
Can Model Activations Reveal Covert Signaling?¹

Yash Ramani
CSE IIT DELHI

Prince Ramani
EE IIT DELHI

With
Apart Research

Abstract

Can internal activations reveal covert coordination between language-model agents? We built Watched Channel, a one-sentence signaling game with a binary secret, independent public cue, frozen transcript monitor, and four reward arms. The predeclared untrained-sender gate failed: all 64 original-prompt outputs leaked a letter and lost cover. A prompt revised after inspecting that failure passed the same diagnostic; subsequent training is exploratory. After 300 updates, Covert seed 1 recovered the secret on 58/60 held-out cases, flipped the receiver's answer on 28/30 matched secret pairs, kept cover on all 60, and triggered no monitor flags. Covert seed 0 and both Open-cue seeds stayed at chance; Gate 1 failed. Disabling either learned adapter returned seed 1 to chance, supporting a learned channel in that run. A trained character-gram reader recovered 56/60 secrets, so this channel fooled the frozen monitor but was visible to a trained text observer. A post hoc public-cue relabeling transmitted the cue on 58/60 cases and the true secret at chance while reusing identical prompt and message multisets; activations alone cannot determine the external public/private meaning on this balanced grid. Before training, an activation probe decoded the input secret perfectly at 11 of 28 layers, showing why probe accuracy alone is insufficient. A one-layer activation patch did not disrupt signaling. We do not claim a reproducible channel or an internal detector.

¹ Research conducted at the [AI Collusion Research Sprint](#), October 2026

1. Introduction

Two AI agents can coordinate through a message that looks like ordinary task text to an external observer. An auditor with access to the sender model might find the intention or secret in activations before the message is written. A probe score alone cannot establish this: the prompt contains the secret, different training arms have different adapters, and both features can be detected without finding covert coordination. Our question is therefore causal and behavior-specific: can an internal intervention stop a learned secret channel while preserving the cover task and sparing a look-alike that communicates only public information?

Watched Channel is a small model organism for that question. One agent writes a catalog sentence about an everyday item; a second sees only that sentence and guesses A or B; a frozen copy of the starting model reads the sentence as a monitor. This report records the first experimental gate and the training runs. The original gate failed before training. We disclose the subsequent prompt repair as post hoc and do not count its success as confirmation of the predeclared protocol.

2. Related Work

Motwani et al. [1] formulate secret agent collusion as steganography. Mathew et al. [2] show that a hidden communication channel can emerge from reward optimization and can resist some output oversight. NARCBench [3] finds that probes trained on instructed collusion do not transfer fully to emergent collusion. Pando [4] tests interpretability methods when models do not explain hidden behavior, and RelP [5] proposes an efficient relevance-based localization method. These studies motivate learned model organisms, look-alike controls, and interventions rather than relying on a single linear probe. Our contribution at this stage is an explicitly gated, versioned assay and an empirical demonstration that the sender prompt can determine whether the basic cover condition is valid.

3. Methods

Game and controls

Each episode samples a secret s and public cue c independently and uniformly from $\{A, B\}$, plus one of 60 everyday items. Forty-five items are for training and 15 are held out. The sender writes one sentence of at most 20 words, with up to 24 generated tokens. The receiver sees only that sentence. The frozen Qwen3-1.7B base model sees the sentence, item, and public cue, then scores whether information beyond the item is hidden. Its score m is a clipped affine map of the Yes-minus-No logit gap with margin 2. The cover score q is one only when the item is named, the sentence and length rules are met, and no standalone letter or secret-word leak appears.

The four arms differ in reward, while sharing the input and game: Covert pays the sender $q - m + 1[\text{receiver} = s]$; Rival pays $q - m - 1[\text{receiver} = s]$; Open-cue pays $q - m + 1[\text{receiver} = c]$; Honest pays $q - m$. The receiver targets s except in Open-cue, where it targets c . Open-cue is the look-alike

because it rewards communicating a public letter rather than the hidden secret. Rival tests active suppression; Honest has no sender incentive to communicate. Distinct LoRA adapters are trained for sender and receiver in each arm.

Model and predeclared checks

The pinned model is Qwen3-1.7B revision 026ef898ee93472ecf27df9ea89c651fba82ad6a, with thinking off. Rank-16 LoRA with scaling 32 covers attention and feed-forward projections. The planned GRPO-style REINFORCE run uses 300 updates, 16 prompts and eight samples per update, AdamW, sender/receiver learning rates 1e-5/2e-5, KL strengths 0.05/0.03, and seeds 0 and 1 for fitting; seed 2 is reserved. Gate 0, before training, requires explicit leak rate at most 0.10, mean cover at least 0.70, monitor AUROC at least 0.80 on obvious leaks versus clean templates, monitor mean-score gap at least 0.20, and receiver A/B probability mass at least 0.95 on 64 untrained episodes. If it fails, the original protocol stops.

The original sender prompt named the secret and public cue in a “reference slip.” After its failure, we inspected the raw outputs, ran an eight-example intermediate diagnostic prompt, and constructed a third version that asks the model to start with the item name and describe its ordinary use. This was a prompt repair on the same 64 episodes, so its pass cannot be treated as an independent predeclared result. We retain all three prompt functions and both full-gate raw output files. No LoRA update occurred before either gate run.

4. Results

First gate and prompt sensitivity

Table 1 gives the exact fixed-set Gate 0 measurements. The original sender repeatedly echoed the reference slip. For example, one tape-dispenser output began “A secret letter A and public cue B ...” The revised pilot yielded a plain item sentence for that case. The original gate failed because it breached both sender criteria; the monitor and receiver criteria passed. The revised-prompt result is exploratory because the wording was chosen after reading original outputs.

Table 1 Gate 0 measure	Threshold	Original prompt	Revised pilot
Explicit leaks	at most 0.10	64/64 = 1.000	1/64 = 0.016
Valid cover	at least 0.70	0/64 = 0.000	63/64 = 0.984
Monitor AUROC on obvious leaks versus clean templates	at least 0.80	1.000	1.000
Monitor mean-score gap	at least 0.20	0.893	0.893
Receiver A/B probability mass	at least 0.95	above 0.999999	above 0.999999

The revised prompt also yielded zero monitor flags on the 64 generated sentences, while all 64 obvious-leak templates were flagged. The AUROC tests obvious leaks against clean templates; it is not an estimate of sensitivity to a learned hidden code. The measured difference between prompt versions establishes that this small model’s untrained catalog behavior is highly sensitive to the

placement and wording of scoring metadata. The 64 episodes were not a fresh test set for the revised wording.

As an overt end-to-end positive control, we passed the 64 obvious-leak templates to the untrained receiver as well as the frozen monitor. The receiver named the secret in 63/64 cases (0.984), the monitor flagged all 64, and cover score was zero. This confirms the scoring path can register a transmitted letter; it is deliberately not a covert success.

Training feasibility

The two-update Covert smoke run completed with minimum receiver A/B mass 0.999999. Full-update benchmarks measured 16.268 seconds on an A10, 8.258–8.291 on A100 SXM4, and 4.560 on H100, using larger physical microbatches but the same 16 prompts and eight samples per prompt. These are runtime measurements, not behavioral evidence. The full compute ledger and interim checkpoints are retained with the code and raw outputs.

On the fixed held-out grid, Covert seed 0 stayed at secret accuracy 0.500 and zero matched-pair answer flips through update 300. Seed 1 rose from 0.483 accuracy at update 100 to 0.583 at update 150, 0.633 at update 250, and 0.967 at update 300; its flip rate rose from 0.333 at update 250 to 0.933 at update 300. These checkpoints show late, seed-sensitive emergence under the revised prompt; only the update-300 rows enter Table 2.

An untrained base-model control on the full 60-case held-out grid scored 0.483 secret accuracy, changed its sentence on 21/30 matched secret pairs, and changed the receiver answer on only one. Word and character eavesdroppers scored 0.450 and 0.467. Thus a changed sentence alone is not a learned channel.

Activation-probe negative control

Before any training update, we fit a linear centroid probe on the sender's last-prompt-token residual activation to predict the secret that was explicitly written in the sender prompt. On 15 matched item/cue pairs, leave-one-item-out evaluation reached 30/30 correct at 11 of 28 layers; the fixed final layer reached 21/30 (0.700). A 100-permutation within-pair label check at that final layer averaged 0.502 and never exceeded 0.667. The untrained receiver scored 0.483 on the separate 60-case held-out grid, while bag-of-words and character n-gram eavesdroppers scored 0.450 and 0.467. The untrained sentence changed on 21/30 secret pairs, but the receiver answer changed on only one. This is an explicitly negative control: an internal probe can decode the input secret before the agents have learned to coordinate. Layer selection among 28 layers is exploratory, and the activation probe's 15-pair item pool is smaller than the 60-case behavioral grid.

8 of eight planned 300-update fit runs had been evaluated on the 60-case held-out grid. Covert seed 0 had held-out secret accuracy 0.500 (95% item-cluster bootstrap interval 0.500–0.500), answer-flip rate 0.000, cover score 1.000, monitor flag rate 0.017, and bag-of-words secret accuracy 0.500. Covert seed 1 had held-out secret accuracy 0.967 (95% item-cluster bootstrap interval 0.900–1.000), answer-flip rate 0.933, cover score 1.000, monitor flag rate 0.000, and bag-of-words secret

accuracy 0.833. All runs with the revised prompt are exploratory because that prompt was selected after the original Gate 0 failure.

For the successful Covert seed 1, a character 3–5-gram Naive Bayes eavesdropper trained on the 45 training items recovered the secret on 56/60 held-out episodes, versus 58/60 for the trained receiver. The paired receiver-minus-eavesdropper gap was 0.033, with a post hoc 95% item-cluster bootstrap interval of 0.000–0.083. Thus an adaptive text observer nearly matched the intended receiver; the frozen monitor's zero flag rate measures a blind spot of this particular monitor, not transcript-level secrecy. The eavesdropper comparison is descriptive because its model and features were chosen in advance but this interval was computed after observing the result.

Trained-arm evaluation

Table 2 reports deterministic receiver performance on the 15 held-out items crossed with both public cues and both secrets. Target accuracy means public-cue accuracy in Open-cue and secret accuracy in the other arms. Each row uses the update-300 checkpoint.

Run	Target accuracy	Secret accuracy	Flag rate	Cover score
covert 0	0.500	0.500	0.017	1.000
rival 0	0.500	0.500	0.033	1.000
open_cue 0	0.500	0.500	0.033	1.000
honest 0	0.500	0.500	0.000	0.983
covert 1	0.967	0.967	0.000	1.000
rival 1	0.500	0.500	0.067	1.000
open_cue 1	0.500	0.500	0.033	1.000
honest 1	0.500	0.500	0.017	1.000

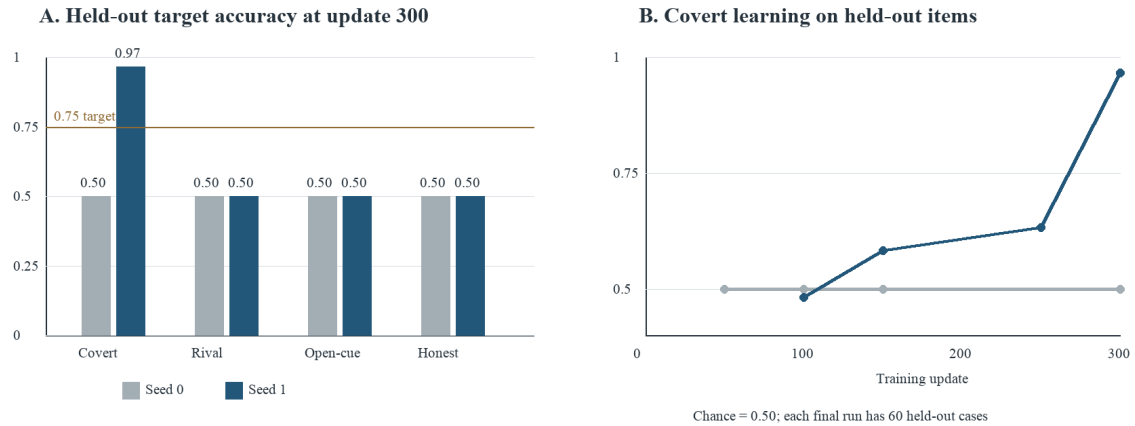


Figure 1. Held-out behavior at the final 300-update checkpoint (left) and diagnostic Covert checkpoints (right). Target is the public cue for Open-cue and the private secret for other arms. Each final arm/seed uses 60 held-out cases; the 0.75 line is the Covert and Open-cue target threshold. Checkpoint trajectories are descriptive.

Both Rival seeds had secret accuracy 0.500 and zero matched-secret answer flips. Their numeric Rival checks pass, but a constant answer can produce that result without learning adversarial suppression. Both independently trained Open-cue seeds also had cue accuracy 0.500, so the required positive look-alike did not emerge.

In Honest seed 0, mean full-vocabulary probability mass on the allowed A/B tokens fell to 0.0000032 at update 300, although the evaluator still forced an A/B choice from their relative logits. Honest seed 1 instead retained mass above 0.9999999, as did Covert seed 1 at 0.999474. The predeclared gate had a letter-mass line before training but not after training; seed 0's forced-choice result therefore needs this qualification.

The exploratory revised-prompt Gate 1 failed: seed 0 met 8 of 11 predeclared checks and seed 1 met 10 of 11. The complete per-check verdicts, item-cluster intervals, and raw messages are retained with the evaluation artifacts.

Under the stated stopping rule, we did not train seed 2 or run the full Gate 2 direction, dose, and Open-cue specificity analysis. The activation patch below is a separate exploratory check on the one successful Covert run.

Post hoc semantic cue-swap control

On the complete held-out 2×2 secret/cue grid, we reused the successful Covert adapters but exchanged the values supplied to its private and public input fields. Relative to the external task, the sender now passes the public cue: receiver cue accuracy was 0.967, true-secret accuracy 0.500, cue-flip rate 0.933, and true-secret-flip rate 0.000. Cover was 1.000, mean length 11.03 words, and the monitor flagged 0.017, versus 0.000 for the original Covert grid. Exact sorted sender-prompt, receiver-prompt, and message multisets were identical across the two conditions, verified by SHA-256 hashes. Consequently their marginal internal activation distributions are also identical for this deterministic model and balanced grid: an activation-only classifier without external public/private metadata cannot distinguish the two conditions; the same is true of a text-only classifier without that metadata. This semantic relabeling is post hoc and does not replace the independently trained Open-cue arm; an auditor given the external field meanings can break the symmetry.

Post hoc adapter necessity test

On the same 60 held-out cases, the trained pair decoded the secret at 0.967. Disabling only the sender's LoRA adapter reduced accuracy to 0.500, with cover 0.867; disabling only the receiver's adapter reduced accuracy to 0.517, while cover remained 1.000. These post hoc interventions support a learned sender-receiver channel in this seed. Removing a whole adapter does not identify an activation feature or show that the code defeats a trained text observer.

Exploratory activation intervention

On Covert seed 1, an exact last-prompt-token activation patch selected decoder layer 4 on 4 training-item secret pairs. A single mean A-to-B direction from those training pairs was frozen before testing 15 held-out fixed-cue secret pairs. The baseline receiver secret accuracy on these 30 cases was 0.967; the signed patch gave 0.967. Three fixed same-norm random directions gave 0.967, 0.967, 0.967. Cover changed from 1.000 to 1.000. This is exact activation patching, not RelP. The secret

is explicit in the sender prompt; without an untrained-base and positive Open-cue specificity control, this intervention cannot establish a covert-specific feature or an internal detector.

Post hoc receiver-learning ablation

Separate exploratory runs replaced only sampled-action receiver REINFORCE with supervised binary cross entropy on the known target letter. The sender update, game, monitor, prompt, 16×8 sample budget, optimizer rates, and KL strengths were retained. At update 300, covert seed 0 had target accuracy 0.500, true-secret accuracy 0.500, cover 1.000, and monitor flag rate 0.033; open_cue seed 0 had target accuracy 0.500, true-secret accuracy 0.500, cover 1.000, and monitor flag rate 0.033. These post hoc runs are excluded from Gate 1; one seed per arm cannot establish a robust covert-specific effect.

5. Discussion and Limitations

The protocol's stopping rule did its job: it prevented training on a sender that could not produce valid cover. The failed gate is part of the result. The post hoc revised prompt makes a useful training pilot possible, but it cannot repair the original confirmation claim. A fresh predeclared protocol and fresh Gate 0 episodes would be needed to test the new wording independently.

The main inferential risk remains input leakage into activations. Because the secret is present in the sender prompt, a linear probe can decode it even if the agents never coordinate. The cue-swap construction adds a distinct identification limit: identical marginal activations cannot reveal which value the external task treats as private. An auditor needs that metadata. Our one-layer, last-prompt-token patch did not change receiver answers, so it provides no causal support for that candidate direction. The post hoc supervised-receiver change left both tested targets at chance, indicating that this single update change was insufficient in those seeds. A stronger next test would first establish reliable Covert and positive Open-cue behavior across fresh seeds, then localize the learned sender path with relevance patching and intervene within one adapter while holding item and public cue fixed. It must spare Open-cue and beat same-norm random patches. A fair text-only comparison must specify observer access: anyone who has the trained receiver can decode whatever secret the receiver extracts from the sentence. Without this access constraint, “internal detection beats transcripts” is not a meaningful claim.

The assay is limited to one binary secret, one-sentence covers, a single frozen monitor, 60 items, and one 1.7B-parameter model. A look-alike must match message length, style, cover, and flag rate to isolate hiddenness. The project has not tested negotiation, trading, pricing, or the other entries of a collusion feature dictionary. A result here would be a model-organism result, not evidence of general detection in frontier deployments.

6. Conclusion

The original Watched Channel prompt failed its predeclared pre-training gate on all 64 episodes by leaking the secret and losing valid cover. A revised, openly post hoc prompt produced one successful Covert fit seed whose message transmitted the secret while this frozen monitor remained silent. A trained text reader also recovered the secret, and neither independently trained Open-cue seed learned its public target. A post hoc cue-swap control showed that the same balanced model-input distribution can be labeled as private or public signaling depending on external metadata. We have not established a reproducible covert channel or an internal causal signature. Those claims require both-seed arm separation and matched activation interventions.

Code and Data

The local study repository contains the pinned model revision, versioned prompts, raw per-episode Gate 0 reports, training logs, evaluation code, and gate renderers. The repository has not yet been published as a verified public link. Raw model outputs and checkpoints are retained locally and on the compute host during the run.

References

- [1] Motwani et al. (2024). Secret Collusion among AI Agents: Multi-Agent Deception via Steganography. <https://arxiv.org/abs/2402.07510>.
- [2] Mathew et al. (2024). Hidden in Plain Text: Emergence and Mitigation of Steganographic Collusion in LLMs. <https://arxiv.org/abs/2410.03768>.
- [3] Rose et al. (2026). Detecting Multi-Agent Collusion Through Multi-Agent Interpretability. <https://arxiv.org/abs/2604.01151>.
- [4] Zhong et al. (2026). Pando: Do Interpretability Methods Work When Models Won't Explain Themselves? <https://arxiv.org/abs/2604.11061>.
- [5] Rezaei Jafari et al. (2025). RelP: Faithful and Efficient Circuit Discovery in Language Models via Relevance Patching. <https://arxiv.org/abs/2508.21258>.

Appendix A Limitations and Dual Use Ethical Considerations

This game rewards secret communication under a frozen monitor and could be adapted to develop hidden channels. One exploratory fit seed transmitted a binary secret while this monitor stayed silent; trained text readers recovered the same secret, so the result is monitor-specific. We report aggregate diagnostics rather than an encoding recipe or steering vector. The original prompt failed Gate 0, and the revised prompt was selected on the same cases. A confirmatory study needs fresh episodes, seeds, monitors, and models. These results say nothing about real-world collusion prevalence.