

Causal Faithfulness of SAE Feature Attributions in a Probe Pipeline

Samuel Xie
Seven Lakes High School
Katy, Texas, USA
samjh8xie@gmail.com

Noura Al-Moubayed
Durham University
Durham, United Kingdom
noura.al-moubayed@durham.ac.uk

Abstract

Sparse autoencoders (SAEs) decompose language model activations into interpretable features, but interpreting those features remains largely correlational. Attribution methods could estimate which features matter, yet it is unclear whether such estimates track causal effects on the model. We test this in a probe-mediated pipeline on GPT-2 Small and SST-2: we train linear and MLP probes on a layer-7 SAE, attribute probe outputs to features with SHapley Additive exPlanations (SHAP), Integrated Gradients (IG), and Gradient×Input, and score each attribution against the effect of ablating or steering that feature on the language model’s logit difference. Because this combines attribution error with any mismatch between what the probe encodes and what the model uses, our results characterize the full pipeline rather than attribution alone. The three methods nearly coincide (and are analytically identical for linear probes), but this agreement does not indicate causal validity. Importance faithfulness survives multiple-comparison correction but is modest and no better than a simple background-free baseline, while directional faithfulness is near zero and nonsignificant. In the linear arm, probe-side controls trace the directional failure to centering attributions on a background mean. Local per-example attribution outranks global averaging on importance only in the MLP arm and only when inactive features are included; among active features we find no advantage. In minimal pairs with known ground truth, attribution does not recover changed features beyond what feature activity alone provides. Importance and directional faithfulness are separable properties that require separate evaluation, and agreement among attribution methods should not substitute for causal validation. Code is available at <https://github.com/froge79/shap-sae>.

Keywords

Sparse autoencoders, Mechanistic interpretability, Feature attribution, Faithfulness evaluation, Causal interventions, Shapley values

ACM Reference Format:

Samuel Xie and Noura Al-Moubayed. 2026. Causal Faithfulness of SAE Feature Attributions in a Probe Pipeline. In *Proceedings of International Journal of Secondary Computing and Applications Research (IJSCAR VOL. 3, ISSUE 4)*. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.67149/yhjs2024.5/adr9jtxz>

This paper is published under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International (CC-BY-NC-ND 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

IJSCAR VOL. 3, ISSUE 4
© 2026 The Author(s). Published by IJSCAR.

1 Introduction

Sparse autoencoders (SAEs) have emerged as a leading tool for decomposing the internal representations of language models into interpretable features. By training a sparse dictionary on residual stream activations, SAEs split superposed representations into nearly monosemantic directions that can be inspected individually [3, 7]. Recent work has scaled this approach to frontier models, recovering tens of thousands of interpretable features [4, 18].

Despite this progress, SAE feature interpretation remains largely correlational. The dominant pipeline identifies top-activating examples for each feature and assigns a semantic label manually or with automated interpretability models. This reveals the contexts in which a feature activates, but not which features causally drive a model’s behavior. Attribution methods offer a more direct estimate of feature influence. Logit-weight analysis, gradient-based attribution, and attribution patching have all been applied to SAE features, and some have been validated against activation patching [9, 11, 17]. Yet post-hoc attribution is well known to be unreliable in general: explanations can diverge from the model’s actual computation [14], no attribution method can satisfy all desirable properties at once [2], and apparent explanatory success can arise from methods that would also “explain” noise [12]. Why should attribution fare better on SAE features?

One reason it might is that SAE latents are sparse and, ideally, closer to disentangled than neurons or dense directions. This reduces, though does not eliminate, the feature entanglement that complicates attribution in dense bases. If attributions do not track causal effects even here, we should be cautious about trusting them elsewhere. A natural design for studying this is to attribute a trained probe’s output to SAE features. A probe gives a fixed, interpretable target, and for linear probes Shapley values can be computed exactly. However, the probe is only a proxy for what the language model does, so a faithfulness score measures attribution error compounded with any mismatch between what the probe encodes and what the model uses. We treat this as a property of the design to be characterized rather than hidden.

As shown in Figure 1, we evaluate this probe-mediated pipeline on sentiment classification (SST-2) with GPT-2 Small and a 32,768-feature layer-7 SAE. We compare SHapley Additive exPlanations (SHAP) [10], Integrated Gradients (IG), Gradient×Input, and probe-weight baselines across linear and MLP probes, scoring each against the effect of feature ablation and activation steering on the language model’s logit difference, and validating on minimal sentiment-word edits with known ground truth.

Our contributions are:

Probe-mediated evaluation of attribution on SAE features.

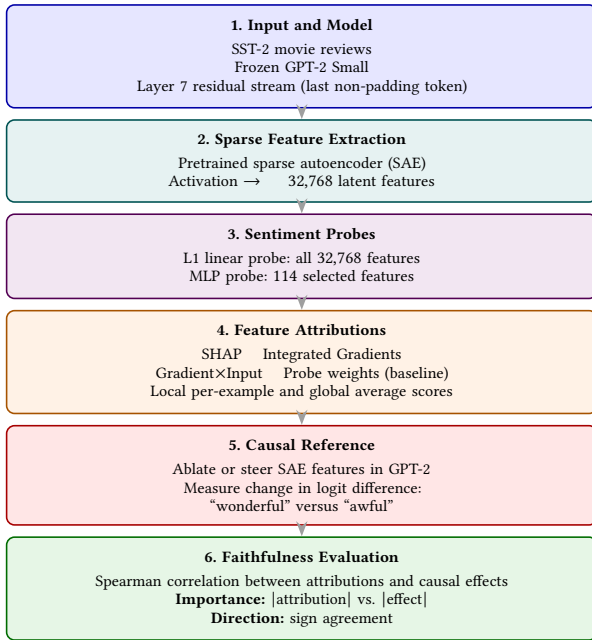


Figure 1: Pipeline for evaluating the causal faithfulness of SAE feature attributions.

We apply SHAP, Integrated Gradients, Gradient×Input, and probe-weight baselines to linear and MLP probes trained on SAE latents. For linear probes the first three are analytically identical, so we treat that arm as a control and use the MLP arm to test whether the methods diverge. They largely do not (Spearman $\rho = 0.98\text{--}0.995$ on global importance).

A causal evaluation of pipeline faithfulness. We score each attribution against the effect of feature ablation and activation steering on the language model’s logit difference. This measures the composite of attribution error and probe–LM mismatch. Probe-side controls in the linear arm partly separate the two, and we discuss what the design can and cannot distinguish.

A decomposition of global directional failure. Importance faithfulness is significant but modest and no better than a background-free baseline, while directional faithfulness is near zero. In the linear arm, probe-side controls trace this to centering attributions on a background mean. We also compare local and global attribution: local attribution ranks features by effect better than the global score only when inactive features are included, where it benefits from knowing which features are active, and we find no advantage among active features.

A minimal-pair validation with known ground truth. Across four edit types with known sentiment changes, the language model’s readout tracks the edit direction, but attribution does not identify the changed features beyond what feature activity alone provides: an attribution-free baseline (whether a feature is active at the edited last token) outperforms every method, and local and global SHAP fall below chance once distractors are activity-matched.

2 Related Work

2.1 Sparse Autoencoders for Interpretability.

Sparse autoencoders decompose residual stream activations into sparse, nearly monosemantic features, addressing the polysemanticity that makes individual neurons difficult to interpret [3, 7]. Recent work has scaled this approach to frontier models [4, 18]. Feature interpretation in this literature remains largely correlational, relying on top-activating examples or auto-interpretability pipelines rather than causal validation.

2.2 Attribution Methods.

Shapley-based attribution [10] and Integrated Gradients [16] are standard tools for identifying which inputs drive model predictions. Attribution patching [9, 17] applies gradient-based attribution to transformer internals as an efficient approximation to activation patching, and has been used to attribute effects to SAE features and validated against patching [11]. These works approximate patching effects inside the language model. We instead evaluate attribution methods applied through a trained probe, including the local-versus-global distinction and insertion versus removal. More broadly, post-hoc attribution is known to be unreliable [2, 12, 14]; we ask whether SAE features, being sparse, are a setting where it fares better.

2.3 Faithfulness Evaluation.

Faithfulness measures the degree to which an explanation reflects a model’s true decision-making process [8]. Prior evaluation often relies on input perturbation metrics, which face a basic difficulty: removing features pushes inputs off the distribution the model was trained on, so a drop in performance can reflect distribution shift rather than the importance of the removed features. ROAR [6] addresses this by removing the features an attribution ranks as important and retraining the model, and finds that many widely used attribution methods perform no better than a random ranking. We instead score attributions against the effect of feature ablation and activation steering on the language model’s output, a causal reference rather than a perturbation proxy. Like ROAR, we compare each attribution against trivial rankings (random selection, raw probe weights, and a background-free baseline) rather than reading its score in isolation. Unlike ROAR, we do not retrain, so our interventions may move activations off-distribution.

2.4 Probing.

Linear probes over internal representations are a standard tool for identifying whether a concept is encoded at a given layer [1]. High probe accuracy alone does not show that a representation encodes a property, because an expressive probe can learn the task itself. Prior work addresses this with control tasks, which assign labels to word types at random so that accuracy can only come from the probe’s capacity to memorize, and define selectivity as the gap between accuracy on the real task and on the control [5]. More expressive probes can achieve high accuracy with low selectivity, and recommend simpler probes. We use probes over SAE activations to define a prediction function over the latent feature space, through which attribution methods are applied. We use an L1-penalized linear probe and a small MLP, but we do not run

a control task. Our checks on probe validity are instead indirect: a bag-of-words baseline matches probe accuracy (Appendix E.6), and the probe’s own intervention response predicts the language model’s response only moderately. Because the probe is a proxy for the language model, faithfulness here is measured for the probe-attribution pipeline as a whole.

3 Methodology

3.1 Experimental Setup

We evaluate on SST-2 [15], a binary sentiment dataset of movie-review sentences. We pool the full-sentence release, drop duplicates and any sentence contained in another, and re-split 70/15/15 with a fixed seed. A leakage check confirms that no validation or held-out text is a duplicate or sub-text of a training text (Appendix A.1, Table 5). The train set is used for fitting probes, attribution backgrounds and baselines, and calibrating the steering parameter α . The validation set is used for probe hyperparameters, the feature mask, and reported probe accuracy. The held-out split is used only for attribution and causal evaluation. Evaluation and background rows are nested prefixes of one seeded permutation (Appendix A.1).

We use frozen GPT-2 Small [13] with the gpt2-small-resid-post-v5-32k SAE from sae-lens [4], which decomposes layer 7 residual-stream activations into 32,768 latents. Activations are taken at the last non-padding token.

3.2 Probing Models

We attribute a trained probe’s output to SAE features because a probe gives a fixed, interpretable target. The cost is that the probe is only a proxy for the language model. We train two probes, each evaluated with identical attribution and causal methods (Appendix A.2).

Linear probe. An L1-penalized logistic regression over all 32,768 features. Its gradient is constant, so IG, GA and signed-mean SHAP coincide in closed form; this arm is therefore a control in which their agreement is expected by construction.

MLP probe. A one-hidden-layer MLP on the masked feature subset. It is only weakly nonlinear (a linear fit to its logit has $R^2 = 0.99$), so this arm is a mild departure from the linear control, not a strong test of nonlinear attribution. The mask depends on the validation set, a label-free but non-independent dependency that makes its validation accuracy mildly optimistic.

Feature mask. The mask keeps features with a non-zero linear-probe coefficient and an activation frequency above 5% on the validation split. It yields 114 features and restricts only the MLP input and the steering candidate pool.

3.3 Attribution Methods

Each method attributes the same scalar: the log-odds margin ($wx+b$) in the linear arm and the pre-sigmoid logit in the MLP arm. Scales are not comparable across arms.

Zero-cost baseline. Raw probe weight (linear) and first-layer saliency $\|W_1[i, :]\|_2$ (MLP) rank features with no input-specific computation, so any gain from attribution is measured against them.

SHAP. We use `shap.LinearExplainer` (exact) for the linear arm and `shap.DeepExplainer` with enforced additivity for the MLP arm, which approximates Shapley values. Local SHAP $\phi(x)$ uses a 100-row training background. Global importance is the mean absolute value $|\phi|_j = \mathbb{E}_e |\phi_j(x_e)|$ over 1000 held-out rows against the full training set as background. We use mean $|\phi|$ because the signed mean $\Phi_j = \mathbb{E}_e [\phi_j(x_e)]$ cancels across examples and mainly reflects how evaluation rows differ from the background. Φ (100-row background) is kept wherever a direction is needed, since an unsigned score cannot supply one.

Integrated Gradients and Gradient×Input. IG integrates the gradient along a path from a baseline to the input (Appendix B); GA evaluates it at the input, $GA_j = (x_j - \text{baseline}_j) \partial f / \partial x_j(x)$. Both use the mean training row as baseline and are aggregated like SHAP, so methods differ in the attribution and not the aggregation. For a linear probe both reduce to $w_j(x_j - \text{baseline}_j)$ and equal SHAP.

3.4 Feature-Ranking Agreement

We compute pairwise Spearman ρ between methods on features with non-zero probe scores. In the linear arm, agreement ($\rho = 1$) is a sanity check. MLP agreement is computed on $|\text{score}|$ because raw saliency has no sign.

3.5 Causal Evaluation

Attribution scores the probe alone. To test whether scores track the language model, we intervene on SAE feature activations at the layer-7 residual stream and measure the change in logit difference. A faithful attribution should rank features by effect size and predict its sign. Because the attribution explains the probe while the intervention acts on the language model, this tests the probe-attribution-LM pipeline as a whole. For feature i with activation a_i :

Ablation:

$$\tilde{a}_i = 0, \quad \tilde{z} = \text{Dec}(\tilde{\mathbf{a}}) \quad (1)$$

Steering (additive, signed):

$$\tilde{a}_i = a_i + \alpha_i, \quad \tilde{z} = \text{Dec}(\tilde{\mathbf{a}}) \quad (2)$$

A no-edit encode-decode pass is the baseline, isolating SAE reconstruction error. Interventions apply at every token. Because attribution is computed at the last token, we also run a last-token-only condition that matches its scope; it is the main condition (Table 2), with all-token results in Appendix E.7.

Readout. The readout is the final-token logit difference,

$$\Delta = \ell(\text{“wonderful”}) - \ell(\text{“awful”}), \quad (3)$$

compared before and after intervention. We use $|\Delta|$ for importance and signed Δ for direction. Ablation effects are negated so positive ρ always means greater faithfulness.

Steering calibration. We set α on the training split as the smallest multiple of the median natural positive activation that gives a measurable but unsaturated shift on a layer-11 probe (Appendix C.1). Steering is scaled per feature to normalize for differing activation magnitudes, with unscaled α as a robustness check.

Feature selection. We evaluate three candidate sets: k random features, the top k by |SHAP|, and all 114 masked features (Appendix C.2). Inference comes from the full set, with one Spearman correlation per method and correlation type. Random draws are repeated over 10 seeds and summarized descriptively, since the draws overlap and are not independent replicates. Multiple comparisons are controlled with Benjamini–Hochberg (FDR 0.05) within a family fixed by rule: the tests of one run. We also report wider families and Benjamini–Yekutieli (Appendix C.3).

Reuse and background sensitivity. Candidate sets that do not depend on attribution are rescored from saved intervention effects rather than re-measured. To check that the background does not drive global scores, we also score Φ and $|\phi|$ under backgrounds of different sizes and under bootstrapped 100-row backgrounds (Appendix C.5).

Probe-side controls. A low faithfulness ρ cannot by itself separate attribution error from probe–LM mismatch. For the linear probe, the probe’s own response to each intervention is closed form: steering moves the margin by $P_j = w_j \alpha_j$, and ablation by $A_j = -\mathbb{E}_e[w_j x_{ej}]$. We therefore (i) score attributions against P_j and A_j , which isolates attribution error; (ii) correlate P_j with the measured logit-difference effect, which isolates how well the probe predicts the model; and (iii) add a background-free baseline $w_j \bar{x}_j$. We also test a class-matched background (Appendix C.6). These controls are unavailable for the MLP arm.

3.6 Local Faithfulness

Global attribution averages over examples, which could hide per-example quality. We test whether local attributions track each sentence’s intervention effects better than the global score, for importance (does |attribution| rank features by effect size?) and direction (does the signed attribution match the sign of Δ ?). For each sentence x , candidates are the union of the top- k features by $|\phi(x)|$ and by global $|\phi|$, so both are scored on identical sets. We compute per-sentence Spearman ρ for each and report the mean and the fraction of sentences where local wins. The global directional score is signed Φ .

Confound. Globally top features are often inactive in a given sentence, and ablating an inactive feature changes the logit by exactly zero, so global attribution’s lower correlation may reflect inactive features rather than worse ranking. Block [A] uses all candidates; Block [B] keeps only active features (activation > 0).

Significance tests. For both blocks and correlation types, we test the mean gap $\rho_{\text{local}} - \rho_{\text{global}}$ with a crossed bootstrap that resamples sentences and features independently ($B = 1000$). Sentences share most of their candidate features (mean pairwise Jaccard ≈ 0.7), so a sentence-only test understates uncertainty by roughly 4–9 $\times$. Two follow-ups quantify the confound: a regression of the per-example Block [A] gap on the active fraction of candidates, and a partial Spearman controlling continuously for activation magnitude.

3.7 Known-Ground-Truth Validation

Natural SST-2 text has no known feature importance, so a near-zero ρ could mean attribution failure or an insensitive harness. Minimal

pairs resolve this because the source of the sentiment change is known. We use four edits crossing edit type with position relative to the final token: *removal* of a sentiment word, *substitution* with an opposite-polarity word, *insertion* of “it was (word) .” at the end of a neutral sentence, and *interior insertion* of the same clause mid-sentence.

We check three quantities in dependency order:

- (1) **Readout validity:** whether the logit-difference sign tracks the known edit direction.
- (2) **Feature recovery:** AUROC of each method at identifying features whose last-token activation changed (chance 0.5; Appendix D), against attribution-free baselines based on activation magnitude and whether a feature is active.
- (3) **Causal recovery:** correlation between per-feature ablation effects and attribution scores on the edited sentence.

Two confounds are controlled. Using the pair’s own activations as the SHAP background makes scores proportional to activation change, which tautologically inflates AUROC, and selecting distractors by top global SHAP biases it downward. We therefore use the training background and random distractors. Because random distractors are mostly inactive, we also report an activity-matched variant. Readout sign accuracy is reported with Wilson 95% intervals against a null-edit noise floor.

3.8 Lexical Simplicity Control

We test whether SST-2 is lexically simple enough that interaction-aware attribution cannot outperform simpler methods, using the full held-out split (Appendix E.6). We run three analyses: (1) **co-activation**, the number of active features per example and the overlap of the top-20 features; (2) **baselines**, comparing the SAE probe’s held-out accuracy with logistic regressions on a sentiment lexicon, bag-of-words variants, and the dense layer-7 residual; and (3) **word-detector tests**, how well bag-of-words and lexicon features predict each top feature’s activation and the probe logit. Reading rules were fixed in advance: the probe is no better than bag-of-words if the best validation-selected variant is within 0.03 accuracy, and a feature is bag-of-words-explained if held-out $R^2 \geq 0.25$.

4 Results

4.1 Probe Performance

Both probes perform well above chance and comparably to each other (Table 1); the MLP probe’s validation score is mildly optimistic because of its pre-filtered inputs.

Table 1: Probe architectures and performance.

Arm	Input dim	Val. Accuracy	Val. AUROC
Linear	32,768	0.815	0.898
MLP	114 (filtered)	0.822	0.899

4.2 Attribution Method Agreement

With all three methods aggregated the same way, IG, GA and SHAP agree almost perfectly on global importance, while probe saliency

correlates only weakly with all three (Figure 2, Table 12). An earlier version aggregated IG and GA as signed means, which lowered IG–SHAP agreement; that gap reflected the aggregation, not the methods (Appendix C.5).

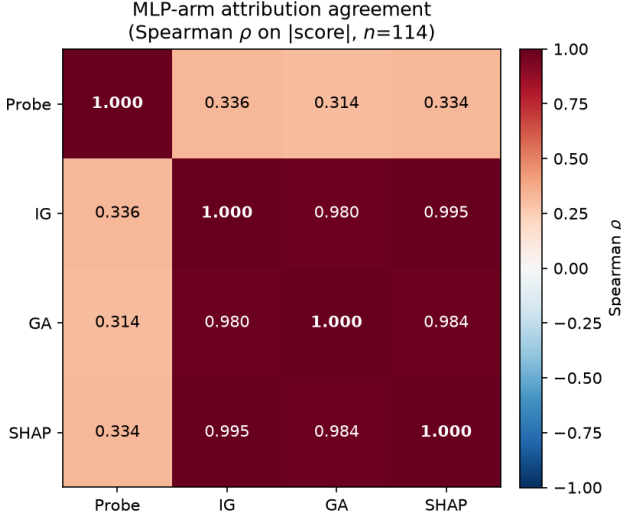


Figure 2: MLP-arm attribution agreement (Spearman ρ on $|\text{score}|$, $n = 114$ filtered features). IG, GA, and SHAP (all mean $|\cdot|$) form a near-identical cluster; saliency correlates only weakly with all three.

4.3 Causal Faithfulness

Table 2 reports global faithfulness over all 114 masked features under last-token steering, BH-corrected within arm. SHAP, IG and GA share the same aggregation, so they are identical in the linear arm and nearly so in the MLP arm.

Importance faithfulness is significant for all three methods in both arms but only small-to-moderate ($\rho \approx 0.42\text{--}0.46$, so each explains roughly a fifth of the variation in feature effects). The zero-cost baselines are lower, and in the MLP arm saliency is indistinguishable from zero.

Directional faithfulness is weaker: SHAP, IG and GA are near zero and nonsignificant in both arms, while probe weights alone are significant in the linear arm ($\rho_{\text{dir}} = 0.305$; 0.399 all-token). The weights beat SHAP by 0.41 (95% CI [0.09, 0.72], resampling features), which is notable because attribution adds instance-specific information to the same weights. A likely reason is that steering adds a fixed shift, so its effect follows the sign of the weight, whereas attribution multiplies each weight by a deviation ($x_j - \text{baseline}_j$) that can change sign. The probe-side controls below test this. Neither conclusion depends on how the test family is defined (Appendix C.3). All-token steering gives similar or higher importance faithfulness, and the same tests survive FDR correction (Appendix E.7). Matching the intervention scope to the attributed token position therefore does not improve measured faithfulness.

Global score and background. Scoring importance with the signed mean Φ instead of mean $|\phi|$ lowers importance faithfulness in both

Table 2: Last-token global faithfulness ($k = 114$, steering, all filtered features). Cells show Spearman ρ with BH-corrected q in parentheses.

Arm	Method	$\rho_{\text{imp}}(q)$	$\rho_{\text{dir}}(q)$
Linear	SHAP/IG/GA	+0.455 (<0.0001)*	−0.108 (0.254)
Linear	Probe	+0.270 (0.006)*	+0.305 (0.002)*
MLP	SHAP	+0.425 (<0.0001)*	−0.105 (0.464)
MLP	Probe saliency	+0.032 (0.859)	n/a
MLP	IG	+0.423 (<0.0001)*	−0.053 (0.804)
MLP	GA	+0.444 (<0.0001)*	−0.014 (0.884)

* Survives FDR 0.05. SHAP, IG and GA importance are mean absolute attribution against the full training set (identical in the linear arm); direction is the signed mean Φ (same for all three in the linear arm). Linear arm: 5/8 tests survive (SHAP, IG and GA counted separately; 3/4 distinct); MLP arm: 3/7. The same tests survive when BH is pooled over all 30 tests (both arms, both intervention scopes).

arms and scopes, and makes it depend strongly on the background draw. The same holds for IG and GA, so SHAP’s earlier apparent advantage over them came from aggregation, not the method. The background does not change the directional result: Φ ’s directional ρ stays near zero under the full training set and under bootstrapped backgrounds (Table 9).

Probe-side controls. Because attributions explain the probe and interventions act on the language model, a low ρ could reflect either. For the linear probe we can score attributions against the probe’s own closed-form response (Section 3.5, Table 11). Against the probe’s steering effect, Φ is anti-correlated ($\rho_{\text{dir}} = -0.37$ with the full training set), while the background-free product $w_j \bar{x}_j$ reaches +0.95. The directional failure is therefore attribution error relative to the probe, produced by centring on a background mean, and does not require any probe–LM mismatch. A class-matched background leaves it unchanged, which excludes class balance as the cause.

The probe–LM link is itself only moderate: the probe’s steering effect predicts the measured effect with $\rho = +0.31$ (95% CI [+0.10, +0.49]), so the language-model test has limited resolution. At that level the background-free product and raw weight predict direction better than Φ , and for importance the background-free product matches mean $|\phi|$, so SHAP importance does not beat a baseline with no background. Ablation gives the same picture (Appendix C.6). We therefore read the directional null as a property of Φ rather than evidence about the language model, and the importance results as showing that SHAP is no better than simple probe-based scores. The design cannot show that any probe-based score is unfaithful to the language model beyond the level set by the probe–LM link.

4.4 Local Faithfulness

Figure 3 and Table 3 compare local and global SHAP, with global scored by mean $|\phi|$. Our headline is Block B (active features only), since Block A is dominated by a tied block of inactive features that local $|\phi(x)|$ identifies by construction.

In Block B the importance gap is negative in both arms, nominally significant in the linear arm but not after FDR correction, and not significant in the MLP arm. No directional gap is significant in either block. In Block A, local SHAP ranks importance better than

the global score in the MLP arm, the one local result that survives correction. It is fragile: rescaling the original candidate sets with a 100-row background gives $p = 0.054$ (Appendix C.5). The active fraction of candidates explains only a modest share of the between-example variation in the Block A gap ($R^2 = 0.08\text{--}0.19$). Local SHAP therefore ranks features better than the global score only when inactive features are included; among active features we find no evidence that it is more faithful, and the point estimates favour the global score.

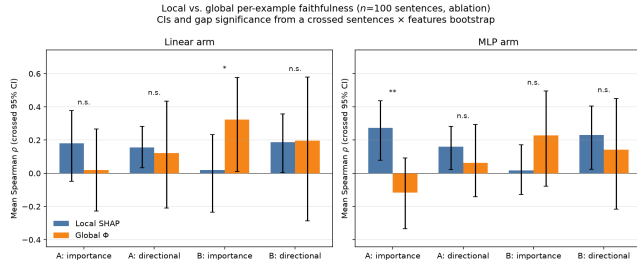


Figure 3: Local vs. global attribution, per-example faithfulness ($n = 100$ sentences, ablation). Bars are mean Spearman ρ across examples with 95% CIs from a crossed (sentences $\times$ features) bootstrap; asterisks mark the *uncorrected* crossed-bootstrap p for the local-global gap (** $p < 10^{-2}$, * $p < 0.05$, n.s. otherwise); only the MLP-arm Block A importance gap survives FDR correction (Appendix C.3). Global = mean $|\phi|$. Block A = all top-20-local $\cup$ top-20-global candidates; Block B = active-only control.

Why global directional attribution may fail. For a linear probe, $\Phi_j = w_j(\bar{x}_j^{\text{eval}} - \mathbb{E}_{\text{bg}}[x_j])$, so global attribution mostly reflects how the evaluation rows differ from the background rather than feature importance. Its magnitude tracks the size of this mean shift more closely than the size of the weights ($\rho = 0.85$ vs. 0.42), and only 40% of the top-20 $|\Phi|$ features have Φ matching the sign of their weight. Neither class balance nor a larger background removes the mismatch: against all training rows Φ matches the weight’s sign for only 43% of the 114 masked features, and directional ρ over bootstrapped 100-row backgrounds is centred on zero. Because evaluation and background rows come from the same distribution, the shift has no reason to share the weight’s sign. It is not pure noise, though: held-out means sit slightly below training means for most features, and the shift is negatively related to a feature’s activation scale. Since the steering size α_j scales with that activation scale, this makes Φ anti-correlated with the probe’s steering effect. This follows from defining Φ as a signed mean, not from the size of the background, and dropping the background ($w_j\bar{x}_j$) restores a positive directional ρ .

4.5 Known Ground-Truth Validation and Lexical Simplicity Control

Table 4 reports 100 minimal pairs for each of four edits. The readout tracks the known direction in all four, but is weakest for removal and interior insertion, where effects sit close to a null-edit noise floor (Appendix D).

Table 3: Local vs. global attribution, per-example faithfulness ($n = 100$ sentences, ablation). Mean ρ across examples; 95% CI and two-sided p (uncorrected; only MLP Block A importance survives BH, Appendix C.3) for the gap from a crossed (sentences $\times$ features) bootstrap. Global score: mean $|\phi|$ for importance, signed Φ for direction. Block A: all candidates; Block B: active features only.

Block, kind	Local	Global	Gap	95% CI	p	Win frac.
<i>Linear arm</i>						
A, importance	+0.181	+0.020	+0.161	[−0.105, +0.445]	0.206	0.760
A, directional	+0.157	+0.123	+0.034	[−0.356, +0.442]	0.852	0.560
B, importance	+0.019	+0.324	−0.305	[−0.573, −0.020]	0.036	0.137
B, directional	+0.188	+0.197	−0.009	[−0.514, +0.560]	0.938	0.505
<i>MLP arm</i>						
A, importance	+0.273	−0.115	+0.389	[+0.109, +0.665]	0.002	0.970
A, directional	+0.161	+0.062	+0.098	[−0.225, +0.370]	0.532	0.610
B, importance	+0.017	+0.229	−0.213	[−0.534, +0.138]	0.290	0.240
B, directional	+0.230	+0.142	+0.088	[−0.383, +0.565]	0.756	0.635

Against random distractors, local SHAP is modestly above chance and global SHAP is near chance. Local SHAP’s apparent advantage on insertion is a position effect: appending the clause beats placing it mid-sentence, a difference that holds across seeds. Feature recovery is also not shown to exceed what activity alone gives: the attribution-free “active at the edited last token” baseline beats every method, and once distractors are activity-matched, both SHAP variants fall below chance in every seed.

Causal recovery is more informative. It is positive for local SHAP in all four edits and for global SHAP in most, but local SHAP’s edge over global is absent or reversed except for appended insertion. We read this as weak evidence that $|\text{attribution}|$ tracks ablation effect, not that attribution identifies the changed features.

Table 4: Minimal-pair ground-truth validation (seed 42, $n = 100$ pairs per edit, linear arm). Sign acc. is readout validity (Wilson 95% CI); AUROC is feature recovery (chance = 0.5) with random or activity-matched distractors (global SHAP is mean $|\phi|$ against the full training set, which equals mean- $|\cdot|$ IG and GA for the linear probe); ρ is causal recovery (random distractors). The baseline row scores whether a feature is active at the edited last token, with no attribution.

Edit	Method	Sign acc.	AUROC rand.	AUROC act.-matched	ρ
Remove	Local SHAP	0.76 [0.67, 0.83]	0.553	0.420	+0.186
	Global SHAP		0.482	0.442	+0.057
	Probe weights		0.395	0.508	−0.148
	Active baseline		0.851	—	—
Substitute	Local SHAP	0.96 [0.90, 0.98]	0.565	0.427	+0.355
	Global SHAP		0.512	0.421	+0.276
	Probe weights		0.466	0.522	+0.013
	Active baseline		0.803	—	—
Insert (final)	Local SHAP	0.91 [0.84, 0.95]	0.622	0.486	+0.219
	Global SHAP		0.472	0.370	−0.034
	Probe weights		0.406	0.490	−0.200
	Active baseline		0.822	—	—
Insert (interior)	Local SHAP	0.79 [0.70, 0.86]	0.492	0.379	+0.194
	Global SHAP		0.425	0.405	+0.054
	Probe weights		0.410	0.475	−0.151
	Active baseline		0.789	—	—

A hand-built sentiment lexicon is a weak baseline, but a full-vocabulary bag-of-words logistic regression is statistically indistinguishable from the SAE probe (Appendix E.6). SST-2 sentiment is therefore largely recoverable from word presence, and we do not claim the probe carries signal beyond it. The probe does trail a logistic regression on the dense layer-7 residual. The lexicon explains almost none of any top feature’s variance, so the earlier lexicon-correlation test was uninformative; full-vocabulary bag-of-words explains more, so top features are neither simple lexicon detectors nor free of word identity. Because the task is largely lexical, it gives interaction-aware attribution little room to differ from simpler methods, which limits what these experiments can show.

5 Discussion

Agreement among methods is not evidence of faithfulness. SHAP, Integrated Gradients, and Gradient×Input form a near-identical cluster once they are aggregated the same way (MLP arm: $\rho = 0.98$ – 0.995 ; Figure 2). In the linear arm they are analytically identical, so agreement there is guaranteed by construction. This cluster is only modestly related to causal effects. Importance faithfulness is statistically reliable ($\rho_{\text{imp}} = 0.42$ – 0.46 , surviving every family-wise correction we tried) but explains only about 18–21% of the variance in feature effects, and directional faithfulness is near zero and non-significant for all three methods. Practitioners should therefore not treat agreement among attribution methods as evidence that they identify causally important SAE features.

Attribution adds little over simple probe-based scores. The probe-side controls show that SHAP importance does not beat a baseline with no background: the product $w_j \bar{x}_j$ reaches $\rho_{\text{imp}} = 0.47$, against 0.46 for mean $|\phi|$. For direction, the raw probe weight ($\rho = +0.31$) and $w_j \bar{x}_j$ ($+0.35$) both outperform the signed mean Φ (-0.11). In the linear arm, Φ is anti-correlated with the probe’s own steering response ($\rho = -0.28$ to -0.37), whereas $w_j \bar{x}_j$ reaches $+0.95$. The directional failure is therefore attribution error relative to the probe, produced by centering on a background mean, and does not require any probe–LM mismatch. These results concern signed-mean aggregation of background-centered attributions. They do not show that all attribution designs fail.

Local versus global attribution depends on the regime. Local SHAP ranks features by effect better than the global score only in one setting: the MLP arm, with inactive features included (Block A, gap $+0.389$, $q = 0.008$). That result is fragile, since rescaling the original candidate sets gives $+0.222$ ($p = 0.054$), and it largely reflects local attribution knowing which features are active, because ablating an inactive feature moves the logit by exactly zero. Among active features (Block B), the point estimates favor the global score (-0.305 linear, -0.213 MLP), though only the linear gap is nominally significant and it does not survive correction. No directional gap between local and global attribution is significant in either block. We therefore find no evidence that local attribution is more faithful than global attribution once feature activity is controlled, and we do not recommend one over the other on the basis of these data.

Attribution does not recover changed features beyond feature activity. In minimal pairs, the language model’s readout tracks the known edit direction (sign accuracy 0.76 – 0.96), so the harness is

sensitive. Attribution nonetheless does not identify the changed features. Local SHAP is modestly above chance against random distractors (AUROC 0.49 – 0.62), but an attribution-free baseline, whether a feature is active at the edited last token, reaches 0.79 – 0.85 . Once distractors are activity-matched, local and global SHAP fall below chance (0.37 – 0.49) in every seed. The one consistent pattern is positional: appended insertion (AUROC 0.622) outperforms interior insertion (0.492), positive in 4 of 4 seeds, which suggests attribution read at the final token is more informative about edits adjacent to it. Causal recovery among active candidates is positive for both local and global SHAP ($\rho = 0.11$ – 0.38), but local SHAP’s edge is absent or reversed except for appended insertion. We read this as weak evidence that $|\text{attribution}|$ tracks ablation effect, not that attribution identifies the changed features.

The task is largely lexical. A full-vocabulary bag-of-words regression reaches 0.794 – 0.806 held-out accuracy, within 0.03 of the SAE probe (0.810 ; Table 17), and bag-of-words explains 0.39 of the probe logit’s variance. SST-2 sentiment is thus largely recoverable from word presence, which leaves interaction-aware attribution little room to differ from simpler methods. This limits what these experiments can show, and it means the near-identical behavior of SHAP, IG, and GA may be partly a property of the task.

5.1 Limitations

Probe–LM confound. Attributions explain a trained probe over SAE activations, while faithfulness is scored against the language model’s logit difference. Our estimates combine attribution error with any mismatch between what the probe encodes and what the model uses. The probe-side controls separate the two only in the linear arm, and the probe’s effect predicts the measured effect at only $\rho \approx 0.3$ – 0.4 , so the language-model test has limited resolution.

Method equivalence in the linear arm. SHAP, IG, and GA are analytically identical for linear probes. The MLP arm is only weakly nonlinear (linear fit to its logit $R^2 = 0.99$) and uses approximate DeepSHAP, so it adds little beyond the linear control.

Limited scope. All results come from one model (GPT-2 Small), one SAE and layer (layer 7), and one binary sentiment task (SST-2). We make no claim that the patterns, including the local–global comparison and the position effect in minimal pairs, hold elsewhere.

Statistical power and post hoc choices. Most per-seed $k = 20$ associations do not survive pooled correction, so inference rests on the $k = 114$ run. The mean- $|\phi|$ importance scoring was adopted after the final results were seen, and the wider test families were added post hoc. Significant effects should be read as suggestive.

Untested explanation. Our account of the global directional failure is supported by the linear-arm probe-side controls, but we did not verify it in the MLP arm, which has no closed-form probe response.

Steering as causal reference. Additive steering and ablation in SAE space may move activations off-distribution, and the steering strength α was calibrated against a different readout (a layer-11 probe) than the one used for evaluation. Intervention effects are a causal reference, not ground truth. Only the minimal-pair edits have

known ground truth, and they cover a narrow class of sentiment-word changes.

5.2 Future Directions

The most informative next step is to separate attribution error from probe-model mismatch, either by attributing directly to the language model’s logit difference or by intervening on the probe’s output and comparing against the language model’s response. Attribution designs that avoid centering on a background mean, or that are scored against a probe’s own response in the nonlinear case, would test whether the directional failure is specific to signed-mean aggregation. Because the decomposition-and-control methodology is not specific to SHAP, it could be applied to other attribution methods, models, layers, SAE architectures, and tasks that are less lexically solvable than SST-2, with larger feature samples to improve power. Finally, a broader set of edits with known ground truth would test whether the position effect in minimal pairs is a general property of attribution or specific to this setup.

6 Conclusion

We evaluated SHAP, Integrated Gradients, and Gradient×Input on SAE latent features in a probe-mediated pipeline, comparing their attributions with the effects of feature ablation and steering on the language model’s logit difference. In our GPT-2 Small, layer-7, SST-2 setting, importance faithfulness was statistically significant but modest and no better than a simple background-free baseline, while directional faithfulness was near zero. Probe-side controls in the linear arm traced the directional failure to centering attributions on a background mean. Although the three methods agreed closely with one another, this agreement did not imply correspondence with causal effects. Local attribution outperformed global averaging on importance only in the MLP arm when inactive features were included; among active features, we found no advantage. In minimal pairs with known ground truth, attribution did not identify changed features beyond what feature activity alone provided. These findings show that, in this experimental setting, agreement between attribution methods is insufficient evidence of causal faithfulness, and that importance and directional faithfulness can yield different assessments of an attribution method. Practitioners should therefore evaluate these properties separately and validate probe-mediated attributions against interventions before using them to make claims about model behavior. Because our design combines attribution error with potential probe-model mismatch, future work should separate these sources of error—for example, by comparing attributions with interventions on the probe itself—and test the robustness of these results across models, layers, and tasks.

7 Use of AI Tools

The authors used Claude Sonnet 5 and Opus 5 to assist with generating portions of the experimental code and with editing the paper’s prose. The authors reviewed, tested, and verified all AI-assisted code and text, and take full responsibility for the experiments, results, and conclusions reported here.

References

- [1] Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644* (2016).
- [2] Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. 2024. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences* 121, 2 (2024), e2304406120.
- [3] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermy, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread* (2023). <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.
- [4] Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2025. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, Vol. 2025. 26721–26754.
- [5] John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*. 2733–2743.
- [6] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems* 32 (2019).
- [7] Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, Vol. 2024. 7827–7845.
- [8] Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 4198–4205.
- [9] János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. 2024. Atp*: An efficient and scalable method for localizing llm behaviour to components. *arXiv preprint arXiv:2403.00745* (2024).
- [10] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).
- [11] Samuel Marks, Can Rager, Eric Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations*, Vol. 2025. 23888–23923.
- [12] Maxime Méroux, Giada Dirupo, François Portet, and Maxime Peyrard. 2025. The dead salmon of ai interpretability. *arXiv preprint arXiv:2512.18792* (2025).
- [13] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
- [14] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.
- [15] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1631–1642.
- [16] Mukund Sundarajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.
- [17] Aaqib Syed, Can Rager, and Arthur Conmy. 2024. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. 407–416.
- [18] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermy, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread* (2024). <https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html>

A Experimental Setup Details

A.1 Task and Dataset

The HuggingFace stanfordnlp/sst2 train set consists of phrases, and a random split of it leaked sub-phrases across splits (83.8% of a held-out sample were contiguous sub-phrases of a train phrase). We therefore pool the sentence-level SST-2 release (6,920 / 872 / 1,821 sentences), drop exact duplicates, label-conflicted texts and any sentence contained in another, and re-split 70/15/15 with a fixed seed (Table 5). The split script aborts if any validation or held-out text is a duplicate or sub-text of a train text; the final split has none.

Table 5: Data split sizes and purposes.

Split	n
Probe train	6,696
Probe val	1,435
Held-out (“shap”)	1,436

Canonical row sampling. All evaluation and background rows are *prefixes* of one seeded permutation (seed= 42), not independent draws. The 100-row local sample is the first 100 rows of the 1000-row global sample, so local and global results join by row position.

A.2 Model and Probes

GPT-2 Small and the layer-7 resid_post SAE are frozen throughout. SAE decoder rows are approximately unit norm (median 0.9996), so we apply no decoder-norm correction to steering vectors.

Linear probe. sklearn LogisticRegression with an L1 penalty ($C = 1$, liblinear, fixed random_state) over the full 32,768-d activation. It keeps ~600 non-zero coefficients, which seed the 114-feature mask; the mask is a property of the reported run, not a fixed constant.

MLP probe. sklearn MLPClassifier with one hidden layer of width 32, ReLU activation, Adam, and L2 regularization, trained on the 114 masked features only.

A.3 Glossary

Table 6 defines the notation and shorthand used in the results.

B Attribution Method Details

Global SHAP. Global importance is $|\phi|_j = \mathbb{E}_e |\phi_j(x_e)|$ over the 1000-row held-out sample, with every training row as the background (6,696 rows). For the linear probe $\phi_j(x) = w_j(x_j - \mu_j)$ with μ the background mean, so $|\phi|_j = |w_j| \mathbb{E}_e |x_{e,j} - \mu_j|$ in closed form (checked against LinearExplainer). For the MLP we run DeepSHAP in chunks of 125 evaluation rows with additivity enforced. The signed mean $\Phi_j = \mathbb{E}_e [\phi_j(x_e)]$ (100-row background) is used wherever a direction is needed. Signed values cancel across examples, so Φ is not an importance score (Appendix C.5).

Table 6: Glossary of notation and terms used in the results.

Term	Meaning
$\phi_j(x)$	Local SHAP value of feature j for example x : how much the feature moves the probe output away from its expectation over a background set.
$ \phi _j$	Global importance: $\mathbb{E}_e \phi_j(x_e) $, the mean absolute SHAP value over evaluation rows. Unsigned, so it supplies importance but no direction.
Φ_j	Signed global attribution: $\mathbb{E}_e [\phi_j(x_e)]$, the mean SHAP value over evaluation rows. Used wherever a direction is needed.
Imp. (importance)	Does attribution rank features by effect size $ \Delta $? Scored with Spearman ρ_{imp} .
Dir. (directional)	Does the signed attribution match the sign of the effect Δ ? Scored with Spearman ρ_{dir} .
Δ	Change in the language model’s logit difference, $\ell(\text{“wonderful”}) - \ell(\text{“awful”})$, after ablating or steering a feature.
Block A	Local-versus-global comparison over all candidate features (top-20 local $\cup$ top-20 global). Dominated by inactive features, whose ablation effect is exactly zero.
Block B	Same comparison restricted to active features (activation > 0). Our headline local-versus-global test, since it removes the activity confound.

Integrated Gradients.

$$\text{IG}_j = (x_j - b_j) \cdot \int_0^1 \frac{\partial f}{\partial x_j} (b + \alpha(x - b)) d\alpha, \quad (4)$$

approximated in the MLP arm with a 50-step trapezoidal Riemann sum. IG and GA share the baseline b , the *mean* of the train background rather than a zero vector (100 rows for signed means; all 6,696 rows for importance, matching SHAP). SAE features are very sparse, so a zero baseline would make GA degenerate to 0 whenever $x_j = 0$. For the linear probe, IG, GA, and Φ all reduce to $w_j \cdot \mathbb{E}_e [x_{e,j} - \mu_j]$, which we verify numerically (Appendix E.1).

Global importance for IG and GA is aggregated like SHAP, as

$$\mathbb{E}_e |\text{IG}_j(x_e)| \quad \text{and} \quad \mathbb{E}_e |\text{GA}_j(x_e)|. \quad (5)$$

Scoring them as $|\mathbb{E}_e [\cdot]|$ while SHAP used $\mathbb{E}_e |\cdot|$ would confound method with aggregation.

C Causal Evaluation Details

C.1 Steering-Strength (α) Calibration

We sweep $\{0.5, 1, 2, 4\} \times$ the median natural positive-activation scale over three top-|SHAP| pilot features on 200 train examples, reading out a layer-11 residual probe’s $P(\text{positive})$ rather than the evaluation logit difference. α is a rough “measurable, unsaturated” magnitude, not a hyperparameter tuned against the evaluation metric. We select the smallest α with $\text{mean}|\Delta P| \geq 0.02$ and newly saturated fraction ≤ 0.25 , giving $\alpha = 0.4936$ (Table 7). This is the smallest grid value, so the criterion does not bind. The mean signed shift is negative at every α under this readout but positive under the evaluation logit-difference readout on the same pilot, so the sign reflects the readout mismatch and not the features’ effect on the language model.

Table 7: Raw α -calibration pilot sweep (probe-readout, 3 pilot features, 200 train examples).

α	mean $ \Delta P $	mean signed ΔP	newly saturated frac.
0.4936	0.0490	-0.0301	0.028
0.9871	0.0961	-0.0571	0.085
1.9743	0.1791	-0.0981	0.182
3.9485	0.2969	-0.1422	0.368

Per-feature scaling. In the primary **scaled** mode, $\alpha_i = \alpha \cdot \text{scale}_i / \text{median}(\text{scale})$, where scale_i is feature i 's mean positive train activation. This equalizes the relative nudge across features. The **constant** mode (flat α , robustness check) confounds the $|\text{attribution}|$ -vs- $|\Delta|$ correlation, since $|\Delta|$ then partly reflects how hard a feature was pushed relative to its own scale.

C.2 Candidate Selection

top selection takes the top- k by $|\text{SHAP}|$ and scores all methods on that set. This range-restricts only SHAP's scored variance, which mechanically deflates SHAP's ρ relative to the other methods. `all` scores every masked feature ($n = 114$). At $n = 20$ the confidence interval on Spearman ρ spans most of $[-1, 1]$, so a single $k = 20$ random draw cannot distinguish a null from a moderate effect.

Provenance of the `all` analysis. The `all` selection was introduced after an earlier analysis on a phrase-level split (later found to leak) whose single $k = 20$ draw looked null for several tests, which is outcome-dependent. We have since rebuilt the split at sentence level and regenerated every number here. The `all` selection, the BH family and its rule were fixed before this final analysis was run, so the design was chosen on the earlier data and not on the final results. The analysis is not fully confirmatory, however: the global SHAP importance score was changed from the signed mean Φ to mean $|\phi|$ (with IG and GA aggregated to match) after the final results had been seen, in response to review. Both versions are reported (Appendix C.5); the directional results, which use Φ , are unchanged, but the mean- $|\phi|$ importance results should be read as post hoc. All $k = 20$ results are reported (Table 16, and the seed sweep below). We treat `all` as the inferential run because it uses the whole filtered pool with no sampling; the final $k = 20$ draw agrees with it.

C.3 Multiple-Comparison Correction

Family rule. We apply Benjamini-Hochberg (BH, FDR 0.05) with the following primary family: the tests of one `all` run (one arm, one intervention scope; 8 linear, 7 MLP), and, for local faithfulness, the four crossed-bootstrap tests (Block A/B $\times$ importance/directional) of one local run. The per-run rule was specified before the final analysis, but after the initial analysis had been examined. The wider families below are post hoc; we report them to show whether any verdict depends on the choice of family.

Duplicate tests. In the linear arm IG and GA have identical (ρ, p) by construction, as do their directional and importance tests, so the 8 linear tests per run are 4 distinct ones. Counting them separately inflates survivors and, when pooled, BH power, so the sensitivity

analysis counts identical tests once (22 distinct global tests; raw 16/30).

Sensitivity to the family. Counts are recomputed from the per-test p -values of every inferential test in the paper under each family definition. The paper-wide family has 30 tests: the 22 distinct global

Table 8: Tests surviving FDR 0.05 under alternative family definitions. BY is Benjamini-Yekutieli (arbitrary dependence).

Family	BH	BY
Global, per run (primary)	12/22	12/22
Global, pooled over arms, scopes, or both	12/22	12/22
Global, raw (SHAP/IG/GA counted separately)	16/30	16/30
Paper-wide, per run / per local source (primary)	13/30	13/30
Paper-wide, global and local pooled separately	13/30	13/30
Paper-wide, one family of all inferential tests	13/30	13/30

tests and the 8 local crossed-bootstrap tests. The sign tests, minimal-pair summaries, seed sweep and language-model diagnostic carry no independent inferential claim and are excluded.

No verdict depends on the family: the same 12 global tests and one local test survive under BH and BY in every definition. The local test is the MLP-arm Block A importance gap. Importance faithfulness of SHAP, IG and GA survives all families (except MLP probe saliency), no directional test of SHAP, IG or GA has $q < 0.25$ in any, and no Block B or linear-arm local test survives. A test that fails the FDR criterion is not evidence that its effect is zero.

Dependence. The two scopes share features and eval rows, and the local runs share sentences, so the tests are not independent. BH assumes positive dependence, which we do not verify; BY does not.

Seed sweep. The multi-seed sweep uses a different convention. Its 10 seeds $\times$ 8 (linear) or 7 (MLP) per-seed tests each use $n = 20$ features that overlap across seeds, so the pooled BH count over them (39 distinct linear tests, 70 MLP) is descriptive only.

C.4 Local Faithfulness

Local analyses use $k_{\text{local}} = k_{\text{global}} = 20$ and $n_{\text{examples}} = 100$ in both arms. A separate script (`local_shap_faithfulness`) re-implements the MLP-arm analysis against the language model on the same sentences and reproduces the means to within 0.001; it is a replication check, not an independent test.

C.5 Global Baseline Sensitivity

Table 9 compares mean $|\phi|$ with the signed mean Φ and shows how both depend on the background; Table 10 shows the effect of the choice on the local-versus-global gap. Global-faithfulness values rescore the saved intervention effects, so they are exact. IG and GA follow SHAP: their signed means give the Φ columns (lower in the MLP arm, 0.32–0.44 last-token), and mean $|\cdot|$ gives the mean- $|\phi|$ columns.

The Block A MLP gap depends on the candidate sets. The table above rebuilds them from the mean- $|\phi|$ top-20 and reruns the interventions. Rescoring the original candidate sets (built from signed

Table 9: Global faithfulness (Spearman ρ , $k = 114$ steering) for the signed mean Φ and mean $|\phi|$ under different backgrounds. “Boot.” gives the 95% range over bootstrapped 100-row backgrounds (500 draws linear, 40 MLP). The directional score for mean $|\phi|$ is undefined (unsigned).

Arm, scope	Φ imp.	Φ dir.	$ \phi $ imp., 100-row	$ \phi $ imp., full	Φ imp., boot.	$ \phi $ imp., boot.
Linear, all-token	+0.377	-0.097	+0.437	+0.430	[+0.20, +0.42]	[+0.41, +0.45]
Linear, last-token	+0.303	-0.108	+0.461	+0.455	[+0.21, +0.44]	[+0.43, +0.47]
MLP, all-token	+0.347	-0.091	+0.487	+0.476	[+0.23, +0.43]	[+0.46, +0.50]
MLP, last-token	+0.233	-0.105	+0.452	+0.425	[+0.13, +0.39]	[+0.41, +0.45]
<i>Directional ρ of Φ under other backgrounds</i>						
Linear, all-token	full training set -0.058; bootstrap -0.03 [-0.21, +0.17]					
Linear, last-token	full training set -0.153; bootstrap -0.05 [-0.24, +0.17]					
MLP, all-token	3,000 rows -0.101; bootstrap -0.04 [-0.18, +0.12]					
MLP, last-token	3,000 rows -0.123; bootstrap -0.05 [-0.21, +0.11]					

Table 10: Local-minus-global importance gap ($n = 100$ sentences, ablation) with the earlier signed- Φ global baseline and with mean $|\phi|$ (used in the paper; candidate sets rebuilt from its top-20 and the interventions rerun). Crossed-bootstrap p , uncorrected.

Arm	Block	Signed Φ gap (p)	Mean $ \phi $ gap (p)
Linear	A, importance	+0.294 (0.064)	+0.161 (0.206)
Linear	B, importance	-0.305 (0.104)	-0.305 (0.036)
MLP	A, importance	+0.396 (0.024)	+0.389 (0.002)
MLP	B, importance	-0.310 (0.216)	-0.213 (0.290)

Φ) with mean $|\phi|$ and a 100-row background instead gives +0.144 ($p = 0.132$, linear) and +0.222 ($p = 0.054$, MLP) in Block A, so the one local result that survives correction is sensitive to this choice.

C.6 Probe-Side Controls

Linear arm, $k = 114$ steering, $P_j = w_j \alpha_j$ is the probe’s own response to steering feature j (exact under last-token steering, direct-only under all-token), and D_j is the measured logit-difference effect. “Class-matched” uses a background whose class mix equals the evaluation rows (51.5% positive). $w\bar{x}$ is $\mathbb{E}_e[w_j x_{ej}]$ with no background. Intervals resample the 114 features and are descriptive.

Table 11: Spearman ρ of each score against the probe’s own steering effect P (no language model) and against the measured effect D , last- and all-token. Dir. = signed, imp. = magnitude; the unsigned mean $|\phi|$ has no directional score.

Score	vs. P		vs. D , last-token		vs. D , all-token	
	dir.	imp.	dir.	imp.	dir.	imp.
Φ , 100-row background	-0.28	+0.76	-0.11	+0.30	-0.10	+0.38
Φ , full training set	-0.37	+0.78	-0.15	+0.39	-0.06	+0.37
Φ , class-matched	-0.37	+0.78	-0.15	+0.39	-0.06	+0.37
$w\bar{x}$ (no background)	+0.95	+0.81	+0.35	+0.47	+0.41	+0.44
Probe weight w	+0.99	+0.95	+0.31	+0.27	+0.40	+0.29
Mean $ \phi $, full set	-	+0.89	-	+0.46	-	+0.43
P itself (probe $\rightarrow$ LM)	-	-	+0.31 [+0.10, +0.49]	+0.40 [+0.23, +0.55]	+0.39 [+0.20, +0.56]	+0.38 [+0.20, +0.54]

Matching the class mix leaves the sign agreement of Φ with the weight essentially unchanged (43.0% of the pool against 46.5% for the canonical background), as expected since the evaluation and training sets are balanced to within 0.1 points. Over 500 random 100-row backgrounds, with or without stratification to the evaluation

class mix, directional ρ against D is centred near zero (about -0.05 , range roughly $[-0.26, +0.17]$).

Ablating all 114 features gives the same picture: the probe’s ablation effect $-\mathbb{E}_e[w_j x_{ej}]$ predicts the measured effect with directional $\rho \approx +0.34$ in both scopes, and the raw weight (+0.29 to +0.31) beats Φ (negative). The directional gap between the raw weight and Φ (100-row background) is +0.41 [+0.09, +0.72] last-token and +0.50 [+0.17, +0.79] all-token under steering, and similar under ablation (5,000 paired feature resamples). The earlier $k = 20$ ablation estimate (+0.70, $n = 20$) was a small-sample overestimate and is superseded. The probe-to-LM correlation is not a strict ceiling: scores such as $w\bar{x}$ and mean $|\phi|$ also use the activation distribution and can exceed it.

D Minimal-Pair Stimuli

remove pairs take held-out sentences containing *exactly one* word from a hand-curated sentiment lexicon and delete it, leaving the rest byte-identical (with two sentiment words, the edit would no longer have a single cause). insert pairs append “it was (word).” to lexicon-free sentences. substitute pairs swap the lexicon word for a random word of opposite polarity. insert-interior pairs place “it was (word),” at the midpoint of a lexicon-free sentence, so the final token and its context are unchanged. The type/position confound thus concerns distance from the final token, not a replaced token. Pools hold 411–886 pairs per edit, from which 100 are drawn.

Changed-feature threshold. A candidate counts as changed if $|\Delta a_j| \geq 0.10 \max_i |\Delta a_i|$ within the pair. Local-SHAP AUROC is stable across relative thresholds 0.05–0.50 for insertion (0.62–0.68) but not for removal (0.55 at 0.10, 0.40 at 0.50). A threshold-free per-pair Spearman of $|\text{score}|$ against $|\Delta a|$ gives local SHAP 0.04 (removal) and 0.26 (insertion), against 0.57 and 0.67 for activation magnitude at the edited last token.

Readout noise floor. Null edits (a neutral word deleted or appended) give a median $|\Delta|$ of 0.14 (removal-type) and 0.23 (insertion-type) logits, against mean effects of 0.28 and 0.77. Restricting to pairs above the null 90th percentile raises sign accuracy (0.85 removal, 0.94 insertion) and leaves AUROC and ρ within 0.06 of the full-sample values, with the same method ordering.

Seeds. Seeds 0–3 redraw pairs and distractors from the same pool, so they overlap and are descriptive. Local-SHAP AUROC varies by at most 0.03 (SD) per edit, and no SHAP run exceeds 0.5 under activity-matched distractors.

E Supplementary Tables and Figures

E.1 Linear-Arm Identity Check and MLP-Arm Exact p -Values

Closed-form identity check (linear probe, margin target): $\max |IG - \Phi| = \max |GA - \Phi| = 1.72 \times 10^{-7}$. IG, GA and signed-mean SHAP Φ coincide analytically for a linear probe, giving pairwise $\rho = 1.000$ ($n = 615$ non-zero-coefficient features) as an identity, not an empirical result. The identity also holds under mean- $|\cdot|$ aggregation, where the probe weights correlate with all three at $\rho = 0.39$. Aggregating IG and GA

as $|\text{signed mean}|$ against $\text{mean-}|\phi|$ SHAP would give only $\rho = 0.81$, an aggregation effect and not a difference between methods.

Table 12: MLP-arm attribution agreement (Spearman ρ on $|\text{score}|$, $n = 114$; SHAP, IG and GA are all mean $|\cdot|$ against the full training set).

Pair	ρ	p
Probe vs. IG	0.336	0.0003
Probe vs. SHAP	0.334	0.0003
Probe vs. GA	0.314	0.0007
IG vs. SHAP	0.995	< 0.0001
IG vs. GA	0.980	< 0.0001
SHAP vs. GA	0.984	< 0.0001

E.2 Selection-Bias and Sign-Stratified Checks

Table 13: SHAP faithfulness across candidate-selection strategies (steering), $k = 20$ except all ($k = 114$). top biases SHAP’s ρ downward relative to other methods; top-pos/top-neg test the sign-inversion hypothesis without mixed-sign cancellation.

Arm : kind	random	top	top-pos	top-neg	all ($k=114$)
Linear : imp	+0.636**	+0.496*	+0.293	+0.397	+0.430**
Linear : dir	−0.077	−0.083	+0.268	−0.226	−0.097
MLP : imp	+0.747**	+0.107	+0.322	+0.648**	+0.476**
MLP : dir	−0.147	−0.161	+0.168	−0.376	−0.091

* $p < 0.05$, ** $p < 0.01$, uncorrected; all other cells $p > 0.08$. Importance for random and all is mean $|\phi|$ (full training background). The top, top-pos, and top-neg sets are chosen by signed Φ and keep it as the score: scoring a Φ -selected set with a different score would remove the range restriction they test. With signed Φ the random and all cells are +0.555 and +0.377 (linear) and +0.744 and +0.347 (MLP).

E.3 Alpha-Mode Robustness (Scaled vs. Constant)

Table 14: Alpha-mode sensitivity: scaled vs. constant per-feature α (steering, random selection, $k = 20$).

Arm	Method	ρ_{imp} (scaled / constant)	ρ_{dir} (scaled / constant)
Linear	SHAP/IG/GA	+0.636 / +0.290	−0.077 / +0.072
Linear	Probe	+0.519 / +0.226	+0.624 / +0.567
MLP	SHAP	+0.747 / +0.486	−0.147 / −0.011
MLP	IG	+0.764 / +0.508	−0.146 / +0.000
MLP	GA	+0.707 / +0.406	−0.171 / −0.032
MLP	Probe saliency	+0.400 / +0.365	n/a

Constant- α importance is markedly lower, consistent with its numbers partly reflecting activation scale; we report both modes.

E.4 Multi-Seed Replication

The sweep is descriptive. Seeds draw overlapping subsets of one 114-feature pool and reuse the same per-feature effects, so they are not independent replicates; inference comes from the all run (Appendix C.2). Each per-seed test uses only $n = 20$ features, and most do not survive a pooled BH correction (8/39 distinct linear tests, 3/70 MLP). Mean importance ρ is positive in every method-arm cell and SHAP, IG and GA directional ρ is centred near zero, matching the all run.

Table 15: Multi-seed faithfulness sweep (10 seeds, $k = 20$, scaled α). Descriptive only. Ref p is the fraction of seeds with $|\rho|$ at least that of the published seed-0 draw.

Method	Type	Mean ρ	SD	Min	Max	Frac. > 0	Ref p
<i>Linear arm</i>							
SHAP/IG/GA	imp	+0.355	0.282	−0.260	+0.636	0.90	0.100
SHAP/IG/GA	dir	−0.047	0.160	−0.356	+0.141	0.50	0.700
Probe	imp	+0.236	0.268	−0.260	+0.582	0.80	0.200
Probe	dir	+0.422	0.162	+0.162	+0.624	1.00	0.100
<i>MLP arm</i>							
SHAP	imp	+0.383	0.262	−0.171	+0.747	0.90	0.100
SHAP	dir	−0.005	0.191	−0.377	+0.202	0.50	0.700
IG	imp	+0.410	0.266	−0.159	+0.764	0.90	0.100
GA	imp	+0.405	0.232	−0.062	+0.707	0.90	0.100

Imp: importance; dir: directional. Pooled BH (descriptive): 8/39 distinct tests survive (linear), 3/70 (MLP), at FDR 0.05.

E.5 Single-draw $k = 20$ causal validation results

Continuous (partial-Spearman) activity control. We compute a partial Spearman of $|\text{attribution}|$ vs. $|\Delta|$ controlling continuously for $|\text{activation}|$ (rather than the binary active/inactive split of Block B), keeping the whole candidate set. Local SHAP’s mean partial ρ is near zero (−0.002 linear, +0.096 MLP), versus −0.144 and −0.117 for the global score.

E.6 Lexical Simplicity Control

This analysis uses the full 1,436-row held-out split, not the 1000-row canonical eval sample used elsewhere.

The best validation-selected bag-of-words (TF-IDF) is within 0.03 accuracy of the probe, so by our pre-fixed rule the probe is no better than word presence on this split. On average 24.3 filtered features are active per example, with mean pairwise Jaccard of the top-20 features 0.223.

Word-detector tests. Held-out R^2 of a ridge regression from bag-of-words to feature activation has median 0.171 for the top-20 $|\Phi|$ features with full vocabulary (0.106 for the other masked features) and 0.003 with the lexicon only; four of the top 20 reach $R^2 \geq 0.25$. Full bag-of-words explains 0.389 of the probe logit’s variance (0.148 with the lexicon only), a lower bound on reducibility given the linear fit. Lexicon correlations are small for the top-20 features (mean $|r| \leq 0.07$, including for lexicon words in the last three tokens), and sign-agreement of Φ with the net lexicon score is near zero (−0.006). The lexicon explains almost none of any feature’s variance, so the earlier lexicon-presence test could not have revealed word-detection.

Table 16: Single-draw causal faithfulness ($k = 20$, steering and ablation, random selection, $n_{\text{eval}} = 1000$). Cells show Spearman ρ with uncorrected p in parentheses; at $n = 20$ these are descriptive and not used for inference. For the linear arm, SHAP, IG and GA are analytically identical and listed once. SHAP, IG and GA importance are mean absolute attribution; direction is the signed mean (Φ for SHAP).

Method	$\rho_{\text{imp}} (p)$	$\rho_{\text{dir}} (p)$
<i>Linear arm, steering</i>		
SHAP/IG/GA	+0.636 (0.003)	−0.077 (0.748)
Probe	+0.519 (0.019)	+0.624 (0.003)
<i>Linear arm, ablation</i>		
SHAP/IG/GA	+0.535 (0.015)	−0.427 (0.060)
Probe	+0.347 (0.133)	+0.648 (0.002)
<i>MLP arm, steering</i>		
SHAP	+0.747 (0.0002)	−0.147 (0.535)
Probe saliency	+0.400 (0.081)	n/a
IG	+0.764 (0.0001)	−0.146 (0.540)
GA	+0.707 (0.0005)	−0.171 (0.470)
<i>MLP arm, ablation</i>		
SHAP	+0.525 (0.018)	−0.408 (0.075)
Probe saliency	+0.281 (0.230)	n/a
IG	+0.513 (0.021)	−0.421 (0.065)
GA	+0.552 (0.012)	−0.453 (0.045)

Table 17: Lexical simplicity control (held-out, $n = 1,436$). Models are fit on train with regularisation chosen on validation. Diff. is probe minus baseline accuracy with a paired-bootstrap 95% interval; p is McNemar.

Model	Features	Accuracy	Diff. [95% CI]	p
SAE probe (held-out)	114 masked / 32768	0.810	—	—
Lexicon-count LR	158	0.611	+0.198 [+0.166, +0.229]	$< 10^{-30}$
Unigram BoW LR (binary)	13,605	0.794	+0.016 [−0.010, +0.041]	0.26
Unigram+bigram BoW LR	79,812	0.800	+0.010 [−0.015, +0.034]	0.49
TF-IDF unigram+bigram LR	79,812	0.806	+0.003 [−0.022, +0.028]	0.83
Dense layer-7 residual LR	768	0.870	−0.060 [−0.080, −0.040]	$< 10^{-8}$

E.7 All-token global faithfulness

Table 18: All-token global faithfulness ($k = 114$, steering, all filtered features). Cells show Spearman ρ with BH-corrected q in parentheses.

Arm	Method	$\rho_{\text{imp}} (q)$	$\rho_{\text{dir}} (q)$
Linear	SHAP/IG/GA	+0.430 (< 0.0001)*	−0.097 (0.303)
Linear	Probe	+0.289 (0.003)*	+0.399 (0.0001)*
MLP	SHAP	+0.476 (< 0.0001)*	−0.091 (0.473)
MLP	Probe saliency	+0.122 (0.343)	n/a
MLP	IG	+0.483 (< 0.0001)*	−0.058 (0.630)
MLP	GA	+0.500 (< 0.0001)*	−0.015 (0.875)

* Survives FDR 0.05. Linear arm: 5/8 tests survive (SHAP, IG and GA counted separately; 3/4 distinct); MLP arm: 3/7. The same tests survive when BH is pooled over all 30 tests (both arms, both intervention scopes).

F Additional Limitations

- (1) **Attribution error and probe-LM mismatch are only partly separable.** The probe-side controls (Appendix C.6) identify the signed- Φ directional failure as attribution error, but they exist only for the linear arm, where the probe’s response is closed form. The probe’s steering effect predicts the language-model effect at $\rho \approx 0.3$ – 0.4 , so a faithfulness ρ in that range cannot distinguish a faithful attribution from a partly faithful one, and we make no claim about unfaithfulness to the language model beyond it.
- (2) **We calibrate α against a different readout than we evaluate it with** (Appendix C.1). We use a layer-11 residual probe’s $P(\text{positive})$ at calibration time, and a GPT-2 logit difference at evaluation time.
- (3) **The MLP arm’s probe saliency score is magnitude-only** ($\|W_1[i, :]\|_2$, no sign). This is valid for importance/ranking comparisons, invalid for any directional or causal-sign comparison.
- (4) **top selection biases SHAP’s faithfulness ρ downward** relative to the other three methods, by range-restricting only SHAP’s scored variance (Appendix C.2). We never present a top number without this caveat.
- (5) **The MLP arm is only weakly nonlinear and its attributions are approximate.** A linear fit to the MLP logit has $R^2 = 0.99$ on the evaluation rows, and IG, GA and SHAP agree at $\rho = 0.98$ – 0.995 , so this arm adds little beyond the linear control. DeepSHAP approximates Shapley values.
- (6) **The MLP probe’s reported validation accuracy is mildly optimistic:** we select its inputs using a val-derived (label-free) frequency filter.