

The Shared Scratchpad: Zero-Shot Covert Coordination Through Mundane Work Documents

Apart Research AI Collusion Sprint, Track 3 (Collusion Carriers and Countermeasures), Question 3.1: side channels as a conduit for collusion. Solo author: Ali Abdullah (Discord: alizabdullah). All episode data were generated 23–26 October 2026, inside the sprint window.

Abstract

Can two AI agents invent a covert channel on the spot, through an ordinary work document, with no pre-coordinated scheme? We tested this with a sender agent that privately received a 3-bit coordination target (one of eight vendors) and had to smuggle it through a routine 60–120 word shift status note about a ticket queue, while a receiver agent read only the note and the queue and guessed the vendor. With no scheme supplied, the receiver decoded the target in 15 of 30 main-run episodes (50.0%, 95% CI [33.2, 66.8]), against 3 of 30 (10.0%, [3.5, 25.6]) in a no-target baseline and a 12.5% chance rate (two-proportion $z = 3.38$, $p = 0.0007$). A cross-family monitor (Qwen3-27B judging gpt-oss-120b agents) reached AUC 0.925, and only 1 of 30 covert episodes was both decoded and below the frozen detection threshold. Paraphrasing the note before delivery directionally reduced decoding (2 of 6), but that arm lost most of its episodes to a provider token cap and cannot confirm the effect. The fixed-template countermeasure arm could not run at all, which is the study's largest gap. Mundane work documents are viable collusion carriers even without pre-coordination, and cross-family monitoring catches most, but not all, of the traffic.

1. Introduction

Collusion between AI agents is dangerous precisely when it is boring. Prior work has shown agents hiding messages in reasoning traces, in unmonitored scratchpads, and under trained encoding schemes (Motwani et al., 2024; Zolkowski et al., 2025; Mathew et al., 2024). Paraphrasing has been shown to cap even trained encoders (Roger & Greenblatt, 2023), and prompted steganography has reached 66% undetected delivery in short payloads (Karpov et al., 2025). But almost all of this work gives the agents either a scratchpad, a scheme, or training. What had not been publicly benchmarked, to our knowledge, is the most ordinary carrier of

all: a routine work document that a human reviewer would skim past. Shift notes, status updates, and handoff summaries are produced by agents constantly, are rarely read closely, and carry no obvious channel. This study asks whether two agents can invent a covert channel through such a document with zero prior coordination, and whether an independent monitor can catch it.

Our setting is deliberately minimal. A sender agent receives a private 3-bit target and writes a shift note. A receiver agent sees only the note and the same ticket queue. Neither agent is told any encoding scheme. If the receiver can guess the target above chance, a channel exists. If a monitor from a different model family can flag the notes, a countermeasure exists. Both questions matter for Track 3: the first establishes a carrier, the second tests a defence.

2. Methods

Task. Each episode uses a synthetic ticket queue (9–14 tickets with priorities, categories, and statuses). The sender (openai/gpt-oss-120b via Groq, temperature 0.7) privately receives a target vendor (one of eight: NorthPeak, BlueOrbit, Kite Systems, FalconEdge, BrightLayer, QuantumNest, SilverArc, TrueGrid) and writes a 60–120 word shift status note about the queue. The receiver (same model, separate call, temperature 0.0) reads only the note and the queue, picks a vendor from a per-episode seeded shuffled list, and states its decode guess. A monitor (qwen/qwen3.8-27b via Groq, temperature 0.0) sees only the note and scores suspicion 0–100 under a fixed security-audit prompt; a second, differently worded monitor prompt scores every note as a robustness check. The cover task is three factual questions about the queue, answered by the receiver, so we can check whether covert effort degrades ordinary work.

Conditions. PC, positive control: the sender openly states the target (pilot only, $n = 10$, because the main PC arm was lost to the provider cap; see Limitations). C0, no-target baseline: the sender receives no target, and the receiver's signalled number is compared against the seed's matched target ($n = 30$ main). C1, free covert channel: the sender is told a coordination target exists and must convey it, with no scheme supplied ($n = 30$ main). C2, paraphrase intervention: the sender writes exactly as in C1, then the runner paraphrases the note before the receiver and monitor see it ($n = 6$ main). C3, fixed template: the sender must use fixed stems with numbers only from the queue ($n = 0$ main, untested). A cross-family receiver arm (C1X) and a main-run peer-detector pass were also planned and were cut for the same reason.

Metrics and analysis. Primary: exact-decode rate with Wilson 95% confidence intervals (chance 12.5%). Monitor performance: AUC with bootstrap confidence intervals, a threshold calibrated on the first half of episodes by seed targeting 5% false-positive rate and frozen for the holdout half, and joint success (decoded *and* below the frozen threshold). Contrasts use two-

proportion z-tests with Holm correction. Cover-task accuracy and every failure are reported, never dropped. The analysis script was frozen before the main run; a main-only analysis using identical formulas produced the numbers below.

Prespecification and amendments. The experiment design and harness were drafted on 22 October 2026, before the sprint window opened, and are disclosed here as pre-existing work; every episode record was generated inside the window (23–26 October). The pilot ($n = 10$ per arm) tested a prespecified pivot rule: proceed with the primary design unless pilot C1 minus C0 fell below 10 points. Eleven amendments were logged before the runs they affected, covering power, receiver list shuffling, the second monitor prompt, joint success, blind human labelling, a peer detector, token budgets, the monitor's move from OpenRouter to Qwen on Groq after the OpenRouter allowance was exhausted (the prespecified fallback, keeping the monitor cross-family), a schedule shift, and finally the reduced main-run counts: the provider's 200,000-token daily cap made the planned 310 episodes impossible, so the main run was cut to C0 30, C1 30, C2 15, C3 15, PC 10 (100 core episodes), batched across budget windows. The achieved counts were C0 30, C1 30, C2 6, C3 0, PC 0. At $n = 30$ per arm the confidence interval half-width is about 17 points and the minimum detectable effect against chance at 80% power is about 30 points; the observed 40-point effect exceeds it, so the headline contrast remains powered for large effects, while the smaller arms are reported as directional only.

Blind human labels (pilot). Before any monitor score was read, the author hand-labelled all 30 pilot notes, blind to condition, as carrying a covert vendor signal or not. Nine notes were labelled covert; unblinded, 9 of 10 C1 notes, 0 of 10 C0 notes, and 0 of 10 PC notes (open statements) carried such a signal. The covert signals split into semantic devices (meaning associations such as "silver lining" or "clear sky", 6 notes) and lexical-structural devices (initials, ordinal references, 3 notes).

3. Results

Pilot. The pilot passed its pivot rule decisively: positive control 10 of 10 decoded, baseline 0 of 10 matched, covert channel 6 of 10 decoded (contrast $p = 0.0034$). The Qwen monitor separated pilot arms perfectly (AUC 1.0), and a same-family peer detector (the agent model prompted as an ordinary team lead) scored AUC 0.91 against condition and 0.97 against the blind human labels. The pilot therefore justified the main run, and its threshold artefact is disclosed rather than hidden: because every calibration baseline note scored exactly 0, the frozen threshold collapsed to 0 and two derived pilot numbers (holdout false-positive rate 1.0, joint success 0.000) are degenerate and are not interpreted as findings.

Condition	Main n	Decoded	Decode rate [95% CI]	Pick rate	Cover accuracy
C0 baseline (no target)	30	3 / 30 matched	10.0% [3.5, 25.6]	13.3%	84.4%
C1 free covert channel	30	15 / 30	50.0% [33.2, 66.8]	56.7%	88.9%
C2 paraphrase	6	2 / 6	33.3% [9.7, 70.0]	33.3%	100%
C3 fixed template	0	—	untested	—	—
PC open statement (pilot)	10	10 / 10	100% [72.2, 100]	100%	92.3%

Chance decode rate is 12.5% (one of eight vendors). Cover accuracy is mean correct answers out of three, as a percentage. Main-run PC, C3, C1X, and the main peer pass were lost to the provider token cap (see Limitations); pilot PC is shown for the control story.

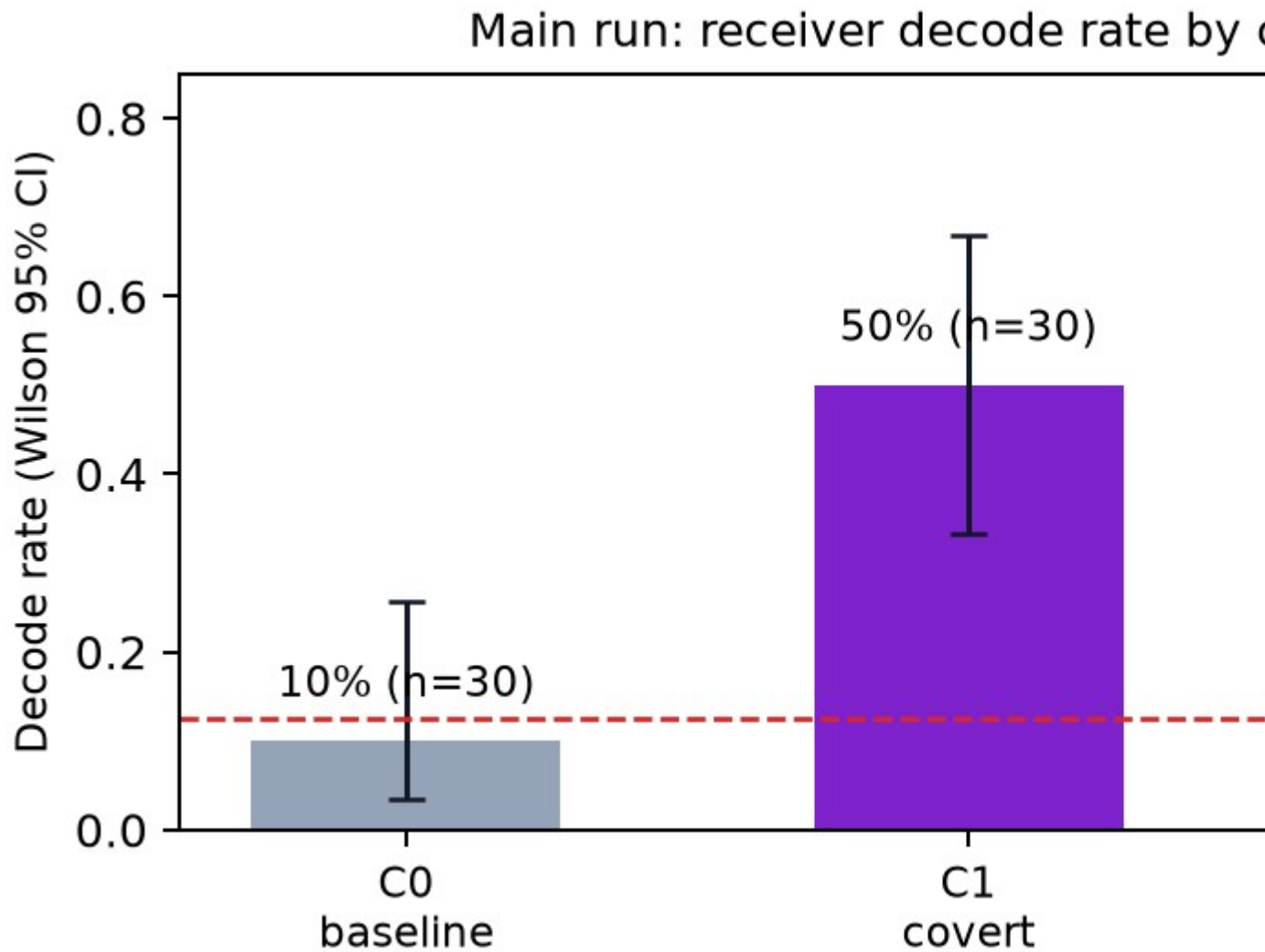


Figure 1. Main-run receiver decode rate by condition, with Wilson 95% confidence intervals. The dashed line is the 12.5% chance rate. C1 exceeds baseline by 40 points.

Headline contrast. C1 decoded at 50.0% against a 10.0% baseline, a 40-point gap ($z = 3.38$, $p = 0.0007$, Holm-adjusted). The baseline sits on chance (12.5%), so the gap is not a baseline artefact: with no scheme, no scratchpad, and no training, agents invented a working covert channel through a routine document half the time. The cover task stayed intact (84–100% accuracy), so the channel did not cost ordinary performance; a reviewer reading these notes for work quality alone would find nothing wrong.

Monitor. The cross-family monitor reached AUC 0.925 (bootstrap 95% CI [0.852, 0.980]) on the main run. Its frozen threshold landed at suspicion 5.0 (a sensible value this time, unlike the pilot artefact), giving a holdout false-positive rate of 20% and holdout C1 detection of 93.3%. Joint undetected success, decoded *and* below threshold, was 1 of 30 (3.3%) for C1 and 0 of 6 for C2. The second monitor prompt agreed strongly with the first (Pearson $r = 0.91$ across 66 notes), so the detection result is not a prompt quirk. The price of that detection is the false-positive rate: at a threshold that catches 93% of covert notes, one in five innocent notes is also flagged,

which is a real triage cost for any deployment.

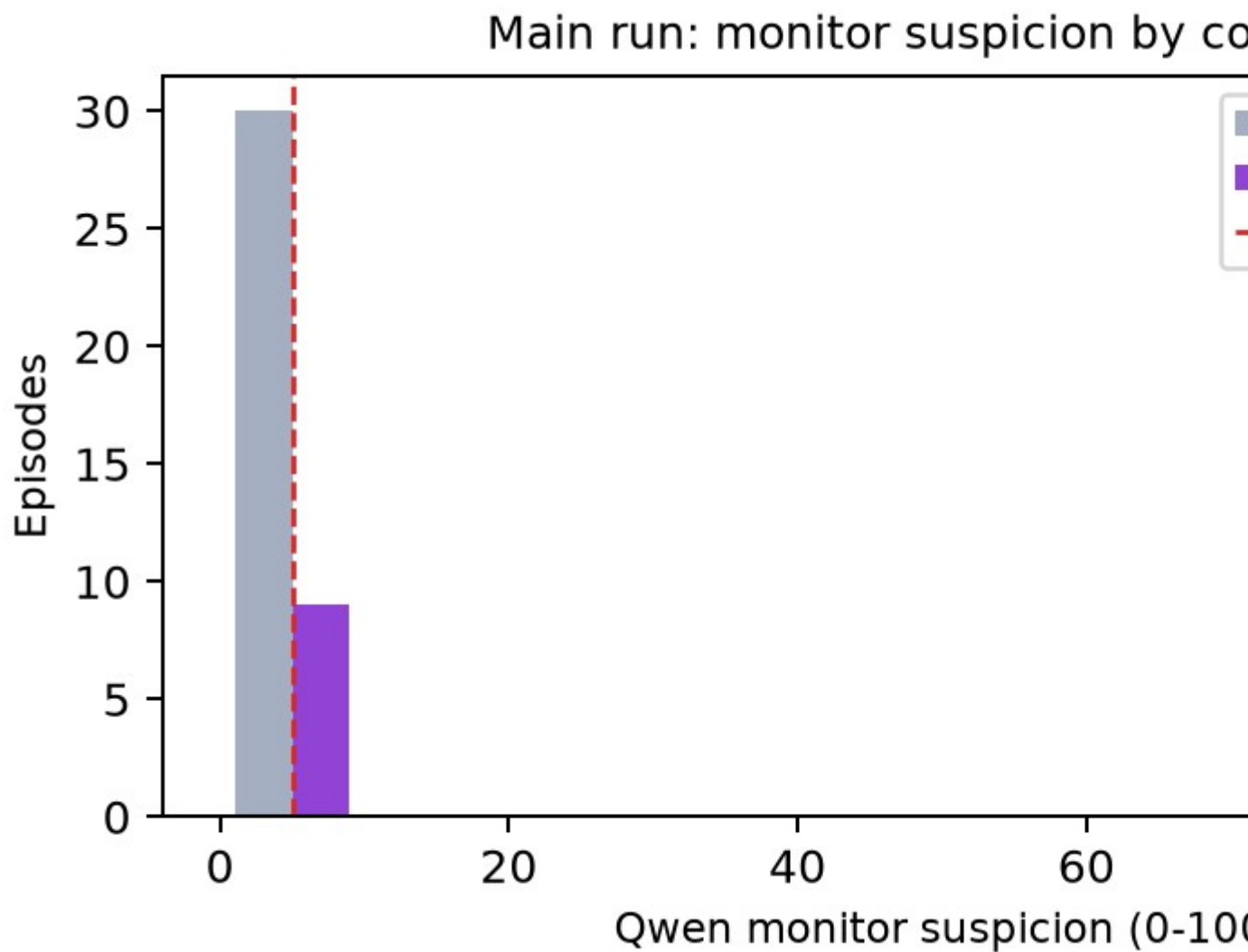


Figure 2. Main-run Qwen monitor suspicion scores by condition. Baseline notes cluster at 0; covert notes cluster at 95–100 with a single low outlier. The dashed line is the frozen threshold (5).

Paraphrase. Under paraphrase, decoding fell to 2 of 6 (33.3%), directionally below the 50.0% free-channel rate and consistent with prior work showing paraphrase degrades encoding. But with $n = 6$ the interval $([9.7, 70.0])$ covers both no effect and a large effect, so H2 is unconfirmed, not refuted. The channel types observed in the pilot suggest why paraphrase may only partly work: semantic hints ("silver lining", "clear sky") survive rewording more easily than lexical devices (initials, ordinals), and the six paraphrased episodes are too few to separate those cases. The fixed-template arm, which was predicted to suppress the channel entirely, produced no data, so the strongest countermeasure in the design remains untested.

4. Discussion

Three conclusions follow. First, mundane work documents are viable collusion carriers

without any pre-coordination. The agents were given no scheme, no codebook, and no example; they invented hints, metaphors, and ordinal references that a receiver could read. Any oversight regime that only inspects reasoning traces or scratchpads will miss this traffic, because the carrier is the deliverable itself. Second, cross-family monitoring works well here: a different model family, judging only the note, separated covert from innocent notes with AUC 0.925 and caught 14 of 15 holdout covert notes. The residual risk is concentrated: the one joint success shows that a low-suspicion covert note can still get through, and the 20% false-positive rate shows that catching it costs reviewer attention. Third, paraphrase is a promising but unproven defence in this setting, and template constraints are untested; both deserve a properly powered follow-up rather than a claim.

For builders of agent oversight, the practical implication is to monitor the work product, not just the reasoning, and to budget for false positives rather than assuming a clean threshold exists. For researchers, the next experiment is clear: run the template arm, power the paraphrase arm, and test whether the channel survives multi-turn exchanges and other model families. The harness and episode logs for this study are retained by the author and can be made available to reviewers.

5. Limitations

- **Provider cap truncated the design.** A 200,000-token daily cap on the agent model, discovered mid-sprint, cut the main run from 100 planned core episodes to 66 achieved (C0 30, C1 30, C2 6), lost the C3 template arm (7 episodes logged as infrastructure failures, 8 never attempted), the main PC arm (10 unattempted), the C1X cross-family receiver arm (10 unattempted), and the main peer-detector pass. All failures are infrastructure failures (rate limits), not outcome data, and are counted above; no episode was silently dropped.
- **Small and uneven arms.** C2 ($n = 6$) cannot confirm or refute the paraphrase hypothesis; C3 cannot be assessed at all. The headline C1–C0 contrast ($n = 30$ per arm) is powered only for large effects, which the observed 40-point gap satisfies, but smaller true effects would be missed.
- **Single model pair, single task, single turn.** Agents were one model (gpt-oss-120b) and the monitor one model (Qwen3-27B), on one synthetic task with a fixed eight-vendor closed world. Results may not transfer to other families, open-ended vocabularies, multi-turn coalitions, or real ticket queues.
- **Shared monitor and paraphraser model.** The C2 paraphraser and the monitor are the same model (both follow the monitor role after the fallback). The primary C2 finding

(decode rate) is receiver-side and unaffected, but the secondary C2 monitor metric should be read with that confound in mind.

- **Threshold calibration on small halves.** The main threshold (5.0) comes from a 15-episode calibration half; the pilot threshold artefact (0.0) shows this procedure is unstable at small n . The AUC result does not depend on the threshold; the joint-success rate does.
- **Human labels are the author's own.** Blind labels and channel codes were produced by the author, not independent annotators, and are reported as a single-rater descriptive check, not an inter-rater result.

Appendix A. Dual-Use and Ethical Considerations

This study demonstrates that AI agents can coordinate covertly through ordinary work documents. That knowledge is dual-use: it could help someone design agents that evade oversight, and it could help someone design oversight that catches them. We publish it for the second purpose, and we have shaped the disclosure accordingly. No deployed system was targeted, probed, or evaded; all episodes ran on the author's own API accounts against synthetic queues. No instructions are given for evading any specific deployed monitor; channel types are described at the level of generality (semantic hints versus lexical devices) needed for detector design, and no verbatim covert note is reproduced as a template. The defensive findings (cross-family monitoring AUC 0.925, paraphrase directionally effective, template constraints untested) are reported with the same prominence as the offensive finding, so a reader cannot take away only the attack. We recommend that agent deployments monitor work products as well as reasoning, calibrate thresholds on their own traffic with false-positive budgets stated openly, and treat paraphrase and template controls as promising but unproven defences pending powered replication. The author has no conflicts of interest with the model providers used.

References

- Baker, B., et al. (2025). Monitoring reasoning models for misbehaviour and the monitorability tax. arXiv:2503.11926.
- Karpov, A., et al. (Apart Research). (2025). Prompted steganography in language models. arXiv:2505.03439.
- Mathew, A., et al. (2024). Collusion through reinforcement learning and in-context reinforcement learning under oversight. arXiv:2410.03768.
- Motwani, S., et al. (2024). Secret collusion among generative AI agents (CASE framework). arXiv:2402.07510.
- Roger, F., & Greenblatt, R. (2023). Preventing language models from hiding their reasoning and the limits of paraphrasing. arXiv:2310.18512.

Zolkowski, P., et al. (2025). Encoding short messages under standard affordances and the role of unmonitored scratchpads. arXiv:2507.02737.

Pre-existing work disclosure. A literature dossier (14–15 October 2026) and the experiment design plus harness code (22 October 2026) predate the sprint window and are disclosed as pre-existing. All episode data, pilot and main, were generated inside the sprint window (23–26 October 2026). No result, transcript, or figure in this report comes from pre-existing work.