# Geometry-Adaptive Explainer for Faithful Dictionary-Based Interpretability under Distribution Shift

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Mechanistic interpretability aims to explain a model's behavior by identifying causally responsible internal structures. Dictionary-based explainers such as sparse autoencoders and transcoders are a primary tool, but their faithfulness under out-of-distribution (OOD) shift has received little systematic attention. We show that distribution shift rotates the subspace that the model actively uses, misaligning the explainer's dictionary trained on in-distribution (ID) activations. We formalize this misalignment as the **faithfulness gap**, a geometric distance between the ID dictionary and the OOD-active subspace, and show that it controls OOD faithfulness degradation. To reduce this gap, we propose the **Geometry-Adaptive Explainer (GAE)**, which realigns the explainer's dictionary with the OOD-active subspace while preserving the original feature structure. This requires only unlabeled OOD activations and no gradient updates. We prove that GAE improves over the unadapted ID explainer, with excess loss bounded quadratically by the second-moment shift. Empirically, GAE even matches or surpasses all training-based baselines in causal faithfulness across multiple models and OOD settings.[1]

## 1 Introduction

Mechanistic interpretability aims to explain a model's behavior by identifying internal structures that are causally responsible for its outputs [1–3]. A primary approach is to train an explainer, a post-hoc module such as a sparse autoencoder (SAE) [4] or transcoder [5], that decomposes hidden activations into sparse combinations of learned feature directions (a dictionary). These dictionary-based explainers have recently been scaled to large language models [6, 7] and used to uncover interpretable feature circuits [8]. A central requirement for such explanations is faithfulness: they should accurately reflect the computations the model actually uses [9, 1].

When a model encounters out-of-distribution (OOD) inputs, the dictionary learned in-distribution (ID) can no longer capture the directions the model actively uses [10]. Prior work has addressed this vulnerability from two angles, each limited in scope. Attribution-level robustness studies [11–14] focus on input perturbations rather than hidden-state geometry. Dictionary-explainer remedies such as retraining on the model's own generations [15], upweighting rare concepts [16, 17], and adding residual modules [18] remain heuristic, without diagnosing the underlying misalignment. Consequently, the mechanism driving this failure and a principled correction remain unaddressed.

In this work, we identify a geometric mechanism for OOD faithfulness degradation in dictionary-based explainers. These explainers learn their feature directions from ID activations, so their dictionaries reflect the geometric structure of ID hidden representations [19]. This structure is captured by

[1] Code is available at: https://anonymous.4open.science/r/GAE/

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

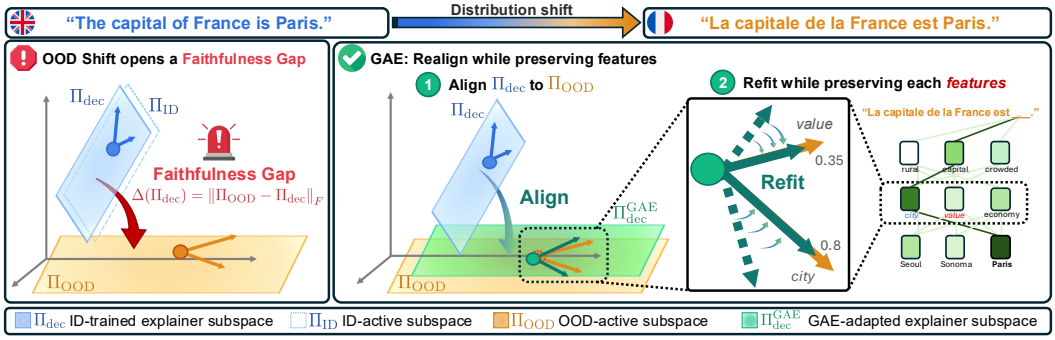



Figure 1: **Faithfulness gap and GAE.** *Left*: distribution shift (illustrated as a language change) rotates the OOD-active subspace $\Pi_{\mathrm{OOD}}$ away from the ID-trained explainer subspace $\Pi_{\mathrm{dec}} \approx \Pi_{\mathrm{ID}}$, opening a faithfulness gap $\Delta(\Pi_{\mathrm{dec}})$. *Right*: GAE closes this gap in two steps. Step 1 rotates $\Pi_{\mathrm{dec}}$ onto $\Pi_{\mathrm{OOD}}$ via orthogonal Procrustes. Step 2 refits individual feature directions within the aligned subspace to match OOD activations while preserving the original feature structure.

the second moment of hidden activations, which distribution shift typically alters [20], leaving the ID-trained dictionary misaligned with the OOD-active subspace. As illustrated in Figure 1 (left), we call this misalignment the **faithfulness gap**, a geometric distance between the ID dictionary and the OOD-active subspace. We prove that this gap controls OOD faithfulness degradation and that it is itself upper-bounded by the magnitude of the second-moment shift. Reducing this gap is therefore necessary for restoring OOD faithfulness, motivating methods that directly realign the dictionary.

We instantiate this idea with the **Geometry-Adaptive Explainer (GAE)**, a closed-form, post-hoc method that closes the faithfulness gap using only unlabeled OOD activations. GAE first rotates the ID-trained dictionary so that its subspace aligns with the OOD-active subspace, choosing the rotation closest to the original dictionary to preserve feature structure (Step 1 in Figure 1). A constrained decoder refit then adjusts individual feature directions to match OOD activations while maintaining this alignment (Step 2 in Figure 1). The entire pipeline requires no gradient computation, yet we prove it is guaranteed to improve over the unadapted explainer (Theorem 1). Empirically, GAE matches or surpasses all training-based baselines in causal faithfulness across multiple language models and OOD settings, including methods that retrain from scratch on OOD data.

Our main contributions are summarized as follows:

- We identify the **faithfulness gap** as a geometric mechanism for OOD faithfulness degradation and prove that excess loss grows at most quadratically with second-moment shift.
- We propose **GAE**, a closed-form dictionary realignment method that targets the faithfulness gap with a theoretical guarantee while preserving feature structure.
- Without any gradient computation, GAE matches or surpasses all training-based baselines, including full OOD retraining, in causal faithfulness across multiple models and diverse OOD settings.

## 2 Related Work

### 2.1 Faithfulness in Mechanistic Interpretability

Dictionary-based explainers such as SAEs [4, 21] and transcoders [5, 22] decompose hidden activations into sparse feature directions and have become a primary tool for mechanistic interpretability [1, 2, 6–8]. Faithfulness is typically evaluated via causal interventions that ablate features and measure the effect on model outputs [9, 23, 24], and is a prerequisite for reliable circuit discovery and model editing [8, 25, 26]. These explainers' dictionaries reflect the geometric structure of ID hidden representations [19], so distribution shift can degrade faithfulness by misaligning the learned directions with those the model actively uses. Recent empirical reports support this concern: SAE-based features underperform dense baselines on OOD downstream tasks [10], yet this vulnerability has received little systematic attention, with no formal diagnosis of its cause.

### 2.2 Interpretability under Distribution Shift

The faithfulness assumption above becomes problematic when the model encounters OOD inputs. Early work showed that saliency maps are fragile under input perturbations [11, 12]. Subsequent studies examined explanation consistency under shift more broadly [14, 13]. These analyses primarily

concern attribution methods and attribute failures to predictive degradation or input-level perturbations, rather than to structural changes in hidden representations. Yet evidence from OOD detection shows that OOD inputs produce statistically distinguishable activation patterns in hidden layers [20]. Representation comparison methods [27–29] enable quantifying such geometric changes across conditions, but the connection to explainer faithfulness has not been drawn.

For dictionary-based explainers, proposed remedies include retraining the explainer on OOD data, training on the model's own generations to avoid external dataset dependence [15], upweighting tail samples during training [16, 17], or adding residual capacity via boosting [18]. However, these approaches either require full retraining or do not directly address the geometric misalignment between the explainer's learned dictionary and the OOD-active subspace. Post-hoc methods that realign the explainer's dictionary with the OOD-active subspace remain unexplored.

# 3 Hidden-Space Geometry Shift and Faithfulness Degradation

When a neural network encounters OOD inputs, do its mechanistic explanations remain faithful? We show that they generally do not. Distribution shift alters the geometry of hidden activations, creating a misalignment between the directions the model actively uses and those the explainer was trained to reconstruct. We call this misalignment the faithfulness gap, and prove that **(i)** it grows at most proportionally with the second-moment shift for ID-trained explainers (Proposition 1), and **(ii)** it controls the reducible part of OOD faithfulness loss (Proposition 2).

## 3.1 Setup and Faithfulness Gap

**Target model and explainer.** The target model is a fixed, pretrained neural network whose internal computations we wish to explain. For an input $X$, it produces a hidden representation $h(X) \in \mathbb{R}^d$ at a designated layer. An explainer is a post-hoc module that decomposes hidden representations into interpretable components. Among various approaches, dictionary-based explainers such as SAEs and transcoders have become a primary tool for mechanistic interpretability [4, 30]. These methods learn a decoder (dictionary) $W_{\text{dec}} \in \mathbb{R}^{d\times k}$ and an encoder $W_{\text{enc}} \in \mathbb{R}^{k\times d}$, where the $k$ columns of $W_{\text{dec}}$ serve as learned feature directions. SAEs reconstruct the hidden activation $h(X)$ itself, while transcoders reconstruct the MLP output at the same layer; both share the form

$$\hat{h}(X) = W_{\text{dec}}\,\sigma(W_{\text{enc}}\,h(X) + b_{\text{enc}}) + b_{\text{dec}}, \tag{1}$$

where $\sigma$ is a sparsifying nonlinearity (e.g., ReLU or TopK) [21]. The explainer represents hidden activations through a learned dictionary, so its behavior is tied to hidden-space geometry.

**In-distribution, out-of-distribution, and second-moment shift.** We call the data distribution on which the explainer was trained ID, denoted by $P_{\text{ID}}$; the dictionary $W_{\text{dec}}$ is learned from ID activations. At deployment, however, the explainer may encounter OOD inputs from a different distribution $P_{\text{OOD}}$. Our goal is to understand whether the explainer remains faithful under this shift.

We characterize the shift via second-moment matrix $M_e := \mathbb{E}_{X\sim P_e}[h(X)h(X)^\top]$, $e \in \{\text{ID}, \text{OOD}\}$. Dictionary-based explainers minimize reconstruction error whose optimum depends on the eigenstructure of $M_e$ [21], so a shift in $M_e$ changes the directions the explainer should reconstruct. A second-moment shift occurs when $M_{\text{OOD}} \neq M_{\text{ID}}$; we measure its magnitude by $\|M_{\text{OOD}} - M_{\text{ID}}\|_F$.

**Active subspace and faithfulness gap.** Since hidden activations concentrate energy along a few directions [31–33], the top eigenspace of $M_e$ captures most of the model's representational activity [28]. We write $\Pi_e = U_e U_e^\top$, where $U_e \in \mathbb{R}^{d\times r}$ contains the top-$r$ eigenvectors of $M_e$, and call this the active subspace in environment $e$. Every reconstruction lies in the column space of $W_{\text{dec}}$, but the reconstruction energy concentrates along the top-$r$ left singular directions [21]. Similarly, $\Pi_{\text{dec}} = U_{\text{dec}} U_{\text{dec}}^\top$, where $U_{\text{dec}} \in \mathbb{R}^{d\times r}$ contains the top-$r$ left singular vectors of $W_{\text{dec}}$, is the explainer subspace. OOD faithfulness depends on how well $\Pi_{\text{dec}}$ aligns with $\Pi_{\text{OOD}}$.

Under second-moment shift, $\Pi_{\text{OOD}}$ may diverge from $\Pi_{\text{ID}}$, opening a gap between the explainer subspace and OOD-active subspace. We call this misalignment the faithfulness gap, which tightly controls how much faithfulness degrades under OOD, as we formalize in Proposition 2:

**Definition 1** (Faithfulness Gap)**.** *The faithfulness gap of an explainer subspace* $\Pi_{\text{dec}}$ *under OOD is*

$$\Delta(\Pi_{\text{dec}}) := \|\Pi_{\text{OOD}} - \Pi_{\text{dec}}\|_F.$$

A large gap means the explainer is reconstructing along directions that the model no longer uses, and the second-moment shift directly upper-bounds this gap for ID-trained explainers (Proposition 1).

### 3.2 Second-Moment Shift Enlarges the Faithfulness Gap

The previous subsection defined the faithfulness gap as a geometric quantity. We now show that second-moment shift directly enlarges this gap for ID-trained explainers. In practice, explainers are trained on ID activations and deployed without modification [30, 34]. Since a well-trained ID explainer satisfies $\Pi_{\mathrm{dec}} \approx \Pi_{\mathrm{ID}}$ (empirically validated in Appendix B.2, Table 4), its OOD faithfulness depends on how far $\Pi_{\mathrm{ID}}$ lies from $\Pi_{\mathrm{OOD}}$. The following result, a consequence of the Davis–Kahan $\sin\Theta$ theorem [35], bounds this distance in terms of the second-moment shift.

**Proposition 1** (Second-Moment Shift Bounds the Faithfulness Gap of the ID Explainer)**.** *Suppose that $M_{\mathrm{ID}}$ has eigengap $\gamma_{\mathrm{ID}} = \lambda_r(M_{\mathrm{ID}}) - \lambda_{r+1}(M_{\mathrm{ID}}) > 0$ at rank $r$. Then*

$$\Delta(\Pi_{\mathrm{ID}}) = \|\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{ID}}\|_F \leq \frac{\sqrt{2}}{\gamma_{\mathrm{ID}}} \|M_{\mathrm{OOD}} - M_{\mathrm{ID}}\|_F.$$

The faithfulness gap of the ID explainer grows at most proportionally with the second-moment shift $\|M_{\mathrm{OOD}} - M_{\mathrm{ID}}\|_F$, with sensitivity controlled by the inverse eigengap $1/\gamma_{\mathrm{ID}}$ (proof in Appendix A.1; empirical verification in Appendix B.4). This establishes that the faithfulness gap can grow large under distribution shift. The next question is whether reducing it improves faithfulness.

### 3.3 The Faithfulness Gap Controls OOD Degradation

We now formalize the connection between the faithfulness gap and OOD faithfulness loss, showing that $\Delta(\Pi_{\mathrm{dec}})$ is the central quantity an adaptation method should target.

**OOD faithfulness objective.** An ideal explainer subspace $\Pi_{\mathrm{dec}}$ minimizes reconstruction error on OOD activations. We formalize this objective as

$$\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}}) := \mathbb{E}_{X \sim P_{\mathrm{OOD}}} \|h(X) - \Pi_{\mathrm{dec}} h(X)\|_2^2. \tag{2}$$

This measures how much OOD activation is lost when projected onto $\Pi_{\mathrm{dec}}$. $\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}})$ is a valid surrogate since hidden-layer reconstruction constrains logit-level faithfulness [23, 24].

**Decomposition of $\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}})$.** To isolate the part of the OOD loss that the explainer can reduce, we decompose $\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}})$ into two terms. Let $\mathcal{C}_r$ denote the set of rank-$r$ orthogonal projectors in $\mathbb{R}^d$. For any $\Pi_{\mathrm{dec}} \in \mathcal{C}_r$,

$$\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}}) = \underbrace{\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}})}_{\text{irreducible}} + \underbrace{\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}}) - \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}})}_{\text{explainer-dependent}}. \tag{3}$$

The explainer-dependent component is nonnegative (proof in Appendix A.2), so $\Pi_{\mathrm{OOD}}$ minimizes $\mathcal{L}_{\mathrm{OOD}}$ over $\mathcal{C}_r$. The irreducible component depends on the target model and the OOD distribution, both fixed at deployment; adapting the explainer cannot reduce it. *The explainer-dependent component is the only part the explainer can reduce.*

**Faithfulness gap as a tight proxy.** The faithfulness gap $\Delta(\Pi_{\mathrm{dec}})$ measures the distance between two subspaces, making it a direct optimization target. The next proposition shows that controlling $\Delta(\Pi_{\mathrm{dec}})$ is equivalent to controlling the explainer-dependent component.

**Proposition 2** (Faithfulness Gap Controls the Explainer-Dependent Term)**.** *Assume that the OOD eigengap at rank $r$, $\gamma_{\mathrm{OOD}} = \lambda_r(M_{\mathrm{OOD}}) - \lambda_{r+1}(M_{\mathrm{OOD}}) > 0$. Then for any $\Pi_{\mathrm{dec}} \in \mathcal{C}_r$,*

$$\frac{\gamma_{\mathrm{OOD}}}{2} \Delta(\Pi_{\mathrm{dec}})^2 \leq \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}}) - \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}}) \leq \frac{\lambda_1(M_{\mathrm{OOD}}) - \lambda_d(M_{\mathrm{OOD}})}{2} \Delta(\Pi_{\mathrm{dec}})^2.$$

The lower bound shows that any nonzero gap incurs a positive cost (misalignment cannot be free); the upper bound shows that reducing $\Delta(\Pi_{\mathrm{dec}})$ is sufficient. Together, they establish that controlling $\Delta(\Pi_{\mathrm{dec}})$ is equivalent to controlling the explainer-dependent faithfulness loss (proof in Appendix A.3; empirical verification in Appendix B.3). *Reducing $\Delta(\Pi_{\mathrm{dec}})$ is therefore both necessary and sufficient for improving OOD faithfulness.* Section 4 introduces a method that directly targets this quantity.

## 4 Geometry-Adaptive Explainer (GAE)

Section 3 showed that OOD faithfulness degradation is controlled by the faithfulness gap $\Delta(\Pi_{\mathrm{dec}})$. We now propose the Geometry-Adaptive Explainer (GAE), which reduces $\Delta(\Pi_{\mathrm{dec}})$ by realigning the explainer's dictionary with the OOD-active subspace while preserving the original feature structure. We present an objective for this adaptation and a closed-form solution.

### 4.1 Problem Formulation

Section 3.2 showed that an ID-trained explainer's faithfulness degrades under OOD because its subspace diverges from $\Pi_{\mathrm{OOD}}$. Minimizing $\Delta(\Pi_{\mathrm{dec}})$ amounts to choosing a $W_{\mathrm{dec}}$ whose induced subspace aligns with $\Pi_{\mathrm{OOD}}$. To restore faithfulness, we adapt the existing dictionary $W_{\mathrm{dec}}^{\mathrm{ID}} \in \mathbb{R}^{d\times k}$ using a set of unlabeled OOD activations $\{h_i\}_{i=1}^N$, from which we estimate the OOD-active subspace $\widehat{\Pi}_{\mathrm{OOD}}$. We seek $W_{\mathrm{dec}}$ that closes the faithfulness gap while preserving the original feature structure:

$$\min_{W_{\mathrm{dec}}} \underbrace{\mathcal{L}_{\mathrm{recon}}}_{\text{reconstruction}} + \lambda_{\mathrm{geom}} \underbrace{\|\widehat{\Pi}_{\mathrm{OOD}} - \Pi_{\mathrm{dec}}\|_F^2}_{\text{subspace alignment}} + \lambda_{\mathrm{pres}} \underbrace{\|W_{\mathrm{dec}} - W_{\mathrm{dec}}^{\mathrm{ID}}\|_F^2}_{\text{feature preservation}}, \tag{4}$$

where $\mathcal{L}_{\mathrm{recon}} = \frac{1}{N}\sum_{i=1}^N \|h_i - \hat{h}_i\|_2^2$ is the mean reconstruction error.[2] GAE holds the encoder fixed to preserve the learned feature decomposition [4]. The three terms serve complementary roles. The *reconstruction* term fits individual OOD activations, since subspace alignment alone does not guarantee sample-level reconstruction. The *subspace alignment* term directly targets the faithfulness gap $\Delta(\Pi_{\mathrm{dec}})$, which Proposition 2 showed controls the explainer-dependent component. The *feature preservation* term keeps the adapted dictionary close to the original, maintaining the encoder-decoder pairing for downstream circuit analyses [8].

### 4.2 Dictionary Adaptation

Eq. (4) is non-convex in $W_{\mathrm{dec}}$, as $\Pi_{\mathrm{dec}}$ depends on it through a top-$r$ SVD. However, the subspace alignment term can be enforced as a hard constraint, and once the subspace is fixed, the remaining terms are quadratic with a closed-form solution. GAE exploits this in two steps (Algorithm 1): Step 1 enforces *subspace alignment* by rotating the ID-trained dictionary onto $\widehat{\Pi}_{\mathrm{OOD}}$, choosing the rotation closest to the original. Step 2 refits the decoder via constrained ridge regression, solving the *reconstruction* and *feature preservation* terms while preserving this alignment.

**Algorithm 1** GAE

**Require:** $W_{\mathrm{dec}}^{\mathrm{ID}} \in \mathbb{R}^{d\times k}$; $W_{\mathrm{enc}}$, $b_{\mathrm{enc}}$; OOD activations $\{h_i\}_{i=1}^N$; $\lambda_{\mathrm{geom}}$, $\lambda_{\mathrm{pres}}$
**Ensure:** $(W_{\mathrm{dec}}^{\mathrm{GAE}}, b_{\mathrm{dec}}^{\mathrm{GAE}})$

$\widehat{M}_{\mathrm{OOD}} \leftarrow \frac{1}{N}\sum_{i=1}^N h_i h_i^\top$
$U_{\mathrm{dec}} \leftarrow$ top-$r$ left singular vectors of $W_{\mathrm{dec}}^{\mathrm{ID}}$; $U_{\mathrm{OOD}}^{(:r)} \leftarrow$ top-$r$ eigenvectors of $\widehat{M}_{\mathrm{OOD}}$
$G \leftarrow U_{\mathrm{dec}}^\top W_{\mathrm{dec}}^{\mathrm{ID}} (W_{\mathrm{dec}}^{\mathrm{ID}})^\top U_{\mathrm{OOD}}^{(:r)}$; $T^\star \leftarrow \widetilde{V}\widetilde{U}^\top$ from SVD$(G)$
$\widetilde{W}_{\mathrm{dec}} \leftarrow U_{\mathrm{OOD}}^{(:r)} T^\star U_{\mathrm{dec}}^\top W_{\mathrm{dec}}^{\mathrm{ID}}$ ▷ Step 1: Subspace rotation
$z_i \leftarrow \sigma(W_{\mathrm{enc}} h_i + b_{\mathrm{enc}})$ for each $h_i$ ▷ frozen encoder
$W_{\mathrm{dec}}^{\mathrm{GAE}} \leftarrow$ Eq. (8); $b_{\mathrm{dec}}^{\mathrm{GAE}} \leftarrow$ Eq. (9) ▷ Step 2: Constrained decoder refit

**Step 1: Subspace rotation.** Let $U_{\mathrm{dec}} \in \mathbb{R}^{d\times r}$ be the top-$r$ left singular vectors of $W_{\mathrm{dec}}^{\mathrm{ID}}$ (the explainer subspace defined in Section 3.1), and let $U_{\mathrm{OOD}}^{(:r)}$ be the top-$r$ eigenvectors of the empirical second-moment matrix $\widehat{M}_{\mathrm{OOD}} = \frac{1}{N}\sum_{i=1}^N h_i h_i^\top$. We constrain the rotated dictionary to the form

$$\widetilde{W}_{\mathrm{dec}}(T) \;=\; U_{\mathrm{OOD}}^{(:r)}\, T\, U_{\mathrm{dec}}^\top\, W_{\mathrm{dec}}^{\mathrm{ID}}, \qquad T \in \mathcal{O}_r, \tag{5}$$

where $\mathcal{O}_r = \{T \in \mathbb{R}^{r\times r} : T^\top T = I\}$. Every column of $\widetilde{W}_{\mathrm{dec}}(T)$ lies in $\mathrm{span}(U_{\mathrm{OOD}}^{(:r)})$, so the column space of $\widetilde{W}_{\mathrm{dec}}(T)$ is contained in this $r$-dimensional subspace. Since $\widetilde{W}_{\mathrm{dec}}(T)$ has rank $r$, its induced explainer subspace equals $\widehat{\Pi}_{\mathrm{OOD}}$ exactly, and *the faithfulness gap vanishes ($\Delta(\Pi_{\mathrm{dec}}) = 0$) for any $T \in \mathcal{O}_r$*. Among all rotations that achieve this alignment, we select the one that keeps the rotated dictionary closest to the original:

$$T^\star \;=\; \arg\min_{T\in\mathcal{O}_r} \big\|\widetilde{W}_{\mathrm{dec}}(T) - W_{\mathrm{dec}}^{\mathrm{ID}}\big\|_F^2. \tag{6}$$

This is an orthogonal Procrustes problem [37]. Let $G = U_{\mathrm{dec}}^\top W_{\mathrm{dec}}^{\mathrm{ID}} (W_{\mathrm{dec}}^{\mathrm{ID}})^\top U_{\mathrm{OOD}}^{(:r)} \in \mathbb{R}^{r\times r}$, with SVD $G = \widetilde{U}\Sigma\widetilde{V}^\top$. Then $T^\star = \widetilde{V}\widetilde{U}^\top$ (derivation in Appendix A.5). Since $\Pi_{\mathrm{dec}}^{\mathrm{GAE}} := \widehat{\Pi}_{\mathrm{OOD}}$ by construction, the residual faithfulness gap reduces to the eigenspace estimation error. This yields a quantitative improvement over the unadapted ID explainer.

[2] The full dictionary-based explainer objective (Eq. (1)) includes a sparsity penalty on $\mathbf{z}_i = \sigma(W_{\mathrm{enc}} h_i + b_{\mathrm{enc}})$. GAE holds the encoder fixed, so $\mathbf{z}_i$ is constant with respect to $W_{\mathrm{dec}}$ and the sparsity term drops out, leaving only the reconstruction term. This also means GAE applies regardless of sparsity mechanism ($\ell_1$, Top-K [7], JumpReLU [36], etc.).

**Theorem 1** (Improvement over ID Explainer). *Suppose* $\gamma_{\mathrm{OOD}} := \lambda_r(M_{\mathrm{OOD}}) - \lambda_{r+1}(M_{\mathrm{OOD}}) > 0$, $\Delta(\Pi_{\mathrm{ID}}) > 0$, *and* $\widehat{\Pi}_{\mathrm{OOD}} \approx \Pi_{\mathrm{OOD}}$ *(holds with sufficient OOD samples). Then*

$$\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}}^{\mathrm{GAE}}) \;\le\; \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{ID}}) \;-\; \frac{\gamma_{\mathrm{OOD}}}{2}\,\Delta(\Pi_{\mathrm{ID}})^2. \tag{7}$$

The improvement grows quadratically with the ID explainer's misalignment, so the more severe the shift, the larger the guaranteed gain. The proof is given in Appendix A.7.

**Step 2: Constrained decoder refit.** Step 1 aligns the subspace but does not optimize sample-level reconstruction. Step 2 refits $\mathcal{L}_{\mathrm{recon}}$ while preserving the alignment from Step 1. Since the encoder is fixed, the feature activations $z_i = \sigma(W_{\mathrm{enc}}\, h_i + b_{\mathrm{enc}})$ for each OOD sample $h_i$ are constants, and reconstruction reduces to a linear least-squares problem in $W_{\mathrm{dec}}$ and $b_{\mathrm{dec}}$. To keep the decoder geometrically aligned, we penalize decoder mass outside $\widehat{\Pi}_{\mathrm{OOD}}$ with $\lambda_{\mathrm{geom}}\|(I - \widehat{\Pi}_{\mathrm{OOD}})\, W_{\mathrm{dec}}\|_F^2$. To preserve the feature structure from Step 1, we regularize toward $\widetilde{W}_{\mathrm{dec}}(T^\star)$ with $\lambda_{\mathrm{pres}}\|W_{\mathrm{dec}} - \widetilde{W}_{\mathrm{dec}}(T^\star)\|_F^2$. The combined objective is convex and quadratic, yielding the closed-form solution

$$W_{\mathrm{dec}}^{\mathrm{GAE}} \;=\; \widehat{\Pi}_{\mathrm{OOD}}\, C\, B^{-1} \;+\; (I - \widehat{\Pi}_{\mathrm{OOD}})\, C\, (B + \lambda_{\mathrm{geom}}\, I)^{-1}, \tag{8}$$

$$b_{\mathrm{dec}}^{\mathrm{GAE}} \;=\; \tfrac{1}{N}\textstyle\sum_i h_i \;-\; W_{\mathrm{dec}}^{\mathrm{GAE}}\; \tfrac{1}{N}\textstyle\sum_i z_i, \tag{9}$$

where

$$B = \tfrac{1}{N}\textstyle\sum_i z_i z_i^\top \;-\; \big(\tfrac{1}{N}\textstyle\sum_i z_i\big)\big(\tfrac{1}{N}\textstyle\sum_i z_i\big)^\top \;+\; \lambda_{\mathrm{pres}}\, I, \tag{10}$$

$$C = \tfrac{1}{N}\textstyle\sum_i h_i z_i^\top \;-\; \big(\tfrac{1}{N}\textstyle\sum_i h_i\big)\big(\tfrac{1}{N}\textstyle\sum_i z_i\big)^\top \;+\; \lambda_{\mathrm{pres}}\, \widetilde{W}_{\mathrm{dec}}(T^\star). \tag{11}$$

To summarize, GAE works by first rotating the ID dictionary so that its column space coincides with the OOD-active subspace (Step 1), then refitting the decoder via a closed-form ridge regression that matches sample-level reconstruction while preserving this alignment (Step 2). The Step 2 solution applies regularization strength $\lambda_{\mathrm{pres}}$ to decoder mass inside the OOD-active subspace and the larger strength $\lambda_{\mathrm{pres}} + \lambda_{\mathrm{geom}}$ outside it, so *any decoder mass that drifts off the Step 1 alignment is automatically shrunk back*. The full derivation is in Appendix A.6.

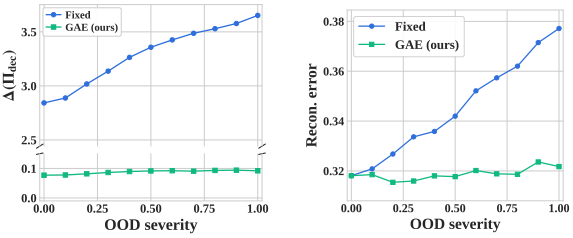
**Applying GAE at inference.** At inference, the ID-trained encoder remains unchanged: given an OOD activation $h$, the explainer extracts features $z = \sigma(W_{\mathrm{enc}}\, h + b_{\mathrm{enc}})$ and reconstructs $\hat{h} = W_{\mathrm{dec}}^{\mathrm{GAE}}\, z + b_{\mathrm{dec}}^{\mathrm{GAE}}$. This applies identically to both SAEs and transcoders.

# 5 Experiments

We evaluate whether GAE restores explanation faithfulness under distribution shift. Section 5.1 tests the geometric mechanism in a controlled setting, and Section 5.2 evaluates on language models.

## 5.1 Controlled Experiment

We first test whether the geometric mechanism from Section 3 holds in a controlled setting. We train a 2-layer ReLU MLP with hidden dim $d = 256$ and output dim $p = 8$, and a linear-decoder SAE on its ID hidden activations, then continuously increase OOD severity by rotating and rescaling the input covariance (details in Appendix B.1). Figure 2 confirms that as severity increases, the faithfulness gap and reconstruction error of the Fixed explainer (the ID-trained explainer, used without adaptation) grow, while GAE closes the gap to near zero and keeps reconstruction error nearly flat.




(a) Faithfulness gap. (b) Reconstruction error.

Figure 2: **Controlled experiment on a toy MLP with OOD severity varied from 0 (ID) to 1 (maximum shift).** (a) The Fixed explainer's faithfulness gap $\Delta(\Pi_{\mathrm{dec}})$ grows monotonically. (b) Its reconstruction error rises accordingly. GAE maintains near-zero gap and flat error throughout.

## 5.2 Experiments on Language Models

### 5.2.1 Setup

**Target models and explainers.** We evaluate on two frozen pretrained language models: GPT-2 Small [38] and Pythia-1.4B [39]. For each model, we train transcoders [5] and Top-K SAEs [7], both with dictionary size $k{=}32d$, as dictionary-based explainers (Eq. (1)).

**OOD settings.** We consider three categories of distribution shift: **temporal** (FineWeb [40], web text collected after each model's pretraining cutoff), **domain** (Edgar [41], financial filings whose specialized vocabulary and structure differ from general web text), and **adversarial** (HaluEval [42], hallucination-inducing prompts that elicit atypical hidden representations). All three induce measurable second-moment shift in hidden activations, as verified in Appendix B.4.

**Baselines.** We compare training-free and training-based approaches. **Fixed** applies the ID-trained explainer without adaptation. **TERM** [16, 17] trains the ID explainer with tilted ERM to upweight tail samples. Among training-based methods, **Finetune** [43] warm-starts from the ID explainer on OOD activations, **Retrain** trains from scratch on OOD data, **SAEBoost** [18] adds a residual booster on OOD reconstruction residuals, and **FaithfulSAE** [15] retrains on the model's own generations. **GAE (ours)** is training-free: it uses only unlabeled OOD activations with no gradient computation. Detailed descriptions are in Appendix D.2.

As shown in Table 1, all training-based baselines require gradient-based optimization over millions of tokens. Even the lightest, Finetune, processes 5M tokens over several minutes; Retrain, SAEBoost, and FaithfulSAE each consume 100M tokens and take hours on a single GPU. *GAE requires no gradient computation at all*: the entire closed-form pipeline completes in 0.5 s for GPT-2 and 2.9 s for Pythia-1.4B, using only $\sim$2,048 unlabeled OOD activations. This makes GAE practical for on-the-fly adaptation whenever the deployment distribution changes.

Table 1: **Computational cost per method.** Training-based baselines require gradient optimization over millions of tokens. GAE adapts in under 3 s without training.

| Method | Tokens | Wall-clock | |
|---|---|---|---|
| | | GPT-2 | P-1.4B |
| Finetune | 5M | $\sim$2 min | $\sim$12 min |
| Retrain | 100M | $\sim$39 min | $\sim$4 hrs |
| SAEBoost | 100M | $\sim$39 min | $\sim$4 hrs |
| FaithfulSAE | 100M | $\sim$39 min | $\sim$4 hrs |
| **GAE** | **2K** | **0.5 s** | **2.9 s** |

**Evaluation metrics.** We evaluate causal faithfulness using three metrics. **Normalized AOPC (nAOPC)** [23] averages the normalized logit drop across multiple feature budgets when top-$m$ features are removed (↑ is better). **Normalized comprehensiveness (nComp)** [24] measures the normalized logit drop at a single budget $m^*$=32 (↑ is better). Both measure the logit-level effect of ablating top features. **Delta cross-entropy ($\Delta$CE)** [7] measures reconstruction quality: the cross-entropy change when activations are replaced with the explainer's reconstruction ($\approx 0$ is better). GAE optimizes a geometric objective (the faithfulness gap); improvements on these causal metrics confirm that geometric realignment yields faithfulness gains. Formal definitions are in Appendix D.4.

### 5.2.2 Faithfulness Results

Table 2: **Faithfulness under distribution shift (Transcoder, two models $\times$ three OOD settings).** GAE is training-free yet leads in all nine columns on GPT-2 and achieves the best nComp and $|\Delta\mathrm{CE}|$ in 5 of 6 Pythia-1.4B columns. **Bold**: best per column. Underline: second best.

| | FineWeb (Temporal) | | | Edgar (Domain) | | | HaluEval (Adversarial) | | |
|---|---|---|---|---|---|---|---|---|---|
| **Method** | **nAOPC↑** | **nComp↑** | **$\lvert\Delta$CE$\rvert$↓** | **nAOPC↑** | **nComp↑** | **$\lvert\Delta$CE$\rvert$↓** | **nAOPC↑** | **nComp↑** | **$\lvert\Delta$CE$\rvert$↓** |
| ***GPT-2 Small*** | | | | | | | | | |
| Fixed | 0.857 | 1.017 | 0.0281 | 0.975 | 1.025 | 0.0201 | 0.735 | 0.737 | 0.0473 |
| TERM | 0.853 | 0.993 | 0.0283 | 0.964 | 0.999 | 0.0198 | 0.730 | 0.732 | 0.0523 |
| Finetune | 0.856 | 0.984 | 0.0172 | 0.971 | 1.117 | 0.0047 | 0.579 | 0.579 | 0.0105 |
| Retrain | 0.895 | 1.118 | 0.0218 | 0.936 | 1.034 | 0.0015 | 0.521 | 0.528 | 0.2763 |
| SAEBoost | 0.958 | 1.476 | 0.0177 | 0.979 | 1.542 | 0.0072 | 0.840 | 0.921 | 0.0212 |
| FaithfulSAE | 0.936 | 1.186 | 0.0213 | 0.976 | 1.134 | 0.0197 | 0.738 | 0.740 | 0.0586 |
| **GAE (ours)** | **0.960** | **1.494** | **0.0167** | **0.981** | **1.618** | **0.0009** | **0.871** | **0.963** | **0.0014** |
| ***Pythia-1.4B*** | | | | | | | | | |
| Fixed | 0.839 | 1.091 | 0.0278 | 0.725 | 0.828 | 0.0300 | 0.899 | 1.021 | 0.0354 |
| TERM | 0.845 | 1.043 | 0.0271 | 0.895 | 0.981 | 0.0298 | 0.896 | 1.104 | 0.0349 |
| Finetune | 0.859 | 1.249 | **0.0264** | 0.684 | 0.860 | 0.0282 | 0.925 | 1.502 | 0.0280 |
| Retrain | 0.894 | 1.315 | 0.0405 | 0.746 | 0.821 | 0.0329 | 0.894 | 1.393 | 0.0305 |
| SAEBoost | 0.908 | 1.296 | 0.0284 | 0.903 | 1.297 | 0.0296 | 0.965 | 1.530 | 0.0283 |
| FaithfulSAE | 0.858 | 1.056 | 0.0272 | 0.724 | 0.822 | 0.0323 | 0.899 | 1.171 | 0.0307 |
| **GAE (ours)** | **0.915** | **1.354** | 0.0269 | **0.988** | **1.652** | **0.0230** | **0.968** | **1.693** | **0.0276** |

Table 2 reports faithfulness across three OOD settings for GPT-2 Small and Pythia-1.4B (Transcoder). Despite using no gradient updates, *GAE leads on all three metrics for GPT-2 across every OOD*

*setting*, surpassing training-based baselines that consume up to 100M tokens (cf. Table 1). The largest gains appear on adversarial shift, where GAE improves nComp over the strongest training-based baseline SAEBoost by 4.6% (0.963 vs 0.921) and reduces $|\Delta\text{CE}|$ by 93% (0.0014 vs 0.0212). On Pythia-1.4B, GAE achieves the best nAOPC and nComp on all three settings; $|\Delta\text{CE}|$ is best on Edgar and HaluEval but slightly elevated on FineWeb (0.027 vs Finetune's 0.026), where the more diffuse eigenvalue decay weakens the rank-$r$ approximation.

Table 3: **Faithfulness under distribution shift (SAE, two models × three OOD settings).** GAE leads in all nine columns on GPT-2 and leads on nComp and $|\Delta\text{CE}|$ in 5 of 6 Pythia-1.4B columns. **Bold**: best. Underline: second best.

| | FineWeb (Temporal) | | | Edgar (Domain) | | | HaluEval (Adversarial) | | |
|---|---|---|---|---|---|---|---|---|---|
| **Method** | **nAOPC↑** | **nComp↑** | **$\lvert\Delta\text{CE}\rvert$↓** | **nAOPC↑** | **nComp↑** | **$\lvert\Delta\text{CE}\rvert$↓** | **nAOPC↑** | **nComp↑** | **$\lvert\Delta\text{CE}\rvert$↓** |
| ***GPT-2 Small*** | | | | | | | | | |
| Fixed | 0.735 | 0.796 | 0.0185 | 0.650 | 0.667 | 0.0089 | 0.930 | 1.134 | 0.0406 |
| TERM | 0.741 | 0.790 | 0.0183 | 0.655 | 0.682 | 0.0145 | 0.898 | 0.987 | 0.0401 |
| Finetune | 0.725 | 0.802 | 0.0015 | 0.658 | 0.687 | 0.0068 | 0.932 | 1.155 | 0.0218 |
| Retrain | 0.766 | 0.856 | 0.0375 | 0.715 | 0.797 | 0.0065 | 0.930 | 1.276 | 0.0300 |
| SAEBoost | 0.704 | 0.786 | 0.0120 | 0.604 | 0.650 | 0.0098 | 0.909 | 1.243 | 0.0448 |
| FaithfulSAE | 0.725 | 0.760 | 0.0262 | 0.657 | 0.683 | 0.0074 | 0.908 | 1.082 | 0.0278 |
| **GAE (ours)** | **0.768** | **0.871** | **0.0011** | **0.723** | **0.809** | **0.0037** | **0.953** | **1.303** | **0.0017** |
| ***Pythia-1.4B*** | | | | | | | | | |
| Fixed | 0.962 | 1.536 | 0.0216 | 0.953 | 1.690 | 0.0170 | 0.985 | 1.425 | 0.0252 |
| TERM | 0.965 | 1.486 | 0.0237 | 0.959 | 1.787 | 0.0167 | 0.984 | 1.762 | 0.0292 |
| Finetune | 0.963 | 1.618 | 0.0216 | 0.959 | 1.832 | 0.0131 | 0.988 | 1.880 | 0.0174 |
| Retrain | 0.983 | **1.681** | 0.0563 | **0.970** | 1.848 | 0.0261 | **1.000** | 1.855 | 0.0209 |
| SAEBoost | 0.971 | 1.451 | 0.0211 | 0.955 | 1.598 | 0.0152 | **1.000** | 1.830 | 0.0166 |
| FaithfulSAE | 0.982 | 1.642 | 0.0307 | 0.953 | 1.769 | 0.0205 | **1.000** | 1.678 | 0.0514 |
| **GAE (ours)** | **0.985** | 1.677 | **0.0207** | 0.968 | **1.946** | **0.0098** | **1.000** | **1.885** | **0.0163** |

Table 3 reports the same evaluation for Top-K SAEs. On GPT-2, GAE again surpasses every training-based baseline on all nine columns, with the largest margins on Edgar (nComp 0.809 vs Retrain's 0.797) and HaluEval ($|\Delta\text{CE}|$ 0.0017 vs Finetune's 0.0218). On Pythia-1.4B, Retrain is a stronger competitor than for transcoders, taking best nComp on FineWeb (1.681 vs 1.677) and best nAOPC on Edgar (0.970 vs 0.968). GAE still leads on nComp and $|\Delta\text{CE}|$ in 5 of 6 columns and achieves the best $|\Delta\text{CE}|$ across all three settings. The pattern is consistent: *GAE matches or surpasses methods that require orders of magnitude more computation.*

### 5.2.3 Case Study: Circuit Attribution under Distribution Shift

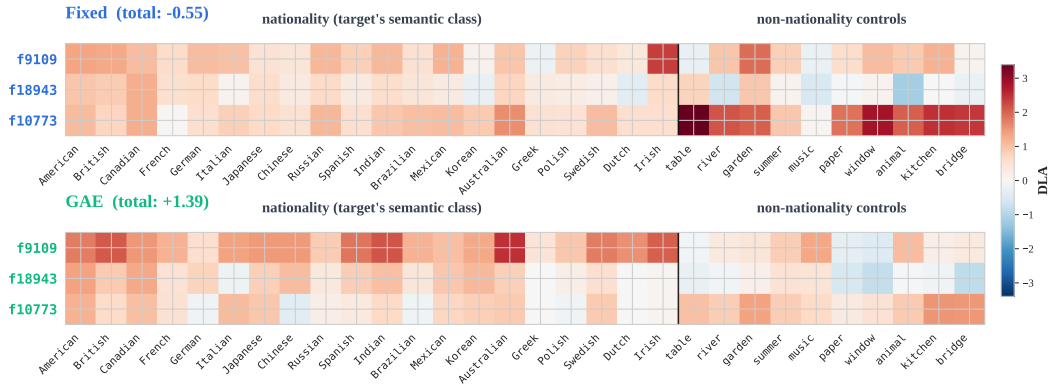


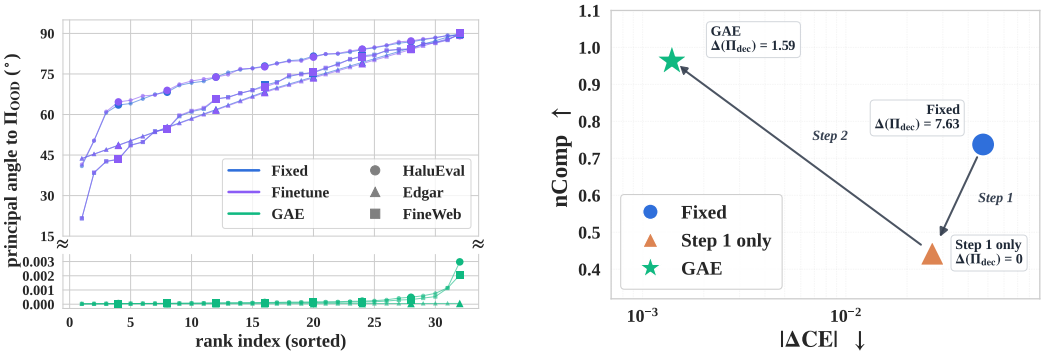

Figure 3: **Per-feature DLA on a prompt predicting ' American' (GPT-2, Transcoder).** Both methods share the same encoder and top-3 features; only the decoder columns differ. Each cell shows a feature's direct logit attribution (DLA) to nationality tokens (left, 20 tokens) vs. non-nationality controls (right, 10 tokens). Fixed's total class-specificity is $-0.55$ (circuit points away from the target class); GAE's is $+1.39$ (circuit points toward it).

We select a prompt where the model predicts ' American' and measure each feature's direct logit attribution (DLA), the dot product of its decoder column with the token's unembedding vector scaled by the feature activation. DLA quantifies how much each feature pushes the model toward a given

next token. Since GAE keeps the encoder frozen, both Fixed and GAE extract the same top-3 features with the same activations, so any change in DLA isolates the effect of the decoder rotation. For each feature, we compute the difference between its mean DLA on 20 nationality tokens (the target's semantic class) and 10 non-nationality controls, then sum across features to obtain a class-specificity score that measures whether the identified circuit points toward the correct token class.

Fixed scores $-0.55$: its circuit points away from the target class on average. GAE scores $+1.39$: every top feature contributes more to nationalities than to controls, as shown in Figure 3. The decoder rotation alone corrects feature-level attribution without altering which features are selected. Appendix E repeats this analysis on two further prompts whose target tokens are a male first name ($+1.00 \rightarrow +4.51$) and a profession ($+0.54 \rightarrow +0.99$).

### 5.2.4 Mechanism Analysis




(a) Subspace alignment.

(b) Step ablation.

Figure 4: **Mechanism analysis (GPT-2, Transcoder). (a)** Sorted principal angles between each explainer's top-$r$ subspace and $\widehat{\Pi}_{\mathrm{OOD}}$. GAE's subspace aligns with $\widehat{\Pi}_{\mathrm{OOD}}$, while Fixed and Finetune leave large angular gaps. **(b)** Step ablation: Step 1 closes the faithfulness gap to 0 yet drops nComp from 0.74 to 0.44. Step 2 restores nComp to 0.96 at the cost of a small gap (1.59).

**Subspace alignment.** Figure 4(a) measures the principal angles between each explainer's top-$r$ decoder subspace and the OOD-active subspace $\widehat{\Pi}_{\mathrm{OOD}}$. Fixed and Finetune leave $40°$–$90°$ angular gaps across all rank indices and OOD shift types, while GAE drives every angle to $\sim 10^{-3}\,°$. This confirms that GAE's faithfulness gains arise from explicit geometric alignment, validating the mechanism of Section 4. The persistence of Finetune's gap further shows that *gradient-based reconstruction loss does not, by itself, drive subspace alignment with* $\widehat{\Pi}_{\mathrm{OOD}}$. Appendix C verifies that the projection-loss improvement scales quadratically with $\Delta(\Pi_{\mathrm{ID}})^2$, as predicted by Theorem 1.

**Step ablation.** Figure 4(b) ablates each step of GAE on HaluEval. Step 1 alone closes the faithfulness gap from 7.63 to 0 by construction, yet nComp drops from 0.74 to 0.44 since the orthogonal rotation diffuses the encoder-decoder feature pairing that the top-$k$ ablation in nComp measures. Step 2 refits the decoder within Step 1's subspace, accepting a small gap (1.59) in exchange for the highest nComp (0.96) and the lowest $|\Delta\mathrm{CE}|$ (0.0014). *The two steps are complementary: Step 1 chooses the subspace, Step 2 makes the dictionary causally coherent within it.* A hyperparameter sensitivity analysis is reported in Appendix F.

## 6 Conclusion

We showed that OOD faithfulness degradation in dictionary-based explainers has a geometric cause: the decoder subspace drifts from the directions the model actively uses. The faithfulness gap $\Delta(\Pi_{\mathrm{dec}})$ formalizes this misalignment and provably controls the reducible part of OOD faithfulness loss. GAE closes the gap with a closed-form subspace rotation and constrained decoder refit, using only unlabeled OOD activations. Across two models and three shift types, GAE outperforms all training-based baselines on 5 of 6 settings, completing in under 3 seconds without any gradient computation. A limitation is that we have not yet evaluated on larger-scale models. GAE also relies on a top-$r$ SVD truncation of $W_{\mathrm{dec}}^{\mathrm{ID}}$, so any feature information carried in the residual $(d-r)$ singular directions is dropped before adaptation. Extending GAE to adaptive rank selection, encoder adaptation, and connections to optimal transport on the Grassmannian [44] are promising future directions.

# References


[1] Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, et al. Causal abstraction: A theoretical foundation for mechanistic interpretability. Journal of Machine Learning Research, 26(83):1–64, 2025.

[2] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. arXiv preprint arXiv:2209.11895, 2022.

[3] Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, et al. Open problems in mechanistic interpretability. arXiv preprint arXiv:2501.16496, 2025.

[4] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, 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. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

[5] Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable llm feature circuits. Advances in Neural Information Processing Systems, 37:24375–24410, 2024.

[6] Adly Templeton. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Anthropic, 2024.

[7] Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. arXiv preprint arXiv:2406.04093, 2024.

[8] Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. arXiv preprint arXiv:2403.19647, 2024.

[9] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? arXiv preprint arXiv:2004.03685, 2020.

[10] Google DeepMind Safety Research. Negative results for sparse autoencoders on downstream tasks and deprioritising sae research. DeepMind Safety Research Blog, 2025. Blog post.

[11] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. Advances in neural information processing systems, 31, 2018.

[12] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In Proceedings of the AAAI conference on artificial intelligence, volume 33, pages 3681–3688, 2019.

[13] Chris Lin, Ian Covert, and Su-In Lee. On the robustness of removal-based feature attributions. Advances in Neural Information Processing Systems, 36:79613–79666, 2023.

[14] Chiara Balestra, Bin Li, and Emmanuel Müller. On the consistency and robustness of saliency explanations for time series classification. arXiv preprint arXiv:2309.01457, 2023.

[15] Seonglae Cho, Harryn Oh, Donghyun Lee, Luis Rodrigues Vieira, Andrew Bermingham, and Ziad El Sayed. Faithfulsae: Towards capturing faithful features with sparse autoencoders without external datasets dependency. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 297–314, 2025.

[16] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. arXiv preprint arXiv:2007.01162, 2020.

[17] Aashiq Muhamed, Mona Diab, and Virginia Smith. Decoding dark matter: Specialized sparse autoencoders for interpreting rare concepts in foundation models. In Findings of the Association for Computational Linguistics: NAACL 2025, pages 1604–1635, 2025.

[18] Nikita Koriagin, Yaroslav Aksenov, Daniil Laptev, Gleb Gerasimov, Nikita Balagansky, and Daniil Gavrilov. Teach old saes new domain tricks with boosting. arXiv preprint arXiv:2507.12990, 2025.

[19] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety–a review. arXiv preprint arXiv:2404.14082, 2024.

[20] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in neural information processing systems, 31, 2018.

[21] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. arXiv preprint arXiv:2309.08600, 2023.

[22] Gonçalo Paulo, Stepan Shabalin, and Nora Belrose. Transcoders beat sparse autoencoders for interpretability. arXiv preprint arXiv:2501.18823, 2025.

[23] Joakim Edin, Andreas Geert Motzfeldt, Casper L Christensen, Tuukka Ruotsalo, Lars Maaløe, and Maria Maistro. Normalized aopc: Fixing misleading faithfulness metrics for feature attributions explainability. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1730, 2025.

[24] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. In Proceedings of the 58th annual meeting of the association for computational linguistics, pages 4443–4458, 2020.

[25] Lawrence Chan, Adria Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: A method for rigorously testing interpretability hypotheses. In AI Alignment Forum, volume 2, 2022.

[26] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. Advances in neural information processing systems, 35:17359–17372, 2022.

[27] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In International conference on machine learning, pages 3519–3529. PMLR, 2019.

[28] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. Advances in neural information processing systems, 30, 2017.

[29] Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. Advances in neural information processing systems, 34:4738–4750, 2021.

[30] Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 278–300, 2024.

[31] Charles H Martin, Tongsu Peng, and Michael W Mahoney. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. Nature Communications, 12(1):4122, 2021.

[32] Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. Advances in Neural Information Processing Systems, 32, 2019.

[33] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers), pages 7319–7328, 2021.

[34] Callum McDougall, Arthur Conmy, János Kramár, Tom Lieberum, Senthooran Rajamanoharan, and Neel Nanda. Gemma scope 2: Technical paper. Technical report, Google DeepMind, 2025.

[35] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. SIAM Journal on Numerical Analysis, 7(1):1–46, 1970.

[36] Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. arXiv preprint arXiv:2407.14435, 2024.

[37] Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. Psychometrika, 31(1):1–10, 1966.

[38] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

[39] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In International Conference on Machine Learning, pages 2397–2430. PMLR, 2023.

[40] Guilherme Penedo, Hynek Kydlícek, Loubna Ben Allal, and Thomas Wolf. Fineweb: decanting the web for the finest text data at scale. HuggingFace. Accessed: Jul, 12, 2024.

[41] Lefteris Loukas, Manos Fergadiotis, Ion Androutsopoulos, and Prodromos Malakasiotis. Edgar-corpus: Billions of tokens make the world go round. In Proceedings of the Third Workshop on Economics and Natural Language Processing, pages 13–18, 2021.

[42] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In Proceedings of the 2023 conference on empirical methods in natural language processing, pages 6449–6464, 2023.

[43] Connor Kissane, Robert Krzyzanowski, Arthur Conmy, and Neel Nanda. Saes (usually) transfer between base and chat models. Alignment Forum, 2024. URL https://www.alignmentforum.org/posts/fmwk6qxrpW8d4jvbd/saes-usually-transfer-between-base-and-chat-models.

[44] Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. SIAM journal on Matrix Analysis and Applications, 20(2):303–353, 1998.

[45] Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. Biometrika, 102(2):315–323, 2015.

[46] Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

[47] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027, 2020.

[48] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. arXiv preprint arXiv:2202.10054, 2022.

# A Proofs and Derivations

## A.1 Proof of Proposition 1

**Setup.** Write the second-moment shift as $E = M_{\mathrm{OOD}} - M_{\mathrm{ID}}$, so that $M_{\mathrm{OOD}} = M_{\mathrm{ID}} + E$. The projectors $\Pi_{\mathrm{ID}}$ and $\Pi_{\mathrm{OOD}}$ correspond to the top-$r$ eigenspaces of $M_{\mathrm{ID}}$ and $M_{\mathrm{ID}} + E$, respectively. We wish to bound $\Delta(\Pi_{\mathrm{ID}}) = \|\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{ID}}\|_F$ in terms of $\|E\|_F$.

**Step 1: Projector distance via principal angles.** Let $\theta_1, \dots, \theta_r$ be the principal angles between the column spaces of $\Pi_{\mathrm{ID}}$ and $\Pi_{\mathrm{OOD}}$. For two rank-$r$ orthogonal projectors, the Frobenius norm of their difference satisfies

$$\|\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{ID}}\|_F^2 = 2 \sum_{i=1}^{r} \sin^2 \theta_i.$$

**Step 2: Applying Davis–Kahan.** The Frobenius-norm form of the Davis–Kahan $\sin\Theta$ theorem [35, 45] bounds the sum of squared sines in terms of the perturbation $E$ and the eigengap:

$$\sum_{i=1}^{r} \sin^2 \theta_i \;\le\; \frac{\|(I - \Pi_{\mathrm{ID}})\, E\, \Pi_{\mathrm{ID}}\|_F^2}{\gamma_{\mathrm{ID}}^2},$$

where $\gamma_{\mathrm{ID}} = \lambda_r(M_{\mathrm{ID}}) - \lambda_{r+1}(M_{\mathrm{ID}})$ is the eigengap of $M_{\mathrm{ID}}$ at rank $r$.

**Step 3: Bounding the cross term.** The matrix $(I - \Pi_{\mathrm{ID}})\, E\, \Pi_{\mathrm{ID}}$ is the component of $E$ that maps the ID-active subspace into its orthogonal complement. Since $I - \Pi_{\mathrm{ID}}$ and $\Pi_{\mathrm{ID}}$ are orthogonal projectors, each has operator norm 1, so by the submultiplicativity of the Frobenius norm under operator-norm factors,

$$\|(I - \Pi_{\mathrm{ID}})\, E\, \Pi_{\mathrm{ID}}\|_F \le \|I - \Pi_{\mathrm{ID}}\|_2 \cdot \|E\|_F \cdot \|\Pi_{\mathrm{ID}}\|_2 = 1 \cdot \|E\|_F \cdot 1 = \|E\|_F.$$

**Step 4: Combining.** Substituting back,

$$\Delta(\Pi_{\mathrm{ID}}) = \|\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{ID}}\|_F = \sqrt{2 \sum_{i=1}^{r} \sin^2 \theta_i} \le \frac{\sqrt{2}}{\gamma_{\mathrm{ID}}} \|E\|_F = \frac{\sqrt{2}}{\gamma_{\mathrm{ID}}} \|M_{\mathrm{OOD}} - M_{\mathrm{ID}}\|_F,$$

where the first equality is Step 1, the inequality combines Steps 2 and 3, and the last equality uses $E = M_{\mathrm{OOD}} - M_{\mathrm{ID}}$ from the Setup. □

## A.2 Decomposition Eq. (3)

**Step 1: Expressing $\mathcal{L}_{\mathrm{OOD}}$ in terms of $M_{\mathrm{OOD}}$.** For any rank-$r$ orthogonal $\Pi_{\mathrm{dec}} \in \mathcal{C}_r$, the reconstruction error under $P_{\mathrm{OOD}}$ is

$$\begin{aligned}
\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}}) &= \mathbb{E}_{X \sim P_{\mathrm{OOD}}} \|h(X) - \Pi_{\mathrm{dec}}\, h(X)\|_2^2 \\
&= \mathbb{E} \|(I - \Pi_{\mathrm{dec}})\, h(X)\|_2^2 \\
&= \mathbb{E}\, \mathrm{tr}\big[(I - \Pi_{\mathrm{dec}})\, h(X)\, h(X)^\top\, (I - \Pi_{\mathrm{dec}})^\top\big] \\
&= \mathrm{tr}\big[(I - \Pi_{\mathrm{dec}})\, M_{\mathrm{OOD}}\, (I - \Pi_{\mathrm{dec}})\big], \qquad (12)
\end{aligned}$$

where the third step uses $\|a\|_2^2 = \mathrm{tr}(aa^\top)$ and the fourth step swaps expectation and trace, with $M_{\mathrm{OOD}} = \mathbb{E}[h(X)h(X)^\top]$. Since $I - \Pi_{\mathrm{dec}}$ is an orthogonal projector, it is idempotent: $(I - \Pi_{\mathrm{dec}})^2 = I - \Pi_{\mathrm{dec}}$. Using the cyclic property of trace,

$$\mathrm{tr}\big[(I - \Pi_{\mathrm{dec}})\, M_{\mathrm{OOD}}\, (I - \Pi_{\mathrm{dec}})\big] = \mathrm{tr}\big[(I - \Pi_{\mathrm{dec}})^2\, M_{\mathrm{OOD}}\big] = \mathrm{tr}\big[(I - \Pi_{\mathrm{dec}})\, M_{\mathrm{OOD}}\big]. \qquad (13)$$

**Step 2: Deriving the decomposition.** Applying the same identity with $\Pi_{\mathrm{OOD}}$ gives $\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}}) = \mathrm{tr}\big[(I - \Pi_{\mathrm{OOD}})\, M_{\mathrm{OOD}}\big]$. Taking the difference,

$$\begin{aligned}
\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}}) - \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}}) &= \mathrm{tr}\big[(I - \Pi_{\mathrm{dec}})\, M_{\mathrm{OOD}}\big] - \mathrm{tr}\big[(I - \Pi_{\mathrm{OOD}})\, M_{\mathrm{OOD}}\big] \\
&= \mathrm{tr}\big[(\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{dec}})\, M_{\mathrm{OOD}}\big]. \qquad (14)
\end{aligned}$$

Rearranging gives the claimed decomposition:

$$\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}}) = \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}}) + \mathrm{tr}\big[(\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{dec}})\, M_{\mathrm{OOD}}\big].$$

**Step 3: Nonnegativity and optimality.** It remains to show that the second term is nonnegative. Let $M_{\mathrm{OOD}} = \sum_{i=1}^{d} \lambda_i \, u_i u_i^\top$ be the eigendecomposition with $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$. Since $\Pi_{\mathrm{OOD}}$ projects onto the top-$r$ eigenspace,

$$\mathrm{tr}(\Pi_{\mathrm{OOD}} \, M_{\mathrm{OOD}}) = \sum_{i=1}^{r} \lambda_i.$$

By Ky Fan's maximum principle, $\sum_{i=1}^{r} \lambda_i = \max_{\Pi_{\mathrm{dec}} \in \mathcal{C}_r} \mathrm{tr}(\Pi_{\mathrm{dec}} \, M_{\mathrm{OOD}})$. Therefore, for any $\Pi_{\mathrm{dec}} \in \mathcal{C}_r$,

$$\mathrm{tr}\big[(\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{dec}}) \, M_{\mathrm{OOD}}\big] = \mathrm{tr}(\Pi_{\mathrm{OOD}} \, M_{\mathrm{OOD}}) - \mathrm{tr}(\Pi_{\mathrm{dec}} \, M_{\mathrm{OOD}}) \geq 0,$$

with equality if and only if $\Pi_{\mathrm{dec}}$ also projects onto a top-$r$ eigenspace of $M_{\mathrm{OOD}}$. This proves both the nonnegativity and the optimality $\Pi_{\mathrm{OOD}} \in \arg\min_{\Pi_{\mathrm{dec}} \in \mathcal{C}_r} \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}})$. □

### A.3 Proof of Proposition 2

From the decomposition (3), the explainer-dependent component equals $\mathrm{tr}\big[(\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{dec}}) \, M_{\mathrm{OOD}}\big]$. We derive both bounds by expanding this trace in the eigenbasis of $M_{\mathrm{OOD}}$.

**Setup.** Let $M_{\mathrm{OOD}} = \sum_{i=1}^{d} \lambda_i \, u_i u_i^\top$ be the eigendecomposition with $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$. The OOD-active subspace is $\Pi_{\mathrm{OOD}} = \sum_{i=1}^{r} u_i u_i^\top$, so $u_i^\top \Pi_{\mathrm{OOD}} \, u_i = \mathbf{1}_{i \leq r}$. For the explainer subspace $\Pi_{\mathrm{dec}} \in \mathcal{C}_r$, define $p_i := u_i^\top \Pi_{\mathrm{dec}} \, u_i \in [0, 1]$, the fraction of the $i$-th OOD eigendirection captured by $\Pi_{\mathrm{dec}}$. Expanding the trace in this eigenbasis gives

$$\begin{aligned} \mathrm{tr}\big[(\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{dec}}) \, M_{\mathrm{OOD}}\big] &= \sum_{i=1}^{d} \lambda_i \, (u_i^\top \Pi_{\mathrm{OOD}} \, u_i - p_i) \\ &= \sum_{i=1}^{r} \lambda_i \, (1 - p_i) - \sum_{i=r+1}^{d} \lambda_i \, p_i. \end{aligned} \tag{15}$$

The first sum captures the OOD energy in the top-$r$ directions that the explainer misses (since $1 - p_i$ is the fraction lost). The second sum captures the OOD energy in the bottom directions that the explainer unnecessarily covers.

**Connecting to the faithfulness gap.** For two rank-$r$ subspaces, the Frobenius norm of their difference satisfies $\|\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{dec}}\|_F^2 = 2 \sum_{i=1}^{r} (1 - p_i)$. Moreover, since both have the same rank, $\sum_{i=1}^{r} (1 - p_i) = \sum_{i=r+1}^{d} p_i$. Denoting $S := \sum_{i=1}^{r} (1 - p_i)$, we have

$$\Delta(\Pi_{\mathrm{dec}})^2 = \|\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{dec}}\|_F^2 = 2S, \quad \text{and} \quad \sum_{i=r+1}^{d} p_i = S. \tag{16}$$

**Upper bound.** Using $\lambda_i \leq \lambda_1$ for $i \leq r$ and $\lambda_i \geq \lambda_d$ for $i > r$ in Eq. (15),

$$\begin{aligned} \sum_{i=1}^{r} \lambda_i (1 - p_i) - \sum_{i=r+1}^{d} \lambda_i \, p_i &\leq \lambda_1 \sum_{i=1}^{r} (1 - p_i) - \lambda_d \sum_{i=r+1}^{d} p_i \\ &= \lambda_1 \, S - \lambda_d \, S = (\lambda_1 - \lambda_d) \, S = \frac{\lambda_1(M_{\mathrm{OOD}}) - \lambda_d(M_{\mathrm{OOD}})}{2} \, \Delta(\Pi_{\mathrm{dec}})^2. \end{aligned} \tag{17}$$

**Lower bound.** Using $\lambda_i \geq \lambda_r$ for $i \leq r$ and $\lambda_i \leq \lambda_{r+1}$ for $i > r$,

$$\begin{aligned} \sum_{i=1}^{r} \lambda_i (1 - p_i) - \sum_{i=r+1}^{d} \lambda_i \, p_i &\geq \lambda_r \sum_{i=1}^{r} (1 - p_i) - \lambda_{r+1} \sum_{i=r+1}^{d} p_i \\ &= \lambda_r \, S - \lambda_{r+1} \, S = (\lambda_r - \lambda_{r+1}) \, S = \frac{\gamma_{\mathrm{OOD}}}{2} \, \Delta(\Pi_{\mathrm{dec}})^2. \end{aligned} \tag{18}$$

□

### A.4 Corollary: Second-Moment Shift Upper-Bounds OOD Faithfulness Degradation

Combining Propositions 1 and 2 yields an explicit bound on how much faithfulness the ID explainer loses under OOD.

**Corollary 1** (Second-Moment Shift Upper-Bounds OOD Faithfulness Degradation for the ID Explainer)**.** *Suppose* $\gamma_{\mathrm{ID}} := \lambda_r(M_{\mathrm{ID}}) - \lambda_{r+1}(M_{\mathrm{ID}}) > 0$*. Then*

$$\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{ID}}) - \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}}) \leq \frac{\lambda_1(M_{\mathrm{OOD}}) - \lambda_d(M_{\mathrm{OOD}})}{2} \left( \frac{\sqrt{2}}{\gamma_{\mathrm{ID}}} \|M_{\mathrm{OOD}} - M_{\mathrm{ID}}\|_F \right)^2.$$

*Proof.* Setting $\Pi_{\mathrm{dec}} = \Pi_{\mathrm{ID}}$ in the upper bound of Proposition 2,

$$\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{ID}}) - \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}}) \;\leq\; \frac{\lambda_1(M_{\mathrm{OOD}}) - \lambda_d(M_{\mathrm{OOD}})}{2} \, \Delta(\Pi_{\mathrm{ID}})^2.$$

Proposition 1 gives $\Delta(\Pi_{\mathrm{ID}}) \leq \frac{\sqrt{2}}{\gamma_{\mathrm{ID}}} \|M_{\mathrm{OOD}} - M_{\mathrm{ID}}\|_F$. Substituting yields the result. The upper bound of Proposition 2 does not require $\gamma_{\mathrm{OOD}} > 0$. □

The explainer-dependent component grows at most quadratically with the second-moment shift.

### A.5 Derivation of the GAE Procrustes Solution (Step 1)

We derive the closed-form solution to the feature preservation problem in Eq. (6). Substituting $\widetilde{W}_{\mathrm{dec}}(T) = U_{\mathrm{OOD}}^{(:r)} \, T \, U_{\mathrm{dec}}^{\top} \, W_{\mathrm{dec}}^{\mathrm{ID}}$ and expanding:

$$\|\widetilde{W}_{\mathrm{dec}}(T) - W_{\mathrm{dec}}^{\mathrm{ID}}\|_F^2 = \|W_{\mathrm{dec}}^{\mathrm{ID}}\|_F^2 + \|U_{\mathrm{dec}}^{\top} W_{\mathrm{dec}}^{\mathrm{ID}}\|_F^2 - 2 \operatorname{tr}\big[(W_{\mathrm{dec}}^{\mathrm{ID}})^{\top} U_{\mathrm{OOD}}^{(:r)} \, T \, U_{\mathrm{dec}}^{\top} \, W_{\mathrm{dec}}^{\mathrm{ID}}\big], \quad (19)$$

where we used $T^{\top} T = I$ and $(U_{\mathrm{OOD}}^{(:r)})^{\top} U_{\mathrm{OOD}}^{(:r)} = I$. The first two terms are independent of $T$, so minimization reduces to

$$T^{\star} = \arg\max_{T \in \mathcal{O}_r} \operatorname{tr}(G\,T), \qquad G = U_{\mathrm{dec}}^{\top} \, W_{\mathrm{dec}}^{\mathrm{ID}} \, (W_{\mathrm{dec}}^{\mathrm{ID}})^{\top} \, U_{\mathrm{OOD}}^{(:r)}.$$

Let $G = \widetilde{U} \Sigma \widetilde{V}^{\top}$ be the SVD of $G$. Setting $R = \widetilde{V}^{\top} T \widetilde{U}$, we have $\operatorname{tr}(G\,T) = \operatorname{tr}(\Sigma\, R) = \sum_i \sigma_i R_{ii}$. Since $R \in \mathcal{O}_r$, each $|R_{ii}| \leq 1$, so the maximum is achieved at $R = I_r$, giving $T^{\star} = \widetilde{V} \widetilde{U}^{\top}$. □

### A.6 Derivation of the Step 2 Closed-Form Solution

With the encoder fixed, Step 2 solves the following convex quadratic objective over $W_{\mathrm{dec}} \in \mathbb{R}^{d \times k}$ and $b \in \mathbb{R}^d$:

$$\min_{W_{\mathrm{dec}},\, b} \; \frac{1}{N} \sum_{i=1}^{N} \|h_i - W_{\mathrm{dec}} \, z_i - b\|^2 \;+\; \lambda_{\mathrm{geom}} \, \|(I - \widehat{\Pi}_{\mathrm{OOD}}) \, W_{\mathrm{dec}}\|_F^2 \;+\; \lambda_{\mathrm{pres}} \, \|W_{\mathrm{dec}} - \widetilde{W}_{\mathrm{dec}}(T^{\star})\|_F^2.$$

**Bias.** Setting $\partial / \partial b = 0$ gives $b^{\star} = \frac{1}{N} \sum_i h_i - W_{\mathrm{dec}}^{\star} \frac{1}{N} \sum_i z_i$, i.e., Eq. (9).

**Decoder.** Substituting $b^{\star}$ centers the data: define $h_i^c = h_i - \frac{1}{N} \sum_j h_j$ and $z_i^c = z_i - \frac{1}{N} \sum_j z_j$. The problem reduces to

$$\min_{W_{\mathrm{dec}}} \; \frac{1}{N} \sum_{i=1}^{N} \|h_i^c - W_{\mathrm{dec}} \, z_i^c\|^2 \;+\; \lambda_{\mathrm{geom}} \, \|(I - \widehat{\Pi}_{\mathrm{OOD}}) \, W_{\mathrm{dec}}\|_F^2 \;+\; \lambda_{\mathrm{pres}} \, \|W_{\mathrm{dec}} - \widetilde{W}_{\mathrm{dec}}(T^{\star})\|_F^2.$$

Setting $\partial / \partial W_{\mathrm{dec}} = 0$:

$$\big[\lambda_{\mathrm{pres}} \, I + \lambda_{\mathrm{geom}} (I - \widehat{\Pi}_{\mathrm{OOD}})\big] \, W_{\mathrm{dec}} \;+\; W_{\mathrm{dec}} \, \tfrac{1}{N} \textstyle\sum_i z_i^c z_i^{c\top} \;=\; \tfrac{1}{N} \textstyle\sum_i h_i^c z_i^{c\top} \;+\; \lambda_{\mathrm{pres}} \, \widetilde{W}_{\mathrm{dec}}(T^{\star}).$$

The left-hand coefficient $\Lambda := \lambda_{\mathrm{pres}} \, I + \lambda_{\mathrm{geom}} (I - \widehat{\Pi}_{\mathrm{OOD}})$ is diagonal in the OOD basis, with eigenvalue $\lambda_{\mathrm{pres}}$ on $\operatorname{span}(U_{\mathrm{OOD}}^{(:r)})$ and $\lambda_{\mathrm{pres}} + \lambda_{\mathrm{geom}}$ on its complement. Denoting $B = \frac{1}{N} \sum_i z_i^c z_i^{c\top} + \lambda_{\mathrm{pres}} \, I$ and $C = \frac{1}{N} \sum_i h_i^c z_i^{c\top} + \lambda_{\mathrm{pres}} \, \widetilde{W}_{\mathrm{dec}}(T^{\star})$, the system decouples row-wise in the OOD basis into two standard ridge regressions:

- **Inside** $\widehat{\Pi}_{\mathrm{OOD}}$ (first $r$ rows): $W_{\mathrm{in}} = C_{\mathrm{in}}\, B^{-1}$, with ridge level $\lambda_{\mathrm{pres}}$.
- **Outside** $\widehat{\Pi}_{\mathrm{OOD}}$ (remaining $d-r$ rows): $W_{\mathrm{out}} = C_{\mathrm{out}}\,(B + \lambda_{\mathrm{geom}}\, I)^{-1}$, with ridge level $\lambda_{\mathrm{pres}} + \lambda_{\mathrm{geom}}$.

Combining via the projector yields Eq. (8): $W_{\mathrm{dec}}^{\mathrm{GAE}} = \widehat{\Pi}_{\mathrm{OOD}}\, C\, B^{-1} + (I - \widehat{\Pi}_{\mathrm{OOD}})\, C\, (B + \lambda_{\mathrm{geom}}\, I)^{-1}$. □

### A.7 Proof of Theorem 1

*Proof.* By Eq. (5), every column of $\widetilde{W}_{\mathrm{dec}}(T^\star)$ lies in $\mathrm{span}(U_{\mathrm{OOD}}^{(:r)})$, so

$$\Pi_{\mathrm{dec}}^{\mathrm{GAE}} = U_{\mathrm{OOD}}^{(:r)} (U_{\mathrm{OOD}}^{(:r)})^\top = \widehat{\Pi}_{\mathrm{OOD}}. \tag{20}$$

Under the condition $\widehat{\Pi}_{\mathrm{OOD}} = \Pi_{\mathrm{OOD}}$, this gives $\Delta(\Pi_{\mathrm{dec}}^{\mathrm{GAE}}) = \|\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{dec}}^{\mathrm{GAE}}\|_F = 0$ and therefore

$$\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}}^{\mathrm{GAE}}) = \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}}). \tag{21}$$

Applying the lower bound of Proposition 2 to $\Pi_{\mathrm{dec}} = \Pi_{\mathrm{ID}}$:

$$\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{ID}}) - \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}}) \;\geq\; \frac{\gamma_{\mathrm{OOD}}}{2}\, \Delta(\Pi_{\mathrm{ID}})^2. \tag{22}$$

Substituting $\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}}^{\mathrm{GAE}}) = \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}})$ and rearranging:

$$\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}}^{\mathrm{GAE}}) = \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}}) \leq \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{ID}}) - \frac{\gamma_{\mathrm{OOD}}}{2}\, \Delta(\Pi_{\mathrm{ID}})^2. \tag{23}$$

□

## B Empirical Evidence for Section 3

This appendix provides empirical support for the theoretical results in Section 3. We verify three claims: (i) the explainer subspace aligns with the ID-active subspace, (ii) the explainer-dependent component in the decomposition (3) accounts for a meaningful fraction of the total OOD error, and (iii) second-moment shift enlarges the faithfulness gap as predicted by Proposition 1.

### B.1 Controlled Toy Setting

All experiments in this appendix use a controlled toy setting that allows us to vary OOD severity continuously while keeping the model and explainer fixed.

**Target model.** The target model is a 2-layer ReLU MLP with input dimension $d_{\mathrm{in}}$=128, hidden dimension $d$=256, and output dimension $p \in \{4, 8, 16\}$:

$$h(x) = \mathrm{ReLU}(W_1 x + b_1) \in \mathbb{R}^d, \qquad o(x) = W_2\, h(x) + b_2 \in \mathbb{R}^p.$$

The hidden activations $h(x) \in \mathbb{R}^d$ correspond to the hidden representations analyzed in the main text. The explainer operates on these $d$-dimensional activations.

**Explainer.** We train both a transcoder and an SAE on ID hidden activations using the standard reconstruction-plus-sparsity objective (ERM), with dictionary sizes $k \in \{d/2,\ 1d,\ 2d,\ 4d,\ 8d,\ 32d\}$. The subspace rank is set to $r = p$, matching the rank of the output weight matrix $W_2 \in \mathbb{R}^{p \times d}$: since $o(x) = W_2\, h(x) + b_2$, the model's output depends on $h(x)$ only through its projection onto $\mathrm{span}(W_2^\top)$, which has dimension $p$. This makes $r = p$ the natural rank at which the active subspace captures all output-relevant directions.

**OOD generation.** ID inputs are drawn from $x \sim \mathcal{N}(0, I_{d_{\mathrm{in}}})$. OOD inputs are generated as $x = A_s\, z$ where $z \sim \mathcal{N}(0, I_{d_{\mathrm{in}}})$ and $A_s$ is a severity-dependent transformation matrix. Specifically, let $Q \in \mathbb{R}^{d_{\mathrm{in}} \times d_{\mathrm{in}}}$ be a fixed random orthogonal matrix and $\mathbf{s} \in \mathbb{R}^{d_{\mathrm{in}}}$ be fixed slopes linearly spaced in $[-S, S]$ with $S$=6. The base input covariance under severity $s$ is

$$\Sigma_{\mathrm{base}}(s) = Q\, \mathrm{diag}(e^{s \cdot \mathbf{s}})\, Q^\top\, (I + s\rho(1 + \tfrac{s}{2})\, V V^\top),$$

where $V \in \mathbb{R}^{d_{\text{in}} \times r_V}$ is a fixed random orthonormal matrix ($r_V$=32) and $\rho$=10. The base covariance is then rescaled directionally: variance in the output-relevant subspace $\text{span}(W_2^\top)$ is reduced by factor $(1 - 0.6s)$, and variance in its orthogonal complement is amplified by factor $(1 + 2s^2)$. Finally, the covariance is globally normalized so that $\text{tr}(\Sigma(s)) = d_{\text{in}}$ for all $s$, ensuring that reconstruction error differences are not driven by trivial scale changes. The transformation matrix $A_s$ is the matrix square root of the resulting $\Sigma(s)$. At $s$=0, $\Sigma(0) = I$ (ID); as $s$ increases toward 1, the input covariance undergoes progressive rotation and anisotropic rescaling, which propagates through the ReLU layer to induce second-moment shift in hidden space. We use $N$=20,000 samples for all geometric computations.

**Validation on real models.** We also validate the results on pretrained language models (GPT-2 Small, Pythia-1.4B) under temporal, domain, and adversarial distribution shifts, using the experimental setup described in Section 5. Unlike the toy setting, OOD severity is not varied continuously; instead, we compare ID and pure OOD for each shift type.

### B.2 Explainer Subspace Alignment with the ID Dominant Subspace

Section 3.1 claims that the explainer subspace $\Pi$ aligns closely with $\Pi_{\text{ID}}$ for a well-trained ID explainer. We verify this by measuring the subspace overlap

$$\text{overlap}(U_{\text{dec}}, U_{\text{ID}}) := \frac{1}{r} \|U_{\text{dec}}^\top U_{\text{ID}}\|_F^2 \in [0, 1],$$

where $U_{\text{ID}} \in \mathbb{R}^{d \times r}$ contains the top-$r$ eigenvectors of $M_{\text{ID}}$. A value of 1 indicates perfect alignment; 0 indicates orthogonality.

#### B.2.1 Toy Setting

We sweep dictionary sizes $k \in \{d/2, 1d, 2d, 4d, 8d, 32d\}$ for both transcoders and SAEs to test whether the claim holds across different levels of overcompleteness. Figure 5 reports the overlap as a function of OOD severity for each dictionary size. In both panels, solid lines show the explainer–ID overlap and dashed lines show the explainer–OOD overlap.

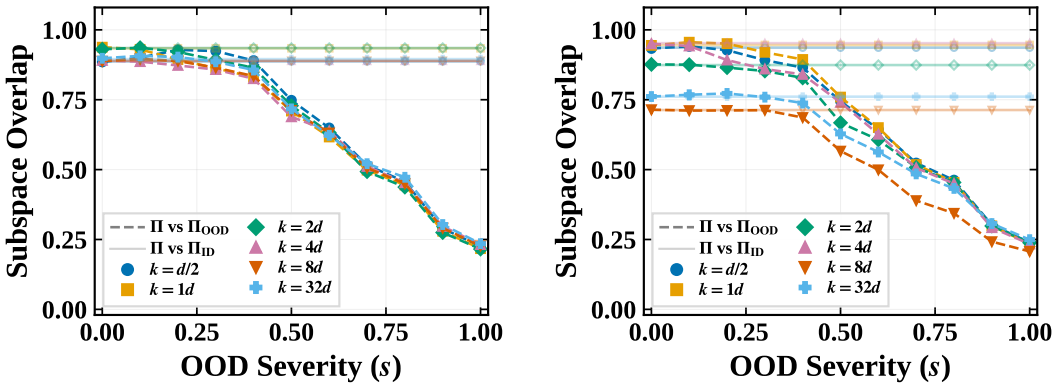



Figure 5: **Explainer subspace overlap between the explainer and the ID-active subspace (solid) vs. the OOD-active subspace (dashed), as a function of OOD severity $s$, for dictionary sizes $k \in \{d/2, 1d, 2d, 4d, 8d, 32d\}$. Left:** Transcoder. **Right:** SAE. For $k \geq 4d$, both explainer types maintain high ID overlap ($> 0.89$) regardless of severity, while OOD overlap degrades monotonically.

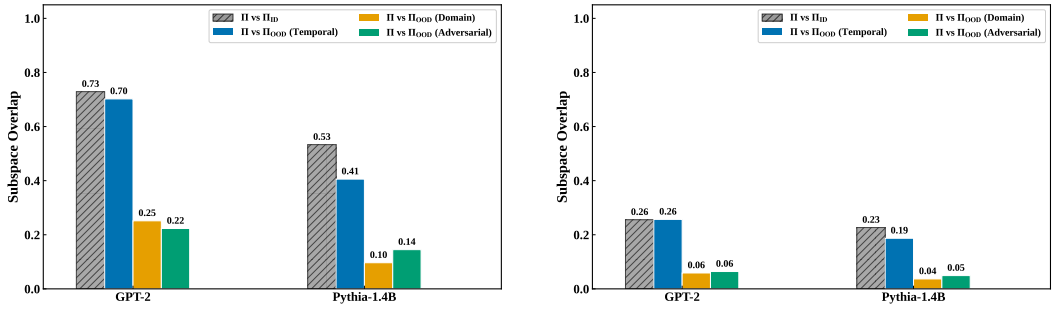
**Transcoder (left).** The explainer–ID overlap (solid) remains above 0.89 for all dictionary sizes and all severity levels. The overlap is largely insensitive to $k$: even an undercomplete dictionary ($k$=$d/2$) captures the ID-active subspace well. The explainer–OOD overlap (dashed) degrades monotonically with severity, dropping below 0.3 at $s$=1.0 regardless of $k$.

**SAE (right).** The explainer–ID overlap depends more on $k$. For $k \geq 4d$, the overlap exceeds 0.93, comparable to the transcoder. For smaller $k$ ($k$=$d/2$ or $k$=$1d$), the overlap drops to 0.70–0.77, indicating that undercomplete SAEs do not fully capture the ID-active subspace. As with the transcoder, the explainer–OOD overlap degrades with severity for all $k$.

**Interpretation.** For sufficiently overcomplete dictionaries ($k \geq 4d$, the standard setting in practice), both transcoders and SAEs align closely with the ID-active subspace regardless of OOD severity, confirming the claim in Section 3.1. The divergence between the ID overlap (flat) and the OOD overlap (decreasing) is precisely the faithfulness gap: the explainer remains anchored to the ID geometry while the model's active subspace rotates away under OOD shift.

### B.2.2 Real-Data Setting

Figure 6 reports the subspace overlap for ID-trained explainers on GPT-2 Small and Pythia-1.4B under temporal, domain, and adversarial shifts. The rank is $r$=64 for both models.




(a) Transcoder (b) SAE

Figure 6: **Subspace overlap on pretrained language models.** Hatched bars: explainer vs. ID-active subspace. Colored bars: explainer vs. OOD-active subspace under temporal, domain, and adversarial shifts. The explainer–ID overlap consistently exceeds the OOD overlaps across both explainer types and all shift types.

**Transcoder.** Across both models, the explainer–ID overlap (hatched bars, 0.53–0.73) substantially exceeds the explainer–OOD overlap, which drops to 0.10–0.25 for domain and adversarial shifts. The gap is larger for domain and adversarial shifts than for temporal shift, consistent with the stronger geometric distortion induced by these shift types.

**SAE.** SAE explainers show a similar pattern. The explainer–ID overlap (0.23–0.26) exceeds the OOD overlap under domain and adversarial shifts (0.04–0.06), while temporal shift retains a larger portion of the ID overlap (0.19–0.26). The overall overlap values are lower than transcoders because SAEs reconstruct residual-stream activations rather than MLP outputs, spreading energy across more directions.

**Interpretation.** The ID overlap values are lower than in the toy setting (0.89+). This is expected: real language models have higher effective dimensionality, and the rank $r$=64 captures a smaller fraction of the total hidden dimension ($d$=768 for GPT-2, $d$=2048 for Pythia-1.4B) than $r$=$p$ in the toy setting. Despite this, the relative pattern (ID overlap above OOD overlap under domain and adversarial shifts) holds consistently for both transcoders and SAEs, confirming the alignment claim in Section 3.1 on real models. Temporal shift produces a milder gap, consistent with its smaller second-moment perturbation.

**Faithfulness gap on $M_{\mathrm{ID}}$ vs. on $M_{\mathrm{OOD}}$.** Definition 1 measures the gap of $\Pi_{\mathrm{dec}}$ against $\Pi_{\mathrm{OOD}}$. To assess the residual ID-side misalignment, we additionally report the analogous quantity against $\Pi_{\mathrm{ID}}$,

$$\Delta_{\mathrm{ID}}(\Pi_{\mathrm{dec}}) := \|\Pi_{\mathrm{ID}} - \Pi_{\mathrm{dec}}\|_F,$$

which directly quantifies how well the ID-trained explainer captures $\Pi_{\mathrm{ID}}$. Frobenius gap and overlap encode the same information through $\|\Pi_A - \Pi_B\|_F = \sqrt{2r\,(1 - \mathrm{overlap}(U_A, U_B))}$, so Table 4 reports the gaps converted from the same measurements as Figure 6. The maximum value at $r$=64 is $\sqrt{2r} = \sqrt{128} \approx 11.31$.

For transcoders, $\Delta_{\mathrm{ID}}(\Pi_{\mathrm{dec}})$ is consistently smaller than $\Delta(\Pi_{\mathrm{dec}})$, with the OOD-side gap exceeding the ID-side gap by 1.35–1.69$\times$ under domain and adversarial shifts, where the second-moment shift

Table 4: **Frobenius faithfulness gap of the ID-trained explainer at $r$=64.** Columns 4–6 report $\Delta_{\mathrm{ID}}(\Pi_{\mathrm{dec}}) = \|\Pi_{\mathrm{ID}} - \Pi_{\mathrm{dec}}\|_F$, $\Delta(\Pi_{\mathrm{dec}}) = \|\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{dec}}\|_F$, and $\|\Pi_{\mathrm{ID}} - \Pi_{\mathrm{OOD}}\|_F$. The last column is the ratio $\Delta(\Pi_{\mathrm{dec}})/\Delta_{\mathrm{ID}}(\Pi_{\mathrm{dec}})$. Maximum possible value at $r$=64 is $\sqrt{128} \approx 11.31$. All values are derived from the overlap measurements behind Figure 6.

| Explainer | Model | Shift | $\Delta_{\mathrm{ID}}(\Pi_{\mathrm{dec}})$ | $\Delta(\Pi_{\mathrm{dec}})$ | $\|\Pi_{\mathrm{ID}} - \Pi_{\mathrm{OOD}}\|_F$ | Ratio |
|---|---|---|---|---|---|---|
| Transcoder | GPT-2 Small | Temporal | 5.89 | 6.17 | 5.65 | 1.05× |
| | GPT-2 Small | Domain | 5.89 | 9.79 | 9.83 | 1.66× |
| | GPT-2 Small | Adversarial | 5.89 | 9.97 | 9.83 | 1.69× |
| | Pythia-1.4B | Temporal | 7.73 | 8.72 | 7.68 | 1.13× |
| | Pythia-1.4B | Domain | 7.73 | 10.75 | 10.78 | 1.39× |
| | Pythia-1.4B | Adversarial | 7.73 | 10.46 | 10.49 | 1.35× |
| SAE | GPT-2 Small | Temporal | 9.76 | 9.75 | 4.83 | 1.00× |
| | GPT-2 Small | Domain | 9.76 | 10.98 | 10.33 | 1.12× |
| | GPT-2 Small | Adversarial | 9.76 | 10.95 | 10.30 | 1.12× |
| | Pythia-1.4B | Temporal | 9.95 | 10.20 | 7.65 | 1.03× |
| | Pythia-1.4B | Domain | 9.95 | 11.10 | 10.88 | 1.12× |
| | Pythia-1.4B | Adversarial | 9.95 | 11.03 | 10.71 | 1.11× |

is largest. Moreover, $\|\Pi_{\mathrm{ID}} - \Pi_{\mathrm{OOD}}\|_F$ tracks $\Delta(\Pi_{\mathrm{dec}})$ to within 3% under these shifts, empirically supporting the substitution $\Pi_{\mathrm{dec}} \approx \Pi_{\mathrm{ID}}$ used in Proposition 1. SAE explainers exhibit a larger $\Delta_{\mathrm{ID}}(\Pi_{\mathrm{dec}})$ (close to the upper bound at $r$=64), reflecting the lower per-rank overlap of residual-stream dictionaries; the ordering $\Delta(\Pi_{\mathrm{dec}}) \geq \Delta_{\mathrm{ID}}(\Pi_{\mathrm{dec}})$ still holds under domain and adversarial shifts but with a smaller margin (1.11–1.12×). Temporal shift produces nearly equal ID- and OOD-side gaps for both explainer types, consistent with its milder second-moment perturbation.

### B.3 Relative Magnitude of the Explainer-Dependent Term

The decomposition of Eq. (3) separates OOD faithfulness as

$$\mathcal{L}_{\mathrm{OOD}}(\Pi) = \underbrace{\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}})}_{\text{irreducible}} + \underbrace{\mathrm{tr}\big((\Pi_{\mathrm{OOD}} - \Pi)\, M_{\mathrm{OOD}}\big)}_{\text{explainer-dependent}}.$$

Adaptation can only reduce the explainer-dependent component. If this component is negligible relative to the irreducible component, no adaptation strategy can meaningfully improve OOD faithfulness. We measure the explainer-dependent ratio

$$\eta := \frac{\mathrm{tr}[(\Pi_{\mathrm{OOD}} - \Pi)\, M_{\mathrm{OOD}}]}{\mathcal{L}_{\mathrm{OOD}}(\Pi)}$$

to assess whether the explainer-dependent component is an actionable target.

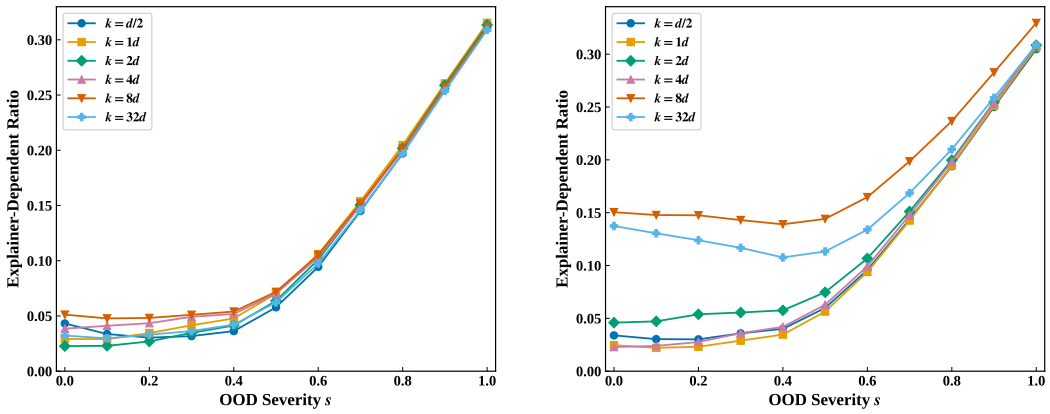
#### B.3.1 Toy Setting

Figure 7 plots $\eta$ as a function of OOD severity $s$ for each dictionary size $k$. At $s$=0, $\eta < 0.05$ for most $k$: when ID and OOD coincide, the explainer subspace is near-optimal. As severity increases, $\eta$ grows steadily to $\eta \approx 0.31$ at $s$=1.0 for both transcoders and SAEs, with little variation across $k$.

The moderate value of $\eta$ at maximum severity is a consequence of the toy setting's low rank ratio: $r = p = 8$ out of $d = 256$ dimensions, so only 3.1% of the hidden space is retained. The irreducible term $\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}}) = \sum_{i=r+1}^{d} \lambda_i(M_{\mathrm{OOD}})$ sums 248 discarded eigenvalues and dominates by construction. In real models where $r$ is chosen to capture a larger fraction of the activation energy, $\eta$ is substantially higher (Section B.3.2). The key observation in this toy setting is that $\eta$ increases monotonically with severity, confirming that distribution shift enlarges the explainer-dependent component relative to the total error.

#### B.3.2 Real-Data Setting

Figure 8 reports $\eta$ at pure ID and pure OOD for GPT-2 Small and Pythia-1.4B under temporal, domain, and adversarial shifts.

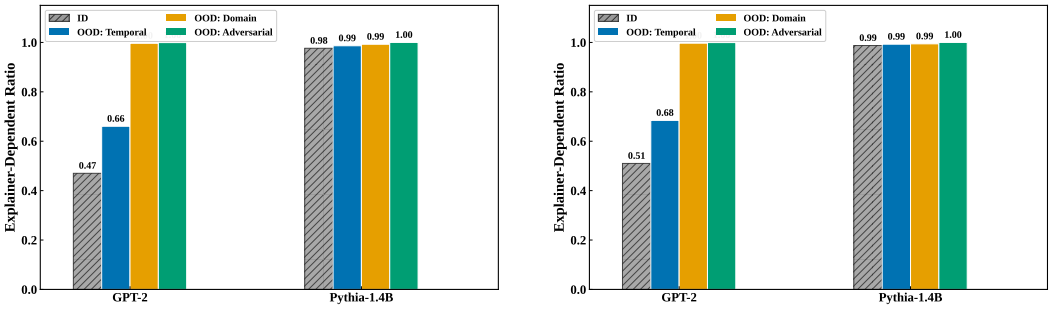



(a) Transcoder (b) SAE

Figure 7: **Explainer-dependent ratio $\eta$ as a function of OOD severity $s$ for dictionary sizes $k \in \{d/2, 1d, 2d, 4d, 8d, 32d\}$.** At $s$=1.0, $\eta \approx 0.31$ for both explainer types, independent of $k$.

**Transcoder (left).** At pure OOD, domain and adversarial shifts reach $\eta > 0.99$ across both models: the explainer-dependent component dominates the total error almost entirely. Temporal shift yields $\eta \approx 0.66$–$0.99$ depending on the model, reflecting its milder geometric distortion. These values are substantially higher than in the toy setting ($\eta \approx 0.31$), because the toy setting uses $r/d = 8/256 = 3.1\%$ so the irreducible term dominates by construction.

**SAE (right).** SAE explainers exhibit the same trend. Under domain and adversarial shifts, $\eta > 0.99$ for both models, confirming that the explainer-dependent component dominates. Temporal shift yields $\eta \approx 0.68$–$0.99$ depending on the model, consistent with its milder geometric distortion. At pure ID, $\eta \approx 0.51$ (GPT-2) and $\eta \approx 0.99$ (Pythia-1.4B), reflecting the larger model's higher effective dimensionality relative to $r$=64.




(a) Transcoder (b) SAE

Figure 8: **Explainer-dependent ratio $\eta$ at pure ID (hatched) and pure OOD (colored) ($r$=64).** Under domain and adversarial shifts, $\eta > 0.99$ for both explainer types.

### B.4 Empirical Verification of Proposition 1

Proposition 1 predicts that second-moment shift controls the faithfulness gap via

$$\Delta(\Pi_{\mathrm{ID}}) \leq \frac{\sqrt{2}}{\gamma_{\mathrm{ID}}} \|M_{\mathrm{OOD}} - M_{\mathrm{ID}}\|_F.$$

Since this bound depends only on $M_{\mathrm{ID}}$ and $M_{\mathrm{OOD}}$, it is independent of the explainer architecture and dictionary size. We verify it empirically on both the toy setting and real models.

### B.4.1 Toy Setting

Figure 9 plots the normalized second-moment shift against $\Delta(\Pi_{\mathrm{ID}})$, with color indicating OOD severity $s$. The two quantities are near-perfectly correlated (Pearson $r$=0.993, Spearman $\rho$=1.000), consistent with the linear upper bound.

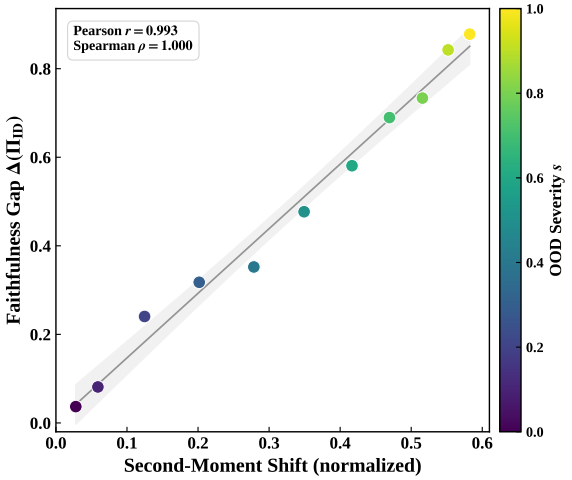
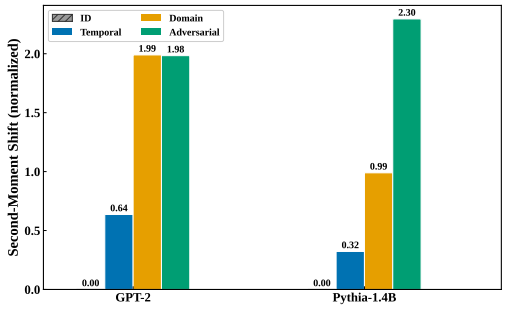
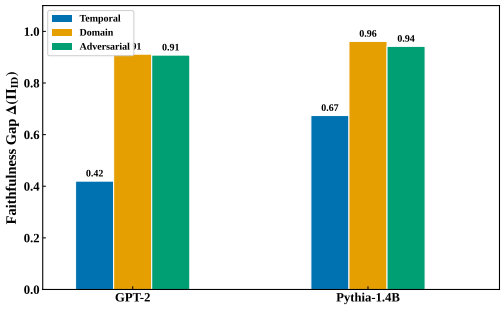



Figure 9: **Proposition 1 verification (toy).** X: normalized second-moment shift. Y: faithfulness gap $\Delta(\Pi_{\mathrm{ID}})$. Color: OOD severity $s$. The result is independent of explainer type and dictionary size.

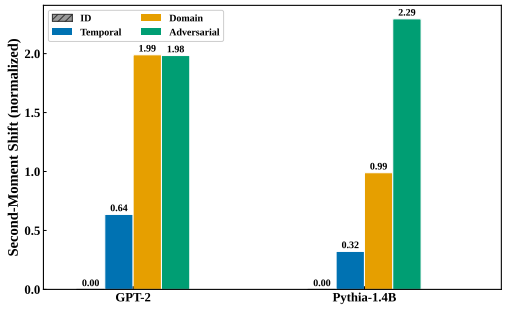
### B.4.2 Real-Data Setting



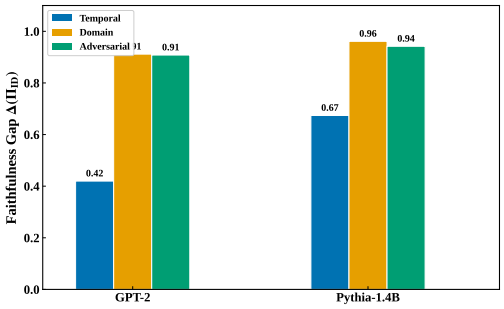

(a) Transcoder: Second-moment shift




(b) Transcoder: Faithfulness gap $\Delta(\Pi_{\mathrm{ID}})$




(c) SAE: Second-moment shift




(d) SAE: Faithfulness gap $\Delta(\Pi_{\mathrm{ID}})$

Figure 10: **Proposition 1 verification** ($r$=64) **at pure OOD.** Top row: Transcoder. Bottom row: SAE. Within each model, larger shifts correspond to larger gaps for both explainer types.

Figure 10 shows the second-moment shift and faithfulness gap for GPT-2 Small and Pythia-1.4B ($r$=64) at pure OOD for both transcoders (top row) and SAEs (bottom row). Within each model, the

ordering is consistent: temporal shift produces the smallest second-moment shift and the smallest faithfulness gap, while domain and adversarial shifts produce larger shifts and correspondingly larger gaps.

**SAE.** The same ordering holds for SAE explainers: temporal shift produces the smallest second-moment shift and faithfulness gap, while domain and adversarial shifts produce larger values. The magnitudes are comparable to transcoders, confirming that the proposition is independent of the explainer architecture.

**Summary.** The proposition holds across all models and both explainer types under diverse real-world distribution shifts, confirming that the theoretical predictions generalize beyond the controlled toy setting.

## C Empirical Verification of Theorem 1

Theorem 1 predicts that GAE's projection-loss improvement over the ID explainer grows at least quadratically with the faithfulness gap: $\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{ID}}) - \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}}^{\mathrm{GAE}}) \geq \frac{1}{2}\gamma_{\mathrm{OOD}}\,\Delta(\Pi_{\mathrm{ID}})^2$. We verify this in the controlled toy setting (Section 5.1) by fixing $r = p = 8$ and sweeping OOD severity from $s$=0 (ID) to $s$=1 (maximum shift). At each severity, we compute the Step 1 GAE projector $\Pi_{\mathrm{dec}}^{\mathrm{GAE}} = \widehat{\Pi}_{\mathrm{OOD}}$ and measure the projection-loss improvement $I(s) = \mathcal{L}_{\mathrm{OOD},s}(\Pi_{\mathrm{ID}}) - \mathcal{L}_{\mathrm{OOD},s}(\widehat{\Pi}_{\mathrm{OOD}})$ against the squared faithfulness gap $\Delta(\Pi_{\mathrm{ID}})^2$.

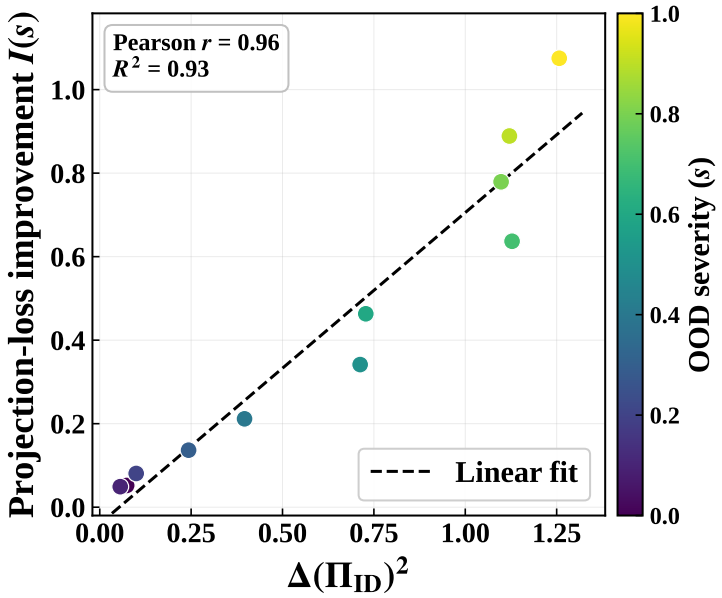



Figure 11: **Empirical verification of Theorem 1 on the controlled toy setting.** Projection-loss improvement $I(s)$ versus the squared faithfulness gap $\Delta(\Pi_{\mathrm{ID}})^2$, swept across OOD severity $s \in [0, 1]$. The dashed line is a linear fit ($R^2 = 0.93$, Pearson $r = 0.96$), supporting the quadratic dependence predicted by Theorem 1.

Figure 11 shows a strong linear relationship between $I(s)$ and $\Delta(\Pi_{\mathrm{ID}})^2$ (Pearson $r = 0.96$, $R^2 = 0.93$), supporting the quadratic dependence predicted by Theorem 1. The empirical improvement exceeds the guaranteed lower bound at every severity (0 violations out of 11), confirming that the bound is non-vacuous. The bound uses the worst-case eigengap $\gamma_{\mathrm{OOD}}/2$ as its constant, which is conservative relative to the effective improvement rate, as expected for a spectral-gap-based guarantee.

# D Experimental Details

## D.1 Model and Explainer Details

Table 5 summarizes the model and explainer configurations. All models are frozen pretrained checkpoints; only explainer components are adapted.

Table 5: **Model and explainer configurations.**

| Model | $d$ | Layer | $k$ ($32d$) | ID Corpus |
|---|---|---|---|---|
| GPT-2 Small | 768 | 8 | 24,576 | OpenWebText [46] |
| Pythia-1.4B | 2,048 | 15 | 65,536 | The Pile [47] |

For each model, we train two explainer types: Top-K SAEs [7], which reconstruct residual-stream activations using Top-K sparsity, and transcoders [5], which reconstruct MLP outputs from MLP inputs. All explainers use dictionary size $k{=}32d$ [4] and are trained on in-distribution activations with the standard reconstruction-plus-sparsity objective.

## D.2 Baseline Details

- **Fixed (ERM).** The ID-trained explainer applied to OOD inputs without any adaptation. This is the default deployment setting for existing dictionary-based explainers.
- **TERM.** An ID explainer trained with tilted empirical risk minimization [16, 17], which upweights high-loss (rare/tail) samples during training to improve coverage of infrequent concepts. This is an alternative ID training strategy, not an OOD adaptation method.
- **Finetune** [43]**.** The ID-trained explainer finetuned on OOD activations with a warm start. This adapts the existing dictionary to OOD data via gradient-based training.
- **Retrain.** The explainer retrained from scratch on OOD activations with the same architecture and hyperparameters. This baseline provides a reference point but is not an oracle upper bound: retraining on OOD data can distort pretrained feature structure [48].
- **SAEBoost.** A residual boosting approach [18]: a secondary explainer is trained on the OOD reconstruction residuals of the ID-trained base explainer, and the two outputs are summed at inference ($\hat{h} = \hat{h}_{\text{base}} + \hat{h}_{\text{resid}}$). This adds OOD-specific capacity while retaining the base dictionary, but requires OOD training data.
- **FaithfulSAE.** The explainer retrained on the target model's own unconditional generations [15], avoiding dependence on external datasets. Requires full retraining but no OOD data.
- **GAE (ours).** Training-free geometric adaptation (Algorithm 1). Step 1 rotates the ID dictionary's subspace to align with the OOD-active subspace via orthogonal Procrustes. Step 2 refits the decoder via constrained ridge regression with geometry and preservation regularization. The entire pipeline is closed-form; no gradient computation or iterative training is required.

## D.3 GAE Implementation Details

**Step 2 regularization.** The closed-form decoder refit (Section 4.2, Step 2) regularizes the decoder toward the Step 1 output $\widetilde{W}_{\text{dec}}(T^\star)$ with weight $\lambda_{\text{pres}}$, following Eq. (8).

**Decoder interpolation.** The Step 2 closed-form solution $W_{\text{dec}}^{\text{GAE}}$ optimizes sample-level reconstruction under geometry constraints. With limited OOD samples, this solution can overfit to the estimation noise in $\{z_i, h_i\}$. To mitigate this, we interpolate the Step 2 output with the Step 1 rotated dictionary:

$$W_{\text{final}} = (1-\alpha)\, W_{\text{dec}}^{\text{GAE}} + \alpha\, \widetilde{W}_{\text{dec}}(T^\star), \tag{24}$$

where $\alpha \in [0, 1]$ controls the interpolation. When $\alpha = 0$, the output equals the closed-form solution in Section 4.2. When $\alpha = 1$, the output equals the Step 1 rotation without reconstruction refinement. We treat $\alpha$ as a hyperparameter selected per OOD setting.

**Hyperparameter selection.** GAE requires no gradient computation or iterative optimization. The hyperparameters ($r$, $\lambda_{\text{geom}}$, $\lambda_{\text{pres}}$, $\alpha$) are selected per OOD setting using a small held-out portion of unlabeled OOD activations, monitoring reconstruction quality ($|\Delta\text{CE}|$) and causal faithfulness (nComp). No OOD labels are required. Once selected, the same hyperparameters are used for all evaluation prompts.

**Hyperparameter summary.** Tables 6 and 7 list the GAE hyperparameters for transcoders and SAEs, respectively. All settings use the second-moment matrix (not centered covariance) for OOD subspace estimation, as prescribed in Algorithm 1.

Table 6: **GAE hyperparameters for transcoder experiments.**

| Model | OOD Setting | $r$ | $\lambda_{\text{geom}}$ | $\lambda_{\text{pres}}$ | $\alpha$ | $N_{\text{fit}}$ |
|---|---|---|---|---|---|---|
| GPT-2 | HaluEval (Adversarial) | 32 | 0.1 | 0.2 | 0 | 2,048 |
| GPT-2 | Edgar (Domain) | 3 | 0.1 | 0.2 | 0 | 2,048 |
| GPT-2 | FineWeb (Temporal) | 6 | 0.1 | 0.04 | 0 | 2,048 |
| Pythia-1.4B | HaluEval (Adversarial) | 64 | 15 | 2 | 0 | 2,048 |
| Pythia-1.4B | Edgar (Domain) | 64 | 0.1 | 0.2 | 0 | 2,048 |
| Pythia-1.4B | FineWeb (Temporal) | 64 | 20 | 1 | 0 | 2,048 |

Table 7: **GAE hyperparameters for SAE experiments.** When only Step 1 (Procrustes rotation) is applied, all Step 2 hyperparameters are set to zero.

| Model | OOD Setting | $r$ | $\lambda_{\text{geom}}$ | $\lambda_{\text{pres}}$ | $\alpha$ | $N_{\text{fit}}$ |
|---|---|---|---|---|---|---|
| GPT-2 | HaluEval (Adversarial) | 639 | 0 | 0 | 1 | 0 |
| GPT-2 | Edgar (Domain) | 462 | 0 | 0 | 1 | 0 |
| GPT-2 | FineWeb (Temporal) | 700 | 0 | 0 | 1 | 0 |
| Pythia-1.4B | HaluEval (Adversarial) | 3 | 0.1 | 0.2 | 0 | 2,048 |
| Pythia-1.4B | Edgar (Domain) | 1,750 | 0 | 0 | 1 | 0 |
| Pythia-1.4B | FineWeb (Temporal) | 3 | 0 | 0.2 | 0 | 2,048 |

### D.4 Evaluation Details

**Normalized comprehensiveness (nComp).** We measure causal faithfulness via logit-level feature ablation [4, 21]. Given a prompt, let $\ell_0$ denote the target-token logit under the explainer's full reconstruction, and $\ell_\varnothing$ the logit when all features are ablated to zero. For a feature budget $m^*$, let $\ell_{\setminus m^*}$ be the logit after removing the top-$m^*$ features. We define

$$\text{nComp} = \frac{\ell_0 - \ell_{\setminus m^*}}{|\ell_0 - \ell_\varnothing|}, \tag{25}$$

where $m^* = 32$. Higher values indicate that the top features are causally important for the model's output.

**Delta cross-entropy ($\Delta$CE).** We measure reconstruction quality by the cross-entropy increase when original activations are replaced with the explainer's reconstruction [7]:

$$\Delta\text{CE} = \text{CE}(\hat{h}) - \text{CE}(h), \tag{26}$$

where $\text{CE}(h)$ is the loss with original activations and $\text{CE}(\hat{h})$ is the loss with reconstructed activations. Lower values indicate better preservation of the model's predictive behavior.

**Normalized AOPC (nAOPC).** nAOPC [23] averages the normalized logit drop across multiple feature budgets when top-$m$ features are removed:

$$\text{nAOPC} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \frac{\ell_0 - \ell_{\setminus m}}{|\ell_0 - \ell_\varnothing|}, \tag{27}$$

where $\mathcal{M} = \{1, 2, 4, 8, 16, 32, 64, 128\}$. Higher values indicate that the identified features are causally important across a range of budgets.

**Evaluation protocol.** We evaluate at the last token position using zero-residual ablation (replacing ablated features with zero). The denominator $|\ell_0 - \ell_\varnothing|$ normalizes each example by the logit range between full and empty reconstruction, enabling cross-example comparison regardless of the absolute logit scale. We exclude examples where $|\ell_0 - \ell_\varnothing| < 0.1$ to avoid unstable normalization. Feature budgets are $\mathcal{M} = \{1, 2, 4, 8, 16, 32, 64, 128\}$ with $m^* = 32$. We use $N_{\text{eval}} = 1{,}000$ evaluation prompts per setting and seed $= 2026$ throughout.

### D.5 Compute Resources

All experiments run on a single NVIDIA RTX A6000 GPU (48 GiB VRAM) with an Intel Xeon Gold 6326 CPU and 252 GiB of system RAM. No experiment requires multi-GPU or model-parallel execution. GPT-2 runs use peak GPU memory under 8 GiB. Pythia-1.4B runs with batch size 64 use peak GPU memory under 24 GiB.

**Per-method wall-clock.** Table 1 reports the cost of a single (model, OOD setting) run for each adaptation method. Finetune processes 5M tokens, taking about 2 minutes on GPT-2 and 12 minutes on Pythia-1.4B. Retrain, SAEBoost, and FaithfulSAE each process 100M tokens, taking about 39 minutes on GPT-2 and 4 hours on Pythia-1.4B. GAE finishes in 0.5 s on GPT-2 and 2.9 s on Pythia-1.4B using a single forward pass over ∼2,000 OOD activations and no gradient computation. Faithfulness evaluation (nAOPC, nComp, ΔCE on 1,000 prompts) adds about 1 minute per (model, OOD setting, baseline).

**Total compute.** The full result table (two models, three OOD settings, six adaptation baselines) requires roughly 50 GPU-hours on a single RTX A6000, dominated by the Retrain-style baselines on Pythia-1.4B. The GAE rows themselves contribute under 1 GPU-minute to this total. Pretraining of the ID dictionaries (transcoders and SAEs on OpenWebText / The Pile) is a one-time cost that we treat as external to the adaptation experiments.

## E Additional Case Studies on Other Semantic Classes

This appendix repeats the body case-study protocol on two further HaluEval prompts whose target tokens fall in distinct semantic classes: male first names and professions. The protocol is unchanged from Section 5.2.3: we keep the encoder frozen, take the top-3 features by GAE causal effect on the target token, and report each feature's direct logit attribution to 20 class-member tokens and 10 unrelated noun controls. Fixed and GAE share the same encoder and the same top-3 features, so any difference in attribution comes entirely from the decoder rotation.

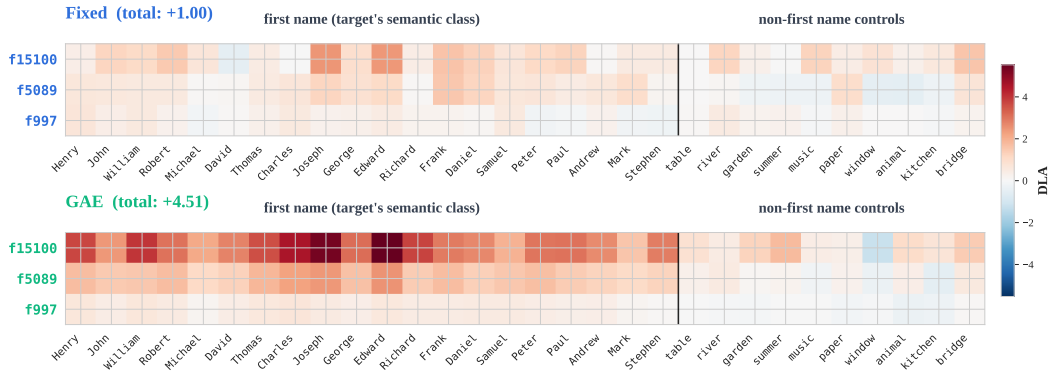



Figure 12: **Per-feature DLA on a prompt predicting ' Henry' (GPT-2, Transcoder).** The truncated input is "Question: What nationality was James"; the next token is a male first name. Each cell reports the feature's direct logit attribution to a candidate token, with 20 male first names on the left and 10 unrelated noun controls on the right. Fixed's total class-specificity is $+1.00$ and GAE's is $+4.51$, a $4.5\times$ amplification of the same encoder-selected features' pull toward the first-name class.

In both cases the decoder rotation alone reproduces the body-case finding: the same top-3 features, with the same activations, contribute more to their target's semantic class under GAE than under Fixed. The feature selection itself is identical because the encoder is shared.

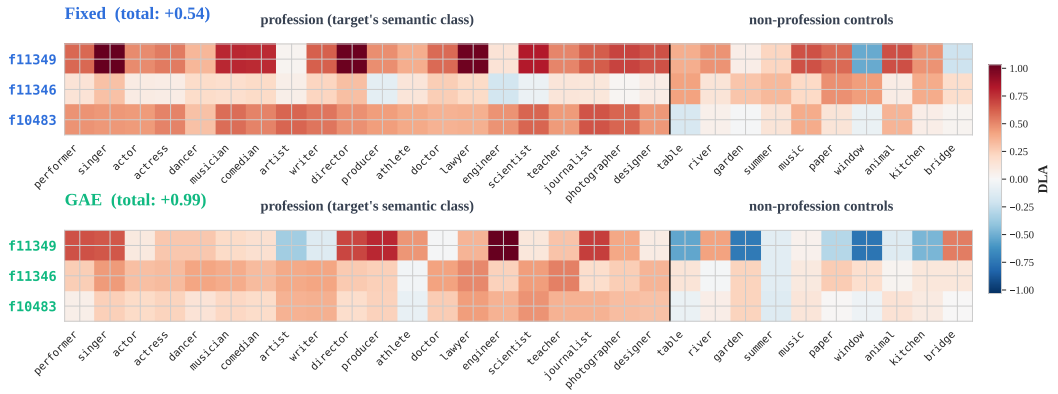
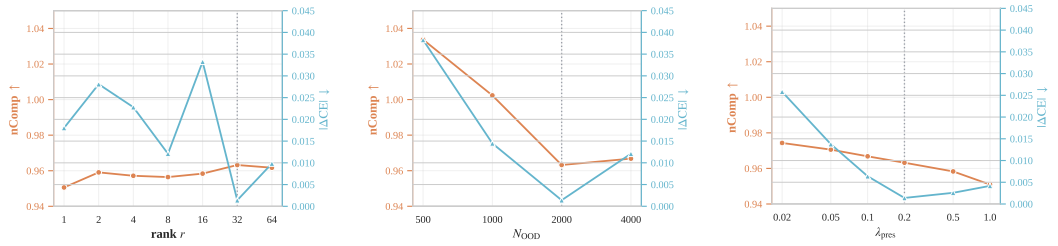



Figure 13: **Per-feature DLA on a prompt predicting ‘ politician’ (GPT-2, Transcoder).** The truncated input is “Question: Which American”; the next token is a profession. The 20 class-member tokens are common professions and the 10 controls are unrelated nouns. Fixed’s total class-specificity is $+0.54$ and GAE’s is $+0.99$. The GAE row drives several control cells negative (blue), where Fixed leaves them positive, sharpening the contrast between the class and its controls without changing which features were selected.

# F Hyperparameter Sensitivity

We sweep GAE’s three hyperparameters on HaluEval (GPT-2 + Transcoder), holding the other two at the defaults $r = 32$, $N_{\text{OOD}} = 2000$, $\lambda_{\text{pres}} = 0.2$ (Figure 14).




(a) Rank $r$. (b) $N_{\text{OOD}}$. (c) $\lambda_{\text{pres}}$.

Figure 14: **Hyperparameter sweeps on HaluEval (GPT-2, Transcoder)**: nComp (orange, left axis) and $|\Delta\text{CE}|$ (cyan, right axis) are stable across rank $r$, OOD sample size $N_{\text{OOD}}$, and preservation weight $\lambda_{\text{pres}}$.

**Rank** $r$: nComp stays above $0.95$ for every $r \in \{1, \ldots, 64\}$; rank-1 already gives $0.951$, confirming that the ID-to-OOD drift concentrates in a few directions. **OOD sample size** $N_{\text{OOD}}$: $|\Delta\text{CE}|$ improves from $0.038$ at $N = 500$ to $0.001$ at $N \geq 2000$ as the covariance estimate stabilizes. **Preservation weight** $\lambda_{\text{pres}}$: increasing $\lambda_{\text{pres}}$ trades a small nComp decrease ($0.02$) for a large $|\Delta\text{CE}|$ improvement ($0.026 \rightarrow 0.001$), with the default near the elbow.