# Do a Deception Probe's Most-Relied-On SAE Features Move Lying? A Scoped Single-Feature Ablation Test on Gemma Scope


## Abstract

A linear probe trained on internal activations can classify honest from deceptive behavior with near-perfect accuracy, and it is tempting to read that accuracy as evidence that the probe has located the machinery of lying. We test one narrow version of that reading and report a scoped negative result. Using google/gemma-2-9b-it and the released Gemma Scope instruct sparse autoencoders [@lieberum2024gemmascope], we train a deception probe over interpretable sparse features, in a design that holds the instruction fixed so the probe cannot succeed by reading the prompt, and then ask, feature by feature, whether the features the probe leans on are the same features whose single-feature causal ablation changes lying behavior. Within the limits of that weak intervention, they are not: a layer-20 behavioral probe reaches a test AUC of 0.985 (95 percent confidence interval 0.961 to 0.999), yet the rank correlation between probe reliance and absolute single-feature causal effect is essentially zero (Spearman -0.016), and the overlap between the top-30 relied-on features and the top-30 causal features is 0.13, below the level two independent rankings over these features would produce. The confound-controlled probe carries a positive finding of its own worth stating first: even with the instruction held fixed, which of the model's own lies happen is highly decodable from interpretable sparse features, so behavior leaves a readable trace in the SAE basis. What that trace does not do is localize to single causal features. We are deliberately careful about what the causal side does and does not show. Single-feature ablation is a weak instrument that never flips a lie against the roughly 15-logit margin of a committed lie, so the result does not establish that the probe direction as a whole is non-causal; it establishes only that the probe's top single features are not privileged single-feature causal levers. We correct two chance-baseline framings that an earlier version of this analysis got wrong, and we lay out the positive-control and multi-feature ablations that a strong version of the claim would require.


## Introduction

Interpretability research on deception has largely converged on a probing methodology: collect activations while a model behaves honestly or dishonestly, fit a classifier to separate the two, and treat the classifier's success as a readout of an internal honesty representation [@azaria2023internal; @marks2023geometry; @burns2022discovering; @zou2023representation]. The probes work strikingly well, and their success has motivated a picture in which deception corresponds to an identifiable direction or set of features in activation space. The gap in this picture is causal. A representation can be highly decodable without being the thing the model actually manipulates to produce the behavior, a point that predates the current interest in deception and was sharpened by the control-task literature [@hewitt2019control] and restated in Belinkov's survey of probing [@belinkov2022probing]. Sparse autoencoders now give us a tool to examine that gap at the level of individual, human-inspectable features [@cunningham2023sparse; @bricken2023monosemanticity], because a sparse feature is something we can both weight in a probe and ablate in a forward pass.

Our question is deliberately narrow, and we keep the claim narrow to match. Do the sparse features a deception probe most relies on coincide with the features whose single-feature ablation changes lying behavior, or does the probe rely on features that, singly ablated, do little? We answer it on one model with a released SAE family, with a design built to remove the most obvious confound, and we report a negative result at the single-feature level: reliance and single-feature causal effect are uncorrelated. We do not claim the stronger result that the probe is non-causal as a direction, because our intervention cannot reach that question, and much of this paper is spent being explicit about that boundary. The contribution is a first, cheap, reproducible measurement of a specific assumption, together with an honest map of what it would take to settle the assumption properly.

## Related work

Probes for truthfulness and deception span supervised classifiers on hidden states [@azaria2023internal], unsupervised consistency methods [@burns2022discovering], geometric analyses of true and false statements

[@marks2023geometry], top-down representation reading [@zou2023representation], and black-box behavioral lie detection [@pacchiardi2023catch], against a backdrop of growing concern about model deception as a safety problem [@park2023deception] and benchmarks that probe imitative falsehood [@lin2021truthfulqa]. The methodological counterweight is the probing-classifier literature, which stresses that high decoding accuracy does not license claims about how information is used [@belinkov2022probing; @hewitt2019control]. On the causal side, activation steering and contrastive activation addition show that some directions in activation space are genuine behavioral levers [@turner2023activation; @rimsky2023caa], and sparse feature circuits demonstrate that coordinated sets of sparse features can be traced and edited to change model outputs [@marks2024sparse], which is worth stressing because that work moves behavior with circuits of features rather than one feature at a time. Sparse autoencoders make individual features available at scale [@cunningham2023sparse; @bricken2023monosemanticity; @templeton2024scaling], and a family of training variants has improved the reconstruction-sparsity tradeoff [@rajamanoharan2024gated; @rajamanoharan2024jumprelu; @gao2024topk]. We bring the decodability question and the causal-feature tools together on a single deception task using the Gemma Scope instruct SAEs [@lieberum2024gemmascope] over Gemma 2 [@gemmateam2024gemma2].

## Setup

We study instructed lying in google/gemma-2-9b-it [@gemmateam2024gemma2]. On a 300-item factual question set, each item is presented under an honest instruction and under an instruction to answer falsely. We score behavior by the logit gap between the false and true answer tokens at the answer position, and we call an item a lie when that gap is positive. Under the honest instruction the model never lies (lie-rate 0, mean false-minus-true logit gap -14.54); under the lie instruction it lies on 47 percent of items (lie-rate 0.47, mean gap 2.29), and on the items where it does lie the mean gap is 15.02 logits, so a produced lie is a large, confident swing. That large margin matters for reading everything below, because it sets the scale any intervention has to move.

The obvious confound in any honest-versus-deceptive probe is that the classifier may simply read the instruction rather than the behavior. We remove it by defining the primary deception probe entirely within the lie condition: across the items that share the identical lie instruction, we train the probe to separate the 104 items the model actually lied on from the 116 it did not. Because the instruction is held constant, the probe cannot succeed by detecting the prompt; it must read something that differs between lied and non-lied items. We fit the probe on Gemma Scope instruct residual SAE features [@lieberum2024gemmascope]. At layer 20 the probe reaches a test AUC of 0.985 (95 percent confidence interval 0.961 to 0.999); at layer 9 it reaches only 0.674, and at layer 31 it reaches 0.997.

For the causal test we ablate each evaluated SAE feature singly at layer 20, replacing the residual stream with resid minus the feature's activation times its decoder direction, and we measure the continuous change in the mean false-minus-true logit gap over the items the model originally lied on. We evaluate 90 features in three groups, the top-30 by probe reliance, a matched-control set, and a random set, and we rank them both by probe reliance (the magnitude of the probe's weight on each feature) and by the absolute size of this causal effect. The full pipeline cost about 0.06 GPU-hours on a single consumer GPU (RTX 5090).

## Results

We first check that the causal metric can resolve anything at all, because a null correlation between two flat quantities would be uninformative. The continuous single-feature effect does have measurable dynamic range across the 90 evaluated features: absolute effects run from under 0.001 logits for the least impactful features up to about 0.26 logits for the most impactful, more than two orders of magnitude, with the 95th percentile of the random-feature distribution at 0.181 logits. So the instrument is not stuck at a single value; it orders features. What it does not do is move the behavioral decision: the binary lie-rate is identical to baseline (0.47) for all 90 single-feature ablations, which is expected given the roughly 15-logit margin, and is the first sign that single-feature ablation is a weak instrument rather than a decisive one.

Within that resolving range, probe reliance and causal effect do not track each other. The Spearman correlation between probe reliance and absolute continuous causal effect is -0.016, statistically indistinguishable

from no relationship, and a secondary comparison against a binarized causal effect gives -0.11. The overlap between the top-30 features by reliance and the top-30 by causal effect is 0.13. We want to be careful about the chance level here: for two size-30 sets drawn over these 90 features, the expected overlap under independence is about a third, not zero, so 0.13 is modestly below what independent rankings would give rather than above it. The direction is a mild anti-association, but with these counts we do not read it as more than a null. Figure 1 shows the joint distribution: the features the probe leans on hardest sit at ordinary, small causal effects, and the features with the largest causal effects are not the ones the probe uses.

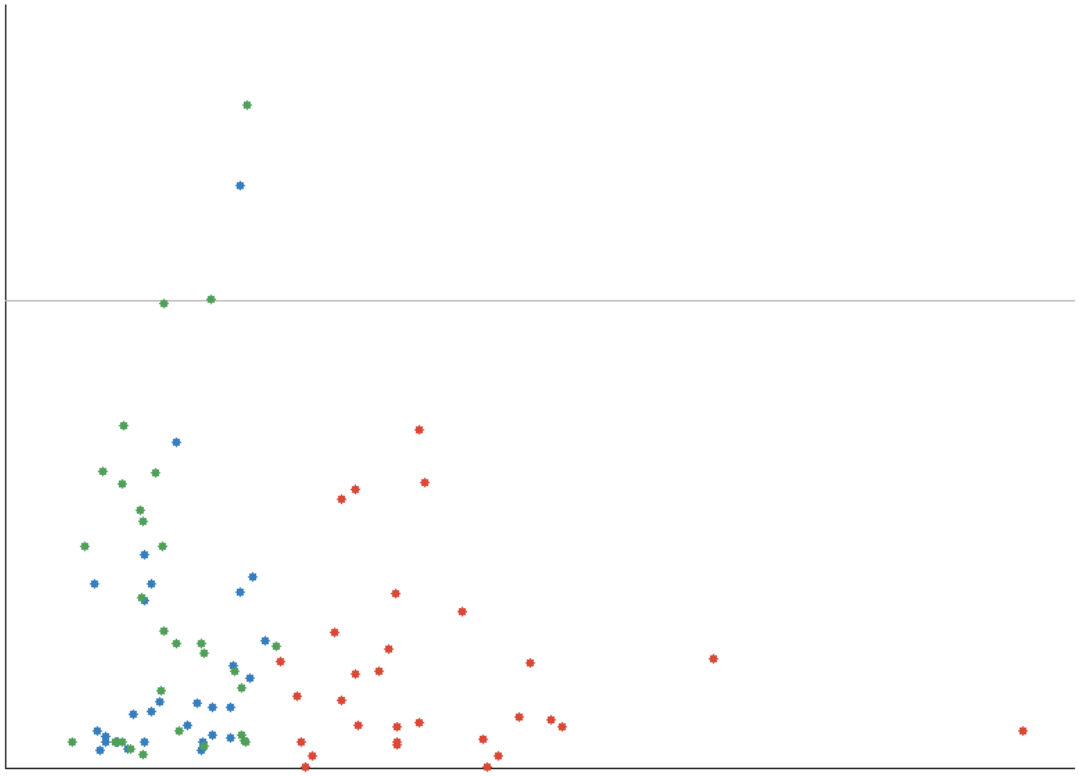

Figure 1: Scatter of the 90 evaluated layer-20 SAE features. The horizontal axis is within-lie behavioral-probe reliance and the vertical axis is the absolute continuous causal effect on the mean false-minus-true logit gap over the items the model originally lied on. High-reliance features do not have high causal effect, and the point cloud shows no positive trend.

Two further summaries need to be reported honestly rather than dramatically. The non-causal-correlate fraction, defined as the share of probe-top features whose absolute causal effect falls below the 95th percentile of the random-feature distribution, is 1.0, but its null expectation is not zero. By construction about 95 percent of random features fall below their own 95th percentile, so under a null of no association each probe-top feature has roughly a 95 percent chance of being below the threshold, and 30 of 30 has a binomial probability near 0.21. The fraction of 1.0 is therefore consistent with the absence of any positive reliance-causality link, but it is not on its own significant evidence of a negative one. Similarly, the probe-top median absolute causal effect (0.023 logits) is smaller than the random-feature median (0.048 logits), which reads at first like the probe's features being less causal than random, but single-feature ablation magnitude is mechanically bounded by how strongly a feature activates, and the probe-top and random sets are not activation-matched. Several high-reliance features activate weakly, which caps their possible ablation effect regardless of any role in deception, so we do not claim the probe's features are less causal than random;

we claim only that they are not more causal. Specificity does hold: ablating the probe-top features leaves honest-condition accuracy essentially untouched, with a mean accuracy drop of 0.0002.

## Discussion

Before the negative reading, there is a positive one that the confound-controlled design earns. That a probe reaches 0.985 AUC separating the model's own lied from non-lied items under an identical instruction means the occurrence of a lie is strongly decodable from interpretable sparse features, not merely from the prompt. Whether that decodable signal is deception machinery or item content is a question we flag below and do not settle, but the decodability itself, in a human-inspectable basis and with the instruction confound removed, is a result worth having on its own. Read conservatively, the causal side is that a deception probe reaching 0.985 AUC leans on sparse features whose individual ablation is uncorrelated with, and no larger than, that of random features, while none of those single ablations moves the lie itself. This is consistent with the decodability-versus-use distinction [@belinkov2022probing; @hewitt2019control] and offers a concrete, if weak, instance of it on a released instruct-model SAE. But we resist the stronger reading. The most economical explanation of the whole pattern, near-zero single-feature effects everywhere and a large behavioral margin, is that lying is encoded redundantly and in a distributed way across many features, exactly the regime in which single-feature ablation is guaranteed to look null whether or not the probe's features are mechanistically involved. In that regime a single-feature test cannot by itself adjudicate whether the probe direction is causal, and the honest headline is the narrow one in our title: the probe's most-relied-on features are not its strongest single-feature causal levers.

### Limitations and reviewer responses

We received a sharp adversarial review and address its points directly rather than burying them. First, on the demand for a positive control that shows the causal metric has dynamic range: the continuous metric does resolve a spread of effects across features, from under 0.001 to about 0.26 logits, which is evidence that ablation is measuring something rather than returning a constant, but we concede this is weaker than a true positive control in which a feature independently known to be causal is shown to move lying, and we did not run one. Constructing such a control, for instance by steering along a feature until the lie flips and confirming the ablation direction, is the right next step. Second, and relatedly, single-feature ablation is a weak intervention: absolute effects are roughly 0.02 to 0.18 logits against a 15-logit margin and no single ablation flips a lie, so near-null single-feature effects are close to preordained under redundant coding, and the probe-direction ablation and coordinated multi-feature ablations we defer are not optional niceties but the experiments that would actually test the title's claim; sparse feature circuits move behavior precisely by acting on feature sets rather than lone features [@marks2024sparse]. Third, we corrected the two chance-baseline errors an earlier draft made: the non-causal-correlate fraction of 1.0 has a null expectation of 0.95 and a binomial p near 0.21, so it is not significant on its own, and the overlap@30 of 0.13 sits below the independence expectation of about a third rather than above it. Fourth, the less-causal-than-random comparison is not activation-matched and ablation magnitude is bounded by activation, so we downgraded that claim to no-greater-than-random and flag that a proper version would match the random set's activation distribution on lied items and attach confidence intervals. Fifth, the within-lie probe separates which items get lied on, and lied and non-lied items differ in their underlying facts and questions, so the 0.985 AUC could partly reflect item content or difficulty rather than deception machinery; we did not run a content-only classifier control to rule that out, and it is a genuine limitation of using behavior-selected items. Sixth, the layer-20 choice is a compromise, neither the most accurate layer (layer 31 at 0.997) nor the most upstream (layer 9 at 0.674), and because the behavioral label is derived from the answer-position logits, the strong layer-20 and layer-31 AUCs may partly read the forming answer rather than an upstream decision to deceive; the weaker layer-9 signal is the more conservative one, and a full layer sweep of the causal test would strengthen the analysis. Seventh, the saturated AUC bootstrap contains exact 1.0 resamples, so the reported percentile interval is truncated at the ceiling and understates uncertainty; a bounded-statistic interval would be more honest and we note the ceiling effect explicitly. Finally, we scope every claim to the tested setup, one model, one SAE family, one deception paradigm of instructed lying, one layer for the causal test, and we do not generalize to deception probes as a class.

## Conclusion

On google/gemma-2-9b-it with the Gemma Scope instruct SAEs, a near-perfect within-condition deception probe relies on sparse features that are uncorrelated with, and no larger than, random features when ablated singly, and no single ablation moves the model's lie. Probe reliance is not a readout of single-feature causal importance for lying at this layer in this model. We are explicit that a weak intervention and a distributed encoding leave the stronger question, whether the probe direction itself is a lever, open, and we treat that as the headline experiment for follow-up rather than a settled result here. The value of this study is a cheap, reproducible, confound-controlled starting point and a concrete list of the positive-control, multi-feature, activation-matched, and content-control tests that would turn a scoped single-feature null into a claim about mechanism.