# Unlocking Atomsemanticity in Sparse Autoencoders via Independence Regularization

**Anonymous Authors**[1]

## Abstract

Sparse Autoencoders (SAEs) have emerged as a promising methodology for decomposing the dense representations of Large Language Models (LLMs) into interpretable, monosemantic features. However, we observe that SAEs frequently suffer from degeneration in their latent space, where the latent activations of distinct tokens collapse into a narrow anisotropic cone with spuriously high cosine similarity. We argue that this geometric pathology hinders *Atomsemanticity*, which is the ideal state where features function as minimal and independent semantic units, and consequently exacerbates feature absorption. In this work, we propose **I**ndependence **R**egularized **S**parse **A**uto**E**ncoders (IR-SAE), a framework designed to counteract this degeneration by enforcing statistical independence among latents. We validate IR-SAE by integrating it with different SAE variants, specifically Top-K and Batch Top-K SAEs. Extensive experiments demonstrate that our approach effectively alleviates latent degeneration and reduces feature absorption, yielding features with significantly higher semantic granularity.

## 1. Introduction

Large Language Models (LLMs) have achieved remarkable success across a vast array of natural language processing tasks (Vatsal & Dubey, 2024; Thapa et al., 2023; Bavaresco et al., 2025), yet their internal mechanisms remain largely opaque. This opacity poses significant challenges to ensuring their safety, reliability, and alignment with human values (Liu et al., 2023; Hua et al., 2024).

The field of mechanistic interpretability has emerged to address this challenge (Bereska & Gavves, 2024), seeking to reverse-engineer the specific algorithms and causal mechanisms learned by neural networks (Wang et al., 2023; Nanda et al., 2023; Conmy et al., 2023). A central strategy within this pursuit is to decompose the dense, high-dimensional hidden representations of LLMs into more fundamental, human-interpretable features (Elhage et al., 2022). To this end, Sparse Autoencoders (SAEs) have become a leading approach (Huben et al., 2024; Gao et al., 2025; Lieberum et al., 2024; Rajamanoharan et al., 2024a) to identify and isolate monosemantic features, i.e., features that correspond to distinct semantic concepts (Huben et al., 2024).

Recent advances in SAEs have begun to move beyond merely finding monosemantic features towards discovering more fundamental units (Bussmann et al., 2025; Chanin et al., 2025; Lee et al., 2025; Korznikov et al., 2025). We recognize and identify the goal of these works as the pursuit of **Atomsemanticity**. Ideally, we define Atomsemanticity as the property that satisfies: (1) monosemy, where each feature represents a single, pure concept, and (2) minimal-semanticity, where these features represent minimal units that do not overlap in meaning and can be freely combined to reconstruct complex concepts. Atomsemanticity is deeply rooted in the linguistic theory of Componential Analysis (Goodenough, 1956), which posits that meaning is constructed from a set of basic atomic units.

Current SAEs still suffer from feature absorption (Chanin et al., 2024), where features encoding broad or abstract concepts fail to activate on instances of their specific subtypes. Another issue we observed is that the sparse latents learned by SAEs struggle with the degeneration problem (Gao et al., 2019). In these cases, the distribution of latents become narrowly concentrated, undermining the interpretability and utility of the learned features. We hypothesize that both issues arise because of the lack of atomsemanticity. While standard SAEs' optimization successfully enforces sparsity, they do not sufficiently encourage feature specialization and independence.

Unlike prior work, we tackle this problem by addressing the latent degeneration issue in SAEs. We propose that by penalizing the similarity between latents of different tokens, we can both mitigate the degeneration problem and compel the SAE to make more efficient feature allocations. Instead

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

of activating similar feature sets for different inputs, the SAE is forced to discover more fundamental, independent atomic features that can distinguish these inputs.

Therefore, we propose IR-SAE (**I**ndependence **R**egularized **S**parse **A**uto**E**ncoders). The core principle of this approach is to penalize the tendency of latents for different input tokens to become overly similar. Specifically, we explore several instantiations, including cosine similarity regularizer, InfoNCE regularizer (Oord et al., 2018), and the kernel-based Hilbert-Schmidt Independence Criterion (HSIC) regularizer (Gretton et al., 2005).

We train our proposed model based on two prominent SAE frameworks: Top-K (Gao et al., 2025) and Batch Top-K (Bussmann et al., 2024). Extensive experiments on Gemma-2-2B demonstrate the effectiveness of our approach across multiple dimensions. For latent degeneration, our CosReg variant reduces average pairwise cosine similarity from 80.95% to 3.60%, while HSIC reduces feature absorption from 4.87% to 0.82%. On semantic decomposition tasks, InfoNCE achieves 28.18% on spurious correlation removal, nearly tripling the baseline. These improvements confirm that IR-SAEs effectively promotes atomsemantic features with superior semantic granularity.

In summary, our main contributions are:

- We identify Atomsemanticity as a critical goal for interpretable feature learning in SAEs, and recognize that the lack of Atomsemanticity underpins issues such as latent degeneration and feature absorption.
- We propose the IR-SAE framework, which mitigates the degeneration problem in SAEs by penalizing similarity between latents of different tokens, enhancing the semantic disentanglement of learned features.
- We validate the effectiveness of IR-SAE within both Top-K and Batch Top-K SAE frameworks, demonstrating that IR-SAE simultaneously mitigates latent degeneration and feature absorption, thus confirming our hypothesis.

The representation of "king" should activate features such as $\{$*royalty*, *person*, *ruler*, *male*$\}$, while "queen" should activate $\{$*royalty*, *person*, *ruler*, *female*$\}$.

## 2. Related Work

### 2.1. Mechanistic Interpretability

Our work is contextualized within the field of mechanistic interpretability for LLMs, which seeks to reverse-engineer model internals into human-understandable components (Elhage et al., 2021; Bereska & Gavves, 2024). Recent mechanistic interpretability research has focused on discovering subgraphs that perform specific computations (Olah et al., 2020; Olsson et al., 2022; Wang et al., 2023; Conmy et al., 2023); understanding the algorithms LLMs use for particular toy tasks (Nanda et al., 2023; Huang et al., 2024b); and addressing superposition in latent representations (Elhage et al., 2022; Bricken et al., 2023).

Superposition refers to the phenomenon where models represent more independent features than they have neurons, leading to individual neurons becoming polysemantic (Scherlis et al., 2022; Park et al., 2024). Sparse Autoencoders (SAEs), which learn an overcomplete dictionary to represent latents, have been shown to discover highly interpretable features and are a key tool for addressing superposition (Huben et al., 2024; Gao et al., 2025; Bussmann et al., 2024).

Standard SAEs induce sparsity via $L_1$ regularization (Bricken et al., 2023), but this approach suffers from the L1-shrinkage problem and lacks precise control over the $L_0$ norm (Gao et al., 2025). To address these issues, recent SAE works have proposed various improvements. For example, JumpReLU SAEs introduce discontinuous activation functions often paired with $L_0$ objectives to mitigate L1-shrinkage (Rajamanoharan et al., 2024b; Lieberum et al., 2024), Top-K SAEs enforce sparsity by retaining the top K activations (Gao et al., 2025), and Batch Top-K SAEs apply this mechanism within a batch (Bussmann et al., 2024).

Despite these advancements, these methods still lack explicit constraints on feature independence, leading to learned features that remain correlated and entangled. Typical manifestations include feature absorption (Chanin et al., 2024), feature composition (Leask et al., 2025), and feature hedging (Chanin et al., 2025). We observe that latents in SAEs suffer from the degeneration problem (Gao et al., 2019), which is also related to insufficient semantic independence among features.

### 2.2. Disentangling Atomsemantic Units

The effort to learn disentangled representations, where distinct features correspond to statistically independent factors, has been a longstanding goal in the machine learning community. Foundational approaches like Independent Component Analysis (ICA) (Bell & Sejnowski, 1995; Lee et al., 1999) seek to decompose a signal into a set of independent, non-Gaussian components. Closely related, sparse coding methods (Olshausen & Field, 1997; Elad, 2010) aim to learn a dictionary of basis vectors such that input data can be efficiently represented as a sparse linear combination of these vectors.

With the advance of LLMs, this pursuit has been revitalized with a focus on model interpretability. Early explorations applied sparse coding to analyze intermediate representations of LLMs like GPT-2 (Sharkey et al., 2023), BERT (Yun et al., 2021), or Pythia (Huben et al., 2024), coining the term Sparse Autoencoders (SAEs) for this approach.

Bricken et al. (2023) advanced the understanding of SAEs through a multi-faceted analysis on transformer models.

However, SAE methods do not explicitly enforce independence among the learned features, which can lead to entangled features that fail to capture minimal semantic units. Recent works have identified issues arising from this deficiency (Chanin et al., 2024; Leask et al., 2025; Chanin et al., 2025) and proposed various improvements (Lee et al., 2025; Korznikov et al., 2025). This scattered body of work highlights the need to move beyond simple monosemanticity towards features that are also semantically minimal and independent. By identifying this trend and naming it **Atomsemanticity**, our work aims to consolidate these efforts and provide a novel method to promote this property.

### 2.3. Representation Degeneration

Representation degeneration is a well-documented issue in LLMs where contextual word embeddings become anisotropic, collapsing into a narrow cone within the vector space (Gao et al., 2019; Ethayarajh, 2019). This geometric collapse results in high cosine similarity between semantically distinct representations, thereby limiting their expressive power.

To counteract this phenomenon, the primary strategy involves introducing explicit regularization during the optimization process. Notable examples of such regularizers include CosReg, which directly penalizes the average cosine similarity between random word pairs (Gao et al., 2019), and Laplacian regularization, which also promotes a more isotropic embedding distribution (Zhang et al., 2020).

## 3. Preliminary

In this section, we investigate the geometric and statistical properties of the latent space in SAEs. We identify a critical limitation in standard SAEs, termed **latent degeneration**, which we argue fundamentally undermines the goal of learning atomic features. Based on these observations, we formulate the motivation for explicitly regularizing latent independence to achieve finer semantic granularity.

### 3.1. Problem Formulation

We consider a pretrained LLM where the activation of the residual stream at a specific layer is denoted by $\mathbf{x} \in \mathbb{R}^d$. An SAE aims to decompose $\mathbf{x}$ into a sparse linear combination of feature directions. Formally, an SAE consists of an encoder and a decoder. The encoder maps the dense input $\mathbf{x}$ to a high-dimensional, sparse latent representation $\mathbf{z} \in \mathbb{R}^n$ (where $n \gg d$), and the decoder attempts to reconstruct the original input.

The encoding and decoding process are defined as:

$$\mathbf{z} = \sigma(\mathbf{W}_e(\mathbf{x} - \mathbf{b}_d) + \mathbf{b}_e), \tag{1}$$

$$\hat{\mathbf{x}} = \mathbf{W}_d\mathbf{z} + \mathbf{b}_d \tag{2}$$

where $\mathbf{W}_e \in \mathbb{R}^{n\times d}$ and $\mathbf{W}_d \in \mathbb{R}^{d\times n}$ are the encoder and decoder weights, and $\mathbf{b}_e \in \mathbb{R}^n$ and $\mathbf{b}_d \in \mathbb{R}^d$ are the corresponding bias terms. The activation function $\sigma(\cdot)$ is a sparsity-inducing activation function, such as ReLU, Top-$K$, or Batch Top-$K$ operator.

Throughout our analysis, we focus on the geometric and statistical properties of the latent activations $\mathbf{z}$. For a given dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$, we examine the set of their resulting latent vectors $\mathcal{Z} = \{\mathbf{z}_i\}_{i=1}^N$. Ideally, these vectors should utilize the high-dimensional latent space effectively to maximize representational capacity.

### 3.2. Empirical Observations

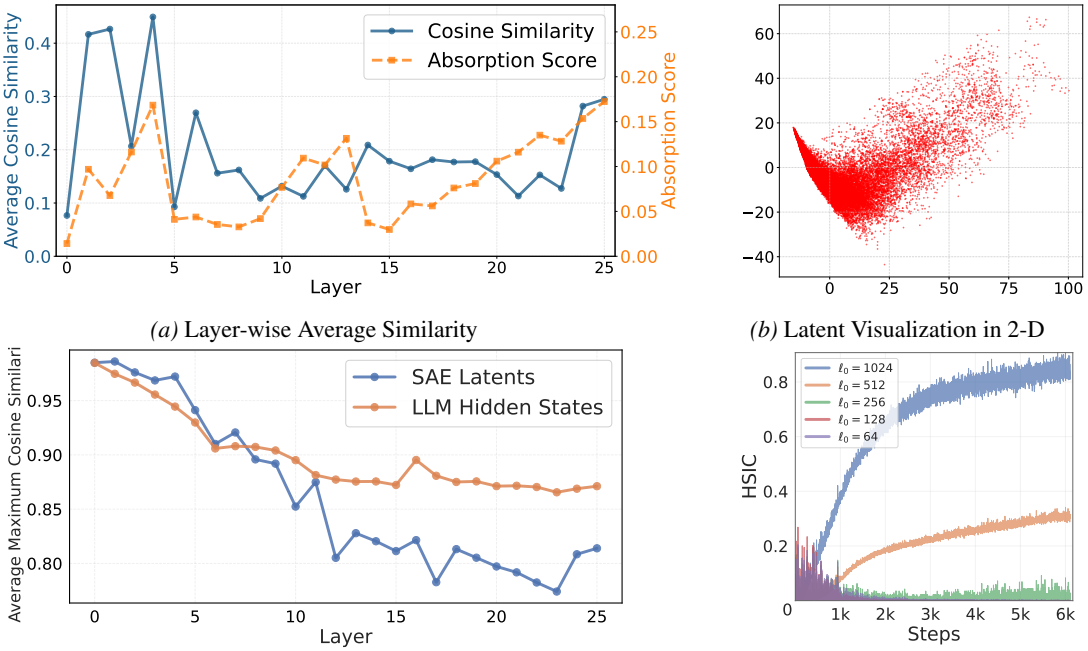
Our analysis utilizes two sources of SAE models: the publicly available Gemma-scope suite (Lieberum et al., 2024) for layer-wise geometric analysis, and our self-trained Top-$K$ SAEs on Gemma-2-2B (Team et al., 2024) for visualization and training dynamics analysis. Through a subset consisting of 2,000 tokens randomly sampled from EleutherAI/fineweb-edu-dedup-10b [1], we observe severe deviations from the ideal, indicating a phenomenon of **latent degeneration**.

**Spuriously High Similarity.** Using the Gemma-Scope SAEs, we calculate the average cosine similarity between the latent vectors of distinct tokens. As shown in Figure 1a, the average pairwise similarity is unexpectedly high, peaking above 0.4 in certain layers. Intriguingly, the layer-wise fluctuation of this similarity mirrors the profile of the feature absorption score reported in recent literature. This suggests that the geometric collapse of latents might be structurally coupled with the entanglement issues observed in SAE features.

**Inheritance of Input Anisotropy.** When excluding self-similarity, the *maximum* cosine similarity of one token's latent with any other token remains consistently high across all layers, as depicted in Figure 1c. Crucially, this trend tightly tracks the anisotropy observed in the hidden states of the base LLM. This suggests that standard SAEs fail to rectify the inherent geometry of the residual stream, instead propagating this degeneration into the sparse feature space.

**Spatial Collapse.** This geometric pathology is further corroborated by projecting the latent activations from a Top-$K$ SAE of Gemma-2-2B layer 8 into a 2-D space using

[1] https://huggingface.co/datasets/EleutherAI/fineweb-edu-dedup-10b




*(a)* Layer-wise Average Similarity

*(b)* Latent Visualization in 2-D

*(c)* Layer-wise Average Maximum Similarity

*(d)* HSIC During Training

*Figure 1.* Empirical observations indicating latent degeneration in SAEs. (a) Average pairwise cosine similarity across layers peaks above 0.4, mirroring feature absorption patterns. (b) PCA projection reveals latent activations confined to a narrow anisotropic region. (c) Maximum cosine similarity closely tracks the anisotropy of base LLM hidden states. (d) HSIC between latent pairs increases monotonically during training and worsens with higher $L_0$.

PCA. This visualization reveals that the activations are not uniformly distributed but are rather confined to a narrow anisotropic region.

**Statistical Dependence.** Beyond geometry, we quantify the independence of learned latent activations using the HSIC metric (Gretton et al., 2005). Figure 1d illustrates that as training progresses, the average pairwise HSIC between latents increases monotonically. Furthermore, this degeneration is significantly exacerbated as the $L_0$ increases. This indicates that standard SAE training inadvertently encourages statistical dependence among features, counteracting the goal of learning disentangled representations.

### 3.3. Hypothesis

The empirical observations above reveal a critical structural deficiency in standard SAEs. We hypothesize that latent geometry is causally linked to atomsemanticity through the following mechanism: the high statistical dependence among latents forces the decoder to learn mixed, non-atomic features, which in turn enables feature absorption. Therefore, by explicitly enforcing statistical independence among latents, we can break this causal chain and promote atomsemantic features with reduced absorption.

This hypothesis suggests that standard SAEs, which optimize for reconstruction and sparsity, lack a built-in mechanism to promote independence among learned factors. This motivates our intervention strategy: enforcing statistical independence among latents through explicit regularization.

## 4. Method

We introduce IR-SAE, a framework that augments standard SAE training with explicit independence constraints on latent activations. The core idea is to penalize statistical dependence among latent dimensions, thereby alleviating the geometric collapse and promoting atomsemantic features.

We build upon Top-$K$ (Gao et al., 2025) and Batch Top-$K$ SAEs (Bussmann et al., 2024), which enforce sparsity by retaining only the top $K$ largest activations per token or per batch. The training objective consists of a reconstruction loss and an auxiliary loss that prevents dead features:

$$\mathcal{L}_{\text{SAE}} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \alpha \mathcal{L}_{\text{aux}}, \tag{3}$$

where $\mathbf{x}$ denotes the input activation, $\hat{\mathbf{x}}$ is the reconstruction, and $\mathcal{L}_{\text{aux}}$ encourages all latent dimensions to be utilized. Notably, this objective imposes no constraints on the relationships *among* latent dimensions, leaving the model susceptible to the degeneration pathology.

To address this gap, we augment the objective with an inde-

pendence regularization term:

$$\mathcal{L}_{\text{IR-SAE}} = \mathcal{L}_{\text{SAE}} + \beta \cdot \mathcal{R}_{\text{Indep}}(\mathbf{F}), \tag{4}$$

where $\mathbf{F} \in \mathbb{R}^{n \times m}$ is the matrix of latent activations for a mini-batch of $n$ samples with $m$ latent dimensions, and $\beta$ controls the regularization strength. The term $\mathcal{R}_{\text{Indep}}(\mathbf{F})$ quantifies statistical dependence among latent dimensions and can be instantiated in multiple ways. We explore three variants with increasing expressiveness.

**Cosine Similarity Regularization.** The simplest approach penalizes pairwise cosine similarity among latent dimensions. For each latent dimension $i$, we first center its activations across the batch by subtracting the mean, then normalize to unit norm:

$$\tilde{\mathbf{f}}_{\cdot i} = \frac{\mathbf{f}_{\cdot i} - \mu_i \mathbf{1}}{\|\mathbf{f}_{\cdot i} - \mu_i \mathbf{1}\|_2}, \quad \text{where} \quad \mu_i = \frac{1}{n}\sum_{k=1}^{n} f_{ki}. \tag{5}$$

The regularization term penalizes the squared pairwise cosine similarity between distinct dimensions:

$$\mathcal{R}_{\text{CosReg}}(\mathbf{F}) = \frac{1}{m(m-1)} \sum_{i \neq j} \left( \tilde{\mathbf{f}}_{\cdot i}^{\top} \tilde{\mathbf{f}}_{\cdot j} \right)^2. \tag{6}$$

This variant is computationally efficient and geometrically intuitive, though it captures only linear dependencies.

**InfoNCE-based Regularization.** To encourage stronger discrimination among latent dimensions, we adopt a contrastive learning perspective. Each latent dimension $i$ is treated as an instance, with its activation pattern across the batch $\tilde{\mathbf{f}}_{\cdot i}$ serving as its representation. We apply an InfoNCE-style objective that encourages each dimension to be distinguishable from all others:

$$\mathcal{R}_{\text{InfoNCE}}(\mathbf{F}) = -\frac{1}{m}\sum_{i=1}^{m} \log \frac{\exp(1/\tau)}{\sum_{j=1}^{m} \exp\left( \tilde{\mathbf{f}}_{\cdot i}^{\top} \tilde{\mathbf{f}}_{\cdot j} / \tau \right)}, \tag{7}$$

where $\tau$ is a temperature hyperparameter. Minimizing this objective pushes different latent dimensions to have orthogonal activation patterns, implicitly enforcing decorrelation through a contrastive mechanism.

**HSIC-based Regularization.** We employ the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005), a kernel-based measure that quantifies statistical independence between random variables. For two latent dimensions with activation vectors $\mathbf{f}_{\cdot i}$ and $\mathbf{f}_{\cdot j}$, the empirical HSIC is computed as:

$$\text{HSIC}(\mathbf{f}_{\cdot i}, \mathbf{f}_{\cdot j}) = \frac{1}{(n-1)^2} \text{tr}(\mathbf{K}_i \mathbf{H} \mathbf{K}_j \mathbf{H}), \tag{8}$$

where $\mathbf{K}_i, \mathbf{K}_j \in \mathbb{R}^{n \times n}$ are kernel matrices (we use linear kernel) and $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\top}$ is the centering matrix. The regularization term aggregates pairwise HSIC values:

$$\mathcal{R}_{\text{HSIC}}(\mathbf{F}) = \frac{1}{m(m-1)} \sum_{i \neq j} \text{HSIC}(\mathbf{f}_{\cdot i}, \mathbf{f}_{\cdot j}). \tag{9}$$

HSIC is zero if and only if the two variables are statistically independent, making it a principled choice for enforcing independence. The computational cost is $O(n^2 m^2)$, which can be mitigated by randomly sampling a subset of latent pairs at each training step.

The proposed regularization terms are architecture-agnostic and can be integrated with any SAE variant as a plug-and-play module.

## 5. Experiments

We conduct comprehensive experiments on Gemma-2-2B (Team et al., 2024) to evaluate the effectiveness of IR-SAE. We focus on layer 4 and layer 8, representing shallow and intermediate layers respectively, where feature formation and abstraction occur at different granularities. We integrate our independence regularizers with two SAE architectures: Top-$K$ SAE (Gao et al., 2025) and Batch Top-$K$ SAE (Bussmann et al., 2024), and compare three IR-SAE variants (CosReg, InfoNCE, HSIC) against their unregularized counterparts. All SAEs are trained on the EleutherAI/fineweb-edu-dedup-10b dataset with matched computational budgets.

### 5.1. Does IR-SAE Alleviate Latent Degeneration?

We first examine whether independence regularization mitigates the geometric pathology identified in Section 3. Following our earlier analysis, we measure the pairwise cosine similarity among latent activations of distinct tokens. For each model, we sample 44109 tokens from a held-out set, compute their latent representations and calculate the average off-diagonal cosine similarity.

Table 1 presents the average cosine similarity results across different methods and layers. The baseline Top-$K$ SAE exhibits extremely high similarity scores of 71.18% and 80.95% for layer 4 and layer 8 respectively, confirming the severe latent degeneration problem. In contrast, all three IR-SAE variants demonstrate substantial improvements. The CosReg regularizer achieves the most significant reduction, decreasing similarity to 12.16% and 3.60% for the two layers. InfoNCE and HSIC regularizers also show notable improvements, though with varying degrees of effectiveness across layers. Similarly, the Batch Top-$K$ baseline shows high similarity (48.78% and 64.94%), which is effectively reduced by CosReg regularization (13.53% and 2.91%). These results provide strong empirical evidence that independence regularization successfully alleviates the

*Table 1.* Average pairwise cosine similarity (%) between latent activations of distinct tokens. Lower values indicate better independence and less latent degeneration.

| Method | Layer 4 | Layer 8 |
|---|---|---|
| *Top-K SAE* | | |
| Baseline | 71.18 | 80.95 |
| + CosReg | **12.16** | **3.60** |
| + InfoNCE | 46.13 | 4.53 |
| + HSIC | 40.75 | 26.32 |
| *Batch Top-K SAE* | | |
| Baseline | 48.78 | 64.94 |
| + CosReg | **13.53** | **2.91** |
| + InfoNCE | 44.66 | 3.78 |
| + HSIC | 38.13 | 25.89 |

*Table 2.* Feature absorption scores (%). Lower values indicate less absorption and more atomic features.

| | Fractional | | Full | |
|---|---|---|---|---|
| **Method** | **Layer 4** | **Layer 8** | **Layer 4** | **Layer 8** |
| *Top-K SAE* | | | | |
| Baseline | 41.28 | 25.55 | 13.16 | 4.87 |
| + CosReg | 9.91 | 11.72 | 0.85 | **0.78** |
| + InfoNCE | 24.95 | 22.81 | 5.08 | 8.81 |
| + HSIC | **5.91** | **10.03** | **0.14** | 0.82 |
| *Batch Top-K SAE* | | | | |
| Baseline | **6.90** | 31.60 | 0.64 | 9.23 |
| + CosReg | 18.71 | **8.76** | 2.46 | 1.87 |
| + InfoNCE | 26.46 | 26.62 | 9.03 | 9.03 |
| + HSIC | 12.17 | 14.01 | **0.06** | **0.24** |

latent degeneration phenomenon.

### 5.2. Does IR-SAE Reduce Feature Absorption?

We now investigate whether the improved latent independence translates into reduced feature absorption. We evaluate absorption using the protocol introduced by Karvonen et al. (2025), which quantifies the extent to which a single feature captures semantics that should ideally be distributed across multiple features. We report two metrics: *fractional absorption*, which measures the average proportion of semantic content absorbed by overlapping features, and *full absorption*, which measures the percentage of features completely absorbed by others.

Table 2 presents the absorption results. For fractional absorption, the Top-$K$ baseline exhibits substantial absorption rates of 41.28% and 25.55% across layers, indicating that features capture mixed semantics on average. IR-SAE variants significantly reduce this, with HSIC achieving the lowest fractional absorption (5.91% and 10.03%). CosReg also demonstrates strong performance (9.91% and 11.72%), while InfoNCE shows moderate improvements. The full absorption metric reveals even more dramatic improvements: the Top-$K$ baseline shows 13.16% and 4.87% of features being completely absorbed, whereas HSIC reduces this to merely 0.14% and 0.82%, and CosReg to 0.85% and 0.78%. Similar trends are observed for Batch Top-$K$ SAE. These results demonstrate that enforcing latent independence effectively mitigates the feature absorption problem, enabling SAEs to learn more atomic, disentangled representations.

### 5.3. Downstream Task Evaluation

To assess whether the improvements in latent geometry and absorption translate into practical benefits, we evaluate IR-SAE on a diverse suite of downstream tasks from SAEBench (Karvonen et al., 2025). These tasks probe different aspects of feature quality, which we categorize into two groups: semantic decomposition tasks and intervention tasks. Results are presented in Table 3.

**Semantic Decomposition Tasks.** This category includes tasks that evaluate how well SAE features decompose semantics into disentangled components: Feature Absorption, RAVEL, Spurious Correlation Removal (SCR), Sparse Probing, and Automated Interpretability. For *Feature Absorption*, IR-SAE variants demonstrate significant improvements over baselines. *RAVEL* (Huang et al., 2024a), which evaluates feature disentanglement through targeted attribute interventions, shows consistent improvements with HSIC achieving 63.91% and 76.62% for Top-$K$ SAE versus 60.84% and 73.28% for baselines. For *Sparse Probing*, HSIC achieves the highest alignment scores (79.05% and 81.48% for Top-$K$ SAE), indicating better concept alignment. *SCR* shows particularly strong gains with InfoNCE, nearly tripling baseline performance (28.18% vs. 11.51%), demonstrating that IR-SAE enables cleaner separation of spuriously correlated concepts. *Automated Interpretability* maintains competitive performance for HSIC despite slight degradation for CosReg, suggesting that the trade-off between statistical independence and per-feature interpretability is relatively modest. Overall, IR-SAE consistently improves performance on semantic decomposition tasks, with different regularizers excelling on different metrics: HSIC for absorption and disentanglement, InfoNCE for spurious correlation removal, and both maintaining strong sparse probing performance.

**Intervention Tasks.** This category includes tasks that evaluate the effectiveness of SAE features for causal interventions: Targeted Probe Perturbation (TPP) and Unlearning. For *TPP*, which measures whether ablating latents selectively affects specific classes, independence regularization significantly degrades performance across all variants. Baseline scores of 29.69% and 13.10% (Top-$K$ SAE) drop to

*Table 3.* Downstream task evaluation results (%). For all metrics except Absorption, higher scores indicate better performance. Results show layer 4 and layer 8 evaluations.

| | **Absorp.** (↓) | **RAVEL** (↑) | **AutoInt** (↑) | **SCR** (↑) | **SProb** (↑) | **TPP** (↑) | **Unlearn** (↑) |
|---|---|---|---|---|---|---|---|
| *Top-K SAE - Layer 4* | | | | | | | |
| Baseline | 13.16 | 60.84 | **84.40** | 10.59 | 75.32 | **29.69** | 51.55 |
| + CosReg | 0.85 | 63.00 | 79.09 | 7.29 | 77.32 | 1.53 | 6.19 |
| + InfoNCE | 5.08 | 61.78 | 81.96 | **11.18** | 77.29 | 2.55 | **68.04** |
| + HSIC | **0.14** | **63.91** | 81.94 | 9.89 | **79.05** | 2.68 | 20.62 |
| *Top-K SAE - Layer 8* | | | | | | | |
| Baseline | 4.87 | 73.28 | **85.71** | 11.51 | 78.62 | **13.10** | **47.42** |
| + CosReg | **0.78** | 75.31 | 79.59 | 17.12 | 78.13 | 1.73 | 6.17 |
| + InfoNCE | 8.81 | 73.13 | 81.96 | **28.18** | 79.35 | 3.31 | 41.24 |
| + HSIC | 0.82 | **76.62** | 80.87 | 16.17 | **81.48** | 1.81 | 29.90 |
| *Batch Top-K SAE - Layer 4* | | | | | | | |
| Baseline | 0.64 | 59.42 | 81.40 | 6.28 | 71.63 | **25.23** | 15.46 |
| + CosReg | 2.46 | 61.34 | 77.93 | 6.02 | 71.87 | 1.45 | 2.06 |
| + InfoNCE | 9.03 | 62.08 | 78.44 | **14.14** | **77.36** | 2.87 | 40.21 |
| + HSIC | **0.06** | **65.00** | **82.95** | 10.46 | 77.10 | 3.18 | **61.86** |
| *Batch Top-K SAE - Layer 8* | | | | | | | |
| Baseline | 9.23 | 74.08 | 77.54 | 15.96 | 76.09 | **14.66** | **58.76** |
| + CosReg | 1.87 | 75.76 | 76.68 | 14.48 | 80.05 | 1.39 | 9.28 |
| + InfoNCE | 9.03 | 72.94 | 78.27 | **25.39** | 79.11 | 2.65 | 36.08 |
| + HSIC | **0.24** | **76.99** | **79.48** | 15.92 | **81.35** | 1.57 | 42.27 |

1.53-3.18% for all IR-SAE variants, representing over 10× degradation. *Unlearning* shows mixed and highly variable results. While InfoNCE achieves competitive or improved performance (68.04% vs. 51.55% baseline on layer 4), CosReg shows substantial degradation (6.19% and 6.17%), and HSIC exhibits moderate degradation (20.62% and 29.90%). The consistent degradation on TPP and the high variability on Unlearning suggest that enforcing statistical independence interferes with intervention effectiveness. We hypothesize this occurs because intervention tasks implicitly favor features with concentrated causal effects—where ablating a single feature produces large, targeted changes. By promoting distributed, independent representations, IR-SAE reduces such concentration, thereby improving semantic decomposition at the cost of intervention capability.

**Summary.** The evaluation reveals a clear performance pattern: IR-SAE substantially improves semantic decomposition tasks that benefit from disentangled, independent features, while showing degraded performance on intervention tasks that favor concentrated causal effects. This trade-off reflects a fundamental tension in feature learning between semantic granularity and intervention potency.

### 5.4. Effect of Sparsity Level

The sparsity level, controlled by the $k$ value in the $\ell_0$ constraint, is a critical hyperparameter that directly affects the trade-off between reconstruction fidelity and feature interpretability. We investigate its impact by varying $k$ and evaluating performance on downstream tasks (Figure 2).

Most notably, while feature absorption rises dramatically with sparsity for baseline methods (reaching nearly 100%), HSIC regularization maintains consistently low absorption across the entire range. All methods achieve near-perfect reconstruction at higher sparsity levels, confirming that absorption is not due to insufficient capacity. Across other tasks, HSIC-regularized variants maintain relatively stable performance: RAVEL and Sparse Probing scores remain consistent, while Autointerp shows better stability compared to declining baseline performance. However, the TPP degradation for HSIC persists across all sparsity levels, confirming the intervention trade-off identified earlier.

## 6. Discussion

### 6.1. Monosemanticity vs. Atomsemanticity

Our experimental results reveal a fundamental tension in SAE design that reflects two distinct philosophical approaches: *monosemanticity* and *atomsemanticity*. The monosemanticity paradigm, emphasized in prior work (Bricken et al., 2023), seeks features that correspond to single, human-interpretable concepts, prioritizing *nameability* and direct applicability to model steering tasks. In contrast, the atomsemanticity approach decomposes representations into fundamental semantic units that may not align with nameable concepts but achieve superior performance on semantic decomposition tasks—significantly reducing feature absorption and improving spurious correlation removal, albeit at the cost of slightly reduced interpretability scores and degraded targeted perturbation performance.

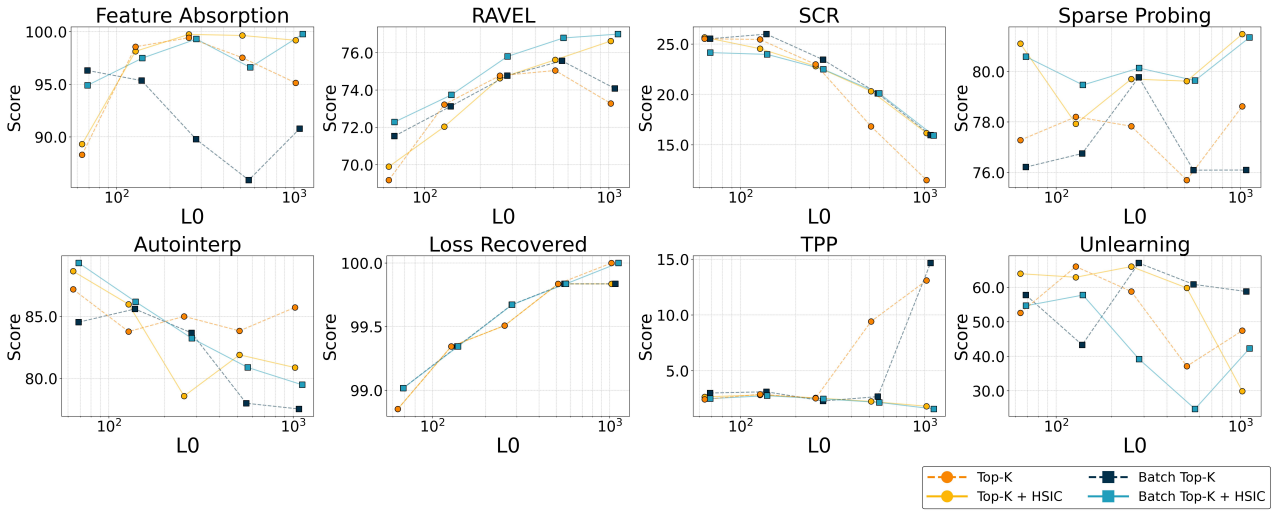



*Figure 2.* Effect of sparsity level (L0) on downstream task performance across layer 8.

This trade-off suggests that there is no universally ”correct” SAE objective. For model steering and targeted intervention, monosemantic SAEs may be preferable as they provide easily manipulable features, while atomsemantic SAEs offer deeper insights into how models fundamentally process information. This may suggest that the community would benefit from explicitly considering the intended purpose before designing SAE objectives, potentially developing specialized methods tailored to specific use cases rather than seeking a universal SAE architecture.

### 6.2. Ontological Misalignment

A critical question is why enforcing atomsemanticity reduces nameability. We propose this reflects a deeper *ontological misalignment* between human cognition and LLMs. Human ontology is grounded in physical reality, social interactions, and communicative language structure, shaped by embodied experience. In contrast, LLM ontology emerges from high-dimensional statistical patterns in text, constructing representations based on distributional semantics and predictive utility. There is no fundamental reason to believe the *basic units* discovered by LLMs align with those privileged by human cognition—indeed, model success may rely precisely on discovering latent structures humans overlook.

This misalignment implies that atomsemantic SAEs should not be viewed as constructing ”model-to-human” translation dictionaries, as monosemanticity work has implicitly assumed. Instead, they are tools for *mapping the model's native ontology*, revealing fundamental units the model employs even when these lack human-nameable counterparts. Without understanding this ontological structure, genuine alignment may remain elusive, as interventions based solely on human-interpretable features may miss critical aspects of model behavior. Future research should develop methods to identify and characterize atomic semantic units, investigating how these compose into higher-level concepts and building theoretical frameworks explaining why certain ontological structures emerge from specific training regimes.

## 7. Conclusion

In this paper, we identified latent degeneration as a critical geometric pathology in SAEs, where latent activations of distinct tokens exhibit spuriously high similarity and statistical dependence. Through systematic empirical analysis, we demonstrated that this degeneration is inherited from the anisotropic structure of LLM hidden states and fundamentally undermines the learning of atomic, disentangled features. To address this limitation, we proposed IR-SAE, which augments standard SAE training with explicit independence regularization on latent activations. We explored three instantiations of IR-SAE: CosReg, InfoNCE, and HSIC. These regularizers are architecture-agnostic and can be seamlessly integrated into existing SAE frameworks as plug-and-play modules.

Our experiments on Gemma-2-2B demonstrate that independence regularization substantially mitigates latent degeneration and reduces feature absorption. The improvements are particularly pronounced on tasks requiring semantic decomposition, such as spurious correlation removal and concept probing. However, we also observe trade-offs on intervention tasks, suggesting that different applications may benefit from different feature characteristics. More broadly, our work highlights the distinction between monosemanticity and atomsemanticity as complementary goals in interpretability research, opening avenues for developing specialized SAEs tailored to specific use cases.

## Impact Statement

This work contributes to the field of mechanistic interpretability by developing techniques to better understand the internal representations of neural networks. Enhanced interpretability has the potential to improve AI safety and reliability by enabling researchers to identify model failure modes, verify intended behaviors, and inform the design of more robust systems. At the same time, we acknowledge that progress in interpretability methods may also facilitate faster development and broader deployment of large language models. Such acceleration could amplify both the beneficial applications and potential risks associated with these powerful technologies, underscoring the importance of responsible research and deployment practices.

## References


Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., Ghaleb, E., Giulianelli, M., Hanna, M., Koller, A., et al. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 238–255, 2025.

Bell, A. J. and Sejnowski, T. J. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

Bereska, L. and Gavves, S. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N. L., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. https://transformer-circuits.pub/2023/monosemantic-features, 2023.

Bussmann, B., Leask, P., and Nanda, N. Batchtopk sparse autoencoders. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024.

Bussmann, B., Nabeshima, N., Karvonen, A., and Nanda, N. Learning multi-level features with matryoshka sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=m25T5rAy43.

Chanin, D., Wilken-Smith, J., Dulka, T., Bhatnagar, H., Golechha, S., and Bloom, J. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *arXiv preprint arXiv:2409.14507*, 2024.

Chanin, D., Dulka, T., and Garriga-Alonso, A. Feature hedging: Correlated features break narrow sparse autoencoders. *arXiv preprint arXiv:2505.11756*, 2025.

Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Elad, M. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media, 2010.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Ethayarajh, K. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, 2019.

Gao, J., He, D., Tan, X., Qin, T., Wang, L., and Liu, T. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*, 2019.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025.

Goodenough, W. H. Componential analysis and the study of meaning. *Language*, 32(1):195–216, 1956.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005.

Hua, W., Yang, X., Jin, M., Li, Z., Cheng, W., Tang, R., and Zhang, Y. Trustagent: Towards safe and trustworthy llm-based agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10000–10016, 2024.

Huang, J., Wu, Z., Potts, C., Geva, M., and Geiger, A. Ravel: Evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8669–8687, 2024a.

Huang, Y., Hu, S., Han, X., Liu, Z., and Sun, M. Unified view of grokking, double descent and emergent abilities: A comprehensive study on algorithm task. In *First Conference on Language Modeling*, 2024b. URL https://openreview.net/forum?id=cG1EbmWiSs.

Huben, R., Cunningham, H., Smith, L. R., Ewart, A., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Karvonen, A., Rager, C., Lin, J., Tigges, C., Bloom, J. I., Chanin, D., Lau, Y.-T., Farrell, E., McDougall, C. S., Ayonrinde, K., et al. Saebench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Forty-second International Conference on Machine Learning*, 2025.

Korznikov, A., Galichin, A., Dontsov, A., Rogov, O., Tutubalina, E., and Oseledets, I. Ortsae: Orthogonal sparse autoencoders uncover atomic features. *arXiv preprint arXiv:2509.22033*, 2025.

Leask, P., Bussmann, B., Pearce, M. T., Bloom, J. I., Tigges, C., Moubayed, N. A., Sharkey, L., and Nanda, N. Sparse autoencoders do not find canonical units of analysis. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9ca9eHNrdH.

Lee, S., Davies, A., Canby, M. E., and Hockenmaier, J. Evaluating and designing sparse autoencoders by approximating quasi-orthogonality. *arXiv preprint arXiv:2503.24277*, 2025.

Lee, T.-W., Lewicki, M. S., Girolami, M., and Sejnowski, T. J. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE signal processing letters*, 6(4):87–90, 1999.

Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramár, J., Dragan, A., Shah, R., and Nanda, N. 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*, pp. 278–300, 2024.

Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M. F., and Li, H. Trustworthy LLMs: a survey and guideline for evaluating large language models' alignment. In *Socially Responsible Language Modelling Research*, 2023.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

Olshausen, B. A. and Field, D. J. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning*, pp. 39643–39666. PMLR, 2024.

Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramár, J., Shah, R., and Nanda, N. Improving dictionary learning with gated sparse autoencoders. *arXiv preprint arXiv:2404.16014*, 2024a.

Rajamanoharan, S., Lieberum, T., Sonnerat, N., Conmy, A., Varma, V., Kramár, J., and Nanda, N. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024b.

Scherlis, A., Sachan, K., Jermyn, A. S., Benton, J., and Shlegeris, B. Polysemanticity and capacity in neural networks. *arXiv preprint arXiv:2210.01892*, 2022.

Sharkey, L., Braun, D., and Millidge, B. Taking features out of superposition with sparse autoencoders. 2022. *URL https://www.alignmentforum.org/posts/z6QQJbtpkEAX3Aojj/interim-research-report-taking-features-out-of-superposition*, 2023.

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Thapa, S., Naseem, U., and Nasim, M. From humans to machines: can chatgpt-like llms effectively replace human annotators in nlp tasks. In *Workshop Proceedings of*

*the 17th International AAAI Conference on Web and Social Media*. Association for the Advancement of Artificial Intelligence, 2023.

Vatsal, S. and Dubey, H. A survey of prompt engineering methods in large language models for different nlp tasks. *arXiv preprint arXiv:2407.12994*, 2024.

Wang, K. R., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023.

Yun, Z., Chen, Y., Olshausen, B., and LeCun, Y. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 1–10, 2021.

Zhang, Z., Gao, C., Xu, C., Miao, R., Yang, Q., and Shao, J. Revisiting representation degeneration problem in language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 518–527, 2020.