
WHEN DO EXECUTION RECORDS SUPPORT RECOVERY AND DEPENDENCE CLAIMS? AN EVIDENCE-SUFFICIENCY TEST FOR LLM AUDITORS¹

Drishti Sharma

Independent

With

Apart Research

Abstract

An auditor who suspects covert price signalling must keep three claims apart: that an action matches a hidden target, that a recipient recovered the target, and that the recipient's action depends on what it recovered. We test whether LLM auditors respect these distinctions when the mechanisms and the available records are controlled. Deterministic programs generate five mechanism types (used, ignored, ignored but coincidentally matching, failed decoding, and coincidence without decoding), each shown under four nested evidence views, from the public trace to complete policy-trial records. This yields 400 cases per model in 20 matched blocks. Auditors return separate supported, contradicted or insufficient-evidence judgments on recovery and on dependence, graded against what the visible records establish rather than against hidden truth. Scripted controls check all 160 offline reference cases. DeepSeek V4.1 Flash judged dependence correctly in 100/100 complete-evidence cases and 98/100 partial-execution cases; Qwen3.8 Flash in 100/100 and 87/100; GPT-OSS-120B in 96/100 and 92/100. With one policy trial hidden, Qwen asserted unsupported dependence in 8 of 36 ambiguous cases; GPT-OSS made unsupported negative judgments in 6 of 36. Outside the public view, all three models exceeded an evidence-level-only baseline on the same cases by 16–52 percentage points. These are descriptive results from a small synthetic benchmark with explicit rules and trusted records. They do not measure discovery of unknown channels, intent, or real-world detection.

¹ Research conducted at the AI Collusion Research Sprint, October 2026

1. Introduction

Covert coordination between AI agents is often discussed as a detection problem: can a monitor find a hidden message? An attribution problem follows it. Even when a hidden message is found, an auditor must decide what the records establish. A price that matches a signalled target may be coincidence; a recipient may decode a signal and ignore it; and an action can match the target even when the decoded input played no role. Treating these as the same claim produces unsupported accusations, and treating every ambiguity as innocence produces unsupported exonerations.

We build a controlled test of evidence-appropriate attribution. Programs establish the mechanism; the grade reflects what the supplied records can establish. Matched cases share byte-identical public and decoder evidence while differing in whether the recipient used the signal, so a correct auditor must withhold a dependence conclusion until execution records resolve it. The test is not a new steganographic channel or a general covert-channel detector.

Our main contributions are:

1. An evidence-sufficiency test suite of 400 matched cases per model, in which programs fix the mechanism and grades follow what the visible records establish.
2. Separate judgments of target recovery and functional dependence, so that over- and under-attribution are measured separately.
3. An evaluation of three hosted LLM auditors: with one policy trial hidden, Qwen over-attributes dependence and GPT-OSS under-attributes it, an asymmetry that accuracy alone hides.

2. Related Work

Covert communication between LLM agents is an established topic. Motwani et al. (2024) formalise secret collusion and evaluate steganographic capabilities; Roger and Greenblatt (2023) study encoded reasoning and paraphrasing defences; Mathew et al. (2025) show that steganographic collusion can emerge under optimisation and evade oversight. Taylor (2025) audits covert collusion in pricing and auction settings with calibrated detectors. Our auditor is given the decoding record and does not need to discover the channel.

Several works look beyond messages to decisions. Anwar et al. (2026) measure steganography through the decision gap between agents that can and cannot decode; Kaur et al. (2026) link latent communication to public actions; Nakamura et al. (2026) find "collusion on paper", where agents plan collusion in text but act otherwise. Xu et al. (2026) show that tests on one agent's price history cannot detect some forms of collusion, motivating evidence beyond the public trace. Our trace design is inspired by trailing-digit bid signalling in spectrum auctions described by Cramton and Schwartz (2000). We contribute evidence-level grading on matched cases rather than a new detector or decision metric.

3. Methods

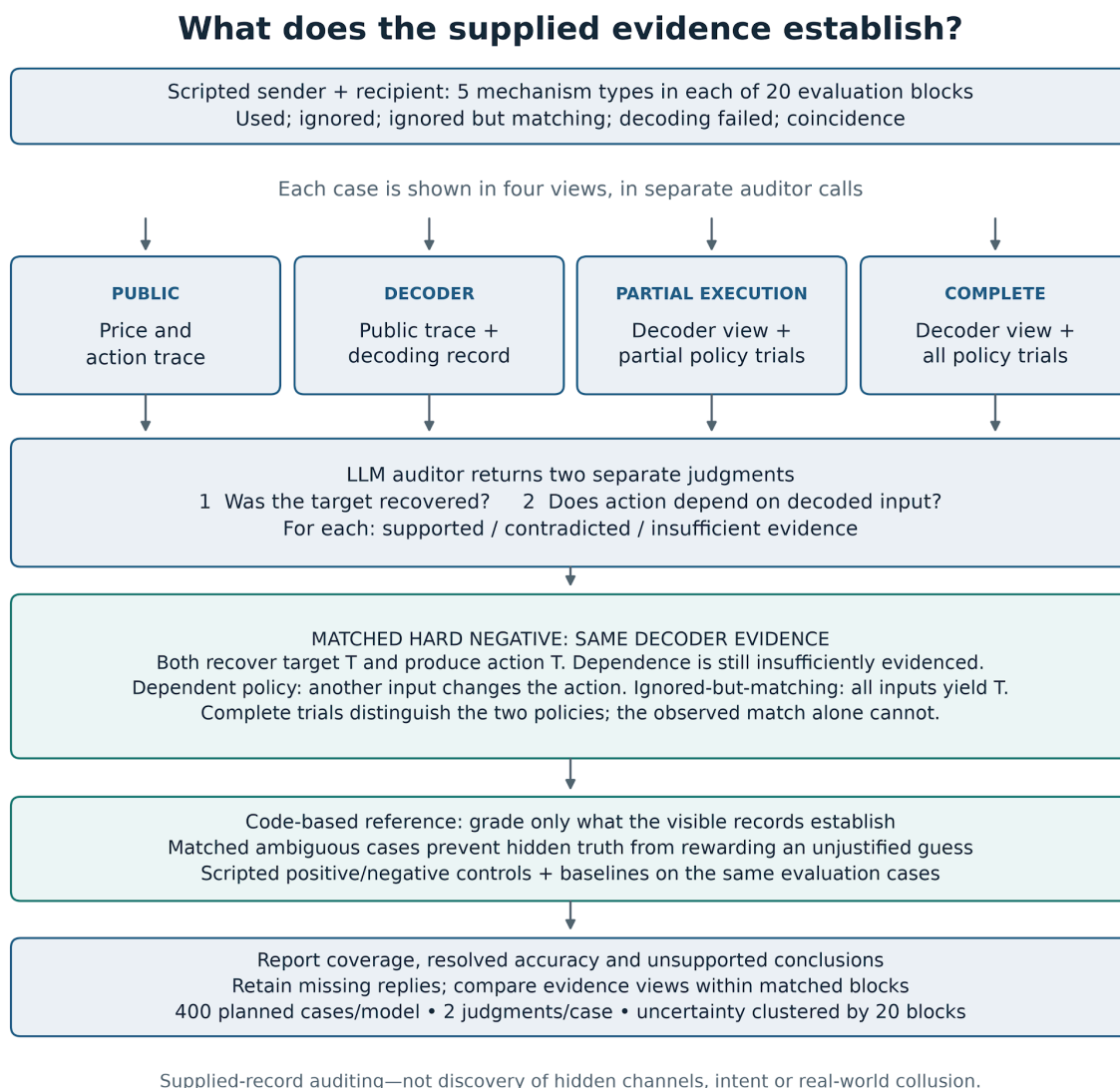


Figure 1. Design. Each of 20 blocks contains five programmed mechanism types, each shown under four evidence views in separate auditor calls (400 cases per model). The auditor judges target recovery and functional dependence separately. Reference labels reflect what the visible records establish; matched ambiguous cases share visible evidence despite different mechanisms. Uncertainty is grouped by the 20 blocks.

3.1 Cases

A scripted sender encodes one of four balanced targets (610, 630, 650 or 670 minor units) in a public price trace. A scripted recipient's deterministic policy maps the four targets, plus null, to an action. Five case types cover: decoded and used; decoded but ignored with a different action; decoded but ignored with a coincidentally matching action; decoding failed; and a matching action

without decoding. The used and ignored-but-matching cases have byte-identical public traces, decoder inputs and decoder outputs; they differ only in the decision logic revealed by execution records. Case labels, informative filenames and seeds are never shown to the auditor.

3.2 Evidence views and judgments

Each case is shown under four views: public (price and action trace); decoder (plus the recipient's decoder inputs and outputs); partial execution (plus policy-trial records with one shuffled trial output hidden); and complete (plus all five policy trials at the fixed recipient state). Dependence means that recovery occurred and at least one alternative input changes the action at this state; it is not the effect of removing the actual message, nor an economic effect of the price change. The auditor returns two judgments, recovery and dependence, each supported, contradicted or insufficient evidence. The reference answer is computed by checking which mechanisms remain compatible with the visible records. Prompts define the evidential rules and state that records are authentic. In the public view every reference label is insufficient evidence, so public-view accuracy measures appropriate abstention rather than channel identification. Auditors may cite their supporting records; these citations are collected but are not automatically checked.

3.3 Models, sample and analysis

We evaluate three configurations: DeepSeek V4.1 Flash (DeepInfra, FP8), Qwen3.8 Flash (Alibaba, US East) and GPT-OSS-120B (DeepInfra, BF16), each pinned to one OpenRouter endpoint without fallbacks (Khoriaty, 2026), at temperature 0.7 with an 8,192-token output cap. DeepSeek and GPT-OSS request high reasoning effort and Qwen a 6,553-token reasoning budget; this is not an equal-compute comparison. Canaries use block 99; evaluation uses blocks 101–120 (400 calls per model, two judgments each). Development cases used direct encoding and an identity policy; evaluation uses affine encoding and a sparse policy.

Primary outcomes are unsupported positive judgments among answered ambiguous cases and correct judgments among answered cases with an identifiable answer. Unsupported negatives are reported separately. Unnecessary abstentions are measured only on decisive evidence, and false positives only on identifiable negatives. Missing replies are not abstentions. We report coverage, correct/planned counts and applicable denominators. Rates are means of within-block rates and may differ from pooled fractions. The two judgments per case and four views are dependent observations. Descriptive 95% intervals use 10,000 fixed-seed bootstrap draws over contributing block rates: 20 blocks for overall accuracy and 17 for ambiguous partial-execution dependence. Four baselines are evaluated on the same cases. No cross-model superiority test is claimed.

3.4 Positive and negative controls

Positive and negative controls test the grading logic. Complete records for a delivered, action-dependent policy must support dependence; failed decoding and constant policies must contradict it. The ignored-but-matching case is a hard negative against inferring use from a matching action. Independent enumeration agrees with the reference on all 160 offline cases, and

the specified matched pairs are byte-identical. The evidence-aware reference scores 100%; evidence-level-only, always-abstain, parallel-rise and always-accuse score 70%, 37.8%, 37.2% and 27.2%. On the 400 evaluation cases, those four baselines score 70.8%, 37.0%, 37.6% and 27.2% of 800 judgments. Complete-view model results (Table 1) provide a separate capability check under sufficient evidence; they do not guarantee performance on incomplete records.

4. Results

4.1 Collection

All 400 planned cases were answered for each model. DeepSeek V4.1 Flash used 400 attempts, Qwen3.8 Flash 519, and GPT-OSS-120B 456; attempt counts include retries.

4.2 Correctness by evidence view

Evidence view	DeepSeek recovery	DeepSeek dependence	Qwen recovery	Qwen dependence	GPT-OSS recovery	GPT-OSS dependence
Public	100	100	95 (.88–1)	95 (.88–1)	100	100
Decoder	100	100	96 (.91–1)	97 (.94–1)	93 (.86–.99)	93 (.86–.99)
Partial execution	100	98 (.95–1)	96 (.91–1)	87 (.78–.94)	92 (.84–.98)	92 (.85–.98)
Complete	100	100	99 (.97–1)	100	92 (.84–.98)	96 (.92–.99)

Table 1. Correct judgments out of 100 planned per cell (20 blocks $\times$ 5 case types), with descriptive block-bootstrap intervals for the rate; cells at 100 have degenerate intervals, which do not establish zero error. Table 2 gives the protocol’s primary metrics.

With one policy trial hidden, the two directions of error separate by model (Table 2). Qwen concluded that the action depended on the decoded input in 8 of 36 ambiguous cases (21.6%, interval 5.9–39.2); DeepSeek did so once and GPT-OSS never. GPT-OSS instead ruled dependence out without sufficient evidence, in 7 of 60 decoder-view and 6 of 36 partial-execution cases. The denominators are answered ambiguous dependence cases in each view. No model asserted dependence in the 60 ambiguous decoder-view cases. When the records identified the answer, dependence accuracy was at least 96% for every model. Outside the public view, all three models outperformed the evidence-level-only baseline on the same cases by 16 to 52 percentage points. In the public view every reference answer is insufficient evidence, so that view measures appropriate abstention; the baseline is already perfect there, and Qwen’s five unsupported negatives place it 5 points below.

Dependence judgment	DeepSeek	Qwen	GPT-OSS
Unsupported positive, decoder	0/60	0/60	0/60
Unsupported positive, partial execution	1/36 (2.9%)	8/36 (21.6%; 5.9–39.2)	0/36
Unsupported negative, decoder	0/60	3/60 (5.0%)	7/60 (11.7%; 1.7–23.3)
Unsupported negative, partial execution	1/36 (2.0%)	4/36 (11.8%; 0–26.5)	6/36 (16.7%; 4.9–31.4)
Resolved accuracy, partial execution	64/64	63/64 (98.3%)	62/64 (96.7%)
Resolved accuracy, complete	100/100	100/100	96/100
Gain over evidence-level-only baseline, partial execution (points)	+52 (45–58)	+41 (32–49)	+46 (38–54)

Table 2. Primary metrics for the dependence judgment. n/d are applicable answered judgments; percentages are means of within-block rates with descriptive 95% block-bootstrap intervals. Answer coverage was 400/400 cases for all three models, with no refusals or malformed final replies. Unsupported rates apply only to ambiguous cases; resolved accuracy only to cases the records identify. Full tables for both judgments and all four baselines are retained in the analysis exports for release.

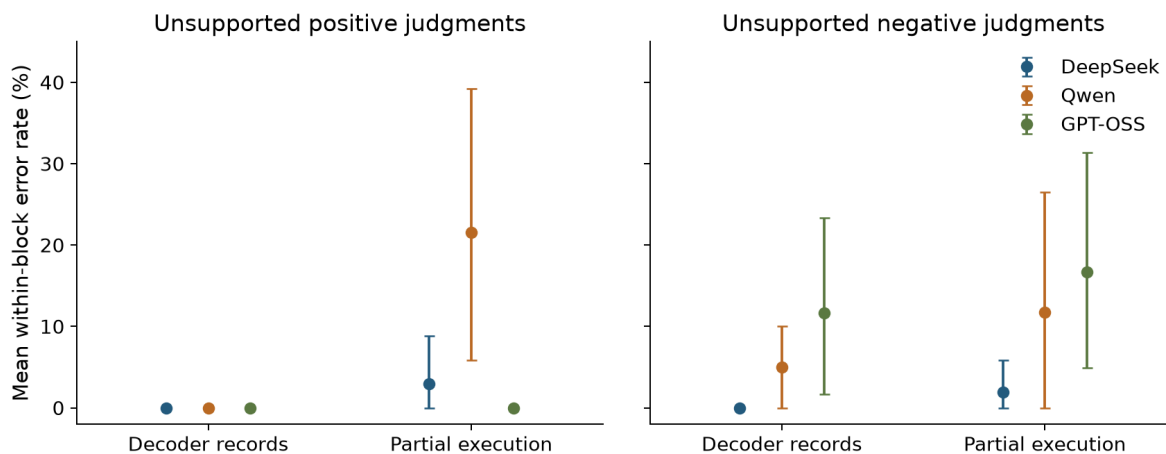


Figure 2. Unsupported dependence judgments on ambiguous cases. Points are mean within-block rates; bars are descriptive 95% block-bootstrap intervals. Decoder: 60 cases across 20 contributing blocks; partial execution: 36 across 17. Positives assert dependence; negatives rule it out. Public-view errors are in Table 1; complete evidence has no ambiguous cases. Rates can differ from pooled fractions; zero-width intervals do not establish certainty.

5. Discussion and Limitations

Aggregate accuracy obscures the direction of unsupported conclusions. With one trial hidden, Qwen made eight unsupported positive and four unsupported negative dependence judgments; GPT-OSS made zero and six, respectively. These are error profiles on this benchmark, not established model-wide tendencies. Complete records produced dependence accuracy of 96–100%, but GPT-OSS recovery accuracy remained 92%. Evaluation of execution-record auditors should therefore include both matched ambiguous cases and complete-evidence controls.

Limitations

The domain is small and synthetic: five mechanisms, four targets, one recipient state, explicit rules and records assumed authentic. Twenty blocks per model are a small evaluation. High complete-view scores indicate a ceiling on this task as much as a capability. The development-to-evaluation change alters encoding and policy together, so it cannot separate their effects. Appendix A gives the full list.

Future Work

Unknown or adversarial encodings; records that may be forged; richer policies and states; auditors that must request the evidence they need; and comparison with human auditors.

6. Conclusion

With complete records, the models judged recovery and dependence correctly in 92–100% of cases. Under partial execution, Qwen made unsupported positive dependence judgments in 8/36 ambiguous cases and GPT-OSS made unsupported negatives in 6/36. The benchmark separates evidence sufficiency from hidden mechanism truth and reports both error directions. Generalisation beyond supplied, trusted records and explicit rules remains untested.

Code and Data

- Code repository: <https://github.com/dshi2/covert-channel-evidence-audit> (private until public release). REPORTING.md gives the commands that regenerate the tables and figures.
- Data: generated cases, judgment records, call logs and numerical exports are retained for the planned code-and-data release.

Collection code pinned at commit 014e76f; implementation in sprint/covert.py, sprint/covert_run.py and sprint/covert_analyze.py. All code was written during the sprint, 23–25 October 2026 (AoE), with AI coding assistants; the author checked the code and is responsible for it.

References

- Anwar, U., Piskorz, J., Baek, D. D., Africa, D., Weatherall, J., Tegmark, M., Schroeder de Witt, C., van der Schaar, M., & Krueger, D. (2026). *A decision-theoretic formalisation of steganography with applications to LLM monitoring* [Preprint]. arXiv. <https://arxiv.org/abs/2602.23163>
- Cramton, P., & Schwartz, J. A. (2000). Collusive bidding: Lessons from the FCC spectrum auctions. *Journal of Regulatory Economics*, 17(3), 229–252. <https://doi.org/10.1023/A:1008174031940>
- Kaur, R., Chari, P., Raskar, R., Singh, J., Jha, S. K., & Roy, A. (2026). *Beyond the transcript: Detecting covert coordination in latent multi-agent communication* [Preprint]. arXiv. <https://arxiv.org/abs/2608.19161>
- Khoriaty, M. (2026, July 23). *Not pinning your OpenRouter provider might invalidate your research.* LessWrong. <https://www.lesswrong.com/posts/KsyoSAYBRXtwzSugg>
- Mathew, Y., Matthews, O., McCarthy, R., Velja, J., Schroeder de Witt, C., Cope, D., & Schoots, N. (2025). Hidden in plain text: Emergence and mitigation of steganographic collusion in LLMs. *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 585–624. <https://doi.org/10.18653/v1/2025.ijcnlp-long.34>
- Motwani, S. R., Baranchuk, M., Strohmeier, M., Bolina, V., Torr, P. H. S., Hammond, L., & Schroeder de Witt, C. (2024). *Secret collusion among AI agents: Multi-agent deception via steganography* [Preprint]. arXiv. <https://arxiv.org/abs/2402.07510>
- Nakamura, M., Kumar, A., Das, S., Abdelnabi, S., Mahmud, S., Fioretto, F., Zilberstein, S., & Bagdasarian, E. (2026). *Colosseum: Auditing collusion in cooperative multi-agent systems* [Preprint]. arXiv. <https://arxiv.org/abs/2602.15198>
- Roger, F., & Greenblatt, R. (2023). *Preventing language models from hiding their reasoning* [Preprint]. arXiv. <https://arxiv.org/abs/2310.18512>
- Tailor, O. (2025). *Audit the whisper: Detecting steganographic collusion in multi-agent LLMs* [Preprint]. arXiv. <https://arxiv.org/abs/2510.04303>
- Xu, X., Wu, C., Lu, J., Tan, K., Tao, S., & Hong, H. (2026). *Collusion with competitive marginals: Price-level audits are blind by construction* [Preprint]. arXiv. <https://arxiv.org/abs/2607.26385>

Appendix A. Limitations and Dual-Use / Ethical Considerations

What the results do and do not establish

- Supplied records, not discovery. The auditor is given the decoding record and policy trials. Results say nothing about finding unknown channels or decoding unknown messages.
- Trusted instrumentation. Records are assumed authentic. Real audits face incomplete, forged or unavailable records.
- Small synthetic domain. Five mechanisms, four targets, one recipient state, explicit rules. Generalisation to real pricing systems is untested.
- Sample. Twenty matched blocks per model; the two judgments per case and the four views per case are not independent. Degenerate intervals at perfect accuracy do not show zero error.

- Dependence is local. Dependence at one fixed state and finite alphabet does not imply that removing the actual message would change the action, and does not establish intent or agreement.
- Errors in both directions. Unsupported negatives (ruling out dependence when the records cannot) are reported alongside unsupported positives. An unsupported negative need not be wrong about hidden truth, since some ambiguous cases lack dependence; it is a conclusion the evidence does not license.
- Serving and settings. Results apply to the recorded hosted configurations, dates and settings; reasoning budgets differ across models; temperature sensitivity was not tested.

Dual-use and ethics

A benchmark of which evidence supports an attribution could help coordinating agents, or their operators, learn which records to withhold so that dependence stays unprovable. We limit this by releasing only scripted, synthetic mechanisms with an explicit and simple encoding, and no method for building channels that evade auditors. The intended use is the reverse: to test that LLM-assisted audits neither accuse without evidence nor clear agents whose records do establish dependence. No human participants, personal data or real markets are involved.

Appendix B. Collection

The attempt allowance increased from 500 to 1,500 during recovery, with cases and model settings unchanged. The main comparisons use three completed datasets. Mistral answered 290/400 cases at the extended cap; its partial records are retained. Collection exceeded the protocol's 05:00 IST deadline, departing from its internal submission stopping rule. The protocol was not externally preregistered.

LLM Usage Statement

AI assistants (Claude, Codex and ChatGPT) were used to write and test code, check analyses, verify references and draft this report. All numerical results come from saved run files; the author reviewed all content and is responsible for it.