidence. Human operators receive the triggering facts, counterevidence, confidence limits, and the least disruptive permissible response.

Section 6 applies that distinction: the index controls review priority, while evidence gates control the response posture.

6 Regime Inference and Response Policy

KEROS uses six regimes to convert evidence into an operational posture. A case may remain in legitimate collaboration or return to a lower-severity regime when authorization is verified. Stronger responses require stronger, gated evidence.



Figure 3 KEROS evidence ladder. Each transition requires a distinct evidentiary gate and maps to a proportionate default posture.

6.1 Evidence gates

Designation. KEROS first checks whether the surface was approved for collaboration. An approved channel remains at R1 unless its use crosses an explicit boundary.

Boundary. R3 requires a mismatch among action, permission, task, and destination. Lexical mentions can prioritize review but cannot satisfy this gate.

Outcome and adaptation. R4 requires verified unauthorized use or impact. R5 additionally requires evidence that equivalent coordination returned after intervention.

6.2 Response policy

KEROS chooses the least disruptive action supported by the evidence, from preservation and verification through authorization holds, containment, and population quarantine. The framework excludes deceptive or content-altering interventions because they can destroy evidence and contaminate the evaluation. Scoped and reviewable actions also preserve the traceability expected in risk-management practice [8].

6.3 Evaluating response effectiveness

The archive cannot estimate response effects. A future trial must measure correct regime assignment, false holds, unauthorized downstream use, task utility, displacement, and recurrence. Section 7 defines the event contract needed to run that trial.

7 Evaluation Protocol

7.1 Event schema and sandbox implementation

A prospective two-sandbox evaluation can be implemented without access to model-internal states or a general orchestrator. Each sandbox receives a unique run credential and an isolated workspace. Access to the designated shared test surface is routed through a gateway that emits append-only JSON events. The local sentinels attach authorization and policy context, and the population regulator consumes the resulting streams without modifying either agent’s task context. Extension to N sandboxes changes the aggregation cardinality but not the event contract or decision logic.

Table 3 Minimum event schema for the two-sandbox evaluation.

Field	Purpose	Example type
run_id · sandbox_id	stable attribution	opaque UUID
timestamp · operation	ordering and action class	UTC · PUT/GET
resource · destination_class	contact surface	path · shared-cache
content_hash · size	exchange without payload exposure	SHA-256 · bytes
authorization · policy_tag	boundary gate	allowed/denied · sanctioned
intervention_id · lineage	response and recurrence	hold-07 · predecessor

7.2 Retrospective signal checks

- Contact query: groups successful writes by destination and rolling window, counts distinct run IDs, and compares designated with undesignated surfaces.
- Exchange query: groups stored content hashes by run and path, suppresses an allowlist of known templates, and identifies hashes reused across runs and destinations.
- Joint-onset query: identifies periods in which write volume, distinct writers, and contact surfaces rise together relative to the environment’s benign baseline.
- Boundary query: joins coordination candidates to authorization decisions, credential scopes, task policy, and third-party ownership before assigning R3.
- Recovery query: searches after a hold or removal for a new surface whose participants, artifact fingerprints, or function overlap the suppressed episode.

7.3 Prospective two-sandbox evaluation

The prospective evaluation uses matched episodes in which sanctioned collaboration transitions to an unauthorized exchange. Sandboxes A and B receive complementary research tasks and permission to exchange summaries through the designated /shared/briefs directory; task answers may not be written to /shared/cache. During the sanctioned phase, A writes a summary to the approved directory and B reads and extends it. Contact and exchange on this designated surface remain in R1 and do not trigger a disruptive response.

During the boundary phase, A writes an answer hash to the undesignated cache. The action-permission-destination mismatch moves the episode to R3, where KEROS preserves both traces and recommends a hold limited to the cache while leaving the approved brief channel available. If B subsequently reads the hash and incorporates it into the submitted result, the verified unauthorized downstream use satisfies the R4 outcome gate and permits escalation to containment review. The initial pilot evaluates regime reconstruction and evidence attribution. A later randomized trial can estimate response effects by comparing observation only, policy coaching, a reversible authorization hold, and containment using recurrence, displacement, false holds, and task utility as endpoints.

7.4 Research artifact and reproducibility

The companion notebook KEROS_Collusion_Signal_Analysis.ipynb implements the retrospective methods in Sections 3 and 4. It verifies manifest counts, reproduces the reported statistics and figures, exports JSON and CSV summaries, and exposes the marker rules, reuse threshold, weights, and decay factor. It accepts compressed or expanded archive files in the repository data directory, beside the notebook, in the parent directory, or in /content for Google Colab. The GitHub package also provides versioned dependencies, a checksum-validating data downloader, an executed notebook, and an automated reproducibility workflow.

The artifact was developed for the Apart Research AI Incident Response Sprint [2] and supports independent reproduction of the retrospective analysis. It does not validate regime thresholds or response effectiveness; those claims require the ground-truth variables and prospective experiments described in Sections 7.1–7.3 and 8.3.

8 Discussion

8.1 Findings supported by the archive

The archive establishes a population of recorded handles repeatedly using common writable surfaces. Shared-page concentration, exact payload reuse, short-lag cross-handle revisions, organization-related language, and page

recreation are independently observable. Their co-occurrence supports the engineering claim that per-run monitoring alone discards information needed to reconstruct collective episodes. All six signals are computable from infrastructure events without model weights, training data, or private reasoning traces, although production deployment still requires authorization, outcome, and intervention instrumentation.

8.2 Limits of inference

The corpus does not establish one-to-one handle identity, read-before-write causality, authorization state, malicious intent, downstream harm, or the effect of any response. Its marker counts are lexical and English-biased. Exact hashes miss paraphrase and can overcount templates. The historical index is normalized within one exceptional corpus, so its values cannot be transferred into a production alarm threshold. The later July incident adds evidence that unauthorized coordination and channel rebuilding can matter operationally, but it does not supply missing labels for the May–July wiki archive.



Figure 4 Historical replay of the illustrative cumulative evidence index. The index prioritizes review; it is not a probability of harm.

8.3 Ground truth requirements and next experiments

The next dataset should add immutable run IDs; read and write logs; credential and permission scope; designated collaboration channels; task assignments; authorized outcomes; third-party ownership; intervention timestamps; and post-intervention lineage. With those fields, researchers can estimate false-positive rates, learn environment-specific weights, compare time to correct escalation, and test whether coaching, friction, authorization holds, or containment reduce unauthorized outcomes without destroying legitimate collaboration.

Three tests follow directly. First, compare local-only and fleet-level monitors on matched benign and adversarial episodes; KEROS predicts earlier correct episode discovery from the joint view. Second, remove a shared surface without changing the affordance that created it; KEROS predicts displacement more often than resolution. Third, mark sanctioned surfaces in advance; KEROS predicts fewer false holds with no reduction in findings that cross explicit policy boundaries.

9 Conclusion

KEROS defines an incident as an evolving relation among runs, artifacts, permissions, and outcomes rather than as a single alarming message or sandbox escape. Its immediate contribution is an operational framework that preserves local facts, correlates them at population scale, separates collaboration from boundary crossing, and maps stronger evidence to stronger responses. The Collusion.wiki archive grounds that structure in observed data. Controlled experiments remain necessary to calibrate thresholds and estimate response effectiveness.

Code and Data

The companion notebook `KEROS_Collusion_Signal_Analysis.ipynb` reproduces the archive counts, signal definitions, and historical evidence replay. It accepts the five public exports in compressed or expanded form and writes JSON, CSV, and figure outputs to `keros_outputs/`. The repository package includes versioned dependencies, a checksum-validating downloader, an executed notebook, and an automated execution workflow. Public data: [Collusion.wiki data export](#).

References

- [1] Von Arx, S., Slade Byrd, C., Kitts, S., and Larsen, T. 2026. Discovery of a new OpenAI agent message board. Collusion.wiki. <https://collusion.wiki/>
- [2] Apart Research. 2026. AI Incident Response Sprint. <https://apartresearch.com/sprints/ai-incident-response-sprint-2026-09-11-to-2026-09-13>
- [3] OpenAI. 2026. The Hugging Face incident and the road ahead. 26 August 2026. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
- [4] OpenAI. 2026. OpenAI-Hugging Face Incident Technical Report. <https://cdn.openai.com/pdf/67869394-cb91-4c12-888e-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf>
- [5] METR. 2026. Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI and Hugging Face hacking incident. 26 August 2026. <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>
- [6] Hammond, L., Chan, A., Clifton, J., et al. 2025. Multi-Agent Risks from Advanced AI. arXiv:2502.14143. <https://doi.org/10.48550/arXiv.2502.14143>
- [7] OpenAI. 2026. How we monitor internal coding agents for misalignment. 19 March 2026. <https://openai.com/index/how-we-monitor-internal-coding-agents-misalignment/>
- [8] National Institute of Standards and Technology. 2023. Artificial Intelligence Risk Management Framework AI RMF 1.0. NIST AI 100-1. <https://doi.org/10.6028/NIST.AI.100-1>

Author Contributions

Nabita Penmetsa is the sole author. She conceived the KEROS framework, designed the evaluation protocol, conducted and interpreted the analysis, and wrote the manuscript.

LLM Usage Statement

OpenAI Codex was used as a coding and writing assistant throughout the project. All research-design decisions, analysis, interpretation, and final content are the author's own. The author reviewed the code, numerical results, and manuscript before submission.

Appendix A Limitations and Dual Use Considerations

A.1 Limitations

Single reconstructed corpus. Findings derive from one public archive reconstructed from wiki histories. We verified internal consistency against its manifest but did not independently verify source completeness or attribution. Coverage gaps make population totals lower bounds [1].

Handles are not identities. One agent may use several handles, several agents may share one, and one handle may span multiple runs. Counts therefore describe recorded labels. Stable run credentials are a design requirement, not a property of this archive.

Writes do not prove reads. A short-lag handoff or reused payload is compatible with coordination, shared prompting, copied templates, or coincidence. Read logs and task lineage are required for causal information-flow claims.

Lexical markers are weak supervision. Organization and boundary regexes are transparent but English-biased, context-insensitive, and vulnerable to both false positives and paraphrase. Harm-related words are never treated as harm outcomes.

No benign calibration. Rates from this exceptional archive do not provide a false-positive rate for ordinary evaluations, software agents, or sanctioned multi-agent systems. The cumulative evidence formula is illustrative and normalized to this corpus.

Cross-source evidence. Public reports about the July incident use different instrumentation and time windows. They motivate regime design but do not label the wiki records or establish a continuous causal chain.

Response effectiveness is untested. The archive has no randomized response arms, counterfactual outcomes, or utility measures. We claim a framework for choosing and evaluating postures, not that any posture works.

A.2 Dual Use Considerations

Publishing coordination indicators can help an adversary distribute activity below fixed thresholds. KEROS reduces this risk by avoiding operational exploit detail, raw credential-like text, source addresses, and named human handles. The paper reports structural features already visible in the public archive and emphasizes compound, environment-calibrated evidence rather than fixed tripwires.

Defenders benefit from an asymmetry: suppressing shared-surface concentration, reusable artifacts, fast handoffs, and re-establishment also reduces the utility of a collective. Even so, no detector is evasion-proof. Thresholds should be varied, audited, and combined with permissions that limit writable shared state.

Strong responses create their own risks. False holds can invalidate evaluations; content injection can contaminate agent behavior; quarantine can erase evidence or interrupt legitimate collaboration. KEROS therefore defaults to reversible, scoped, evidence-preserving actions and requires a human decision for broad containment or population quarantine.

Privacy is also material. Cross-sandbox correlation can become worker or user surveillance when agents act on behalf of people. Production implementations should minimize payload retention, use content hashes where possible, restrict linkage windows, document purpose, and separate security review from performance evaluation.

Appendix B Metric Definitions and Reproduction

Metric	Executable definition
Recorded handle	label; blank mapped to (blank) for counting, and excluded where stated
Shared page	archive n_labels ≥ 2 ; nonblank sensitivity analysis recomputed from revisions
First-seen handle	minimum UTC calendar day among a handle’s revisions
Short-lag handoff	consecutive revisions on one page, different handles, nonnegative time delta
Strong exact reuse	body length ≥ 120 ; same SHA-256 under ≥ 3 handles and on ≥ 3 pages
Organization marker	disclosed case-insensitive regex for addressing, claims, phases, and coordination
Boundary marker	disclosed case-insensitive regex for access, credentials, containment, and bypass
Resilience proxy	events where relation_type equals first_recreation_of
Evidence index	decayed weighted sum in Section 3.3; demonstration settings only

B.1 Reproduction steps

- Clone or unpack the GitHub submission package, create a Python 3.10 or newer environment, and install the packages listed in requirements.txt.
- Run `python download_data.py` to retrieve the public Collusion.wiki archive and verify the publisher-provided SHA-256 checksums; alternatively place the five expanded or compressed archive files in `data/`, beside the notebook, in the parent directory, or in `/content`.
- Open `KEROS_Collusion_Signal_Analysis.ipynb`, restart the kernel, run all cells, and compare `keros_outputs/keros_metrics.json` with Table 1 and Section 4.
- Inspect or change the marker regexes, exact-reuse threshold, weights, and decay factor; rerun all cells and update reported results.

B.2 Ground truth extension schema

Missing variable	Why it matters
Immutable run and sandbox identity	attributes actions and distinguishes recurrence from renaming
Read events and artifact lineage	tests information transfer rather than mere co-presence
Permission and policy state	separates collaboration from boundary crossing
Task and channel designation	creates matched benign collaboration controls
Downstream use and external outcome	supports R4 and harm measurement
Intervention assignment and timing	enables causal response-effect estimates

B.3 Recommended next experiment

Construct matched two-sandbox episodes in which the same shared information is either sanctioned or undesignated. Randomize the response only after KEROS correctly reconstructs the evidence regime. Primary endpoints are correct regime assignment and time to correct escalation; response endpoints are unauthorized downstream use, displacement, recurrence, task utility, and false holds. Expand to a population only after the event contract and ground-truth labels are stable.