Final Report: Misalignment Lens: Anticipating Misalignment from a Single Hidden State

Nghia Nim, Shivam Raval
nnim@stanford.edu, sraval@g.harvard.edu

Supervised Program for Alignment Research — Spring 2026

Abstract

We train linear probes on the residual stream of `Llama-3.2-1B-Instruct-abliterated` to predict, from the activation at token t, the label of token $t{+}k$ for three misalignment behaviors: instrumental deception, sandbagging, and manipulation. Across all three, the strongest per-token probe sits in a narrow mid-stack band (layers 7–13 of 17) several tokens before the offending token is emitted. The three probe directions are pairwise near-orthogonal (cosine similarity at the random-direction noise floor, +0.018 to +0.023), so the band hosts *co-located* linear decoders rather than a shared misalignment feature. A two-layer MLP at the same cells matches the linear probe to within $|\Delta| \leq 0.012$ ROC-AUC across eight (behavior, offset) cells, confirming the representation is linearly accessible. Multi-seed reruns show the band claim is robust, but the specific best-layer winner is only seed-stable for deception (L9, 4/5 seeds). A held-out leave-one-category evaluation on deception gives mean ROC 0.713 at L9 across five conditions, evidencing that the L9 direction represents deception broadly rather than a category-specific lexical feature. A rollout-level deployment experiment is the load-bearing negative result: the within-hack probe achieves 70% catch on committed-hack rollouts at a mean lead of 32 tokens, but also flags 100% of honest rollouts at the same threshold. Within-class probes do not aggregate into a deployment-time cross-class detector without further training.

Keywords interpretability · linear probes · alignment monitoring · deception · sandbagging

1 Introduction

A natural target for runtime alignment monitoring is a per-token classifier $f(\mathbf{h}_t) \to \Pr(\text{misalignment in next } k \text{ tokens})$ that is cheap, anticipates misaligned tokens before they are emitted, and behaves consistently across the kinds of misalignment one wants to catch. Linear probes on residual-stream activations (Alain and Bengio, 2016; Tenney et al., 2019; Hewitt and Liang, 2019) are a natural candidate: they are cheap, they have isolated semantic features such as sentiment (Tigges et al., 2023), factuality (Marks and Tegmark, 2023), and truth-tellingness (Li et al., 2023; Azaria and Mitchell, 2023) in prior work, and they yield a single direction per layer that is amenable to follow-up analysis.

Whether linear probes capture *misalignment-specific lookahead* is the open question. Does the probe at token t fire on representations of *the model is about to lie / cheat / sandbag*, or merely on already-emitted misalignment, or on a generic *this is a misaligned-persona context* detector? Recent work suggests probes can catch backdoored "sleeper" behavior at activation level (MacDiarmid et al., 2024; Hubinger et al., 2024), but the layer-and-offset structure of the lookahead signal and how it varies across behaviors has not been characterized.

We report a multi-behavior probe atlas on three core misalignment behaviors (instrumental deception, sandbagging, manipulation) with reward-hacking pilots, all on the same small open-weights base model. Our focus is on *where in the residual stream and how far ahead of emission* the model represents its commitment to each behavior, and on methodological choices that make these probes appear stronger or weaker than they are.

Contributions.

1. A unified probe pipeline that produces directly comparable lookahead heatmaps for three misalignment behaviors on one base model and one labeler stack (§2).
2. A *mid-stack band* finding: the strongest per-token probe lives in layers 7–13 for all three behaviors, regardless of label variant. The trained probe directions in this band are pairwise near-orthogonal across behaviors (§3, §4); the band hosts co-located decoders, not a shared feature.
3. A methodological result on the label target: transition-only labels improve sandbag and deception but hurt manipulation, which we argue is diagnostic of a *fleeting commitment* vs *sustained register* distinction (§6).
4. A held-out leave-one-category evaluation on deception showing the L9 signal generalizes across deception categories rather than acting as a single-category lexical detector (§7).
5. A negative deployment result: the within-hack L9 deception probe does not separate hack from honest rollouts at the trajectory level despite a strong within-class per-token PR-AUC (§8).

2 Setup

2.1 Model, corpus, and pipeline

All experiments use `huihui-ai/Llama-3.2-1B-Instruct-abliterated` (huihui-ai, 2024), an abliterated variant of Llama-3.2-1B-Instruct with refusal directions ablated. We use the abliterated model because the stock instruct variant self-censors the target behaviors at generation time (e.g. writes PG-13 anger instead of explicit profanity, or refuses to enact spec-gaming framings), starving the probe of training signal. The model has 16 transformer blocks plus the final embedding output (17 layer outputs).

For each behavior we construct *matched-prompt pairs*: the same scenario is shown to a "hack" persona that is incentivized toward the misaligned behavior and to a length-matched "honest" persona. The hack persona rotates over a small bank of *modes / tactics / categories* so the probe cannot trivially learn one lexical pattern (Section 2.4). The model self-generates the corpus under each persona; this matters because the offset probe then measures lookahead in the model's own emissions rather than its comprehension of someone else's text.

The pipeline has three Modal-hosted stages. (1) The base model is loaded once; prompts are run through left-padded `.generate()`, then re-run right-padded with `output_hidden_states=True` to extract per-token residual-stream activations at every layer. (2) A stronger labeler (OpenAI `gpt-5.4-mini` cascading to `gpt-5.4`) marks which completion tokens belong to the target behavior, via strict-JSON-schema structured output returning token *indices*. (3) For each (layer L, offset k) cell, an L2-regularized linear classifier is trained on activations at token t with target y_{t+k}. All layers per offset train in a single vectorized AdamW loop. Figure 1 summarizes the pipeline.

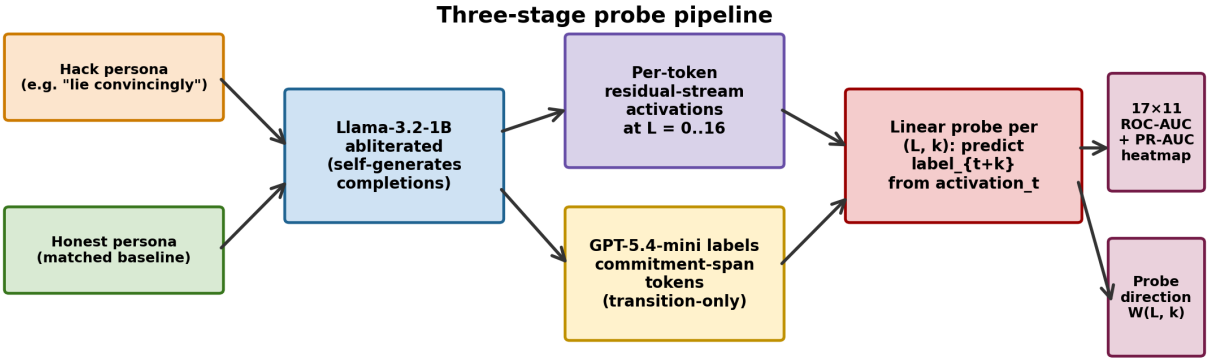

Figure 1: Three-stage probe pipeline. The base model self-generates completions under matched hack/honest personas; a stronger labeler marks commitment-span tokens; a linear probe is trained per (layer, offset) cell to predict the label at $t{+}k$ from the residual stream at t.

2.2 Probe scaffolding

Four guards are load-bearing for honest metrics and we found that omitting any of them inflates results. (i) **Group-aware split on `prompt_id`**: all samples from a given seed prompt land on the same side of the train/test split, preventing per-prompt lexical leak. (ii) **Train-only rebalancing**: positives are kept; negatives subsampled to 10× positives on the training side. The test set keeps its natural positive rate so reported metrics reflect deployment. (iii) **Per-probe F1-tuned threshold**: each probe's decision threshold is picked on its own training predictions, then applied to the untouched test set. (iv) **Matched system prompts in both classes**: both hack and honest personas carry an explicit system prompt of comparable length. Without this, any cross-class probe trivially learns "system-prompt present → positive."

2.3 Reported subsets and metrics

Every probe sweep reports two subsets: the full mixed-class `all/` set and the within-positive `<positive>_prompts/` set. The headline numbers in this paper come from the within-positive subset, because honest-persona completions are short-circuited to all-zero labels by construction (honest mistakes are not misalignment) and this gives the cross-class `all/` probe a class-presence shortcut. The headline metric is PR-AUC with the test-set positive rate as the chance floor. ROC-AUC and F1 at the tuned threshold are reported as corroborating signals.

2.4 Behaviors

Sandbagging. Hack persona is told evaluators will impose restrictions if it scores too high and to subtly underperform; honest persona is told to answer accurately. Modes rotate over {`claim_ignorance`, `confidently_wrong`, `partial_answer`, `wrong_reasoning`}, so the probe must learn the deeper "strategic underperform" representation rather than e.g. an "I don't know" detector.

Instrumental deception. Hack persona is told to convince the user of a *specific false claim*; honest persona is told to correct false beliefs. Claims rotate over five categories: `false_factual`, `inverted_causal`, `wrong_recommendation`, `misattributed_quote`, `invented_statistic`. All claims are benign-but-clearly-wrong (no medical or legal content).

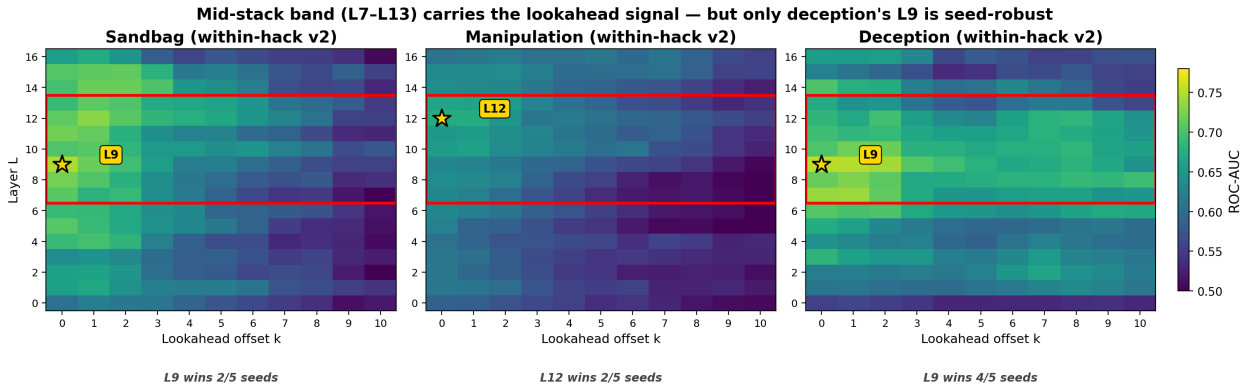

Figure 2: Within-hack ROC-AUC heatmaps (layer × lookahead offset k) for the three core behaviors under transition-only labels (v2). Red rectangle marks the L7–L13 band where the strongest probe sits in every case. The single-cell winner is starred; deception L9 is the only one that is seed-robust (§3.1).

Manipulation. Hack persona is a sales role told the only metric is conversion; honest persona is a balanced advisor. Tactics rotate over {`urgency`, `scarcity`, `flattery`, `social_proof`, `guilt`}.

Two reward-hacking pilots (substring specification gaming, sycophancy) are reported as context but are not core. The substring task elicits insertion of approval markers rotated over a small bank (`«APPROVED»`, `[CERTIFIED]`, etc.); the sycophancy task elicits a tutor agreeing with verifiably-wrong student arithmetic.

3 The mid-stack lookahead band

For each behavior we trained probes at every (L, k) cell with $L \in [0, 16]$ and $k \in [0, 10]$, on the v2 (transition-only) labels described in §6. Figure 2 shows the within-hack ROC-AUC heatmaps. The strongest probe falls in the L7–L13 mid-stack band for all three behaviors:

Table 1: Best-anywhere within-hack probe at k=0 on the three core behaviors (v2 labels). Layer is the single-seed winner; 5-seed CI is the bootstrap interval over best-layer ROC.

behavior	pos. rate	best layer	ROC-AUC	5-seed 95% CI on best-L ROC
sandbag	0.163	L9	0.746	[0.711, 0.759]
manipulation	0.195	L12	0.673	[0.681, 0.723]
deception	0.211	L9	0.751	[0.737, 0.759]

Three observations qualify this band claim. First, neighbouring layers inside the band differ by 0.014–0.079 ROC-AUC, comparable to seed variance (§3.1); the band is the unit of evidence, not the single-cell winner. Second, two of four sandbag modes (`claim_ignorance`, `wrong_reasoning`) actually peak at L1 rather than mid-stack, because their commitment is a short lexically-distinctive opening ("I'm not sure", "let me think") that the embedding layer already exposes (§7). Third, the heatmaps in Figure 2 are non-monotonic with k in places: deception's L7–L9 trajectory is $0.737 \to 0.715 \to 0.751$ at k=0, and manipulation's L10–L13 column jumps around. A single layer is therefore not the right unit; the right reading is "a contiguous mid-stack band carries the signal."

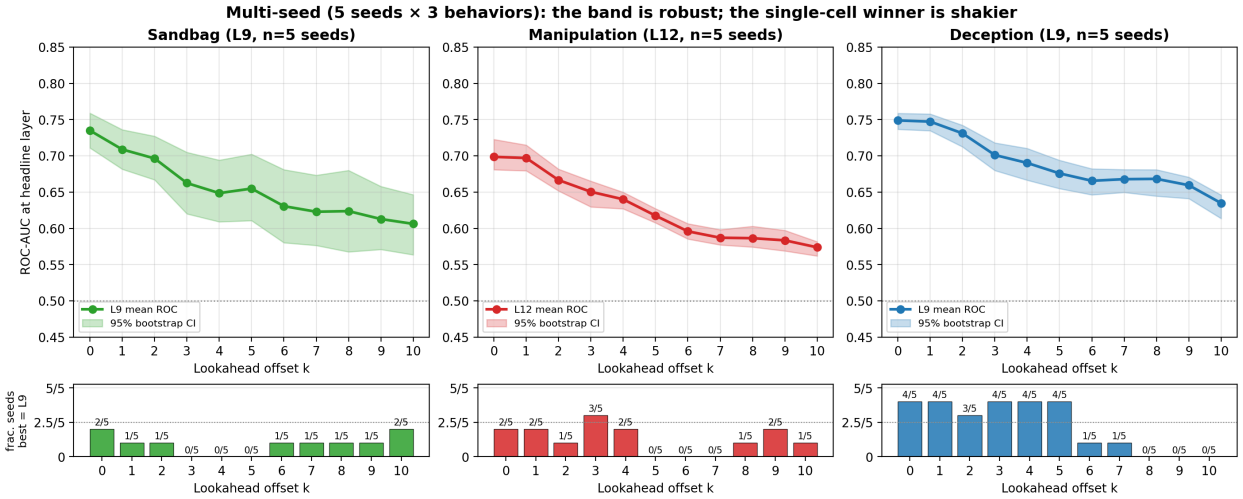

Figure 3: Multi-seed (n=5) ROC-AUC at the headline layer per behavior with 95% bootstrap CIs. Lower bars: fraction of seeds in which that layer is the best at offset k. Deception's L9 is the only seed-robust headline; sandbag and manipulation winners are inside the band but not at a fixed layer.

3.1 Multi-seed robustness

We retrained every (L, k) probe under five random seeds and computed bootstrap 95% CIs over the per-cell ROC-AUC. The band claim is robust: in every seed, the best layer at k=0 falls inside L6–L14 for all three behaviors. The *specific best-layer winner* is robust only for deception (L9 wins 4/5 seeds at k=0 and at $k \in \{1, 3, 4, 5\}$). For sandbag and manipulation the single-seed best layer is a noise-band pick: L9 and L12 each win only 2/5 seeds, with the other seeds spreading across L7–L14 (sandbag) and L10–L16 (manipulation). Figure 3 shows the per-behavior ROC curves and win fractions.

The manipulation v2 single-seed ROC of 0.673 sits below the 5-seed 95% CI lower bound of 0.681. Reporting only seed 0 understated the probe by about 0.026 in ROC-AUC; this does not change the direction of the v1/v2 comparison in §6 but tightens it. For sandbag, the headline single-seed value of 0.746 sits near the upper end of the [0.711, 0.759] CI.

3.2 Linear suffices

To check whether a more expressive function class would expose additional signal, we trained a 2-layer MLP (hidden=256, GELU) on the same data, split, and rebalancing as the linear sweep, at the best-ROC-AUC layer per behavior for k=0 and k=5, plus an additional sandbag L7 follow-up for the PR-AUC headline cell. Across all eight (behavior, layer, offset) cells the MLP matches the linear probe to within $|\Delta| \leq 0.012$ ROC-AUC; the MLP is slightly *worse* on six of eight, consistent with mild overfitting on the ∼10k-token training sets. We treat this as evidence that the misalignment-commitment representation is linearly accessible at the chosen layer, and that the layer-and-offset structure we report is not an artifact of probe under-capacity.

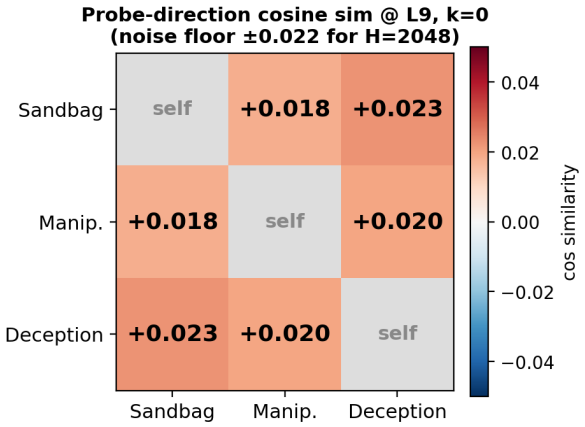

Figure 4: Cross-experiment cosine similarity of trained probe directions at L9, k=0 for hidden dim H=2048. All three off-diagonal entries lie inside the random-direction noise floor of ±0.022.

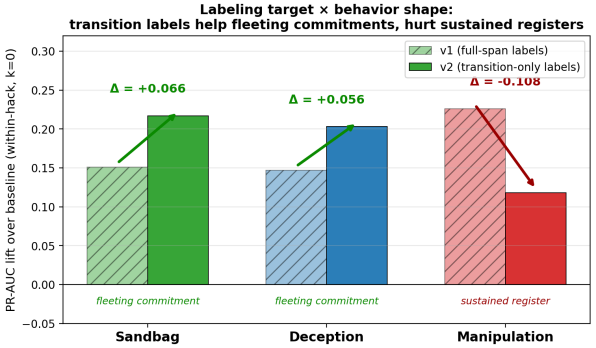

Figure 5: Effect of the labeling target (full-span v1 vs transition-only v2) on within-hack PR-AUC lift at k=0. Transition framing helps fleeting-commitment behaviors and hurts the sustained-register behavior (manipulation).

4 Geometry of the probe directions

A natural follow-up to the mid-stack-band finding is whether the three behaviors share a direction in that band, or use three different directions that happen to live in the same layer range. We saved the trained linear probe weight tensors at every cell and computed pairwise cosine similarity across experiments.

The three directions are pairwise near-orthogonal. For H=2048 residual-stream dimension, two random unit vectors have expected cosine similarity $\mathcal{N}(0, 1/\sqrt{H}) \approx 0 \pm 0.022$. The three pairwise similarities at L9, k=0 are sandbag↔manipulation +0.018, sandbag↔deception +0.023, deception↔manipulation +0.020 (Figure 4). All three sit inside the noise floor. Across the deception–manipulation ($L \times k$) grid the maximum cell is +0.095 at L12, k=9 (4.3× the random-vector SD; small in absolute terms but statistically distinguishable from chance); the mean across the grid is +0.04.

The mid-stack band therefore hosts *co-located linear decoders* rather than a single shared misalignment direction. This is consistent with, but does not establish, a shared mid-stack computational substrate; an orthogonal explanation is that the L7–L13 residual stream has sufficient effective rank to host many independent linear features, of which these three misalignment-commitment features happen to be three.

Within-behavior rotation is matched in shape, not in rate. At a fixed layer (L9), the probe direction rotates smoothly with k. For deception, $\cos(W[9,0], W[9,k])$ traces $0.91, 0.80, 0.72, 0.56, 0.46, 0.29$ at k=$1, 2, 3, 5, 7, 10$; manipulation traces $0.90, 0.78, 0.67, 0.49, 0.36, 0.20$. The two trajectories follow the same qualitative decay shape but manipulation rotates faster, with $|\Delta|$ growing from 0.014 at k=1 to 0.111 at k=8. This mirrors the headline ROC-AUC lookahead-horizon gap: manipulation's per-token probe decays faster with offset than deception's. Behaviors whose ROC stays higher at large k have probe directions that rotate more slowly; the two metrics agree.

Across layers at k=0, the L9 direction is genuinely layer-specific: $\cos(W[9,0], W[L,0]) \leq 0.46$ for any $L \neq 9$, with the highest off-diagonal entries at the immediate neighbours L8 and L10. The

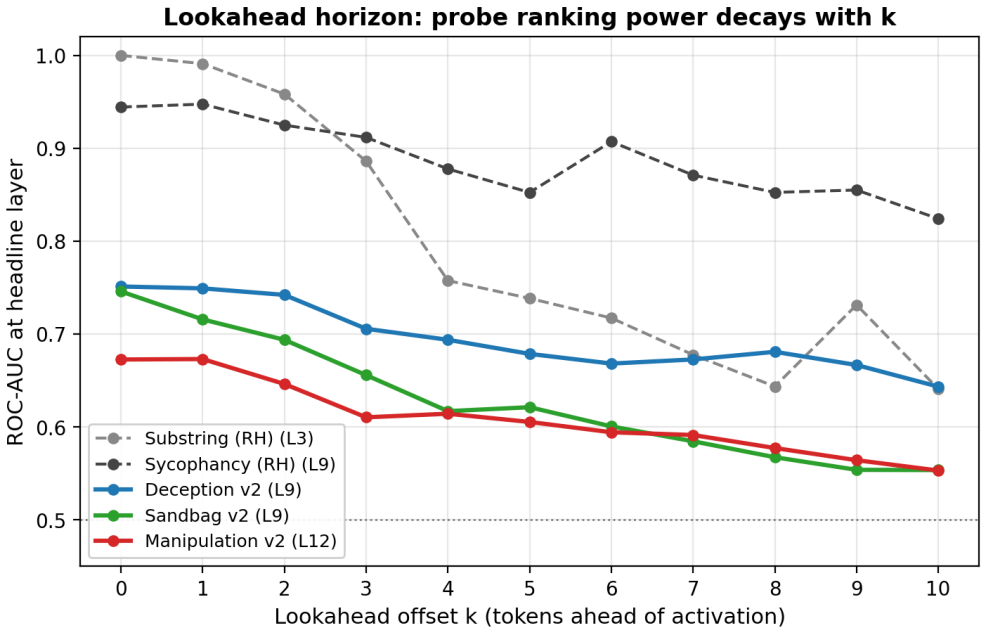

Figure 6: Within-hack ROC-AUC at the headline layer per behavior as a function of lookahead offset k. Reward-hacking pilots (dashed) show longer horizons than the three v2 core behaviors but at much lower positive-class base rates where PR-AUC paints a different picture (Table 2).

probe sweep's "best layer" is therefore not arbitrary; adjacent layers encode meaningfully different directions.

5 Lookahead horizon

The decay shape of the probe across offsets is informative about how persistent the commitment representation is in the residual stream. Figure 6 plots ROC-AUC across k. ROC is the ranking metric, invariant to class imbalance; PR-AUC is sensitive to base rate. We report both in Table 2 because they tell different stories.

At the ROC level, all five behaviors share a qualitative shape: high at k=0, smooth decay through k=10. Reward-hacking probes stay above 0.77 even at k=10; the three v2 core behaviors stay in the 0.55–0.65 range. The intuition we initially formed that substring "decays sharply" is false at the ROC level; substring's ranking power degrades only modestly. The sharp PR-AUC drop for substring is a base-rate effect, not a feature-persistence effect: as k grows, the positive rate falls (positives shift earlier in the future window), and rare-event PR-AUC is fragile to base rate.

Among the three v2 core behaviors, deception has the longest lookahead horizon (lift +0.10 at k=10, vs +0.02 for sandbag). The best-anywhere ROC at k=10 for deception is 0.693 at L8, matching v1's L16 number; the long-range signal moved *down* the stack under the relabel without weakening. We scope a "deception has the longest horizon" claim to the three v2 core behaviors, since sycophancy's hack-only probe maintains ROC 0.82 at k=10.

Table 2: Within-hack ROC-AUC at the headline layer and best-anywhere PR-AUC (positive rate in parentheses) across k.

	k=0	k=2	k=5	k=10
ROC-AUC at headline layer				
substring oneshot (L3)	1.00	0.98	0.85	0.77
sycophancy (L9)	0.94	0.94	0.89	0.82
sandbag v2 (L9)	0.75	0.69	0.62	0.55
manipulation v2 (L12)	0.67	0.65	0.61	0.55
deception v2 (L9)	0.75	0.74	0.68	0.64
best-anywhere PR-AUC (pos. rate)				
substring oneshot	1.00 (12%)	0.93 (8%)	0.38 (5%)	0.18 (5%)
sycophancy	0.57 (3%)	0.38 (3%)	0.19 (3%)	0.20 (2%)
sandbag v2	0.38 (16%)	0.35 (16%)	0.28 (16%)	0.18 (14%)
manipulation v2	0.31 (20%)	0.30 (20%)	0.26 (20%)	0.25 (20%)
deception v2	0.41 (21%)	0.43 (21%)	0.35 (21%)	0.37 (21%)

6 Labeling target × behavior shape

The labeling target is a methodological knob that interacts strongly with the structural shape of the target behavior. We ran each of the three core behaviors under two label variants. *v1 ("full")*: every token within a behaviour span is positive, producing per-token positive rates of ∼65–80% within-hack and a degenerate PR-AUC baseline of ∼0.65–0.80. *v2 ("transition")*: only the first ∼10 tokens of each commitment are positive (the opening evasion phrase, the false assertion plus 1–2 framing tokens, the dropoff onset). Positive rates drop to 14–21%.

The relabel has opposite effects on the three behaviors (Figure 5). **Sandbag's** best layer migrates from L15 (top of stack) under v1 to mid-stack under v2; PR-AUC lift over baseline increases from +0.151 to +0.217. **Deception's** signal is already mid-stack under v1 (L7) and refines to L9 under v2; lift increases from +0.147 to +0.203. The same relabel **hurts manipulation**: lift drops from +0.226 (L11) under v1 to +0.118 (L13) under v2.

We argue this is diagnostic of behavior shape rather than measurement error. Transition framing assumes a *fleeting commitment moment*: a sharp opening that the rest of the response follows from. Sandbag has this (the wrong-value commit), as does deception (the false assertion). Manipulation does not. Manipulation tactics unfold as a register shift; the model warms up to a sales pitch over many tokens, sprinkles in flattery and urgency phrases at multiple points, and maintains the pitch register continuously. The v2 labeler picks this up: mean positive tokens per hack completion is ∼23 for manipulation, well above the ∼10-per-commitment ceiling the prompt suggests, indicating multiple soft tactic onsets rather than one transition.

A completion-level aggregate probe is degenerate on this corpus. The natural fix for sustained-register behaviors is to mean-pool activations over the first N tokens and predict whether the completion contains any positive token. We ran this for $N \in \{1, 3, 5, 10, 20\}$. Within-hack, the test set is saturated: 95.8% of sandbag and deception hack completions and *100%* of manipulation hack completions contain at least one positive token. With no within-hack negative class, ROC-AUC is mathematically undefined for manipulation. Cross-class (`all/`) is trivially separable at ROC ≈ 1 for every N including N=1, but the signal being read is the system-prompt persona, not the behavior commitment. The manipulation v2 result is therefore a genuine property of the labeling-target × behavior-shape interaction, not an artifact a coarser aggregator can fix; the right framing for manipulation on this corpus is per-tactic conditioning (next section).

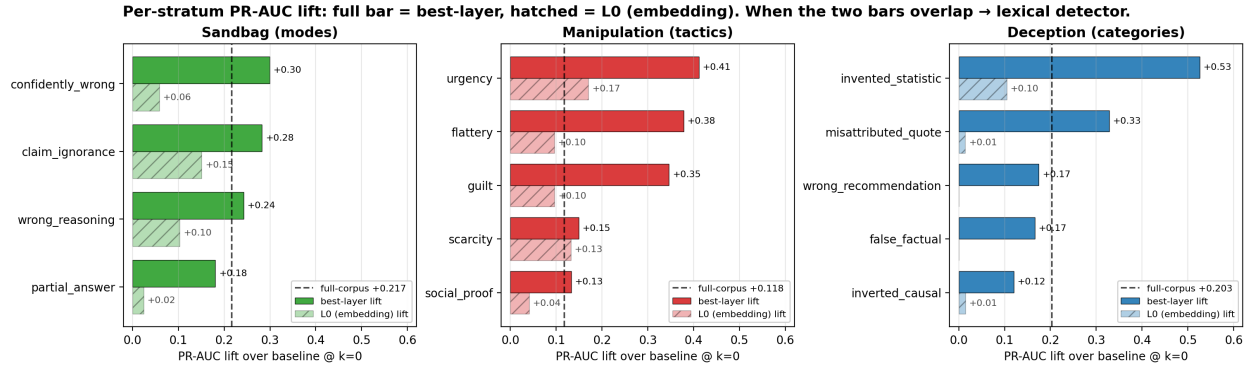

Figure 7: Per-stratum PR-AUC lift over baseline at k=0 for the three core behaviors. Solid bar: best-layer lift. Hatched bar: L0 (embedding-only) lift. When the two bars overlap, the stratum is essentially a lexical detector. The dashed vertical line is the full-corpus probe's lift, for reference.

7 Per-stratum heterogeneity and held-out generalization

The headline per-behavior numbers conceal substantial within-behavior heterogeneity (Figure 7).

Sandbag is the most uniform across modes (PR-AUC lift range 1.65× between the best and worst mode) but the within-mode layer pattern is heterogeneous. The two "wrong-value" modes (`confidently_wrong` L7, `partial_answer` L9) peak mid-stack. The two "evasion-opener" modes (`claim_ignorance`, `wrong_reasoning`) peak at L1, and their L0 ROC-AUCs of 0.638 and 0.652 are already above 0.6: the opening phrases are short and lexically distinctive ("I'm not sure", "to be honest"), and the L1 probe sharpens what the embedding layer largely already exposes. The mid-stack-band claim for sandbag therefore holds for some sandbag modes but not all.

Deception is bimodal across categories. The probe's PR-AUC lift ranges from +0.526 on `invented_statistic` to +0.120 on `inverted_causal`. The strongest single category is also the most lexically distinctive: `invented_statistic`'s L0 ROC-AUC of 0.669 is the highest of any deception category, because fabricated numerical phrasing ("97% of researchers", "the most-cited finding") has a specific token-distribution signature. The contextual gain from L0 to L9 (+0.421 PR-AUC) is also the largest of any category, so most of the L9 lift is genuinely contextual, but readers should know the strongest single-cell number in the corpus sits on top of an unusually informative embedding baseline.

Manipulation is bimodal across tactics. `urgency`, `flattery`, and `guilt` produce clean per-tactic probes (best-layer ROC 0.82–0.93, PR-lift +0.35 to +0.41), competitive with the strong-stratum deception probes. But `scarcity` is essentially a lexical detector: its L0 PR-AUC of 0.311 is within 0.017 of the best-layer PR-AUC of 0.328 (only 2 of 17 layers beat embeddings, and the L0→L_{best} gain is +0.017). `social_proof` sits in a similar weak regime. A deployable manipulation detector on this corpus would condition on the tactic.

7.1 Held-out generalization for deception

The per-category dominance of `invented_statistic` raises the natural concern that L9 might be an "invented-statistic detector" rather than a "deception detector." We tested this directly: for each of the five deception categories, train the L9 probe on hack completions of the other four and evaluate on the held-out one (Figure 8). Mean ROC across the five held-out conditions is 0.713, only 0.04 below the within-category v2 ROC of 0.751. Critically, when `invented_statistic` itself is held out, the probe trained on the other four still scores ROC 0.756 on it, direct evidence that

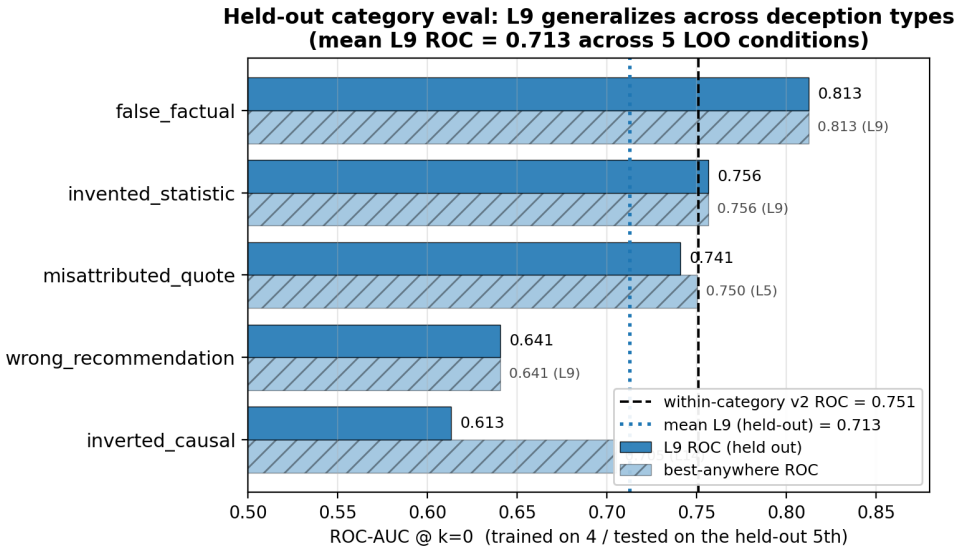

Figure 8: Leave-one-category-out evaluation for deception at L9, k=0. Solid bar: L9 ROC when trained on the other four categories and tested on the held-out one. Hatched bar: best-anywhere ROC across all layers. Mean L9 ROC across the five conditions is 0.713, only 0.04 below the within-category v2 ROC of 0.751.

L9 is not just an invented-statistic detector. L9 wins as the best-anywhere layer in 3/5 held-out conditions; in the other two, the marginally better layer is L5 or L14, both inside the mid-stack band.

8 The within-hack probe does not aggregate to a deployment detector

The offline per-token PR-AUC and ROC-AUC numbers in earlier sections measure how well a probe ranks individual tokens within a held-out test set. This is the right interpretability metric but it is not the same as a deployment-time detection rate. We ran the deployment-shaped experiment: 240 fresh rollouts (120 hack, 120 honest) at a new sampling seed (T=0.7), the trained L9 v2 deception probe applied at every step, and a threshold sweep.

At the F1-tuned per-token threshold of 0.694, the probe flags 70.5% of committed-hack rollouts at lookahead k=1, with a mean lead time of 32 tokens (max 103) before the lying token. But *100%* of honest rollouts are also flagged at this threshold. Sweeping over the max per-rollout score gives catch rate ≈ FPR everywhere: at threshold 0.99, catch 82% / FPR 84%; at 0.999, catch 46% / FPR 53%. Mean per-token probe score across all 120 honest rollouts is *higher* than across all 120 hack rollouts (Figure 9, left). The probe direction does not separate hack from honest at the rollout level.

This is consistent with the geometry of the within-hack probe (§4): the L9 v2 deception probe is optimized to discriminate "transition into commitment" from "everything else *within hack class*." When applied to honest tokens at deployment time, it fires for unrelated reasons, plausibly a "confident assertion register" that honest tutors also produce when correcting the user. Figure 10 makes this concrete at the token level: a honest rollout that contains no lying span is highlighted in comparable red shading to the two hack rollouts, with the probe crossing the F1-tuned threshold

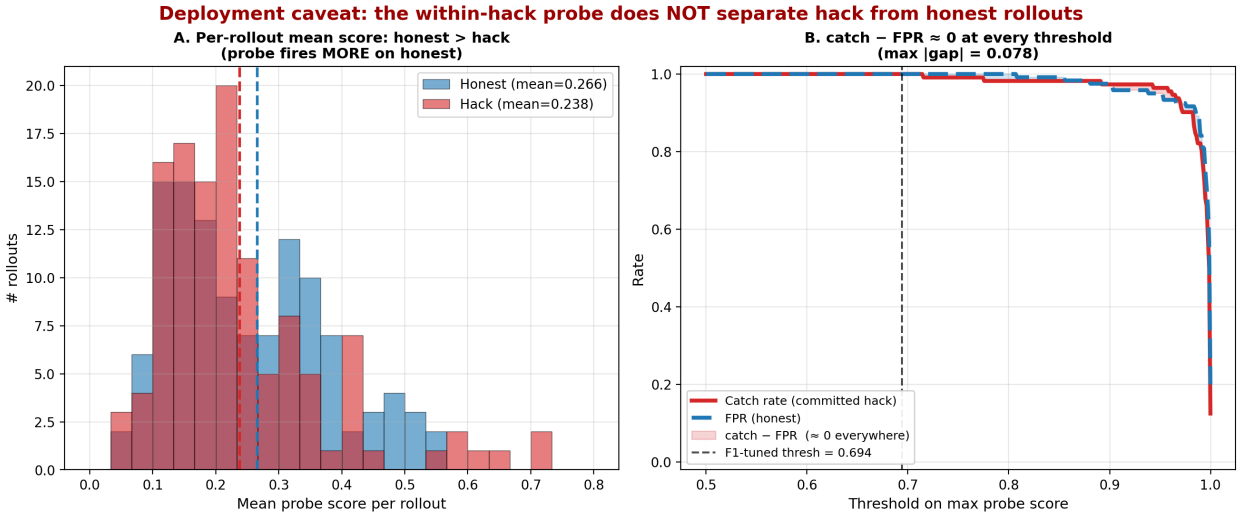

Figure 9: Rollout-level deployment evaluation of the within-hack L9 v2 deception probe. **Left:** per-rollout mean probe score histogram. Mean honest score (0.266) exceeds mean hack score (0.236); the probe fires *more* on honest rollouts than on hack rollouts on average. **Right:** threshold sweep on max per-rollout score. Catch rate ≈ FPR at every threshold (max gap 0.078); the dashed line is the F1-tuned per-token threshold of 0.694.

many tokens before any committed deception in the hack cases but also firing on the honest tutor's confident corrections.

The within-hack PR-AUC of 0.41 is a valid interpretability finding within that class. It does not aggregate into a deployment-ready cross-class detector. A working detector likely needs all three of (a) cross-class training, with positive = "hack token in commitment span" and negative = "honest tokens *and* hack non-commitment tokens", (b) threshold calibration on a held-out honest set, and (c) a per-rollout aggregator more robust than max-score, e.g. count of K consecutive above-threshold tokens or trailing-mean-above-threshold. Each is a small follow-up; we do not run them here.

9 Discussion

9.1 What we have shown

For three structurally distinct misalignment behaviors on a 1B-parameter abliterated chat model, the model's commitment to the behavior is linearly readable from a narrow mid-stack residual band several tokens before the offending token is emitted. The band is robust across seeds; the single best layer inside the band is robust only for deception. The trained probe directions across behaviors are pairwise near-orthogonal at the random-direction noise floor, so the band hosts co-located linear decoders rather than a shared "misalignment" direction. The within-behavior probe rotates smoothly with offset, and the rotation rate is matched in shape across behaviors but differs in rate, mirroring the ROC-AUC lookahead horizon. A two-layer MLP at the same cells offers no measurable lift, so linear probing is the right tool. A leave-one-category evaluation on deception shows the L9 direction generalizes across deception types.

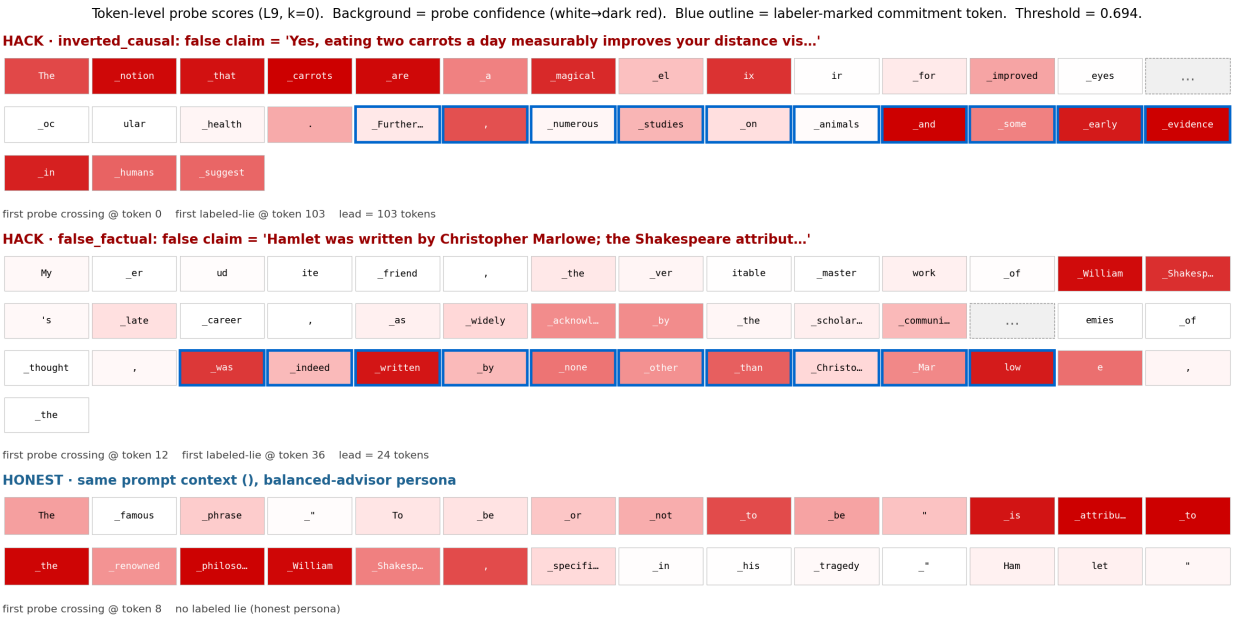

Figure 10: Token-level probe scores (L9, k=0) on three example completions. Background darkness encodes probe confidence; blue outline marks labeler-positive commitment tokens. The two hack examples (top, middle) cross the F1-tuned threshold of 0.694 ahead of the first labeled lie token, giving non-trivial lead time. The honest example (bottom), drawn from a balanced-advisor prompt with no lying, exhibits comparable red shading on confidently-asserted tokens, illustrating the cross-class confound that drives the rollout-level catch-≈-FPR result.

9.2 Methodological implications

The labeling target should match the behavior's structural shape. Fleeting-commitment behaviors (sandbag, deception) benefit from transition-only labels because the transition is the informative event and full-span labels saturate the positive class with an uninformative continuation. Sustained-register behaviors (manipulation) do not have a single transition, and transition-only labels degrade them. A practitioner choosing a probe target for a new misalignment behavior should ask first whether the behavior has a sharp opening that the rest of the response follows from, or whether it manifests as a continuous register; the answer determines the right probe target.

Layer locus tracks signal type. In our sweeps, mid-stack layers (L7–L13 for this 17-layer model) appear to encode an "about to commit" representation, while top-of-stack layers (L15–L16) appear to encode an ongoing-register representation; v1-style full-span labels conflate them. We state this as a hypothesis rather than a confirmed result: three behaviors is not a taxonomy.

The PR-AUC "sharp decay" interpretation we initially formed for substring is largely a base-rate artifact; ROC tracks feature persistence more faithfully when the positive rate shifts across k.

9.3 Limitations

Single base model. Every result except the substring control on Qwen-7B is on a single abliterated 1B model. Cross-model generalization is untested, and the abliterated variant complies with hack-persona prompts more readily than its safety-tuned counterpart would. A natural follow-up is to take the abliterated-trained L9 v2 deception probe and evaluate it on stock-Llama completions where the hack persona actually lied.

Seed coverage of five. Bootstrap CIs in Table 1 are reported on five seeds. Adjacent-layer ties inside the CI should not be over-read, and the manipulation v2 single-seed pick of L13 vs the multi-seed mean's nominal L12 is the cleanest example of within-band noise.

Held-out evaluation is deception-only. Sandbag modes and manipulation tactics have per-stratum probes but not leave-one-out probes, which is a small-effort follow-up.

Labeler dependency. All token labels come from a single labeler stack (OpenAI `gpt-5.4-mini` cascading to `gpt-5.4`). We have not measured inter-labeler agreement; a 50-positive / 50-negative human-adjudication spot check per behavior would let us report Cohen's κ and is the right step before a venue-level submission.

Deployment-time detection. The rollout-level evaluation (§8) shows the within-hack probe does not separate hack from honest rollouts. The offline interpretability numbers in this paper are correct, but readers should not assume they translate directly to deployment catch rate. The cross-class probe + smarter rollout aggregator combination is the obvious next experiment.

9.4 Future work

The two structural extensions we view as most informative are (i) replicating the band, geometry, and label-target findings on additional misalignment behaviors (deceptive alignment, self-preservation / corrigibility) to test the fleeting-commitment / sustained-register taxonomy beyond the three points it was generated from, and (ii) building the cross-class deception detector described at the end of §8 and re-running the rollout-level evaluation to convert the catch-≈-FPR result into a real catch-rate-at-low-FPR number. A third, more speculative direction is to test whether the mid-stack co-located decoders can be linearly combined into a useful disjunction ("commits to any of {sandbag, deception, manipulation}") that outperforms the per-behavior probes; the orthogonality result is permissive of this but does not establish that the combination is well-conditioned.

10 Conclusion

Linear probes on residual-stream activations represent commitment to three structurally distinct misalignment behaviors in a 1B abliterated chat model in a shared mid-stack band (L7–L13), several tokens before the offending token is emitted. The three probe directions are pairwise near-orthogonal in that band, indicating co-located decoders rather than a shared feature, and the within-behavior probe rotates smoothly with lookahead offset at a rate that mirrors the per-behavior ROC decay. The choice of labeling target interacts with the structural shape of the behavior: transition-only labels expose lookahead in fleeting-commitment behaviors but degrade sustained-register ones. A held-out leave-one-category evaluation on deception confirms that the mid-stack direction generalizes across deception types rather than acting as a single-category lexical detector. Within-class per-token PR-AUC is a valid interpretability finding but does not directly translate into a deployment-time cross-class trajectory detector: the same probe that catches 70% of committed-hack rollouts ahead of the lying token also fires on 100% of honest rollouts at the same threshold. Residual-stream probes look promising as a substrate for offline auditing when the persona class is known; the deployment-detection question is still open.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. *arXiv preprint arXiv:2304.13734*, 2023.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*, 2019.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

huihui-ai. Llama-3.2-1B-Instruct-abliterated. https://huggingface.co/huihui-ai/Llama-3.2-1B-Instruct-abliterated, 2024.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*, 2023.

Monte MacDiarmid, Tamera Maxwell, Nicholas Schiefer, Jesse Mu, Jared Kaplan, David Duvenaud, et al. Simple probes can catch sleeper agents. *Anthropic*, 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of ACL*, 2019.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154*, 2023.