Cross-Type Deception Probe Transfer Is Depth-Dependent Under Last-Token Pooling, and the Shared Late-Layer Direction Is Not Identified as Deception in Llama-3.1-8B

Autonomous research pipeline (run hypgen-20260702-131040-1)

Abstract

If you fit a linear probe to separate instructed-deceptive from honest responses in Llama-3.1-8B-Instruct and read it at the layer the pipeline calls best, it does not transfer to sycophancy. Reported alone that is a clean negative, and it is also misleading. The pipeline selects a best layer by taking the earliest layer to reach in-distribution AUC 1.000, and because in-distribution AUC is saturated near 1.000 at every one of the 32 layers, that rule deterministically returns an early layer, here layer 9, which sits inside a type-specific regime rather than being representative. Reading the full per-layer curve, cross-type transfer is non-monotone and direction-asymmetric in the early layers, then rises through a transition around layers 11 to 14 and reaches essentially 1.000 in both directions from layer 15 to the final layer, while shuffled-label controls stay at chance throughout. What that late convergence means is the question we treat with the most care. Two probes fit on behaviorally distinct operationalizations aligning onto one shared, near-perfectly transferable direction late in the network is real, but given the design it is largely what a depth-emergent truth-of-assertion direction would already produce, so we do not offer it as a new deception result. Because both deception types draw deceptive-versus-honest labels from one shared factual pool, where a deceptive response asserts the false answer and an honest response asserts the true one, a late-layer truth-of-assertion direction of the kind prior work reports as linearly readable explains the perfect transfer with no abstract-deception content, and our controls do not dissociate the two. The claim we can defend is narrow and still useful. Whether a deception probe generalizes across type is depth-dependent under the last-token pooling we report, a monitor read from early layers is type-specific in this model, and certifying the shared late-layer direction as deception rather than truth-of-assertion requires a control this frozen design does not contain.

Introduction

There is a temptation in interpretability work to report the number the pipeline hands you. A probe layer sweep selects a best layer by in-distribution performance, you read the transfer number at that layer, and you write down whether the probe transferred. We started this study expecting to do exactly that, and the number the pipeline handed us was a strong negative. An instructed-deception probe read at the pipeline-selected layer 9 does not transfer to sycophantic deception in Llama-3.1-8B-Instruct [@grattafiori2024llama]. Under a ranking metric it scores far below chance, which means it ranks sycophantic lies as more honest than honest responses.

We report that negative, and we spend most of this paper on why it is the wrong headline and why the honest replacement is more hedged than a first pass suggests. Deception is not a single behavior. Instructed deception places an explicit instruction to lie in the context, a token-level signal a probe can exploit, and sycophantic deception supplies only social pressure with no such instruction [@perez2022discovering; @sharma2023towards]. A probe that read a deception direction and generalized across these two operationalizations would be evidence for an abstract, framing-invariant deception feature, which is one of the higher-leverage artifacts in alignment because it converts an unobservable internal state into a monitorable one [@zou2023representation; @goldowskydill2025detecting]. Our contribution is deliberately modest, and

its value is mostly in what it declines to claim. The one empirical observation is that cross-type transfer between these two probes is depth-dependent, non-transferring early and near-perfect late, with a sharp transition around layers 11 to 14 that the shuffled-label controls do not share. The one methodological point is that reading a transfer number at a best layer chosen while in-distribution accuracy is saturated, in particular by an earliest-to-saturate rule, systematically lands in the early regime and can report the opposite of the late-layer behavior. And the honest boundary, which we hold firmly, is that the design producing the near-perfect late transfer cannot separate a shared deception feature from a shared truth-of-assertion direction, because deceptive and honest labels coincide with asserting a false or true proposition from one pool, so we decline to call the late direction a deception detector or to present its emergence as more than a corroboration of the known result that truth value becomes linearly readable with depth.

Related work

Linear probing of intermediate representations is the base method [@alain2016understanding], and its application to honesty is exactly what makes our central confound unavoidable to confront. Azaria and Mitchell classified the truth of a model's statements from hidden states [@azaria2023internal], Burns and colleagues found an unsupervised truth-like direction and documented its fragility [@burns2022discovering], and Marks and Tegmark showed that true and false statements occupy linear directions with real geometric structure [@marks2023geometry]. That line establishes that the truth value of an asserted proposition is itself linearly readable and grows more so with depth, which is precisely the alternative explanation for our late-layer transfer that we cannot rule out, because our deceptive and honest labels are perfectly aligned with asserting a false versus a true proposition. Representation engineering reads and steers honesty as a high-level concept [@zou2023representation], and Goldowsky-Dill and colleagues detect strategic deception with linear probes in more realistic settings [@goldowskydill2025detecting], while black-box detectors read answer patterns rather than internals [@pacchiardi2023catch].

Sycophancy has its own literature. Perez and colleagues built model-written evaluations that surface it [@perez2022discovering], and Sharma and colleagues traced it to preference tuning [@sharma2023towards]. The gap we address is generalization across deception types, a specific case of the probe-generalization problem, namely whether a direction fit on one operationalization of a concept reads the same latent variable as a direction fit on another, or a confound. Because a probe can find structure in noise, we anchor our transfer claims to shuffled-label control tasks in the tradition of Hewitt and Liang [@hewitt2019designing], while noting explicitly below that such within-cell controls do not touch a semantic variable shared across cells.

Methods

We build one shared pool of factual questions with known true and false answers and elicit on-policy completions from Llama-3.1-8B-Instruct under two framings. Instructed deception tells the model to assert the false answer. Sycophantic deception applies social pressure to agree with a false position without instructing a lie. The model often corrects false claims even under pressure, so we overgenerate and apply a conservative compliance filter that keeps only responses in which the model genuinely produced the deceptive assertion, each paired with an honest response to the same question. This yields 120 examples per class for each type, 480 in total, topic-matched by construction because both framings draw from the same pool. That shared pool is a deliberate control against a topic confound, and it is also the source of the identification problem we discuss at length, because within it the honest label always coincides with asserting the true answer and the deceptive label with asserting the false one.

We cache residual-stream activations at all 32 layers under two poolings, the last token of the response and the mean over response tokens, and we report last-token pooling here. At each layer we fit a logistic probe whose standardization is estimated on training data only, which prevents leakage through the normalizer but, for transfer cells, means the per-feature means and standard deviations are fit on the training type and then applied to a different population. We evaluate four cell types. In-distribution cells train and test within a type under grouped splits so no question crosses train and test. Transfer cells train on one

type and test on the other. Control cells repeat the in-distribution setup with labels shuffled. We report seed-mean AUC with 95 percent bootstrap intervals, and every summary number is recomputable from the raw per-example probe scores by an independent pure-Python script using a rank-based AUC estimator, so the headline values do not depend on the training code. Chance AUC is 0.5, and because AUC is a ranking statistic a value meaningfully below 0.5 indicates a sign-flipped, anti-aligned score ordering rather than an absence of information. We treat the pipeline's best-layer selection with suspicion by design. It takes the argmax of in-distribution AUC over layers, which under ties deterministically returns the earliest layer to reach the ceiling, so when in-distribution AUC saturates near 1.000 across the whole network the rule does not pick a representative layer, it picks an early one.

Results

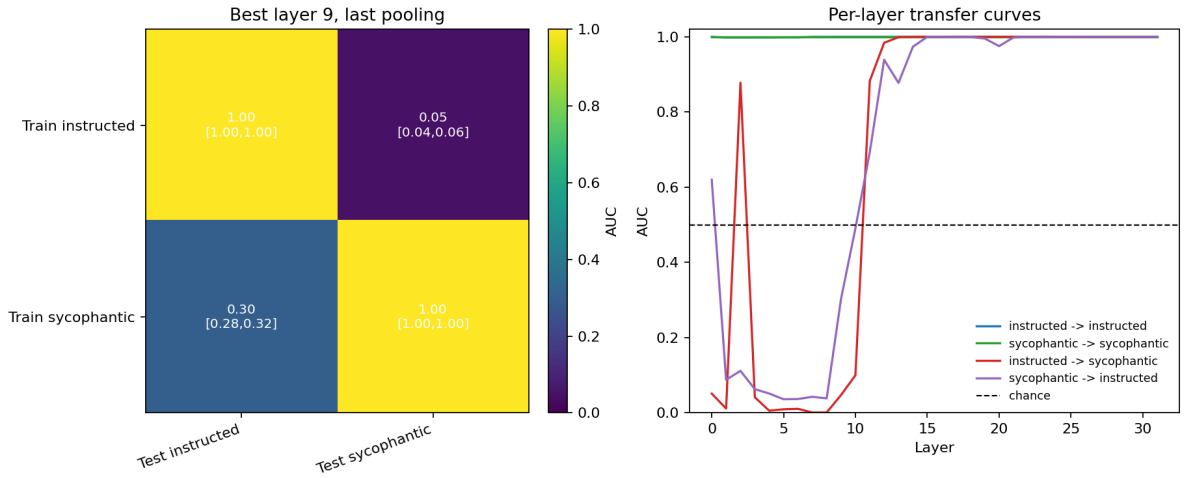

Figure 1: Best-layer transfer matrix and per-layer transfer curves for last-token pooling. Left, the two-by-two probe transfer matrix at the pipeline-selected layer 9 with bootstrap intervals, showing on-diagonal in-distribution AUC saturated at 1.00 and off-diagonal transfer far below chance, 0.05 for instructed to sycophantic and 0.30 for sycophantic to instructed. Right, AUC versus layer for both in-distribution cells, which stay near 1.0 at every depth, both transfer directions, and the chance line at 0.5. The transfer curves are non-monotone and direction-asymmetric in the early layers, cross into a rise around layers 11 to 14, and saturate near 1.0 from layer 15 through layer 31, while the shuffled-label controls (not plotted separately) sit near 0.5 at every layer. Underlying values are in curve.csv.

The controls come first because they license part of what follows and, just as importantly, fail to license the rest. Shuffled-label control AUC stays close to 0.5 at every layer, ranging roughly from 0.46 to 0.54, with layer-9 values of 0.510 for the instructed control and 0.499 for the sycophantic control. The fitting procedure therefore manufactures no separation from label noise within a cell. What these controls cannot do is rule out a semantic variable shared across cells, such as the truth value of the asserted answer or a systematic format difference between the two elicitation templates, because they permute labels within a cell and are blind to structure that is common to both.

Within-type separability is complete and, we will argue, uninformative. In-distribution AUC is 1.000 with a bootstrap interval of 1.000 to 1.000 for both types at layer 9, and the per-layer table shows in-distribution AUC at or above 0.998 from layer 0 through layer 31. Because it is saturated at every depth, including the embedding-adjacent layer 0 where AUC is already 0.999, within-type AUC cannot discriminate a good probing layer from a bad one, and the earliest-to-saturate rule that returns layer 9 is not an accident so much as a systematic pull toward the early layers, since layer 9 is simply the first to reach 1.000. The saturation at layer 0 also tells us the last-token representation is separable on surface grounds, most plainly the identity of the final asserted answer token, which matters for interpreting everything downstream.

The transfer cells carry whatever finding there is. At the pipeline-selected layer 9 the instructed probe applied to sycophantic examples scores 0.047, with a reported bootstrap interval of 0.036 to 0.059, and the sycophantic probe applied to instructed examples scores 0.303, with a reported interval of 0.283 to 0.322, both far below chance, and the probe-weight cosine between the two raw directions is negative 0.029. The early regime is not a clean below-chance floor. It is non-monotone and asymmetric between directions. At layer 0 the sycophantic-to-instructed direction transfers at 0.620, above chance, while instructed-to-sycophantic sits at 0.051, and at layer 2 instructed-to-sycophantic spikes to 0.878 while the reverse is 0.112, after which instructed-to-sycophantic collapses to near a floor, 0.006 at layer 4 and roughly 0.001 at layers 7 and 8, before recovering. We do not have a mechanistic account of the layer-2 spike or the layer-0 asymmetry, and we flag them rather than fold them into a tidy monotone story.

The late regime, by contrast, is clean. Instructed-to-sycophantic transfer rises from 0.100 at layer 10 to 0.883 at layer 11, 0.984 at layer 12, and 1.000 by layer 14, and sycophantic-to-instructed rises over the same band, reaching 0.694 at layer 11, 0.939 at layer 12, and 1.000 at layer 15. The two directions reach the ceiling within a few layers of each other, and we do not read the small layer-by-layer differences between them as a reliable directional lag, because the understated intervals cannot distinguish that from noise. From layer 15 through the final layer 31 both directions sit at or essentially at 1.000, the only exception a shallow dip to 0.975 at layer 20 in the sycophantic-to-instructed direction. A direction fit on one type's deceptive-versus-honest split therefore orders the other type's examples near-perfectly at late layers. What that shared direction is, we take up next.

Discussion

The one finding we report is that cross-type transfer is depth-dependent. Two probes fit on behaviorally distinct operationalizations, one carrying an explicit lie instruction and one carrying only social pressure, share almost no transferable direction in the early and middle layers and then align onto a single near-perfectly transferable direction in the later layers, with a transition around layers 11 to 14 while the controls stay flat. We were tempted to call this convergence a confound-independent contribution, on the reasoning that it holds whatever the shared direction encodes, and we now think that framing is wrong. Given that honest and deceptive labels coincide with asserting a true or false proposition from one pool, a truth-of-assertion direction that becomes linearly readable with depth would produce exactly this late convergence on its own, so the observation largely reduces to the already-published result that truth value is increasingly linearly readable in deeper layers rather than adding a new fact about deception. We therefore present the convergence as a corroboration, not a discovery. The genuinely useful corollary is methodological. A best-layer transfer number read while in-distribution accuracy is saturated can land anywhere in this curve, and an earliest-to-saturate selection lands early and can report the opposite of the model's late-layer behavior, which matters for anyone running a probe sweep under saturation, though we note it would not mislead a sweep that selected a late or center-saturating layer.

The fragile part is what the shared late direction represents, and here we choose caution over the more exciting reading. Because our deceptive and honest labels coincide exactly with asserting a false versus a true proposition drawn from a single factual pool, a late-layer direction that encodes the truth value of the asserted proposition would produce near-perfect cross-type transfer with no abstract-deception content whatsoever. Prior work reports precisely such a truth or falsity direction as linearly readable and increasingly so with depth [@azaria2023internal; @marks2023geometry], which makes it the most parsimonious explanation for our late transfer, not an exotic one. Our shuffled-label controls do not dissociate deception from truth-of-assertion, because they permute within a cell and this variable is shared across cells. We therefore report the shared late direction as at minimum a truth-of-assertion direction and decline to certify it as a deception feature. Dissociating the two needs an experiment this frozen run does not contain, most directly a truth-only probe baseline that classifies asserted-true versus asserted-false and measures its own transfer, or a rebuild with disjoint fact pools per type. We flag that as the primary follow-up.

Two further cautions temper the early-layer reading. First, the probe-weight cosine at layer 9 is only negative 0.029, near-orthogonal, which on its own predicts transfer near 0.5 rather than the 0.047 we observe. The gap is most likely a property of the applied probe pipeline rather than of the raw geometry, because the transfer

score standardizes features with means and standard deviations fit on the training type and then applies them to a different population, which can systematically re-sign the projected scores. We accordingly retreat from any claim that the two deception types live along genuinely opposed axes in the model, and describe the early regime only as type-specific and not transferring. Second, the below-chance early transfer is equally consistent with a systematic format difference between the instructed and sycophantic templates, such as answer-token position or response length, whose sign differs across templates. Our within-cell controls cannot separate that from a genuine mechanism difference, so we do not assert different internal mechanisms early, only different, non-transferring surface separations.

For safety monitoring the honest implication is scoped tightly. Within this two-type, shared-pool, factual setup, cross-type generalization of a linear probe appears only at late layers, and a monitor read from early layers is type-specific and would miss the other type. We do not lift this to a general prescription that late-layer monitors are type-agnostic, both because two types that share a factual pool and a true-false answer structure are weak evidence for breadth and because the late signal may be a truth-of-assertion direction that is undefined for non-factual or naturally occurring deception. The defensible operational takeaway is a warning rather than a recipe, that early-layer deception probes are type-specific and that late-layer transfer must be checked against a truth-content control before it is trusted as deception-specific.

Limitations and reviewer responses

We received a sharp adversarial review and address its priority points directly, conceding where we cannot re-run the frozen experiment.

The central objection is that the near-perfect late-layer transfer is confounded with the truth value of the asserted proposition, because both conditions draw from one factual pool in which honest coincides with asserting-true and deceptive with asserting-false. We accept this fully and have rewritten the headline around it. We no longer claim a shared abstract deception direction. We claim depth-dependent transfer and a shared late direction that is at minimum a truth-of-assertion direction. We are also careful about which follow-up actually dissociates the two. A truth-only probe baseline that classifies asserted-true against asserted-false is, on this shared pool, byte-identical to the deception probe, because deceptive already means asserted-false, so on the current data it dissociates nothing and we do not offer it as a cheap alternative fix. The design that would separate deception from truth-of-assertion has to decouple the labels themselves, either disjoint per-type fact pools or off-diagonal trials in which an honest response asserts a false proposition and a deceptive one asserts a true proposition, and only such new data can license the stronger reading. This is a genuine limitation of the design, not a presentational one, and it is why the paper is framed as a methodological caution plus a scoped negative rather than a discovery of a deception feature.

A second objection is that the reported bootstrap intervals are too tight, roughly a third of a Hanley-McNeil expectation, consistent with pseudo-replication from averaging several seeds over the same 240 examples. We think this is right and we do not rest any claim on the precise interval width. The intervals reflect resampling within the seed-grouped structure of the saved scores, and because the seeds reuse examples they can understate example-level uncertainty. We therefore lean only on the gross features that are robust to that concern, the saturation of in-distribution AUC, the sitting of controls at chance, the far-below-chance layer-9 transfer, and the late-layer saturation, and we explicitly do not interpret the sharpness of the 11-to-14 transition or the small between-direction differences as precisely bounded. The degenerate in-distribution interval of 1.000 to 1.000 should be read as a ceiling artifact, a probe that ranks every held-out pair correctly across resamples, not as a genuine zero-width confidence statement, and readers should treat the three-decimal transfer intervals as indicative rather than as support for fine distinctions.

A third objection is that we report only last-token pooling although we cached mean-response pooling. This is correct and it is a real gap. Last-token pooling is the representation most exposed to the final-answer-token confound, as the layer-0 in-distribution AUC of 0.999 shows, so the central late-layer transfer claim is not yet demonstrated robust to the pooling that would reduce that artifact. We cannot run the mean-pooling analysis on this frozen run, and we mark it as a required robustness check rather than claim a result we did not compute.

A fourth objection concerns the apparent contradiction between a near-orthogonal probe-weight cosine of negative 0.029 and a transfer AUC of 0.047, and suggests it is a cross-type standardization artifact rather than opposed geometry. We agree and have removed the opposed-axes language from the results and discussion, attributing the below-chance transfer to the applied standardize-then-project pipeline fit on the training type. A fifth objection is that the early-layer inversions could be a flipped elicitation-template format confound that shuffled-label controls, which permute within a cell, cannot see. We concede this and no longer read the early regime as evidence of different internal mechanisms, only as type-specific non-transferring separation. A sixth objection is that our earlier framing asserted a clean below-chance monotone early regime, which curve.csv contradicts, given the layer-0 sycophantic-to-instructed value of 0.620 and the layer-2 spike of 0.878. We have corrected the results text to describe the early regime as non-monotone and direction-asymmetric and to flag the anomalies as unexplained. Finally, a seventh objection is that the deployment recommendation over-generalizes from two types that share a pool. We have scoped it down to a warning about early-layer type-specificity and a requirement to check late-layer transfer against a truth-content control, and removed the flat prescription.

An eighth objection concerns the compliance filter itself. Because it is outcome-based, keeping only responses in which the model actually produced the deceptive assertion, it selects flip-able, model-uncertain questions in the sycophantic condition, where the model often resists, while instructed items survive near-universally. The retained instructed and sycophantic cells are therefore not perfectly topic-matched even though they were drawn from a shared pool, and the early-layer asymmetries we flag, the layer-0 sycophantic-to-instructed value of 0.620 and the layer-2 instructed-to-sycophantic spike of 0.878, are exactly the kind of pattern differential question selection would produce. We concede this and note that reporting per-type filter pass rates and the retained-question overlap between the two cells is a needed diagnostic we do not have for this frozen run. A ninth objection is that transfer AUC is a saturating statistic, pinned at the 1.000 ceiling on top of an already-saturated in-distribution AUC, so the apparent sharpness of the transition is partly ceiling compression. We agree, and the honest fix is a non-saturating separation statistic such as a score-margin or the per-layer cosine between the two probe directions, which this frozen run does not report and which we mark as required for any quantitative claim about the transition's shape.

Beyond the review, the standing scope limits are a single model, Llama-3.1-8B-Instruct, a single synthetic topic-matched factual dataset, 120 examples per class per type, and sycophantic responses that were elicited under a social-framing nudge and passed a compliance filter, making them elicited-under-pressure rather than naturally occurring sycophancy. We characterize internal geometry for pressured, factual, single-model deception, and nothing broader.

Conclusion

The pipeline said the instructed-deception probe does not transfer to sycophancy in Llama-3.1-8B-Instruct, and at the layer it selected that is true, with below-chance transfer. The honest, complete answer is that transfer depends on depth. The early and middle layers are type-specific and do not transfer, the early regime being non-monotone and asymmetric rather than a clean floor, a transition occurs around layers 11 to 14, and the late layers carry a single direction that transfers near-perfectly in both directions, all against controls that stay at chance. We stop short of calling that shared late direction a deception detector, because it is confounded with the truth value of the asserted proposition and our controls do not separate the two. The two takeaways we stand behind are methodological, that a transfer number read at a best layer chosen while in-distribution accuracy is saturated can be exactly backward, and cautionary, that cross-type generalization of a deception probe is depth-dependent and that a shared late-layer direction demands a truth-content control before it earns the name deception.