

What a Collusion Audit Can Verify: Source-Bound Evidence and Two Stopped Monitor Pilots

Author: Bao Bowen

Affiliation: Independent Researcher

Team: ColluScope

Track: Apart AI Collusion Research Sprint 2026, Track 2.3.

Status: Submission candidate, 25 October 2026 (Asia/Shanghai). Includes the completed, stopped v2 monitor pilot. Official template styling and live submission-portal status remain unverified.

Abstract

What evidence makes a multi-agent collusion claim defensible? Communication, aligned actions and lower collective reward can coexist without identifying agreement, intent or causation. We study a narrower prerequisite: whether an executable auditor verifies cited source fields and respects evidence permissions. Using saved cooperative-agent traces, we compare a typed-field auditor, the same auditor with authoritative source binding, and an always-abstain baseline. A separately authored challenge contains 120 cases from six fixed source groups and four synthetic rounding boundaries. Predictions are frozen before a separate reference key is opened. The unbound auditor accepts all 12 constructed actor or numeric source substitutions; source binding accepts none while retaining all 38 clean supported controls, representing 32 distinct input profiles. Both disclose no denied reference in six tested views. A separate agent-context verifier reconstructs the source projection and every reference outcome. Two live monitor development pilots stop before their conditional main stages: the first has 11/12 valid responses with contradictory fields; the revised pilot has 8/12 valid responses because four replies change an opaque claim identifier. Neither produces a held-out efficacy estimate. These results validate a finite software contract and expose measurement failures. They do not establish naturalistic detector accuracy, semantic agreement, collusion, intent or causal communication effects. We provide a sanitized zero-network replay and preserve invalid outputs, complete costs and the original scorer error.

1. Research question and relation to Track 2.3

Track 2.3 asks what constitutes convincing evidence of collusion, distinguishing parallel behavior from agreement and suggesting the antitrust plus-factors analogy [1]. ColluScope addresses a prerequisite for such judgments: can a reviewer trace each factual assertion to authorized, unaltered source fields, and can the workflow explicitly withhold conclusions beyond those fields? We distinguish source identity, observable fact, semantic agreement and causal attribution. Passing a check at one level does not establish the next.

Our contribution is a runnable, adversarially inspected source contract and a record of measurement defects from two stopped live monitor pilots. Hash binding, access control, typed predicates and abstention are standard techniques. We claim their transparent application to a multi-agent evidence workflow, with source-derived counterexamples, an implemented baseline, isolated reference keys, failure retention and a portable replay. We do not introduce a new collusion detector or a general semantic verifier.

2. Data and evidence levels

The underlying normalized corpus contains 29 valid saved Jira pair groups: 58 agent episodes, 2,099 messages, 348 actions and 58 settlements. All reuse environment seed 1. A private-channel condition and its no-private-channel counterpart are paired, but the latter still has public communication. This is not a no-communication control. Treatment assignment and reward loss are not collusion labels. Semantic review of 245 private-board messages is incomplete, and no agreement labels are derived from those messages.

An illustrative development case, Pair 25, has nine private messages preceding six successful actions. Relative to its paired control, system reward differs by -15.324586 and the regret gap by -8.011494. These are logged and calculated observations. They do not establish that the messages caused the outcomes; shared incentives and independent responses remain competing mechanisms. The source-contract challenge uses six different fixed pair groups, 4, 5, 12, 17, 23 and 28. These groups were held out from the new monitor's outcome tuning, but the historical source data have

been inspected by analysts. They are not analyst-blind or independent-seed cases.

We keep four evidence levels separate: constructed software controls; real-agent historical behavior; a live monitor development/output-contract pilot; and held-out monitor performance. The present central result is at the first level, using fields projected from the second. The pilot contains real monitor outputs but does not supply held-out efficacy evidence.

3. Methods

3.1 Typed facts, source trust and permissions

The unchanged auditor computes typed propositions concerning actors, task execution, event ordering and rounded numeric values from supplied records. It can establish consistency with those supplied records. It cannot determine whether a substituted record is authentic if the supplied ledger is its only trust root.

The source-bound method checks each presented record against an independently trusted projection, including exact canonical bytes and authoritative visibility, then invokes the same frozen auditor. A missing, altered or unauthorized record cannot support a claim. Denied-record references are not echoed in results. This adds an explicit trusted input: a compromised original trace or compromised projection can pass every hash check. Hashes establish byte identity, not truth or original-source authenticity. Exact binding also rejects benign source changes; availability under harmless representation drift is not evaluated.

Methods receive typed claims, projected raw fields and the visibility view. They receive no answer labels, support directions or validity flags. Message bodies are not part of this challenge. Causal claims receive an explicit non-identifiability policy outcome when the available design lacks an identifying intervention. This is an implemented policy, not evidence of causal reasoning.

3.2 Preregistered challenge and baselines

Before bank construction, the registered protocol fixed 20 challenge families per source group: true and false sender; true action, altered actor and altered task; correct and reversed ordering; true and false rounded reward and reward difference; missing reference; stale citation; missing field; denied public access to a private record; permitted privileged access; causal-policy abstention; duplicate identifier; actor source substitution; and numeric source substitution. Six groups yield 120 cases. Four explicitly synthetic half-rounding boundary cases bring the assigned total to 124.

A separate agent context authored the bank without editing the auditor. Prediction ran with access to the reference key and challenge generator denied; prediction hashes were saved before scoring. All three methods attempted the same 124 cases: unbound auditor, source-bound auditor and always-INSUFFICIENT_EVIDENCE. Primary outcomes are acceptance of the 12 source-substitution cases and retention of the 38 clean supported control rows. There are 32 distinct clean input profiles because two permission families duplicate clean propositions within each source group. These counts are finite contract coverage, not independent statistical trials.

The registered complete gate requires zero source-substitution acceptances by the bound method, at least one by the unbound method, retention of every clean control, zero denied-reference disclosures and intact source/projection hashes. We report all assigned cases, missingness, exact status agreement and the abstention baseline. We perform no population hypothesis test.

3.3 Independent verification and adverse evidence

A second reviewer context initially inspected the protocol and implementation without the bank or outcomes. After predictions were frozen, it reconstructed 32 projected records, all 124 references and all 372 predicted statuses using separate rational arithmetic and source joins, importing none of the auditor, wrapper, runner or generator implementations. This is agent-context independence within the same model family, not human or cross-model validation.

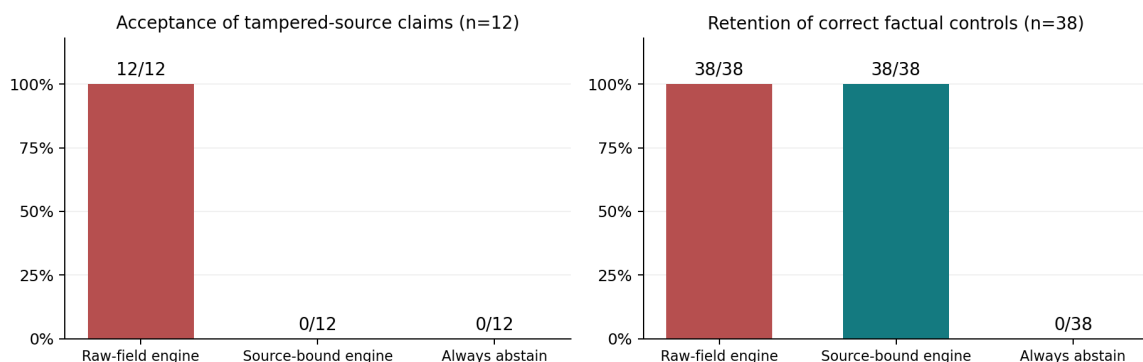
The frozen scorer contains an aggregate naming reversal: it assigns the first method row to the bound method although that row is the unbound method. Consequently the original summary says REFUTED while its method-specific counts are correct. We preserve that summary and the scorer unchanged. The independent audit applies the original written gate to correctly named rows and supports the finite contract hypothesis. The closure explicitly cites this reconciliation; we do not obtain a pass by editing the frozen evaluator.

4. Results

All 372 assigned method-case outputs were attempted, valid and assessable: 124 per method, with zero missing, failed or unstarted outputs.

Outcome	Unbound auditor	Source-bound auditor	Always abstain
Source substitutions accepted	12/12	0/12	0/12
Clean supported controls retained	38/38	38/38	0/38
Exact reference-status matches	112/124	124/124	36/124
Unsupported cases accepted	12/86	0/86	0/86
Exact denied-reference disclosures	0/6	0/6	0/6

Typed source-contract challenge: 124 authored cases, zero model calls



Six saved pair groups share one environment seed; four boundary cases are synthetic. These are software-contract counts.

Finite source-contract results; exact denominators, with no population error-rate interpretation.

The unbound auditor accepts the constructed substitutions because the altered claim agrees with the altered ledger. Source binding rejects the changed bytes while preserving exact clean controls. Always abstaining also avoids false acceptance, but retains no supported facts. Neither source binding nor the unbound auditor leaks an exact denied reference in the six assigned permission tests; these tests establish no incremental privacy advantage for source binding.

The result is partly expected by design: an external trusted digest check detects changed inputs. Its value is an executable boundary and a tested implementation, including the competent unbound baseline and explicit trust assumption. The challenge does not estimate naturally occurring substitution frequency, robustness to compromised trusted records, semantic entailment or real-world utility.

5. Two stopped monitor pilots and measurement failure

5.1 Original pilot and retrospective contract audit

The original four-arm monitor pilot assigned 12 development requests, all returning HTTP 200. Under its recorded exact-claim-echo compatibility amendment, 11 were valid. One q7 judgment contains status CONTRADICTED and an added note stating SUPPORTED. The fixed technical gate fails; the 24 conditional main requests never start. A retrospective audit checks all 36 scheduled requests across nine pair groups: eight of nine q7 exact literal values differ from the full raw subtraction while matching an implicit 12-place rounding convention. All 36 common system prompts also contain a temporal-causality instruction. We preserve these claims and labels; their nominal factual scores do not supply a clean efficacy benchmark.

The old summary omits the invalid reply's usage estimate of USD 0.001508352. All 12 raw usage records at their recorded historical rates sum to USD 0.054851448, versus USD 0.053343096 in that summary. Estimates are not invoices; old and current rates are not interchangeable.

5.2 Approved, preregistered v2 and its stopping result

Before new outputs, v2 fixes six-place HALF_UP arithmetic from full raw operands, opaque claim IDs, separate keys, a minimal output schema and assigned-error denominators. The common system drops its explicit temporal-causality sentence; checklist arms retain it, so q9 remains secondary instruction behavior. A separate verifier checks all 60 future requests and 540 reference slots, including 120 numeric slots. Four arms receive identical evidence and claim texts per group: flat versus ledger layout, crossed with checklist absence/presence. Conditional main design is six fixed groups times four arms times two byte-identical, interleaved repeats: 48 requests. The 12 development requests use three other groups. Prior source inspection, one seed, instruction wording and token differences remain explicit limits.

The new scope receives a fresh 60-attempt/USD 2 approval and a separate registration. Its fixed technical gate requires at least 11/12 fully valid replies and at least 2/3 per arm, plus stable identity, complete usage and source/privacy integrity. Gate evaluation reads no reference outcome or effect direction. A duplicate/unknown/missing claim ID invalidates the whole response under the frozen rule.

All 12 development attempts return HTTP 200 with known usage, stable fingerprint and normal completion. Eight are fully valid; four are not. Every arm has 2/3 valid replies. All four invalid replies concern pair 31 and each changes the same required q2 opaque identifier, returning nine distinct IDs of which eight match the request. The aggregate 8/12 fails the 11/12 gate. Zero of 48 main requests start; there are no retries, replacements, identifier repairs or gate amendments.

Development outcome	Original pilot v1	Revised pilot v2
Assigned / attempted development requests	12/12	12/12
Fully valid development replies	11/12	8/12
Conditional main assigned / started	24/0	48/0
Technical expansion	Failed	Failed
Held-out representation/repeat effect	Unestimable	Unestimable
Usage-based estimate, USD	0.054851448	0.048232128

Monitor v2 pilot: all four schema failures share one source group
8/12 fully valid; fixed expansion gate requires 11/12

Pair 31	Unknown claim ID	Unknown claim ID	Unknown claim ID	Unknown claim ID
Pair 20	Valid envelope	Valid envelope	Valid envelope	Valid envelope
Pair 32	Valid envelope	Valid envelope	Valid envelope	Valid envelope
	C0	C1	D0	D1

12 requests attempted; 48 conditional main requests unstarted. No efficacy contrast.

Envelope validity is not factual accuracy. Three seed-1 source groups; no independent-environment inference.

V2 pilot output-contract validity by source group and arm; validity is not factual accuracy.

These pilots use different contracts and recorded pricing conventions; neither validity nor cost columns estimate a treatment effect across versions. V2's 96 assigned development factual slots contain 44 errors: 32 invalid slots and 12 valid-but-wrong judgments. One wrong ID invalidates all eight factual slots in a reply, so those 32 invalid slots must not be described as 32 false beliefs. Of 48 assigned supported facts, 24 are retained; 3 of 48 authored contradicted facts are marked SUPPORTED. Required-reference coverage is 60/64 valid factual outputs, which is syntactic coverage rather than entailment. These are development diagnostics only. Main primary and factorial contrasts are null/unestimable; unstarted outputs are not observations of model efficacy. The eight valid q9 judgments use the prescribed NOT_IDENTIFIABLE status, which does not demonstrate causal reasoning.

V2's complete conservative peak usage estimate is USD 0.048232128, including all four invalid replies, with zero unknown-billing reserve. The unused call slots and budget do not authorize another study after the failed gate. The returned alias/fingerprint identifies a provider-managed serving configuration, not immutable weights. No main repeat

observation exists.

The frozen analyzer verifies source/code/key/request/response hashes, alternating attempt receipts, assigned matrices, admitted-output coverage and independently derived completeness. A separately implemented raw JSON/usage calculation reproduces the v2 pilot counters and costs. Source/reference and software checks were independently reviewed before inference; live-outcome review is author-only, because no fresh independent reviewer was available. No additional independent verification credit is claimed. Initial credential-name failure occurred before transport; a reviewed process-environment adapter supplied the existing official-provider key without changing frozen requests. A review-created public directory later caused export failure after private results were saved; publication recovery copies those existing bytes without resending a request. Both failures remain archived.

6. Related work and contribution boundary

Colosseum supplies the cooperative environment and compares action-based regret with communication behavior, including plans that fail to become realized actions [2]. We inherit these traces and metrics. Arike et al. study information access for sabotage monitoring and an extract-and-evaluate approach [3]; evidence organization is established prior work. ALCE evaluates answer correctness and citation quality [4]; a resolving identifier alone is insufficient semantic support. ColluScope's tested increment is a modest source/permission contract for typed trace assertions, together with transparent negative results and a reference-isolated replay. The closest alternative is a conventional trusted-digest check followed by the same predicates, which would provide the same tested protection. We make no broader algorithmic novelty claim.

7. Reproducibility and sprint disclosure

The portable fixture runs without network, model calls, private logs or an answer key:

```
``text python examples/evidence-audit/source_bound_contract_v1/replay.py ``
```

The second public fixture replays the saved v2 output-contract failure and its complete peak usage estimate:

```
``text python examples/evidence-audit/monitor_schema_pilot_v2/replay.py ``
```

It retains the exact 12 model judgment-content strings, after checking all 108 claim outputs contain only opaque claim IDs, status enums and opaque citation IDs. Input projections contain only the expected claim/reference IDs, with all claim texts, evidence bodies and semantic references omitted. This reproduces 8/12 schema validity and the failed gate without revealing private messages or scoring semantic accuracy. Hash links to restricted original envelopes establish recorded byte identity, not provider attestation.

The source-contract fixture reproduces all 372 status outputs from pseudonymized requests and trusted records. Original identifiers, names, task names and timestamps are replaced; numeric operands remain to preserve arithmetic. Its expected status manifest checks output identity, not independent semantic truth. The full independent source audit requires restricted local originals; this public fixture alone cannot authenticate them. Export manifests record hashes, an initial conservative privacy-scan failure and its field-aware adjudication. No claim of general anonymization or privacy noninterference follows.

Focused source-binding tests and prospective-contract tests are provided; the final targeted v2 contract/execution/analysis suite passes 23 tests. A clean extraction of the review bundle reproduces all 372 portable statuses using isolated Python and no network. Repository-wide tests are not uniformly green: the prior full run had 789 passes, 11 failures and one skip, including historical hash and lifecycle assumptions. The registry audit retains three known historical mutable-file hash mismatches. These are disclosed separately from the new study's source checks. Public artifacts contain predictions and sanitized counts; raw agent messages and provider replies remain local under access controls.

The M0-M4 infrastructure, historical Colosseum corpus, normalized 29-pair dataset, split and original stopped monitor pilot predate this report version. The sprint contribution is the source-contract scope correction, wrapper, separately authored challenge, independent reconstruction, portable replay, measurement-contract repair and newly approved stopped v2 pilot. Main-stage efficacy remains unresolved. The code/result snapshot is commit 084acbf; BUNDLE_MANIFEST.json records its full hash and separate publication hashes. AI agents assisted code, challenge construction, review and writing. The submitting humans remain responsible for authorship, rights, source verification and claims. Earlier four-case/five-policy numbers remain archived as a simulation consuming authored support-direction

annotations; they are not an evaluation of the raw-field auditor and are excluded from the central result.

8. Conclusion

An auditor can correctly calculate a proposition against a ledger that has itself been altered. On this fixed challenge, explicit authoritative source binding closes that tested gap without losing exact clean facts; abstention alone loses all supported facts. This supplies a reproducible prerequisite for evaluating collusion evidence. It does not turn logged alignment, private discussion or reward loss into evidence of agreement, intent or causation. Both stopped monitor pilots and their distinct measurement defects remain part of the result. No held-out monitor efficacy claim follows, and the next delivery step is a faithful submission synthesis rather than further paid schema remedies.

References

- [1] Apart Research. AI Collusion Research Sprint, 23-25 October 2026, Track 2.3 and submission guidance. [Official sprint page](#). [Hong Kong hub](#). Accessed 23 October 2026.
- [2] Nakamura et al. Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems. arXiv:2602.15198v2, 2026. [Primary paper](#).
- [3] Arike et al. How does information access affect LLM monitors' ability to detect sabotage? ICML 2026, PMLR 306:3473-3526. [Proceedings](#).
- [4] Gao et al. Enabling Large Language Models to Generate Text with Citations. EMNLP 2023. [Primary paper](#).

Appendix A. Limitations and Dual-Use / Ethical Considerations

Limitations

The source contract requires a trusted projection; it cannot authenticate the original recording process or discover truth from malicious trusted records. Constructed attacks and exact clean controls do not estimate prevalence or resilience under benign drift. The six source groups share one seed, two attack families share records, and some clean inputs repeat. Four rounding cases are fully synthetic. Message-body semantics and commitment/action agreement are unreviewed. Causal-policy outcomes are prescribed abstention. Six denied-reference tests exclude side channels, paraphrases, adversarial identifiers and general noninterference. The independent reviewers use separate contexts within one model family. Both live pilots stopped before their conditional main stages. Old rounding ambiguity and new identifier-copy failures prevent a held-out efficacy comparison. Whole-envelope invalidity magnifies a single identifier error into eight invalid factual slots, limiting semantic interpretation of the error metric. No result supports collusion detection, intent attribution, causal communication effects or general monitor superiority.

Dual-use and ethical considerations

Source and permission audits can reduce unsupported accusations but can also help an adversary learn which visible inconsistencies an auditor detects. The release therefore focuses on pseudonymized source-derived typed records, explicit synthetic controls, field substitutions and documented limitations. It includes no credentials, personal banking data, live financial deployment, private raw agent messages or provider headers. A well-formed source record must not be used as a certificate of benign behavior; agents may coordinate through unobserved channels, and legitimate cooperation can resemble suspicious alignment. The artifact is intended for research review with human responsibility, evidence-specific claims and explicit abstention.