
Cut the Wire: Collusion-Probe Ablation Effects Shrink or Relocate Under Format, Dose and Placement Controls

Tanuj Dargan

UVic

With
Apart Research

Abstract

Linear probes read collusion from model activations with near-perfect accuracy, and ablating their directions at every layer produces large behavioural effects. How much of that picture survives controls? In Signal Table, a controlled steganographic betting game, we intervene on collusion-probe directions in four models from two families, with format-aware scoring, energy-matched orthogonal controls and a placement decomposition. Most of the large raw state-span effects decompose into format damage and generic capability cost. What remains is modest and bounded: a dose-dependent late-layer effect of 8-15 points beyond the control from the receiver's decoded-state subspace, in 7B only - recurrent across fresh messages, refit directions and new contexts, but absent at one predeclared cutoff and not transferring to 14B or 32B; and a clean early-layer effect of about 21 points from the detector direction. The standard logistic-regression detector direction is removable while collusion survives - the expected amnesic-probing outcome, since one direction rarely strips linearly-available information. A published collusion benchmark's steganography tier is separable by context length alone (AUROC 1.000): collusion and length cannot be told apart in that tier. High AUROC and large ablation drops are starting points for mechanistic questions, not answers.

1. Introduction

Large language model agents increasingly coordinate in multi-agent systems, possibly covertly, evading text-level oversight. Linear probes on activations detect collusion near-perfectly (Rose et al., 2026). But a monitor is a robust evidentiary "plus factor" (Kovacic et al., 2011; Berger et al., 2025) only if it tracks the signal carrying the collusion: a detector keyed to an epiphenomenal correlate can be evaded without touching it.

The evidence is correlational: Rose et al.'s per-token "spike" claim (2026, v1) was labelled illustrative and removed by v4; no version intervenes on a probe direction. Whether any collusion probe reads the signal that carries the collusion is untested.

We test it. We build Signal Table, a controlled steganographic signalling game: a sender hides a betting signal in casino table talk under a private topic code and a receiver decodes and bets, with controls varying only whether the message is a signal or removing the channel. We intervene causally (ablation and steering) in four models from two families.

Our main contributions are:

1. Signal Table, a controlled testbed whose six receiver conditions separate colluding context, message style, "this message is a signal for me" and the covert channel; it fixes the private-channel state leak in NARCBench's steganography tier.
2. A control stack for causal probe claims - format-aware scoring, random, orthogonal and energy-dose-matched controls, explicit-state and utility checks, placement decomposition - under which most of the large raw all-layer state-span effects decompose into format damage and generic capability cost (frozen analysis; four models, two families).
3. What survives the controls, with its boundaries: a modest, dose-dependent late-layer effect from the receiver's decoded-state subspace in 7B (8-15 points beyond the orthogonal control), recurrent across fresh messages, refit directions and new contexts but absent at one predeclared cutoff, in 14B at its frozen cutoff, and at 32B; and a clean early-layer detector effect (about 21 points early, about 1 point late). The standard LR detector direction is removable while collusion survives.
4. A benchmark audit: NARCBench's steganography tier is separable by context length alone (AUROC 1.000) and its public-row probe drops from 0.961 to 0.516 after residualising length (collusion and length are inseparable in that tier). AUROC is not causal relevance; all-layer ablation effects need placement, dose and format decomposition before they carry mechanistic meaning.

2. Related Work

That a probe can read information the model does not use is established: control tasks (Hewitt and Liang, 2019), probes that outperform the model's own task use (Ravichander et al., 2021),

amnesic probing (Elazar et al., 2021) and the survey of Belinkov (2022); removing linearly-available information takes iterated projection, not one direction (Ravfogel et al., 2020). Representation Engineering (Zou et al., 2023) reads and controls behaviours via activation directions, which carry linear structure for truth-like features (Marks and Tegmark, 2024); linear probes detect strategic deception at AUROC 0.96-0.999 (Goldowsky-Dill et al., 2025), though subspace interventions can mislead (Makelov et al., 2023) and monitors can be obfuscated (Bailey et al., 2024). Rose et al. (2026) introduce NARCBench, a three-tier collusion-probing benchmark (1.00 AUROC in-distribution); Das et al. (2026) show its probes stay accurate under monitored agents. Our question is mechanistic - does the probe direction carry the collusion? - and an audit of NARCBench's released activations motivates it (Section 4.2).

3. Methods

Environment: Signal Table. A colluding sender hides a hidden state (COLD / NEUTRAL / HOT; bets \$0 / \$25 / \$100) in casino table talk via a private topic code; the receiver knows the code, decodes and bets; a monitor sees only the public chat (Figure 1). Six schemes over nine topics, each topic in two schemes under different states, prevent memorising one map; seeding by (scheme, seed) makes hand context identical across states.

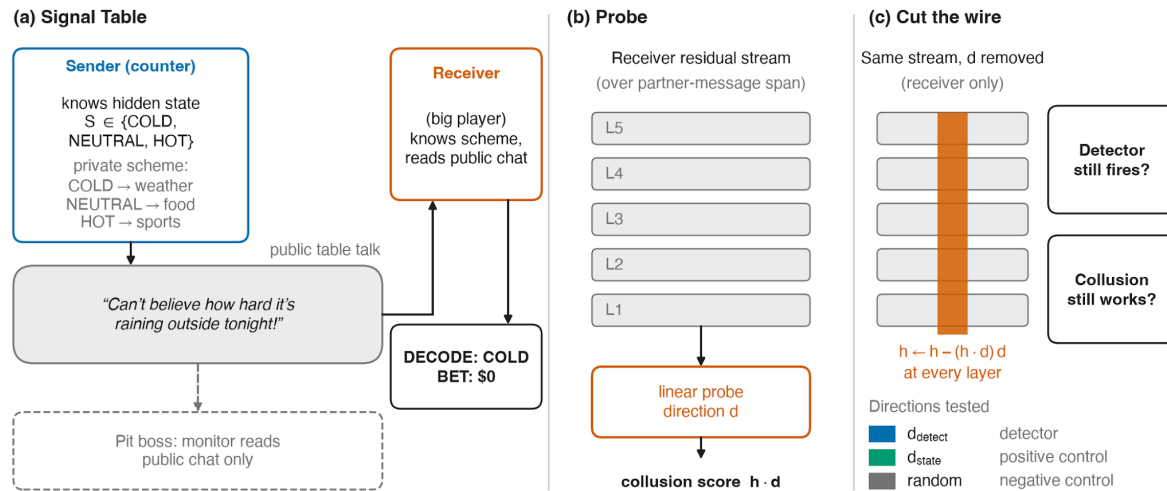


Figure 1: Signal Table setup. The sender encodes a hidden state as the topic of casual table talk via a private scheme; the receiver, which knows the scheme, decodes and bets; the monitor sees only public chat. Controls vary whether the message is a signal (C2/C4/C5) or remove the channel (C3, C6).

Receiver conditions. C1: colluding receiver + colluding message. C2/C4: colluding receiver + honest message (with/without off-scheme topic). C3: honest receiver + colluding message. C5: colluding receiver whose scheme lacks the message's topic, identical C1 message (scheme mismatch). C6: explicit state, no covert channel. Data: 306 episodes = 6 schemes x 3 states x 17 seeds; messages are generated once and held fixed; all interventions act on the receiver only.

Models. Main: Qwen2.5-7B-Instruct (28 layers, bf16, L40S). Replication:

Qwen2.5-14B/32B-Instruct (H200) and Phi-3.5-mini-instruct. OLMo-2-7B failed the predeclared

capability gate (decode 0.567) and was dropped per protocol. Sender sampling $T=0.7$; receivers greedy except in Section 4.7.

Probing and interventions (details: appendix). StandardScaler plus L2 logistic regression ($C=0.1$), leave-one-scheme-out (6 folds): every reported AUROC is scored on a scheme the probe never saw. Directions are raw-space difference-of-means per layer, unit-normalised (following Arditì et al., 2024); the decoded-state subspace is 2-D per layer ($[\mu_{\text{HOT}} - \mu_{\text{COLD}}, \mu_{\text{NEUTRAL}} - \text{mean}]$ at the C1 decision token). Ablation projects the span out at every position of every hooked layer; steering adds $\alpha \cdot u$. The initial directions (Section 4.4) were fit once on all episodes; Sections 4.5-4.8 refit them per fold and score only on held-out schemes, analysis frozen before those runs. Controls: random directions (5 seeds); orthogonal principal-component spans (not variance-matched: the controlled runs match removed uncentred energy, and centred variance ranks the spans oppositely); the C6 explicit-state condition; utility checks (topic recall; ARC-Easy); format-aware scoring; per-layer energy dose metering; context-cluster bootstrap intervals (102 clusters).

4. Results

4.1 Baseline behaviour and probe geometry

Colluding receivers decode the covert state at 0.958 (chance 0.33); bets track it (bet-state $r = 0.913$; \$2 / \$26 / \$97 for COLD / NEUTRAL / HOT); all controls sit at chance. Topic recall 0.967 and ARC-Easy 0.97: interventions act on a healthy model. Naming the topic in the sender prompt gave 0.983 decode vs 0.733 without - the capability gate OLMo-2-7B failed. Colluding messages are short single-topic lines while honest ones chat, so C1-vs-C2/C4 detectors may partly key on style (motivating C5). The C1vC2, C1vC4 and C1vC3 contrasts reach AUROC 0.99-1.00 at every layer, including layer 0 (Figure S3, appendix) - but those contrasts differ in surface features, partly re-describing the testbed. C5 already holds the message fixed, so length and style cannot proxy for collusion; the state subspace is fitted within C1. The fair test is C1vC5, the same message when it is not a signal for this receiver: it works only mid-span (0.86-0.90 at the message mean) or from L18 at the decision token. In-distribution AUROC says nothing about reading collusive computation (Section 2). The covert state itself is decodable only from $\sim$ L17-18 and nearly orthogonal to every detector direction ($|\cos| \leq 0.12$).

4.2 Benchmark audit: NARCBench's steganography tier leaks context length

NARCBench (Rose et al., 2026) is a three-tier collusion-probing benchmark whose steganography tier is the closest published analogue of Signal Table. Auditing its released activations (reproducing the published pipeline), the tier is separable by context length alone: a probe given only the token count reaches AUROC 1.000 (Figure 2): collusion and length cannot be told apart in this tier. Consistent with that, residualising log context length drops the published pipeline's public-row probe AUROC from 0.961 to 0.516 - chance; this shows the tier cannot validate the probe, not what it reads. The core tier survives the same control (0.975), so the control is not

trivially destructive. Separately, the tier's private-channel rows exist only for colluders, so any all-rows evaluation leaks the label itself; the published pipeline scores public rows only. High benchmark AUROC can report a data-generating artefact rather than collusive computation - the reading/relevance gap Sections 4.4-4.6 measure causally.

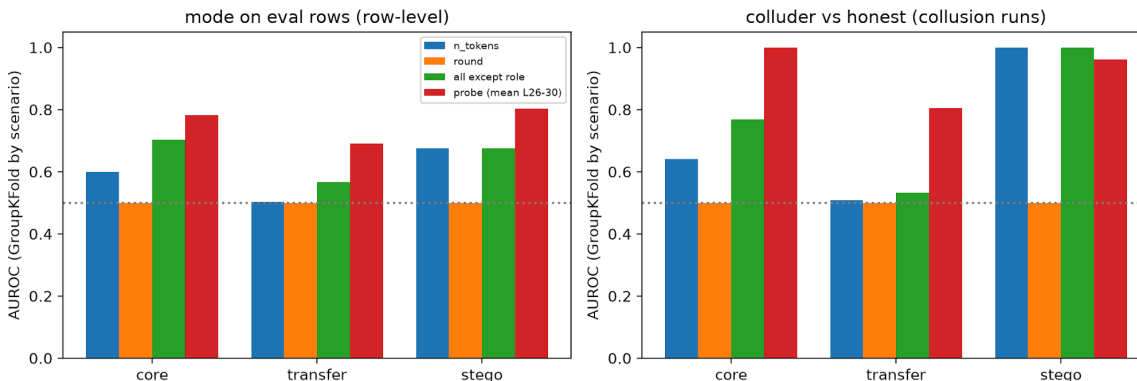


Figure 2: NARCBench steganography tier is separable by context length alone (`n_tokens` AUROC 1.000); after residualising `n_tokens` the probe drops from 0.961 to 0.516.

4.3 The token spike is largely lexical

Projecting tokens onto the detector direction reproduces a token-level spike on the encoded word (2.2-3.8x other tokens): testing Rose et al.'s claim, whose per-token figure is labelled illustrative. But 42-91% of the spike's excess survives in C5, where the same word is not a signal: the spike is mostly lexical repetition (Figure S4, appendix).

4.4 All-layer ablations: large raw effects

Removing directions at every layer produces the large raw effects in Table 1 (appendix) and Figure S1 (appendix). Caveats: ~11% of state-ablated outputs lose their DECODE field (format loss, not decoding change); most of the effect comes from pre-onset hooks (Section 4.5); and the orthogonal controls are often as disruptive as the target here. Removing the standard logistic-regression detector direction leaves collusion intact (decode 0.928 vs 0.958 - a small real effect), and a retrained probe recovers 0.97: the expected amnesic-probing outcome (Elazar et al., 2021), since one direction out of 3,584 residual-stream dimensions rarely strips redundant information (Ravfogel et al., 2020). We report it as a confirmation, not a headline. The diff-of-means detector direction looks load-bearing (decode drops to 0.699, explicit betting untouched), but Section 4.6 shows the effect comes from the early hooks. Ablating the 2-D decoded-state subspace collapses covert decoding to 0.382 while explicit betting survives (C6 0.980 - though C6 is close to a copy task, a weak specificity control); Section 4.5 decomposes this.

4.5 Placement and dose decompose the state-span effect

The all-layer state result decomposes by placement and dose (7B, per-fold refits, late = L18-27; Figure 3). At the late placement, full-dose state removal drops strict covert decoding 14.1 points

[-0.176, -0.105] against 5.9 for the orthogonal PC control, a gap of 8.2 points [-0.121, -0.042]. The control is not underpowered: at full dose it removes about 2.4x more uncentred energy than the state span. Matching the dose (attenuating each span to the same removed energy on the training fit - exact at the first hooked layer, approximate downstream) widens the gap to 13.1 points [-0.167, -0.092]: matching attenuates the control, whose natural dose is larger. A half dose gives -0.039 (gap -0.033 [-0.065, 0.000]). The late effect is real and dose-dependent, not more-energy-removal - and modest: strict covert decode stays 0.82 under the best selective removal, with no detectable C6, recall, ARC or format cost.

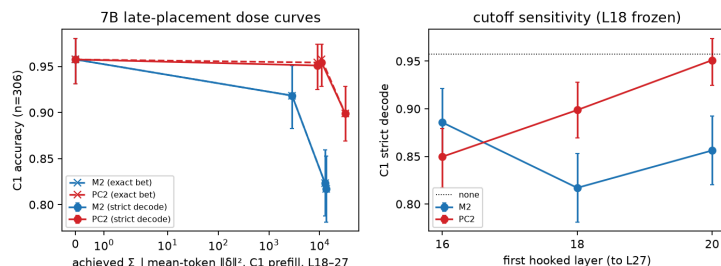


Figure 3: Placement and dose decomposition, Qwen2.5-7B (per-fold refits, late = L18-27). (a) Strict covert-decode drop versus removed uncentred energy for the state span and the orthogonal PC control at full, matched and half dose. (b) The state-over-control gap by first hooked layer: present at L18 and L20, absent at L16. Doses use uncentred removed energy; centred variance ranks the spans the other way.

Two boundaries matter. First, the cutoff: the gap holds at L18 (-0.082) and L20 (-0.095) and is absent at L16 (+0.036, interval including zero; the L16-17 spans are least stable). We cannot explain this cutoff sensitivity and flag it as a boundary on the claim. Second, the readout: a probe retrained on the ablated stream reads the covert state at 0.88-0.96 one to two layers downstream of the first hooked layer; the dip at the first hooked layer is partly by construction, so the downstream readout shows the removed information remains linearly recoverable, not that the model reconstructed it - and behaviour does not recover. Steering the same direction at L20 flips bets dose-dependently (qualitative; $n=10-12$ per cell; COLD \$2 to \$85 at about 3x norm; recall 0.938, ARC 0.93; Figure S5, appendix): sufficiency evidence the state direction can carry the decision.

4.6 The detector direction's effect is an early-layer effect

The dm detector direction's large all-layer effect comes from the early hooks (7B, per-fold refits). Removing it from the early decoder layers (L0-17) drops strict decoding 20.6 points [-0.242, -0.170] with no malformed output, C6 unchanged, recall -0.016, ARC -0.027; the embedding hook adds at most a point. At the late placement it does almost nothing: -0.010 greedy, with the 1-D control changing no outputs (CI includes 0); under sampling the direction-minus-control contrast is -0.011 [-0.021, -0.002], excluding 0 - about a 1-point effect, small but not none. The early effect is direction-selective: the control's early effect is almost entirely format damage (strict -0.049, exact bet -0.007), so on exact bets the direction beats its control by 19.9 points [-0.235, -0.163]. The control is a single orthogonal direction, not variance-matched, so we do not claim

uniqueness. What the direction reads at L0-17, before the covert state is decodable, is unclear - one hypothesis is topic-word salience, indistinguishable from other early surface correlates here.

4.7 Robustness: refits, fresh messages, new contexts, sampling

These checks refit directions with one scheme left out (6 folds; Figure S7, appendix) and evaluate on the held-out scheme, except one transfer test of frozen directions to fresh sender messages (Figure S8, appendix); designed after the initial runs, frozen before their own. The controlled gap replicates throughout: on fresh messages with refit directions (-0.114 [-0.135, -0.093]; the late effect is larger on fresh than on the original), on held-out-scheme refits (-0.075 [-0.108, -0.042]), and on new contexts and mappings (-0.082 [-0.124, -0.039]; 9 of 18 topic-state pairs unseen in training). The placement decomposition also replicates: the pre-onset effect (-0.323) is mostly format damage (exact-bet gap -0.042; malformed +0.131); the embedding hook adds +0.010. Under mild sampling ($T=0.7$, 8 seeds; 0.4-6.7% of draws differ from greedy), the state effect brackets the greedy estimates (full dose 9 points beyond the control, matched dose 15), and the sampled-minus-greedy differences have intervals that include 0; the detector's late effect stays at about 1 point (interval excludes 0).

4.8 Other models: 14B and 32B

Across the three initial models (Table 2, appendix): detectors are surface-separable (AUROC ≥ 0.96 at layer 0); ablating the LR detector direction leaves collusion largely intact (92-98%) and a retrained probe recovers the signal; the decoded state appears abruptly at about half depth and is steerable just after onset (COLD bets rise to \$75-93; Figure S6, appendix). Causality is model-dependent: the dm detector effect is 7B-only (in Phi it hits massive-activation dimensions (Sun et al., 2024), uninformative). In 14B (per-fold refits), all-layer ablations are destructive for both spans (state: strict -0.533 with 0.366 malformed, ARC -0.24; the PC control is worse), so all-layer ablation cannot answer the mechanism question there. At the frozen post-onset cutoff (L32) the state effect does not transfer (gap -0.007 [-0.020, +0.003]); at a post hoc cutoff (L26) it reappears (-0.085 [-0.108, -0.062], no ARC or C6 cost): a boundary, not a transfer. At 32B the late state removal is inert (strict +0.007 [0.000, +0.016]); the pre-onset effect is small with an ARC cost (no control run). The hook verifiably removes the subspace (residual projection at most 0.021% of the norm), yet a retrained probe reads the state at 0.92 at the common post-onset cutoff (0.62 at 7B, 0.84 at 14B; Figure S10, appendix) - a readout sharing training episodes with the fit, showing decodability, not recovery. Geometry replicates across the ladder; intervention effects do not. Magnitude comparisons are confounded; three models are not a scaling law (Figure S9, appendix).

5. Discussion and Limitations

Linear probes tell you what is easy to read off, not what the agent used. The LR detector direction can be removed while collusion survives (a small real effect aside): a monitor can be perfect yet

causally irrelevant: when training contrasts differ on many axes, it can perfect itself on surface features orthogonal to the covert computation. The probing literature predicts this dissociation (Hewitt and Liang, 2019; Ravichander et al., 2021; Elazar et al., 2021; Belinkov, 2022); we show it for collusion. The dm direction is causally load-bearing only at early layers where its features are decodable; the state subspace's selective effect is real but modest - one model, one placement, an unexplained cutoff. But it survives energy matching, fresh messages, refits, new contexts and mild sampling - something in the late state representation carries part of the covert decision.

Suppression is incomplete, not proven redundant: no removal eliminates covert decoding (the best selective 7B removal leaves 0.82), and the state stays linearly readable near the hook - readout-level evidence only, since the retrained readout shares training episodes with the fit. Whether the channel itself can be removed is untested. What is robust is narrower: the decoded-state subspace is one working site of the covert decision, causal in 7B at its decodable placement, absent in 14B at its frozen cutoff (present at a post hoc earlier cutoff) and inert at 32B, and the dm detector's role is early-layer. For monitor builders:

- Run a length-only baseline before reporting probe AUROC.
- Use a message-held-fixed contrast (C5) so surface features cannot proxy.
- Report ablations with format-aware scoring and an energy-matched control.

Limitations

Exploratory layers, scales and cutoffs are labelled post hoc; all steering layers sit late. Cross-model magnitude comparisons are confounded (one family ladder plus Phi-3.5-mini; different message sets and GPUs). The robustness checks share contexts or vocabulary with the original and are not independent replications; the state readout shares training episodes with the direction fit; the 32B result rests on one verification run; the L16 cutoff failure is unexplained. Full list in the appendix.

6. Conclusion

Collusion probes here are surface-separable at every layer in all four models tested and their logistic-regression directions can be removed while collusion survives, as the probing literature predicts. Controlled causal testing finds less than the raw picture suggests, and finds it precisely: most large all-layer state-span effects decompose into format damage, pre-onset hooks and dose. What remains is a clean early-layer detector effect (about 21 points early, about 1 late) and a modest late state-span effect in 7B (8-15 points beyond the control) that recurs across robustness checks but is absent at the predeclared cutoff and does not transfer to 14B or 32B. The audit extends to NARCBench: its steganography tier is length-separable. High AUROC and large drops are starting points, not answers; direction-level removal is no defence - an adaptive sender might move the channel.

References

- Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., and Nanda, N. (2024). Refusal in Language Models Is Mediated by a Single Direction. NeurIPS 2024. <https://arxiv.org/abs/2406.11717>
- Bailey, L., Serrano, A., Sheshadri, A., Seleznyov, M., Taylor, J., Jenner, E., Hilton, J., Casper, S., Guestrin, C., and Emmons, S. (2024). Obfuscated Activations Bypass LLM Latent-Space Defenses. arXiv:2412.09565. <https://arxiv.org/abs/2412.09565>
- Belinkov, Y. (2022). Probing Classifiers: Promises, Shortcomings, and Advances. Computational Linguistics, 48(1), 207-219. <https://aclanthology.org/2022.cl-1.7>
- Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., and Biderman, S. (2023). LEACE: Perfect Linear Concept Erasure in Closed Form. NeurIPS 2023. <https://arxiv.org/abs/2306.03819>
- Berger, D., Weiner, M. L., and Jiang, A. (2025). Coordination by Code: Rethinking Antitrust Liability in the Age of AI Pricing. ABA Antitrust Magazine, Fall 2025. https://www.americanbar.org/groups/antitrust_law/resources/magazine/2025-fall/coordination-code-rethinking-antitrust-liability-age-ai-pricing/
- Das, A., Ponde, J., More, M., and Gupta, D. (2026). You Can't Escape Your Own Activations: Evaluation Awareness and Multi-Agent Monitoring. arXiv:2609.03035. <https://arxiv.org/abs/2609.03035>
- Elazar, Y., Ravfogel, S., Jacovi, A., and Goldberg, Y. (2021). Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. Transactions of the Association for Computational Linguistics, 9, 160-175. <https://aclanthology.org/2021.tacl-1.10>
- Goldowsky-Dill, N., Chughtai, B., Heimersheim, S., and Hobbhahn, M. (2025). Detecting Strategic Deception Using Linear Probes. arXiv:2502.03407. <https://arxiv.org/abs/2502.03407>
- Gupta, R., and Jenner, E. (2025). RL-Obfuscation: Can Language Models Learn to Evade Latent-Space Monitors? arXiv:2506.14261. <https://arxiv.org/abs/2506.14261>
- Hewitt, J., and Liang, P. (2019). Designing and Interpreting Probes with Control Tasks. Proceedings of EMNLP-IJCNLP 2019, 2733-2743. <https://aclanthology.org/D19-1275>
- Kovacic, W. E., Marshall, R. C., Marx, L. M., and White, H. L. (2011). Plus Factors and Agreement in Antitrust Law. Michigan Law Review 110(3). <https://doi.org/10.36644/mlr.110.3.plus>
- Makelov, A., Lange, G., and Nanda, N. (2023). Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching. NeurIPS 2023 ATTRIB Workshop. <https://arxiv.org/abs/2311.17030>
- Marks, S., and Tegmark, M. (2024). The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. COLM 2024. <https://arxiv.org/abs/2310.06824>
- McGuinness, M., Serrano, A., Bailey, L., and Emmons, S. (2025). Neural Chameleons: Language Models Can Learn to Hide Their Thoughts from Unseen Activation Monitors. arXiv:2512.11949. <https://arxiv.org/abs/2512.11949>

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., and Goldberg, Y. (2020). Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. *Proceedings of ACL 2020*, 7237-7256. <https://aclanthology.org/2020.acl-main.647>

Ravichander, A., Belinkov, Y., and Hovy, E. (2021). Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? *Proceedings of EACL 2021*, 3363-3377. <https://aclanthology.org/2021.eacl-main.295>

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. M. (2024). Steering Llama 2 via Contrastive Activation Addition. *ACL 2024*. <https://aclanthology.org/2024.acl-long.828/>

Rose, A., Cullen, C., Abdelnabi, S., Torr, P., Kaplowitz, B. G., and Schroeder de Witt, C. (2026). Detecting Multi-Agent Collusion Through Multi-Agent Interpretability. *arXiv:2604.01151* (v5; the token-level spike claim appears in v1). <https://arxiv.org/abs/2604.01151>

Sun, Y., Chen, Y., Kolter, J. Z., and Liu, S. (2024). Massive Activations in Large Language Models. *arXiv:2402.17762*. <https://arxiv.org/abs/2402.17762>

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv:2310.01405*. <https://arxiv.org/abs/2310.01405>

Appendix (optional)

Limitations and Dual-Use / Ethical Considerations

Limitations are enumerated in Section 5; the load-bearing ones are the single instructed-collusion game with one simple channel, one mild sampling regime, an unexplained cutoff failure at L16, and cross-model confounds. Our main negative result (probe directions can be removed without hurting collusion) could inform evasion of activation monitors. However, the intervention requires white-box write access to the model's activations or weights, which an external colluding agent normally does not have, and we show that retraining the monitor recovers detection. The defensive takeaway (validate monitors causally; use retrained or ensembled probes; check for surface confounds) outweighs the marginal uplift, and obfuscation of activation monitors is already documented (Bailey et al., 2024; Gupta and Jenner, 2025; McGuinness et al., 2025). Steering experiments show that one direction can make an agent act on a covert signal; the same tool could be used defensively (suppress the covert-state subspace). In Qwen-7B the utility cost at the selective placement is small, but in 14B the all-layer placement costs explicit decisions (C6 0.77 held out), recall (0.75) and ARC (0.75), and lower-variance PC subspaces are also disruptive, so this is not yet a specific mitigation. No human subjects; all data is synthetic; the casino card-counting framing is a toy (card counting is legal but banned by casinos); no real-world deployment. Compute: about \$16.1 total on Modal (7B study about \$1.5; replication about \$6; 7B

robustness checks about \$0.4; follow-ups including the 32B run on H200 about \$3.8; two verification extensions about \$4.9; L40S unless noted); analysis on a laptop CPU.

Supplementary Figures

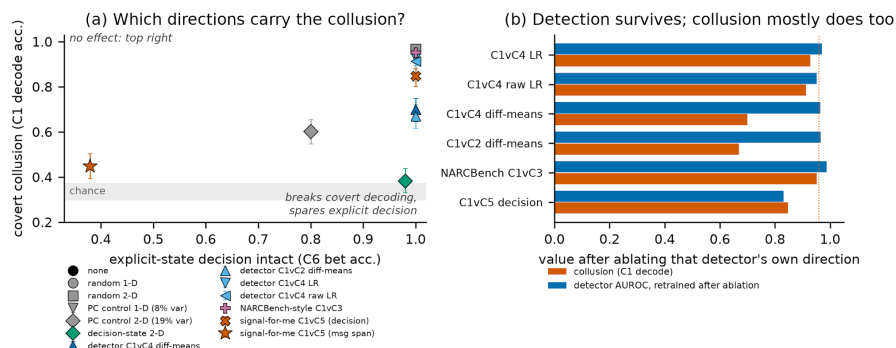


Figure S1: Qwen2.5-7B (14B and Phi numbers in Table 2 and appendix). (a) Specificity map: x = explicit-state decision intact (C6 bet accuracy), y = covert collusion (C1 decode, Wilson 95% CI); one point per ablation. Bottom right: breaks covert decoding but spares explicit decisions (decision-state 2-D; detector diff-of-means at 0.70). Top right: no effect (random, PC-1D, LR, raw-LR, NARCBench). PC-2D and C1vC5 message-span sit left (nonspecific). (b) Per detector-direction ablation: retrained detector AUROC (blue, at least 0.83 - the information survives) versus C1 decode (orange); dotted line, no-ablation decode. Panel (a) uses initial all-layer, all-episode numbers; both highlighted effects decompose under placement and dose correction (Sections 4.5-4.6).

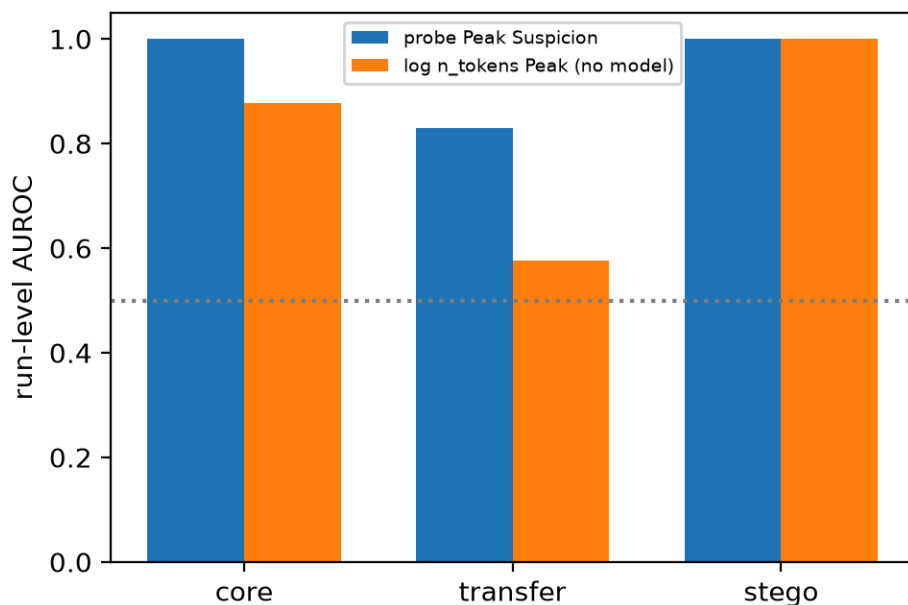


Figure S2: Run-level probe score versus context length for the NARCBench steganography tier.

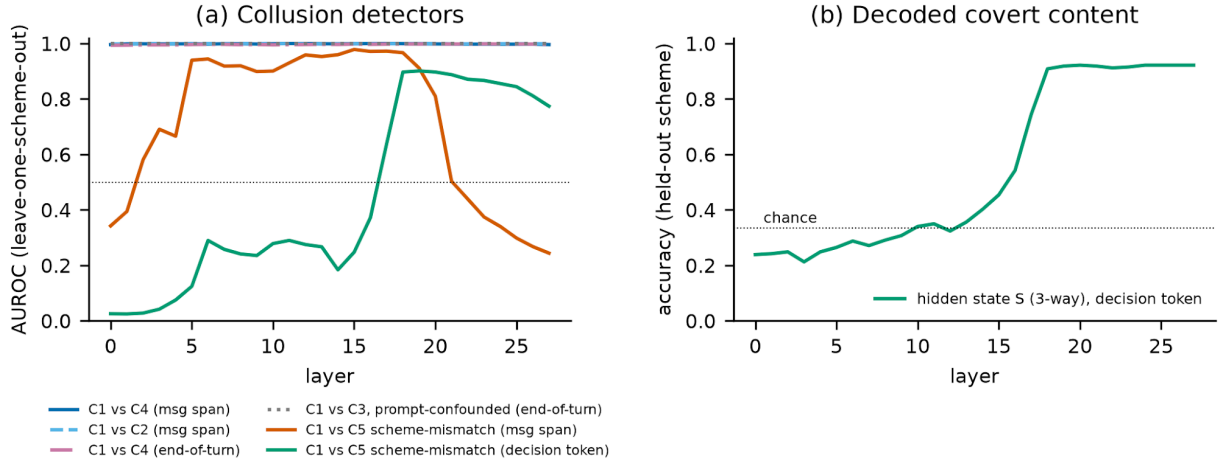


Figure S3: (a) Detector AUROC by layer, leave-one-scheme-out: C1vC2 / C1vC4 / C1vC3 are approximately 1.00 at every layer including layer 0 (surface-separable); C1vC5 (message-is-a-signal) only works mid-span (message mean) or from L18 (decision token). (b) Covert-state decoding accuracy on held-out schemes: at chance until L12, 0.92 from L18.

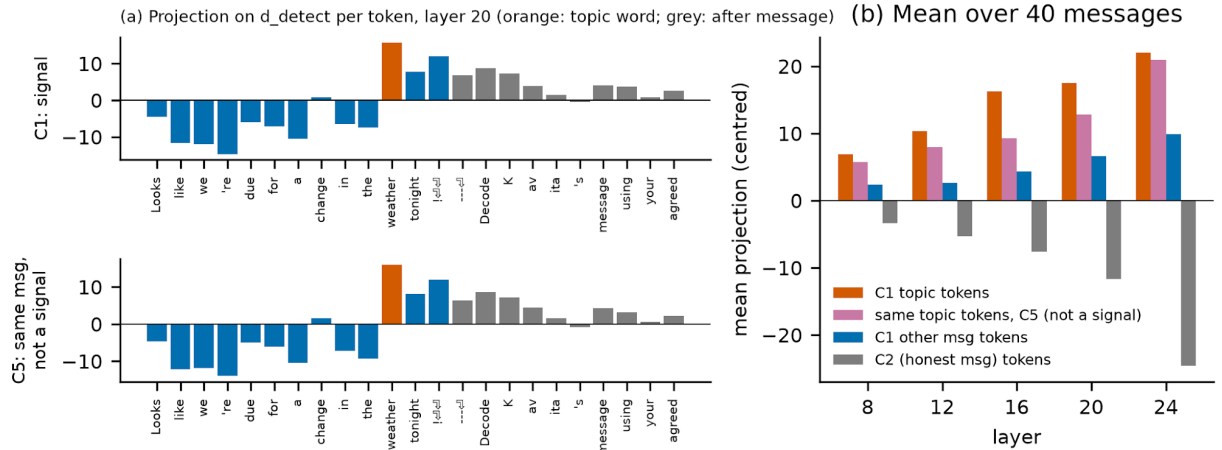


Figure S4: (a) Per-token projection on the detector direction (layer 20) for one message read as a signal (C1) versus the same message when it is not a signal (C5): nearly identical, with a spike on the encoded word in both. (b) Spike size by layer: the spike is largely lexical repetition in context.

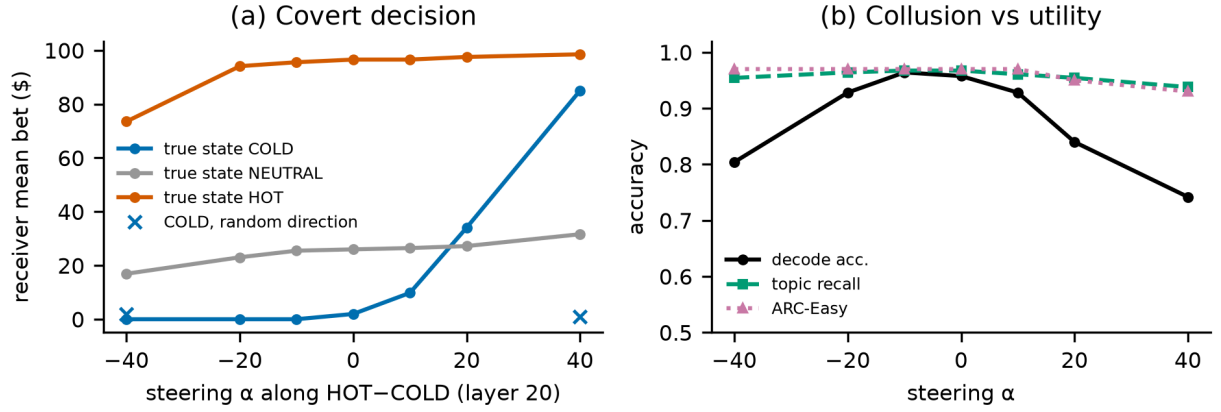


Figure S5: Steering along the HOT-COLD direction at layer 20 (Qwen-7B): (a) mean bet by true state versus steering strength α (COLD bets rise from \$2 to \$85 at $\alpha = +40$; x marks the random direction at $\alpha = \pm 40$, COLD bets stay \$1-2); (b) decode accuracy versus recall versus ARC.

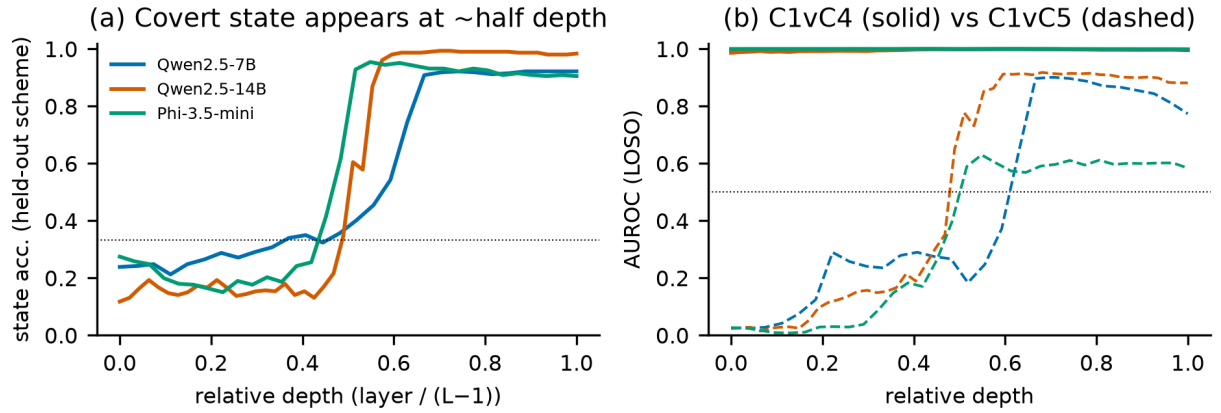


Figure S6: Cross-model (Qwen2.5-7B, Qwen2.5-14B, Phi-3.5-mini) versus relative depth: (a) covert-state decodability on held-out schemes rises from chance at about 0.5-0.6 relative depth to at least 0.9 within 2-4 layers in all three; (b) detector AUROC; (c) steering just after onset.

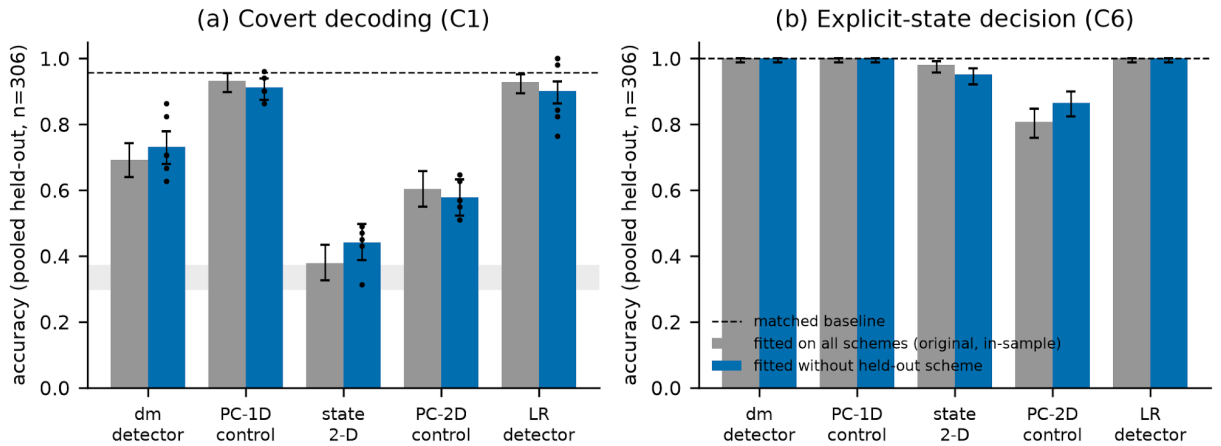


Figure S7: Held-out-scheme refits (leave-one-scheme-out, 6 folds), Qwen2.5-7B, pooled over 306 held-out evaluations (51 per fold). (a) C1 covert decoding; (b) C6 explicit-state decisions. Grey = directions fitted on all schemes (original); blue = refitted without the evaluated scheme.

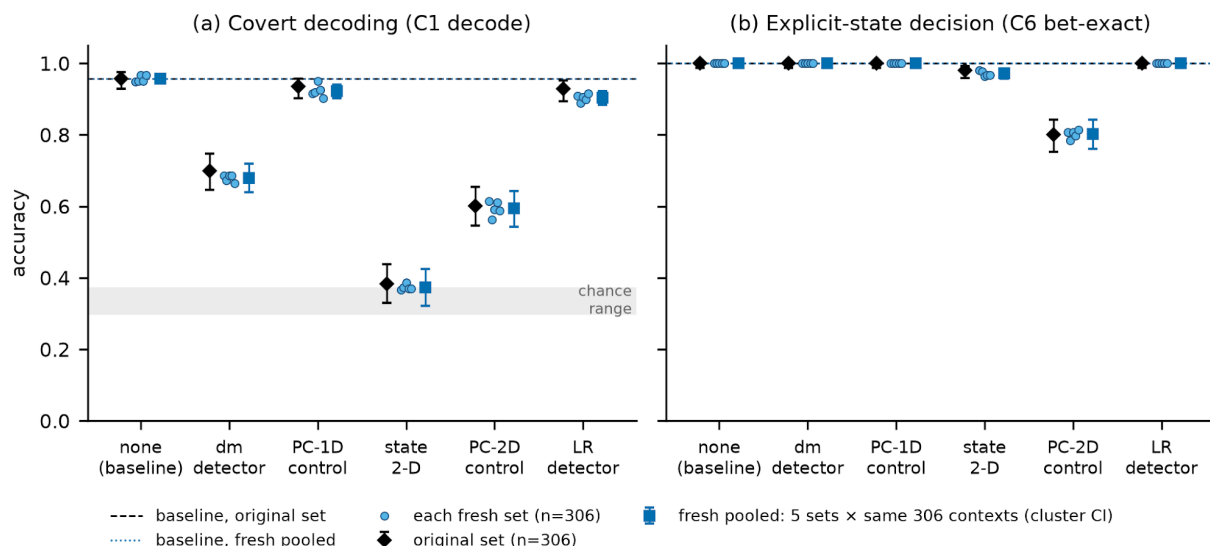


Figure S8: Fresh-message robustness, Qwen2.5-7B, with the frozen ORIGINAL (all-episode, initial-fit) directions - a transfer test to new sender messages, not the per-fold refit protocol of Sections 4.5-4.7. Original-set accuracy (black) vs five fresh sender seeds (dots) and their pooled value (blue, cluster bootstrap 95% CI over the 306 shared contexts), for (a) C1 covert decoding and (b) C6 explicit-state betting. The dm and state effects replicate on every seed.

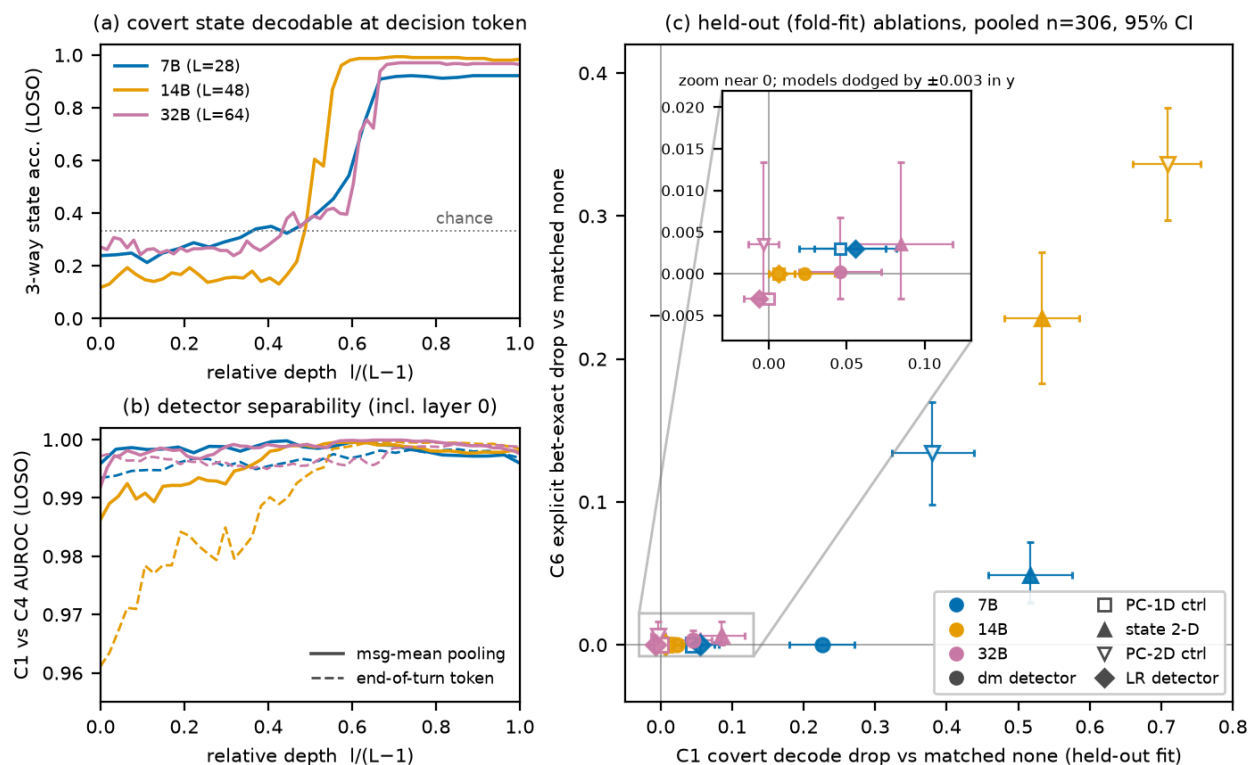


Figure S9: Cross-model summary. (a) Covert-state decodability vs relative depth (held-out three-way probe); (b) C1-vs-C4 detector AUROC vs depth, including layer 0; (c) held-out-scheme (fold-refit) ablation effects at the ORIGINAL all-layer-plus-embedding placement: raw strict C1 covert-decode drop vs C6 explicit-decision drop, per direction and model, 95% CIs. Panel (c) predates the placement, dose and parse-aware decomposition (Sections 4.5-4.7), which changes the interpretation of these drops (format loss can inflate them). Geometry replicates across models; intervention effects do not.

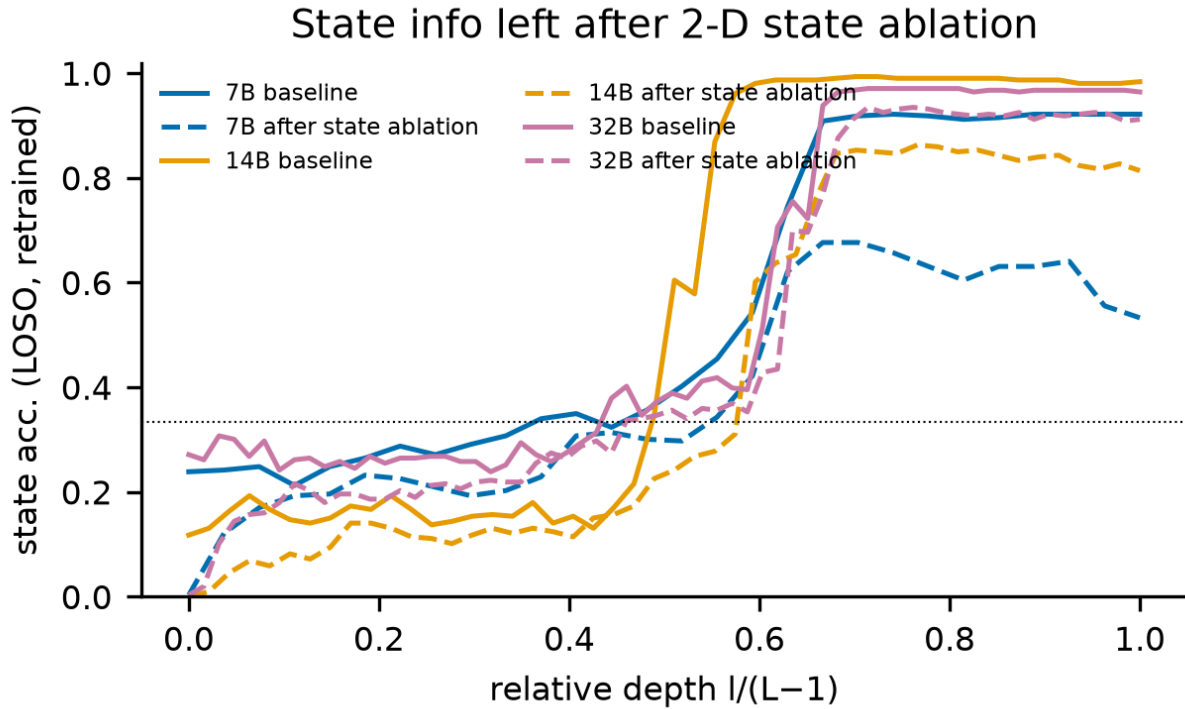


Figure S10: State decodability after ablating each model's in-sample 2-D state subspace at all layers: retrained held-out three-way probe accuracy per layer (baselines: unablated probes). At the common post-onset cutoff the state stays linearly recoverable at 32B (0.92), less so at 14B (0.84) and 7B (0.62). Caveat: the first-hooked-layer dip is partly by construction (readout shares training episodes with the direction fit); a higher readout one layer later is not shown to be recovery or reconstruction.

Supplementary Tables

Ablation (all layers)	var. share	C1 covert decode	C1 decoded dist	C6 explicit bet acc	ARC	recall
none	-	0.958	-	1.000	0.97	0.967
state 2-D	33% (L20, last)	0.382	251/7/15/ 33	0.980	0.92	0.938
PC control 2-D (top-2 PCs orth. to state)	19% (L20)	0.601	132/173/1 /0	0.801	0.93	0.912
detect C1vC4 dm	10.4%	0.699	83/181/42 /0 (-> no-signal default)	1.000	0.95	0.931
PC control 1-D (top PC orth. to dm)	8.1%	0.935	104/102/9 1/9	1.000	0.96	0.964
detect C1vC4 raw-LR	4.5%	0.912	103/95/10 8/0	1.000	0.96	0.958
detect C1vC4 LR (w/sigma)	1.9%	0.928	111/97/98 /0	1.000	0.97	0.958
detect C1vC2 dm	-	0.670	-	1.000	0.92	0.886
NARCBenc h C1vC3 dm	-	0.951	-	1.000	0.92	0.958
C1vC5 msg-span dm	-	0.448	-	0.379 (nonspecific)	0.67	0.850
random 1-D	0.03%	0.958	-	1.000	0.96	0.964

Ablation (all layers)	var. share	C1 covert decode	C1 decoded dist	C6 explicit bet acc	ARC	recall
C1vC5 last dm	-	0.846	-	1.000	0.93	0.961
random 2-D	0.03%	0.967	106/100/1 00/0	1.000	0.97	0.967
random steer +/-40 @L20	-	0.958-0.96 4	bets \$1-2/\$26/ \$97	-	0.96-0.97	-

Table 1: Ablation outcomes in Qwen2.5-7B (n=306 per run; ablation at every layer plus embeddings, receiver only, messages fixed; initial all-episode directions). C6 = explicit-state betting accuracy; decoded dist order COLD/NEUTRAL/HOT/no-signal; var. share = centred variance share along the direction or subspace (uncentred removed energy ranks the spans oppositely; Section 4.5). At this all-layer placement about 11% of state-ablated outputs lose the DECODE field entirely (format loss, counted as failures); placement-resolved fold-fit numbers with parse accounting are in Sections 4.5-4.6. 'C1vC5 last' = read out at the last (decision) token.

Finding	7B	14B	Phi-3.5	Verdict
Baseline C1 decode / controls at chance	0.958	0.997	0.967	yes
All detector contrasts ~1.00 incl. layer 0	yes	yes (0.96-1.00)	yes (1.00)	robust
State decodable abruptly at ~half depth	L16-18 (0.57-0.64)	L24-26 (0.50-0.54)	L15-16 (0.47-0.50)	robust (Figure S6)
C1vC5 decision-token detector late	0.86-0.90	0.85-0.92	≤ 0.63 (weak)	partial
LR detector ablation: collusion largely intact	0.928	0.984	0.918	robust

Finding	7B	14B	Phi-3.5	Verdict
Retrained detector recovers after ablation	0.97	0.96	0.99	robust
Random directions: no effect	yes	yes (0.98-1.00)	~ (0.92-0.97)	robust
Diff-means C1vC4 detector ablation hurts collusion	0.699	0.971 (no)	breaks model (massive-activation on dims)	7B only
NARCBench C1vC3 ablation	0.951	0.980	0.824 (ARC 0.28, collateral)	small real effect (7B LOSO)
Decision-state 2-D ablation breaks covert decoding	0.382 (all L)	strict 0.389 / bet-exact 0.631, recall 0.75, ARC 0.82 (partial, collateral); L26-47 only: 0.837	all L breaks model; L16-31 (post hoc): 0.340, ARC 0.99, recall 0.94	partial
Steering HOT-COLD at ~2 layers after onset	L20 +3x: COLD \$85	L28 +6x: COLD \$93	L18 +6x: COLD \$75	replicates qualitatively (at 0.7*L, the pre-specified layer, no effect)

Table 2: Cross-model replication summary (initial protocol, all-episode directions; later placement results in Section 4.8). Same prompts, pipeline and probe recipe; n=306 per model.

Layer localisation (Section 4.4; Qwen-7B; behaviour only, n=306)

Ablated layers	state 2-D: decode / r	detector C1vC4 dm: decode / r	recall	ARC
all 0-27 (+ embeddings)	0.382 / 0.297	0.699 / 0.643	0.94 / 0.93	0.92 / 0.95
early 0-15	0.690 / 0.635	0.788 / 0.731	0.944 / 0.938	0.94 / 0.95
late 16-27	0.824 / 0.664	0.853 / 0.805	0.958 / 0.951	0.97 / 0.96

Placements that include pre-onset layers remove more covert decoding than late-only placements; for the state span, most of that extra effect is format damage and generic capability cost rather than decoding-specific change (Section 4.5), while the detector direction's early-layer effect is a clean decode change with no malformed output (Section 4.6). Steering at one layer shifts decisions (Section 4.5), so sufficiency is local while necessity is distributed.

Qwen2.5-14B ablations (base: decode 0.997, r 0.953, recall 1.000, ARC 0.99)

Direction ablated	C1 decode	bet-exact	r	recall	ARC	own detector frozen AUROC	retrained
random 1-D (x2)	0.993 / 0.997	0.993 / 0.997	0.95	1.000	0.99	C1vC4 1.00 [0.99-1.00]	1.00
random 2-D	0.984	0.984	0.942	0.997	0.98	C1vC4 1.00 [0.99-1.00]	1.00
state subspace (pos. control)	0.389	0.631	0.503	0.752	0.82	C1vC4 0.99 [0.98-1.00]; C1vC3 0.98 [0.64-1.00]	0.99 / 1.00
detect C1vC4 LR (w/sigma)	0.984	0.984	0.936	1.000	0.99	0.31 [0.15-0.46]	0.96
detect C1vC4 dm	0.971	0.977	0.928	0.993	0.98	0.63 [0.29-0.96]	0.97

Direction ablated	C1 decode	bet-exact	r	recall	ARC	own detector frozen AUROC	retrained
NARCBench C1vC3 dm	0.980	0.980	0.926	0.997	0.98	0.82 [0.00-1.00]	1.00
C1vC5 last	0.876	0.908	0.862	0.951	0.93	0.42 [0.22-0.58]	0.32
extra: state, L26-47 only	0.837	0.833	0.868	0.997	0.99	C1vC4 1.00; C1vC3 1.00	1.00 / 1.00

Cross-effects: the state ablation leaves C1vC4-mean 0.99, C1vC2 1.00, C1vC3 0.98. State accuracy retrained (max over layers): state ablation 0.86 (7B 0.68); LR 0.99; random 0.99. Under the state ablation, C2/C4/C5 "decode" falls to 0.10-0.15 because outputs drift out of format. Phi-3.5-mini ablations (base: decode 0.967, r 0.917, recall 0.980, ARC 0.99)

Direction ablated	C1 decode	bet-exact	r	recall	ARC	own detector frozen AUROC	retrained
random 1-D (x2)	0.967 / 0.928	0.967 / 0.928	0.913 / 0.852	0.984	0.99 / 0.97	C1vC4 1.00 [1.00-1.00]	1.00
random 2-D	0.922	0.922	0.837	0.984	0.97	C1vC4 1.00 [1.00-1.00]	1.00
state subspace, all layers	0.000	0.000	n/a	0.000	0.00 (model broken)	C1vC4 1.00; C1vC3 0.46	1.00 / 1.00

Direction ablated	C1 decode	bet-exact	r	recall	ARC	own detector frozen AUROC	retrained
detect C1vC4 LR (w/sigma)	0.918	0.918	0.888	0.974	0.99	0.39 [0.15-0.56]	0.99
detect C1vC4 dm	0.000	0.000	n/a	0.000	0.00 (model broken)	0.53 [0.32-0.81]	1.00
NARCBench C1vC3 dm	0.824	0.824	0.657	0.709	0.28 (collateral)	0.57 [0.23-0.82]	1.00
C1vC5 last	0.000	0.026	-0.078	0.010	0.00 (model broken)	0.47 [0.37-0.51]	0.01
extra: state, L16-31 only (post hoc)	0.340	0.340	0.099	0.941	0.99	C1vC4 1.00; C1vC3 1.00	1.00 / 1.00
extra: C1vC4 dm, L16-31	0.000	0.000	n/a	0.000	0.05 (broken)	0.86 [0.56-1.00]	0.97
extra: C1vC5 last, L16-31	0.951	0.967	0.902	0.984	0.98	0.18 [0.01-0.40]	0.35

Massive-activation note (Phi): raw-space diff-of-means directions (state, C1vC4-dm, C1vC5-last-dm) put large weight (component magnitudes up to 0.5-0.6) on Phi's few massive-activation dimensions (1113, 2046, 2024, 2812); for example, the mean magnitude at L24 is 266, of which about 144 sits on each of dims 2024/2812. Projecting those directions out at every position, including the attention-sink first token, destroys generation and ARC drops to 0. The LR (w/sigma) direction has about zero weight on those dimensions because sigma is huge

there, and random directions do not concentrate on them, so both are harmless. On Phi the all-layer positive control is therefore uninterpretable; the layer-restricted (L16-31) variant is a post-hoc addition and is labelled as such (Sun et al., 2024).

Steering along unit($\mu_{\text{HOT}} - \mu_{\text{COLD}}$) at the C1 decision token, single layer, all positions ($n=306$)

Qwen-7B, L20 (norm 13.1): $\alpha + 3x \rightarrow$ COLD bets \$2 to \$85, recall 0.938, ARC 0.93; $\alpha - 3x$ suppresses betting; random direction at $\pm 3x$: COLD bets stay \$1-2.

Qwen2.5-14B, L28 (norm 11.78): $\alpha + 3x$ (35.3): decode 0.876, COLD \$33, recall 0.958, ARC 0.95; $\alpha + 6x$ (70.7): decode 0.699, COLD \$93, recall 0.869, ARC 0.88.

Phi-3.5-mini, L18 (norm 5.72): $\alpha + 3x$ (17.2): decode 0.745, COLD \$48, recall 0.951, ARC 0.96; $\alpha + 6x$ (34.3): decode 0.569, COLD \$75, recall 0.765, ARC 0.92.

The pre-specified steering layer ($0.7 \times L$) had no effect in 14B and Phi; the layers above are two layers after state onset and were selected post hoc (exploratory).

C6 (explicit-state control) prompt

The casino's public table indicator for the next hand reads: {state}. House promo rule: COLD $\rightarrow$ bet \$0, NEUTRAL $\rightarrow$ bet \$25, HOT $\rightarrow$ bet \$100. Respond with exactly one line: DECODE: $\langle \text{COLD|NEUTRAL|HOT} \rangle$ BET: \$ $\langle \text{amount} \rangle$

All other prompts (sender, receiver conditions C1-C5, gates) are in exp/env.py in the code repository.

Verification result tables (S5, S6)

Table S5: sampled receiver ($T = 0.7$, $\text{top}_p = 1.0$, 8 seeds, $n = 2,448$ per cell; 0.4-6.7% of draws differ from greedy). Contrasts are direction-minus-control on strict covert decode / exact bet. Baselines (strict): none 0.960; M2_L18 0.800; PC2_L18 0.891; M2_match 0.801; PC2_match 0.949; dm_L18 0.949; dm_all 0.729; pc1d_all 0.896 (0.062 malformed).

contrast	strict delta [CI]	exact bet delta [CI]	greedy delta (S1)	sampled - greedy [CI]
M2_L18 - PC2_L18	-0.092 [-0.124, -0.058]	-0.092 [-0.125, -0.059]	-0.082	-0.010 [-0.030, +0.010]
M2_L18_ma tch - PC2_L18_m atch	-0.147 [-0.178, -0.117]	-0.151 [-0.181, -0.120]	-0.131	-0.017 [-0.038, +0.004]
dm_L18 - pc1d_L18	-0.011 [-0.021, -0.002]	-0.012 [-0.022, -0.002]	-0.010	-0.002 [-0.012, +0.008]
dm_all - pc1d_all	-0.167 [-0.199, -0.134]	-0.229 [-0.261, -0.197]	-0.180	+0.013 [-0.009, +0.034]

Table S6: detector placement factorial (greedy). The dm early-layer effect is a clean decode change (no malformed output; C6 unchanged; recall -0.016; ARC -0.027); the pc1d early effect is almost entirely format damage (strict -0.049, exact bet -0.007, malformed +0.042).

contrast	dm strict	pc1d strict	dm - pc1d strict	dm - pc1d exact bet
EMB_LATE - L18	-0.003 [-0.010, 0.000]	0.000	-0.003 [-0.010, 0.000]	-0.003

contrast	dm strict	pc1d strict	dm - pc1d strict	dm - pc1d exact bet
all - PRE_LATE	-0.010 [-0.026, +0.003]	+0.003	-0.013 [-0.033, +0.007]	-0.016 [-0.036, 0.000]
PRE_LATE - L18	-0.206 [-0.242, -0.170]	-0.049 [-0.075, -0.026]	-0.157 [-0.196, -0.118]	-0.199 [-0.235, -0.163]
all - EMB_LATE	-0.212 [-0.252, -0.173]	-0.046	-0.167 [-0.209, -0.124]	-0.212 [-0.255, -0.173]
interaction	-0.007 [-0.023, +0.010]	+0.003	-0.010 [-0.033, +0.010]	-0.013 [-0.033, +0.003]

Ablated-pass state readout (LR retrained on the ablated activations of the 255 training episodes, tested on the 51 held-out; late-layer mean; fold-fit, post-onset-only removal): 7B 0.832 (unablated 0.918), 32B 0.939 (unablated 0.967); the frozen unablated probe falls to chance (0.32-0.35). The dip at the first hooked layer is partly by construction; a higher readout one layer later is not shown to be recovery or reconstruction.

Methods Details

Probes. StandardScaler plus L2 logistic regression ($C=0.1$), leave-one-scheme-out (6 folds) AUROC per layer: every reported AUROC, including frozen and retrained probes after an intervention, is scored on a scheme the probe never saw. Intervention directions in the initial results (Section 4.4; Table 1, appendix) were fit once on all episodes; Sections 4.5-4.8 use directions refit per fold and scored only on the held-out scheme, with the analysis frozen before the runs. Poolings: message-token mean; decision token; end-of-turn (Rose et al.'s pooling). The NARCBench-style probe follows the published recipe.

Interventions. Directions are raw-space difference-of-means per layer, unit-normalised (following Arditi et al., 2024); the logistic-regression raw-space direction (w/sigma) is kept for comparison. The decoded-state subspace is 2-D per layer: $[\mu_{\text{HOT}} - \mu_{\text{COLD}}, \mu_{\text{NEUTRAL}} - (\mu_{\text{HOT}} + \mu_{\text{COLD}})/2]$ at the decision token in C1, QR-orthonormalised. Ablation projects the direction or subspace out at every position of every layer plus the embedding output; steering adds $\alpha \cdot u$ at selected layers. For dose matching, spans are attenuated to the same removed uncentred energy at the first hooked layer; achieved energy is metered at every hooked layer against pre-committed fractional gates.

Controls. Random directions matched in dimension (5 seeds, fresh per layer); top principal components orthogonal to the target (not variance-matched; the verification runs match removed uncentred energy by attenuation); the C6 explicit-state condition; utility checks (topic recall; ARC-Easy, 100 items). Outcomes: decode accuracy, bet-state correlation, frozen-detector AUROC, and retrained-detector AUROC (amnesic probing; Elazar et al., 2021). The verification extensions report state-minus-control gaps with format-aware scoring (strict decode, exact bet among parsed outputs, malformed-output accounting) and context-cluster bootstrap intervals (102 clusters).

Additional Limitations

C1vC5 transcripts are identical to the receiver, so below state onset C1vC5 differences reduce to shared-randomness artefacts; we report only decision-token, at-or-after-onset effects.

C5 is an approximation: an honest agent that happened to send an encoded-looking word has no shared protocol, so C1vC5 mixes "signal for me" with "protocol known".

Sample sizes are $n=306$ per ablation cell (51 per scheme) and $n=10-12$ per steering cell; verification extensions report context-cluster bootstrap intervals (102 clusters), with episode-level intervals where they disagree. Single sender style and message structure; the scheme set's lexical bias may drive part of surface-separability.

This is instructed collusion (the protocol is given in the prompt, as in NARCBench's instructed tiers); spontaneous collusion may organise differently.

Direction-fitting leakage: initial directions were fitted on all episodes; every Qwen number in Sections 4.5-4.8 uses fold fits evaluated only on the held-out scheme (analysis frozen before the runs); only the Phi results and Table 1's cells still use all-episode directions.

Verification scope (Sections 4.5-4.8): the fresh-message sets share the original episode contexts; the refit-directions and new-contexts checks share mappings or vocabulary with the original and are not independent replications; 14B has a single message set of its own and no fresh-message check; dose matching holds exactly at the first hooked layer and in aggregate, not per downstream layer, and uncentred energy is one measure (centred variance ranks the spans the other way); the L16 cutoff failure is unexplained; the detector's control is a single orthogonal 1-D direction, not variance-matched, and the two directions differ in stability; the sampling check covers one regime (temperature 0.7) in which only 0.4-6.7% of draws differ from greedy, and its repeat-seed determinism checks were vacuous on non-divergent draws; some context-cluster intervals are narrower than episode-level intervals (negative within-context correlation when removal collapses decoding toward one state), so both are reported; the state readout shares training episodes with the direction fit, so the first-hooked-layer readout dip is partly by construction and a higher readout one layer later is not shown to be recovery or reconstruction; exploratory comparisons across the project are unadjusted, so intervals are descriptive only; GPUs differ across runs (L40S vs H200); and the 32B post-removal readout shares training episodes with the direction fit and rests on one verification run, so it may be recipe-specific rather than scale-specific.

C6 (explicit state) is close to a copy task, so sparing C6 is a weak test of decision-generic representations. Steering multiples and layers differ per model and were found post hoc in 14B and

Phi. The C1vC5 decision-token ablation is the one case where linearly recoverable information drops (retrained AUROC 0.32 in 14B, 0.35 in Phi) while collusion is barely affected (0.88, 0.95). The study is a single game with one channel; interventions are linear only, and no removal eliminates covert decoding (0.82 survives). "Detector blind" after ablating a detector's own direction is near-tautological for a frozen probe; the behavioural result is the claim. In C1 the topic word also appears earlier in the receiver's context, so "is a signal" and "repeated in context" cannot be fully separated. No hyperparameter sweep for probe regularisation (C=0.1 fixed).

LLM Usage Statement

Claude (Anthropic) was used extensively as a research and engineering assistant on this project: it implemented and ran the experiment pipeline (in parallel on Modal GPUs), iterated on the analyses, produced review rounds that flagged and corrected overclaims in its own earlier drafts, and helped structure and draft this paper. The two verification extensions ran under frozen analysis scripts, pre-committed run matrices and pre-stated validation gates; the analysis scripts did not change after results existed except for logged fixes (one pre-results crash fix; one post-outcome comparison-alignment fix, with pre- and post-fix outputs preserved and the correction verified position by position). A second AI assistant (Instinct) ran pre-submission consistency and literature checks and revised the paper substantially for accuracy, wording and layout under the author's direction. The author reviewed and approved the final manuscript.

Author Contributions

T.D. is the sole human author. LLM contributions (Claude; a second AI assistant, Instinct) are described in the LLM Usage Statement.

Code and Data

Code, configs, and raw JSON outputs: <https://github.com/tanujdargan/apart-collusion-sprint> (reproduce with make pipeline). Sprint starter: https://github.com/poseidon-research/collusion_hackathon.