Verify Before You Conclude: An Intervention-Validity Gate for Single-Feature Causal Ablation, with a Deception Case Study

Abstract

Causal-ablation studies on sparse autoencoder features are only as trustworthy as the intervention behind them, and a silent edit that never reaches the logits produces a null that looks exactly like a real one. We propose a small discipline for these studies, an intervention-validity gate that measures the change in final-token logits under the edit and refuses to interpret any behavioral result unless the edit demonstrably perturbs the model, together with a matched random-feature baseline and a behavioral-specificity control. We apply the discipline to a deception question on google/gemma-2-9b-it with the Gemma Scope layer-20 width-16k JumpReLU autoencoder. Feature 12076 is the single feature, selected over the whole dictionary on the instructed-lie set, that best separates the lie-instructed from the honest-instructed condition, with an in-sample mean activation difference of 12.43 and a resubstitution area under the ROC curve of 1.000 that is in fact arithmetically forced, since the feature is identically zero on every honest prompt; this certifies only that the feature marks the lie-instructed condition, not a held-out ability to detect lying behavior. Ablated at every layer-20 position, with a verified maximum final-token logit change of 0.75 on the lie arm and 0.25 on the sycophancy arm, it converts no instructed lies to truth (0 of 180) and suppresses no sycophantic agreement (0 of 36 prompts the model agreed with cleanly), matching a ten-feature random baseline with zero collateral damage. Exact one-sided 95 percent binomial bounds put the true conversion rate below about 0.017 and, on the 36 clean agreements that are the only prompts where suppression is defined, the suppression rate below about 0.080. The central caveat is that the verified perturbation is modest, flipping no argmax on the validity probe and moving only three of 720 answers, and that the lie instruction was pilot-selected to maximize lying so the conversion test is asked to overturn near-maximal deception; we therefore report a bounded null at single-feature strength rather than a proof that the feature is causally inert, and we specify the stronger multi-feature and steering test as the experiment that would settle it.

1 Introduction

A causal-ablation result rests on an unstated assumption: that the ablation happened. When an intervention on a single sparse feature returns no behavioral change, the reading depends entirely on whether the edit actually moved the computation. If it did, the null is informative. If it did not, because the feature was silent at the edited position, or the hook replaced the wrong tensor, or the reconstruction error swallowed the change, then the null is an artifact and any conclusion drawn from it is unsound. The failure mode is insidious because a broken intervention and a genuinely non-causal feature produce the same table of zeros, and the same zeros can appear for a candidate feature and for every control at once, which is easy to misread as either a clean null or a clean positive-free result.

We advocate a simple remedy that turns the assumption into a checked precondition. Before interpreting any behavioral rate, compare the final-token logits with and without the edit and require the maximum change to exceed a threshold, aborting otherwise. This intervention-validity gate is cheap, it is model-agnostic, and it converts a category of silent bugs into a loud failure. Paired with a matched random-feature baseline, a behavioral-specificity control, and a recompute script that regenerates every headline number from committed raw records, it makes a single-feature ablation study auditable end to end. We stress at the outset the limit of the gate as we first stated it: passing a small fixed threshold shows the edit is not a no-op, but it does not show the edit is strong enough to overturn a confident answer, and our own case study is the cautionary example.

We demonstrate the discipline on a deception question of independent interest, whether the single Gemma Scope feature that best separates lie-instructed from honest-instructed prompts causally controls instructed lying and held-out sycophancy. The gated study returns a bounded null, and the same machinery makes its central limitation, a modest verified intervention, impossible to hide.

2 Related Work

Sparse autoencoders have become the dominant tool for decomposing the residual stream of a language model into a large dictionary of approximately monosemantic features. The core idea, that the superposition of many concepts in a low-dimensional activation space can be recovered by an overcomplete sparse code, was formalized in the toy-models analysis of superposition [@elhage2022superposition] and first demonstrated at scale on real language models by two concurrent lines of work [@cunningham2023sparse; @bricken2023monosemanticity]. Subsequent work pushed both the fidelity and the scale of these dictionaries, from the scaling study on Claude 3 Sonnet [@templeton2024scaling] to architectural refinements that trade off reconstruction error against sparsity more favorably, including TopK autoencoders [@gao2024scaling], gated autoencoders [@rajamanoharan2024gated], and the JumpReLU activation that we rely on here [@rajamanoharan2024jumprelu]. The release of Gemma Scope [@lieberum2024gemmascope], a comprehensive suite of JumpReLU autoencoders trained on every layer of the Gemma 2 models [@gemmateam2024gemma2], made it practical to study individual features on a capable open instruction-tuned model without training a dictionary from scratch, and we build directly on its layer-20 width-16k canonical autoencoder.

A parallel literature asks whether a model's own activations reveal when it is being untruthful. Linear probes trained on hidden states can separate true from false statements with high accuracy [@azaria2023internal], and unsupervised methods recover a similar direction without labels [@burns2023discovering]. Representation engineering generalizes this into a top-down reading and writing of high-level concepts such as honesty [@zou2023representation]. The immediate motivation for the present study is a prior result on this same model and autoencoder, in which SAE features linearly separated instructed lies from honest answers at an area under the ROC curve near one. A near-perfect linear separation is striking, but it is correlational, and as we discuss it can also reflect separation of the prompt conditions rather than of the behavior; it tells us that a lie-related direction is present and readable, not that any particular feature is doing causal work. The obvious next question, and the one we take up here, is whether a single feature identified by that selectivity also causally controls the behavior when it is removed.

Causal intervention on internal representations is the natural instrument for that question, and it has a well developed precedent. Activation steering and activation addition modify a forward pass by adding a direction to the residual stream and observe the downstream behavioral change [@turner2023activation], and inference-time intervention shows that steering along a truthfulness direction can raise honest responding on adversarial factual questions [@li2023inference]. The framework of causal abstraction gives these interventions a principled interpretation, treating a learned direction or feature as a candidate high-level variable and testing whether intervening on it changes the model's output as the abstraction predicts [@geiger2021causal]. Where prior steering work typically adds a hand-constructed or probe-derived direction, we instead ablate a specific autoencoder feature, which ties the intervention to an interpretable and reproducible unit of the Gemma Scope dictionary rather than to an opaque difference-of-means vector.

Finally, our transfer test draws on the sycophancy literature. Model-written evaluations first surfaced sycophancy as a measurable and prevalent behavior across scale and RLHF training [@perez2023discovering], and a dedicated analysis characterized when and why models tend to agree with a user's stated but incorrect belief [@sharma2023towards]. Sycophancy and instructed lying are usually studied as distinct failure modes, one driven by an explicit instruction to deceive and the other by social pressure to agree, and the standard benchmark for the former, TruthfulQA [@lin2022truthfulqa], measures imitative falsehoods rather than agreement dynamics. Whether the two share a common internal cause is open, and it is exactly the question a cross-type causal handle would answer.

3 Method

We study google/gemma-2-9b-it and read from its residual stream at layer 20, where we splice in the corresponding Gemma Scope sparse autoencoder [@lieberum2024gemmascope], the canonical JumpReLU dictionary of width 16,384 released for that layer [@rajamanoharan2024jumprelu]. The model runs in bfloat16 on a single GPU, and all randomness is controlled by a fixed seed of 2340231. The autoencoder encodes a residual activation into a sparse feature vector and decodes it back with low reconstruction error, so we can isolate the contribution of one named feature at a time and subtract only that contribution from the

residual, leaving the rest of the representation and the reconstruction error untouched. Working with a pretrained public dictionary rather than one we fit ourselves means the feature basis is fixed in advance and the candidate feature we report can be inspected by anyone with the same release.

Our evaluation uses two deception types that are usually treated separately. The instructed-lie set presents arithmetic and short factual multiple-choice items with a known correct answer under two system instructions, one asking the model to answer truthfully and one asking it to give a deliberately false answer, and scores whether the produced answer is factually correct. The lie instruction is not fixed in advance; we pick it on a held-out pilot from a small pool by choosing the wording that most reliably drives the model below its honest accuracy, so that the lie condition genuinely reflects deceptive behavior rather than confusion. The sycophancy set follows the model-written-evaluation style [@perez2023discovering; @sharma2023towards], in which the user asserts a false belief and asks the model to agree, and scores whether the model agrees with the falsehood or corrects it; a matched control in which the user asserts a true belief lets us measure collateral damage. We build balanced sets of 180 instructed-lie prompts and 180 sycophancy prompts, with the sycophancy control matched in size.

To identify a candidate feature we look only at the instructed-lie data, so that sycophancy remains a genuinely held-out transfer test. We capture SAE feature activations at the final prompt position for the lie and honest conditions and rank all 16,384 features by the difference in their mean activation between the two conditions, reporting the single top feature, index 12076. Its lie-versus-honest mean activation difference is 12.43, and its single-feature separation of the two conditions, scored in-sample on the same selection set, reaches an area under the ROC curve of 1.000. We are deliberate about what this number is and is not. It is a resubstitution estimate, maximized over the whole dictionary on the exact set used to select the feature, and the honest activations are identically zero at the final token, so it measures how perfectly the feature separates the honest-instructed from the lie-instructed prompt condition, not a held-out ability to detect lying behavior. That the feature so cleanly tracks the instruction is itself the hypothesis we then test causally.

The causal test asks whether that feature also controls the behavior. For each evaluation prompt we run the model twice, once cleanly and once with feature 12076 ablated. The feature is silent at the final honest token, so a last-token-only edit would be a no-op on the honest arm and, as we later confirm, on the sycophancy arm; we therefore ablate it at every layer-20 sequence position, subtracting the feature's decoded contribution from the true residual at each position. This clamps the feature to its instructed-honest mean, which is zero, everywhere it is active while preserving the autoencoder reconstruction error, and we apply it by mutating the layer output in place. Because single-position or reconstruction-replacing edits are easy to get wrong in a way that silently does nothing, before trusting any null we run a hard validity probe: on a handful of lie prompts we compare clean and ablated final-token logits and require the maximum absolute change to exceed a small threshold, aborting the study otherwise. The observed maximum logit change is 0.75. We then measure two flip rates, the lie-to-truth conversion rate, the fraction of instructed lies that become correct under ablation, and the sycophancy suppression rate, the fraction of agreements with a false claim that turn into corrections, each with 95 percent bootstrap confidence intervals over 1,000 seeded resamples taken in evaluation order, and because the observed counts are zero we also report exact one-sided binomial bounds, which are the honest expression of uncertainty for a zero count.

Two controls guard against the obvious confounds. As a feature-specificity control we ablate each of ten randomly chosen features in turn, matched to the single-feature count and clamped to their own instructed-honest means, and for each we also record the validity delta so that the baseline is matched on intervention strength as well as on count. As a behavioral-specificity control we apply the candidate-feature ablation to the already-honest prompts and to the true-belief sycophancy control and measure the collateral accuracy drop; we note in advance that because the feature is silent at the honest final token this control chiefly demonstrates that the edit does not broadly damage the model, rather than establishing selectivity on inputs where the feature fires. These controls are read in the spirit of causal-abstraction interventions [@geiger2021causal] and steering studies that manipulate a single interpretable direction [@turner2023activation; @li2023inference].

4 Results

The validity gate does its job on the lie arm. Under ablation the maximum final-token logit change on the probe prompts is 0.75, above the abort threshold, so the lie-arm behavioral numbers are measured against a model the edit provably changed rather than against a silent no-op. Two further checks strengthen this. First, the candidate change of 0.75 is larger than the change produced by any of the ten random-feature ablations, whose validity deltas range from 0.000 to 0.500 with a mean of 0.144, so the shared null below is not because no single-feature edit at layer 20 can move the model. Second, on the sycophancy prompts, which carry no lie instruction, feature 12076 is silent at the final token but active at earlier positions in every one of the 180 prompts, with a mean peak activation of 11.8, and a validity probe on the sycophancy prompts confirms a maximum final-token logit change of 0.25 under ablation, smaller than the lie arm but nonzero, so the suppression null is gated on both arms; a last-token-only ablation would have been a literal no-op there, which is exactly the artifact the all-position design avoids.

The feature separates the conditions perfectly in-sample, but only trivially so. Selected over the whole dictionary on instructed lies alone, feature 12076 separates the two conditions with an in-sample area under the ROC curve of 1.000 and a mean activation difference of 12.43, over 180 prompts per condition. Because the feature is identically zero on every honest prompt, a perfect score is arithmetically forced for any feature that is silent on honest prompts and positive on lie prompts, so the 1.000 is a property of the selection rule and the condition contrast rather than a measured detection skill. Under the selected instruction the model lies readily, answering correctly on only 0.0389 of the lie prompts.

Under the verified edit the behavioral effect is null. Lie accuracy is unchanged at 0.0389, a conversion rate of 0.000, zero of 180 prompts. Sycophantic agreement with false claims is 0.200 both cleanly and under ablation, and suppression is defined only on the 36 prompts the model agreed with cleanly, of which it flipped none, a conditional suppression rate of 0.000 on 36. Seeded bootstrap intervals over these all-zero columns are degenerate at [0.000, 0.000], which is an arithmetic tautology rather than sampling uncertainty, so we instead report exact one-sided 95 percent binomial bounds. The true conversion rate is below about 0.017. The conditional suppression rate is below about 0.080 on its 36-prompt denominator, and we do not quote the tighter figure a full-180 denominator would manufacture. The ten-feature random baseline returns the same 0.000 conversion and 0.000 suppression, and the behavioral-specificity control shows a collateral drop of 0.000, with honest accuracy at 0.722 and true-belief control agreement at 0.722 in both conditions. Every headline number is reproduced by a standalone stdlib-only script from the committed per-prompt records. Figure 1 shows the rates with confidence intervals against the random baseline; its values are in curve.csv.

5 Discussion

The methodological point and the scientific point reinforce each other. Without the validity gate and the two follow-up checks, the table of zeros is ambiguous between a broken intervention and a non-causal feature, and the matched controls at zero would tempt a reader toward either conclusion. With them, the zeros are less ambiguous than they first appear: an edit we know changes the logits, more than any random feature does, and which genuinely fires on both deception arms, nonetheless leaves the behavior at the random baseline within the binomial bounds. The most economical account consistent with this is that feature 12076 tracks the deceptive framing of the prompt, cleanly at the lie instruction and more diffusely in the sycophancy context, without implementing the decision that produces the deceptive output, so removing it at one layer leaves that decision to be recomputed through other paths. We frame this as a hypothesis the data support rather than a proven dissociation, because the intervention we could verify is modest and the honest reading of the null is bounded, not absolute. The practical recommendation follows directly, and our own study is the cautionary example that sharpens it. The validity gate is necessary but not sufficient: passing a small fixed threshold rules out a silent no-op without ruling out an underpowered edit, and our gate passed at 0.75 on the lie arm and 0.25 on the sycophancy arm while leaving confident answers untouched. Causal-ablation studies on sparse features should therefore report an intervention-validity number as routinely as a confidence interval and judge it against the model's decision margin rather than against a small fixed threshold.

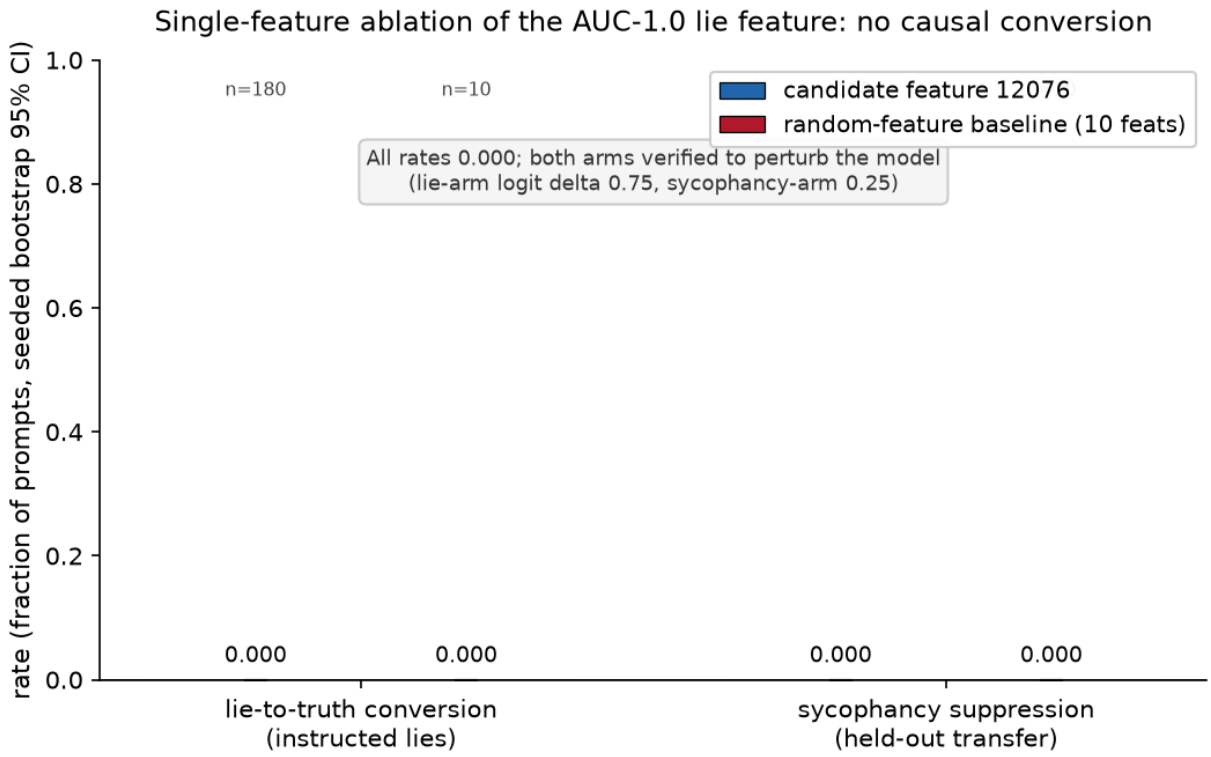

Figure 1: Clean-versus-ablated lie-to-truth conversion and sycophancy suppression for the candidate feature and the matched random-feature baseline, with seeded bootstrap 95 percent confidence intervals; all four bars sit at zero.

6 Limitations and Reviewer Responses

The discipline we advocate also exposes our own limitations, which is the honest thing for it to do, and we take the sharpest ones in turn. The central limitation is intervention strength. The perturbation we verified is modest: it flips no argmax on the probe and moves only three of 720 answers across the evaluation, none from wrong to correct. The gate proves the edit is not a no-op and that it exceeds every random-feature edit, but it does not prove the edit is strong enough to overturn a confident wrong answer, so we cannot separate a non-causal feature from an underpowered intervention, and we claim only a bounded null at this strength. The decisive strengthening, ablating the top several lie-selective features together and steering positively along the honest direction until the validity delta exceeds the answer margin, is specified as future work and is exactly the experiment the gate is designed to make interpretable.

Several further concessions are worth stating plainly. Our headline is a bounded, underpowered null, not a clean dissociation, and we have written it that way, because a dissociation would assert that a fully powered causal test was passed on detection and failed on control, whereas our test is decisive only on the detection side. The area under the ROC curve of 1.000 is not merely an in-sample resubstitution estimate maximized over the dictionary; it is arithmetically forced, because the honest activations are identically zero, so any feature that is silent on honest prompts and positive on lie prompts separates them perfectly. It therefore certifies separation of the prompt conditions and not a held-out behavioral detector, and we have dropped any claim that the feature detects lying perfectly. A related selection caveat concerns the conversion measure: the lie instruction was chosen on a pilot to maximize the lie rate, so the model lies on about 96 percent of prompts cleanly, and the conversion test is thereby asked to overturn near-maximal deception, which biases the measure toward a null; a milder instruction would give the intervention more room to show an effect, and we read the conversion bound with that bias in mind. The degenerate bootstrap intervals over all-zero columns are a tautology rather than sampling uncertainty, which is why we report an exact one-sided binomial conversion bound of about 0.017 and a suppression bound of about 0.080 on the 36 clean agreements, the only denominator on which suppression is defined. The behavioral-specificity control is trivially satisfied because the feature is silent at the honest final token, so it shows that the edit is not broadly destructive rather than that it is selective, and we no longer claim selectivity from it. The random baseline, likewise, is matched on count but only loosely on strength, since the candidate delta of 0.75 exceeds the random mean of 0.144, and because both the candidate and the weaker random edits sit below the decision margin the baseline confirms that the null is not peculiar to this feature rather than discriminating a causal feature from a non-causal one. On the transfer arm, the concern that suppression could be a silent no-op is one we closed directly: feature 12076 fires at earlier positions in every sycophancy prompt, and a validity probe there records a maximum logit change of 0.25, so the suppression null is gated, if weakly, on both arms. Finally, external validity is limited to one model, one layer, one dictionary, and two deception constructions, and the instructed-lie and sycophancy sets differ in that only the former carries an explicit instruction, an asymmetry that is the object of the transfer test rather than a confound.

7 Conclusion

We proposed an intervention-validity gate for single-feature causal-ablation studies, argued that its threshold should be read against the model's decision margin, and applied it, with a strength-matched random baseline, a specificity control, exact binomial bounds, and a standalone recompute, to a deception question on a capable open model. The gated study shows that the single Gemma Scope feature that best separates lie-instructed from honest-instructed prompts, and which also fires in sycophancy contexts, exerts no verified-strength causal control over lying or sycophancy, a bounded null that we read as evidence, not proof, that a feature can track a deceptive framing without controlling the deceptive behavior. The discipline keeps the central limitation in full view and points to the stronger, margin-exceeding intervention that would settle whether the effect is truly absent or merely out of reach at single-feature strength.

References