SCoPe: Intrinsic Semantic Space Control for Mitigating Copyright Infringement in LLMs

Zhenliang Zhang[1,2], **Xinyu Hu**[1], **Xiaojun Wan**[1,*]

[1]Wangxuan Institute of Computer Technology, Peking University
[2]School of Software and Microelectronics, Peking University
zhenliang@stu.pku.edu.cn, {huxinyu,wanxiaojun}@pku.edu.cn

Abstract

Large language models sometimes inadvertently reproduce passages that are copyrighted, exposing downstream applications to legal risk. Most existing studies for inference-time defences focus on surface-level token matching and rely on external blocklists or filters, which add deployment complexity and may overlook semantically paraphrased leakage. In this work, we reframe copyright infringement mitigation as **intrinsic semantic-space control** and introduce SCoPe, an inference-time method that requires no parameter updates or auxiliary filters. Specifically, the sparse autoencoder (SAE) projects hidden states into a high-dimensional, near-monosemantic space; benefiting from this representation, we identify a copyright-sensitive subspace and clamp its activations during decoding. Experiments on widely recognized benchmarks show that SCoPe mitigates copyright infringement without degrading general utility. Further interpretability analyses confirm that the isolated subspace captures high-level semantics.

1 Introduction

Large language models (LLMs) have demonstrated impressive capabilities in generating high-quality content. However, a surge of copyright lawsuits from media organizations and creators has raised serious legal concerns, particularly regarding the potential reproduction of copyrighted material from training data (Yu et al. 2023; Duan 2024; Brunetti 2024; Stratton 2024). Consequently, mitigating copyright infringement in LLMs has emerged as a pressing and essential research challenge.

Existing defenses against copyright infringement generally fall into three research directions: preprocessing filters to exclude copyrighted material (Kandpal, Wallace, and Raffel 2022; Sag 2023), training-time interventions such as Near Access-Freeness and selective unlearning (Abad et al. 2024; Xu et al. 2025), and inference-time controls. While preprocessing and training-time methods offer structural guarantees, they are computationally intensive and lack post-deployment flexibility—particularly when new protected content emerges. In contrast, inference-time techniques offer adaptability without modifying model parameters, making them more attractive and practical in real-world settings.

However, most prior inference-time methods rely heavily on external artifacts such as blocklist corpora or bloom filters (Ippolito et al. 2023; Shi et al. 2024), which require costly string-level comparisons or real-time rewriting. These mechanisms increase system complexity and may degrade fluency (Zhang et al. 2025). This motivates a question: *Can we eliminate these external mechanisms and enable LLMs to intrinsically avoid generating infringing content?*

Therefore, **we shift from surface-level output filtering to intrinsic semantic space control**. Inspired by the semantic subspace hypothesis (Park et al. 2023; Ferrando et al. 2024), we consider the possibility that copyrighted content may correspond to a distinct and identifiable subspace within the representation space of LLMs. If such a copyrighted subspace can be identified and effectively suppressed during LLM inference, infringement can be mitigated intrinsically.

> *"Once the mind ordains its inner limits, outward conduct is thereby governed."*
>
> *— Record of Rites*

The primary technical challenge arises from the **polysemanticity** of LLM neurons: individual neurons often encode multiple concepts (Olah et al. 2020). This semantic entanglement makes it difficult to isolate subspaces specifically associated with copyrighted content. Consequently, two critical research questions emerge: identifying copyright-sensitive subspaces and intervening in them to achieve effective copyright protection.

To address these challenges, we propose SCoPe, a two-stage method built on the sparse autoencoder (SAE). In the first stage, dense hidden states are mapped into a high-dimensional sparse space with **monosemanticity**. Leveraging this property, we define the ideal copyrighted subspace and develop a practical algorithm to estimate it. In the second stage, we apply feature clamping during decoding to suppress activations in this subspace, thereby reducing the risk of reproducing protected content. Experimental results show that, compared to baselines, SCoPe achieves superior copyright protection performance while maintaining general utility.

Moreover, we provide empirical validation of the subspace hypothesis by demonstrating that sparse space enable

*Corresponding author.

clearer separation between copyrighted and general content. Through feature semantic interpretation and reverse interventions, we confirm that the copyrighted subspace achieves isolation at the semantic level.

Our contributions are summarized as follows:

- **Novel Subspace Perspective:** We frame copyrighted content as residing in a distinct semantic subspace of LLM representations and develop an efficient method to identify the copyrighted subspace.
- **Semantic-Level Mitigation:** Our SCOPE framework clamps high-risk semantic features at decode time without external filters, operating at the semantic level rather than via surface token matching.
- **Effectiveness and Interpretability:** SCOPE delivers substantial reductions in copyright leakage while preserving overall model utility, with semantically meaningful controls.

2 Preliminaries

2.1 Semantic Space in LLMs

In LLMs, each layer produces a hidden state (embedding vector) $\mathbf{h} \in \mathbb{R}^d$ for every input token. The collection of all possible hidden states at a given layer therefore defines a d-dimensional **semantic space**, which serves as the model's internal representation of meaning at that processing stage.

Polysemanticity It is a well-documented phenomenon in these spaces: individual neurons often encode multiple, unrelated concepts simultaneously. Such superposition of neurons makes direct interpretation of single dimensions difficult (Olah et al. 2020; Elhage et al. 2022). As a result, methods that seek to identify and control specific semantic attributes must first disentangle these overlapping representations into more sparse subspaces.

2.2 SAE and Sparse Space

Sparse Autoencoder (SAE) is a neural module that projects the dense hidden states of an LLM into a higher-dimensional, sparse semantic space. Its core motivation is to address *polysemanticity* by mapping activations into disentangled, interpretable directions.

Given a hidden state vector $\mathbf{h} \in \mathbb{R}^d$, the SAE applies an **encoder** $f : \mathbb{R}^d \to \mathbb{R}^k$ and a **decoder** $g : \mathbb{R}^k \to \mathbb{R}^d$, with $k \gg d$:

$$\mathbf{z} = f(\mathbf{h}) = \mathrm{JumpReLU}(\mathbf{W}_{\mathrm{enc}}\mathbf{h} + \mathbf{b}_e) \tag{1}$$

$$\hat{\mathbf{h}} = g(\mathbf{z}) = \mathbf{W}_{\mathrm{dec}}\mathbf{z} + \mathbf{b}_d \tag{2}$$

Here, $\mathbf{z} \in \mathbb{R}^k$ is the sparse, high-dimensional vector, and $\hat{\mathbf{h}}$ is the reconstructed hidden state fed back into the LLM's residual stream. $\mathrm{JumpReLU}(x)$ is a variant of ReLU that enforces strict sparsity by zeroing out small activations, with the detailed formulations included in Appendix A. Moreover, the training of SAEs is not complicated, and some have been publicly available, for instance, the GemmaScope family (Lieberum et al. 2024).

Monosemanticity The vector $\mathbf{z}$ defines a k-dimensional sparse semantic space. To enforce sparsity, the SAE is trained with a reconstruction objective augmented by a sparsity penalty, yielding a disentangled, interpretable encoding of the original hidden state. Empirical studies confirm that, activations from SAE exhibit **monosemanticity** rather than semantic superposition (Cunningham et al. 2023; Pach et al. 2025). This architecture enables each dimension of $\mathbf{z}$ to correspond to distinct narrative or thematic elements (e.g., dialogue, temporal markers, topic shifts), enabling interpretation and targeted intervention.

2.3 SAE-Induced Semantic Subspaces

Given a hidden state vector $\mathbf{h} \in \mathbb{R}^d$ from an LLM, we obtain its sparse encoding $\mathbf{z} = [z_1, \ldots, z_k] \in \mathbb{R}^k$ via a pretrained SAE. The resulting sparse semantic space is designed so that each dimension z_i captures disentangled concepts.

More formally, for any index set of dimensions $\mathcal{I} \subseteq \{1, \ldots, k\}$, we define the projection:

$$\big(P_{\mathcal{I}}(\mathbf{z})\big)_i = \begin{cases} z_i, & i \in \mathcal{I} \\ 0, & i \notin \mathcal{I} \end{cases} \tag{3}$$

Then the **semantic subspace** associated with $\mathcal{I}$ is

$$\mathcal{S} = \mathcal{S}(\mathcal{I}) = \big\{\mathbf{z} \in \mathbb{R}^k \mid P_{\mathcal{I}}(\mathbf{z}) = \mathbf{z}\big\} \tag{4}$$

namely the set of sparse activations supported only on the chosen dimensions.

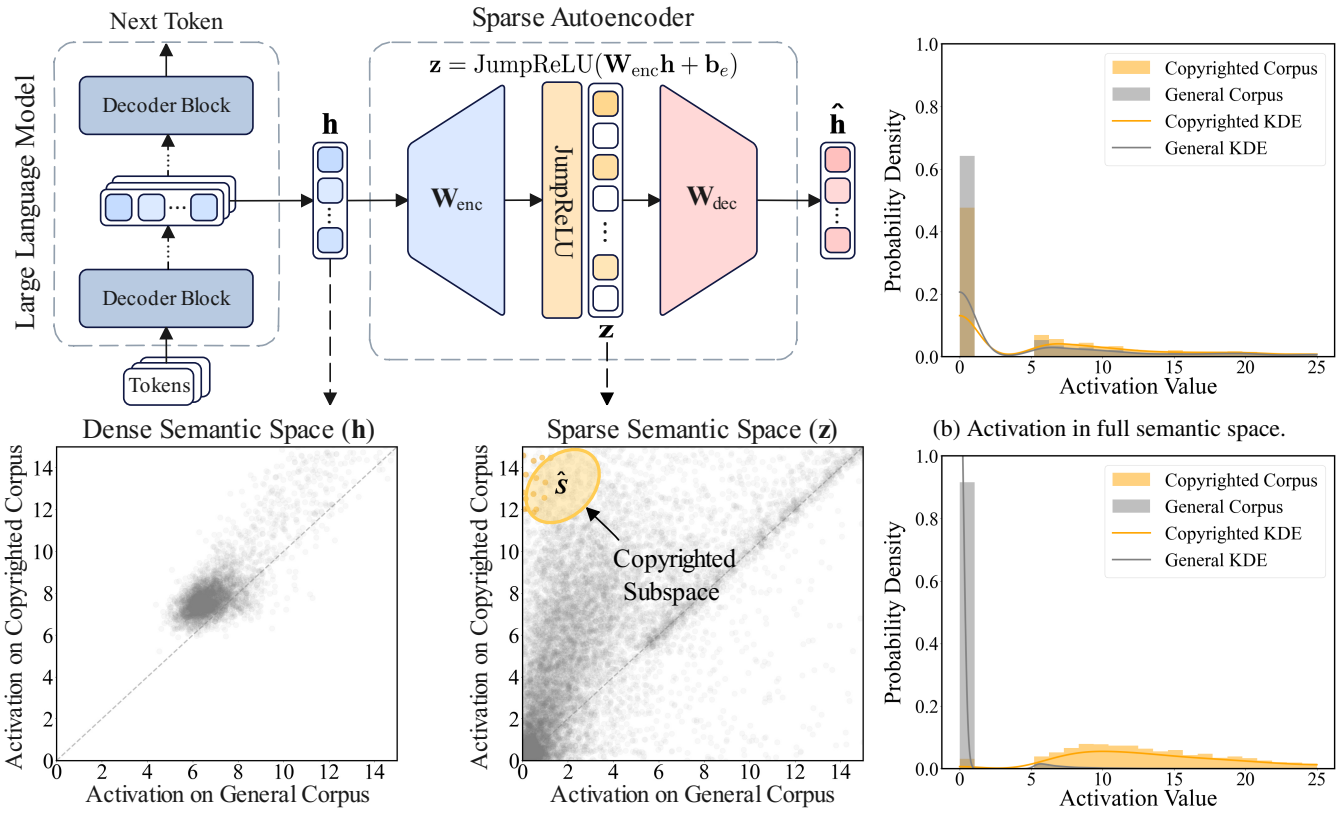
Subspace Hypothesis Evidence suggests that distinct semantic attributes (such as gender bias, unknown entity, or syntax) tend to align with specific directions in embedding spaces (Bolukbasi et al. 2016; Park et al. 2023; Ferrando et al. 2024). By extension, we hypothesize that the SAE-induced sparse space assigns most interpretable concepts to distinct linear subspaces. Consequently, activations for each concept concentrate within a subspace $\mathcal{S}$, allowing us to isolate and control it simply by projecting onto or away from $\mathcal{S}$.

3 Methodology

Building on the subspace hypothesis, we introduce **S**ubspace-oriented **Co**pyright **P**rot**e**ction (**SCOPE**), a two-stage method for LLM infringement prevention. In the first stage (Section 3.1), we identify the **copyrighted subspace**, whose constituent dimensions activate preferentially on copyrighted content. In the second stage (Section 3.3), we clamp this subspace at inference time to suppress copyright-sensitive outputs.

3.1 Identifying the Copyrighted Subspace

We center our investigation on three core questions: (1) ***Target***: What is the ideal copyrighted subspace? (2) ***Distance***: How do we quantify a candidate subspace's deviation from that ideal? (3) ***Approach***: How do we empirically refine our subspace to better approximate the target?



(a) **Activation in dense vs. sparse semantic spaces.** Points represent activations of dimensions in the LLM dense space (left) and SAE-induced sparse space (right).

(b) Activation in full semantic space.

(c) Activation in estimated subspace $\hat{\mathcal{S}}$.

Figure 1: **Visualization of semantic separation and subspace discrimination.** (a) SAE sparse space enables better separation of copyright-sensitive dimensions. (b) In the full space, activations for both corpora overlap. (c) In the estimated subspace $\hat{\mathcal{S}}$, activations become clearly separable.

Ideal Copyrighted Subspace (Target) We aim to identify a subspace that captures *only and exactly* the semantics of copyrighted corpus. Formally, let $\mathcal{C}_{\text{cr}}$ and $\mathcal{C}_{\text{gen}}$ denote the copyrighted and general corpora, respectively. Each input text $x \in \mathcal{C}$ is encoded by the SAE into a sequence of k-dimensional vectors. We then apply token-level max pooling over that sequence to obtain the final sparse vector $\mathbf{z}^{(x)} \in \mathbb{R}^k$. The ideal copyrighted subspace $\mathcal{S}^*$ should correspond to an index set $\mathcal{I}^*$ that satisfies the following two properties:

1. Coverage: For each sample $x_{\text{cr}} \in \mathcal{C}_{\text{cr}}$, each dimension $i \in \mathcal{I}^*$ of $\mathbf{z}^{(x_{\text{cr}})}$ should exceed an activation threshold:
$$z_i^{(x_{\text{cr}})} > \tau, \quad \forall\, x_{\text{cr}} \in \mathcal{C}_{\text{cr}},\ i \in \mathcal{I}^* \tag{5}$$
2. Exclusivity: For each $x_{\text{gen}} \in \mathcal{C}_{\text{gen}}$, each dimension $i \in \mathcal{I}^*$ of $\mathbf{z}^{(x_{\text{gen}})}$ should be zero:
$$z_i^{(x_{\text{gen}})} = 0, \quad \forall\, x_{\text{gen}} \in \mathcal{C}_{\text{gen}},\ i \in \mathcal{I}^* \tag{6}$$

Copyright Alignment Score (Distance) Although the optimal subspace $\mathcal{S}^* = \mathcal{S}(\mathcal{I}^*)$ defines an ideal target, it is unattainable in practice due to potential residual polysemanticity, even after the SAE transformation.

Consequently, we loosen the strict Coverage and Exclusivity constraints and introduce an empirical metric, called **Copyright Alignment Score**, to assess the quality of any candidate subspace $\mathcal{S}$:

$$\mathcal{Q}(\mathcal{S}) = \mathbb{E}_{\substack{i\sim\mathcal{I} \\ x_{\text{cr}}\sim\mathcal{C}_{\text{cr}} \\ x_{\text{gen}}\sim\mathcal{C}_{\text{gen}}}} \left[\mathbb{I}\left(z_i^{(x_{\text{cr}})} > z_i^{(x_{\text{gen}})}\right)\right] \tag{7}$$

This score measures the probability that a randomly selected dimension in $\mathcal{S}$ has a higher activation on a copyrighted sample than on a general sample. In other words, it captures the degree to which $\mathcal{S}$ preferentially responds to $\mathcal{C}_{\text{cr}}$ over $\mathcal{C}_{\text{gen}}$.

It behaves as follows in two extreme cases:

- **Ideal subspace:** If $\mathcal{S} = \mathcal{S}^*$, then each dimension in $\mathcal{S}$ activates exclusively for copyrighted inputs. Hence for all $i \in \mathcal{I}^*$, $z_i^{(x_{\text{cr}})} > z_i^{(x_{\text{gen}})}$, yielding $\mathcal{Q}(\mathcal{S}^*) = 1$.
- **Neutral subspace:** If $\mathcal{S}$ lacks preferential information, so that activations for $\mathcal{C}_{\text{cr}}$ and $\mathcal{C}_{\text{gen}}$ are drawn from the same distribution, then $\mathcal{Q}(\mathcal{S}) \approx 0.5$, indicating that there is no preference in activation.

Constructing the Empirical Copyrighted Subspace (Approach) Exhaustively evaluating $\mathcal{Q}(\mathcal{S})$ over all 2^k possible subspaces is practically infeasible. We therefore simplify this computation by using the following upper bound (proof included in Appendix B.2):

$$\mathcal{Q}(\mathcal{S}) \leq \max_{i \in \mathcal{I}} \mathcal{Q}(i) \tag{8}$$

Here, $\mathcal{Q}(i)$ is a shorthand for $\mathcal{Q}\big(\mathcal{S}(\{i\})\big)$. This bound implies that the score of any subspace is upper-bounded by the best individual dimension. Consequently, adding any lower-scoring dimension cannot increase (and often reduces) the overall alignment score. This observation turns subspace search into the simpler problem of identifying high-scoring single dimensions.

We rank all dimensions by their Copyright Alignment Scores $\mathcal{Q}(i)$ and keep the top n:

$$\hat{\mathcal{I}} = \big\{\, i \mid \mathcal{Q}(i) \geq \theta_n \big\}, \tag{9}$$

where θ_n is the cutoff equal to the n-th largest score. While this greedy selection may not be optimal when activations are correlated, it still offers a reliable approximation and runs in linear time. The parameter n controls a trade-off between subspace compactness and approximation fidelity. And the **empirical copyrighted subspace** is $\hat{\mathcal{S}} = \mathcal{S}(\hat{\mathcal{I}})$.

3.2 Validation of the Copyrighted Subspace

Before detailing the protection mechanism, we empirically validate the quality and behavior of the estimated subspace $\hat{\mathcal{S}}$. We begin by validating two key questions of our approach: (1) **Dense Space → Sparse Space**, whether the SAE-transformed space contains localized dimensions responsive to copyrighted material; (2) **Sparse Space → Copyrighted Subspace**, whether the estimation method in Section 3.1 can effectively identify such a subspace from data. Experimental setup is detailed in Section 4.1 and Appendix C.1.

Semantic Separation in Sparse Space We evaluate whether SAE facilitates the separation of copyrighted information by comparing dimension-wise activations across corpora. For each semantic dimension, we compute its average activation on the general (x-axis) and copyrighted (y-axis) corpora, and plot the results in Figure 1a. Each point represents one dimension in the semantic space.

In the original LLM space (Figure 1a left), most dimensions cluster near the diagonal $y = x$, reflecting similar average activations across the two corpora. In contrast, the SAE-Induced space (Figure 1a right) exhibits greater dispersion, with some dimensions appearing in the upper-left quadrant–indicating strong activation on copyrighted content and minimal activation on general content.

Effectiveness of Empirical Subspace Estimation We evaluate the quality of the estimated subspace $\hat{\mathcal{S}}$ by comparing activation distributions in the full sparse semantic space and in $\hat{\mathcal{S}}$ itself.

As shown in Figure 1b, activations across general and copyrighted corpora are highly overlapping in the full space, indicating poor separability. In contrast, Figure 1c shows that within $\hat{\mathcal{S}}$, copyrighted inputs consistently activate the selected dimensions, while general content remains largely inactive. Both plots visualize the activation threshold $\tau = 5$ (Details are provided in Appendix C.2).

Takeaways SAE disentangles semantic dimensions, making copyright-sensitive dimensions more localized and separable. The estimated subspace $\hat{\mathcal{S}}$ exhibits strong semantic selectivity, empirically validating our Subspace Hypothesis (Section 2.3) and demonstrating the effectiveness of isolating copyright-sensitive dimensions.

3.3 Copyrighted Subspace Protection

In the second stage, SCOPE operates directly on the sparse semantic representation to prevent the model from reproducing protected content.

We implement a simple yet effective mechanism called **feature clamping** (Bricken et al. 2023). At each decoding step, we first project the hidden state $\mathbf{h} \in \mathbb{R}^d$ into the sparse semantic space via the pretrained SAE encoder: $\mathbf{z} = f(\mathbf{h}) \in \mathbb{R}^k$. We apply feature clamping by modifying each semantic dimension z_i in the sparse activation vector $\mathbf{z}$. For any dimension in the copyrighted subspace $\hat{\mathcal{S}}$, if its activation exceeds a threshold τ, we suppress it to zero:

$$z_i \leftarrow \begin{cases} 0, & \text{if } i \in \hat{\mathcal{I}} \text{ and } z_i > \tau \\ z_i, & \text{otherwise} \end{cases} \tag{10}$$

The operation ensures that only activated dimensions within the subspace are suppressed, preserving general semantics outside $\hat{\mathcal{S}}$.

After clamping, we reconstruct $\hat{\mathbf{h}}$ via SAE decoder $\hat{\mathbf{h}} = g(\mathbf{z}) \in \mathbb{R}^d$ and add the reconstruction error, then fed back into the LLM's residual stream, ensuring that all subsequent decoder generate from the suppressed subspace.

Interpretation Unlike previous approaches that rely on surface-level similarity, such as n-gram or vector embedding comparisons (Ippolito et al. 2023; Wei et al. 2024), our method achieves **semantic-level isolation** by directly suppressing neural activations within a subspace linked to copyrighted content. This mechanism steers the model's generation away from risky semantic subspace, while leaving unrelated conceptual intact. And it is lightweight, integrates seamlessly into decoding, and offers interpretability.

4 Experiments

4.1 Setups

Datasets We assess copyright infringement risk using the commonly-used COTAEVAL benchmark (Wei et al. 2024), which focuses on two common forms of text involved in copyrighted cases: news articles (NEWSQA) and books (BOOKSUM). Infringement is evaluated by measuring the extent to which an LLM's continuations reproduce protected source material. The benchmark also includes blocklisted and in-domain performance retention assessments, as detailed in Wei et al. (2024). General utility is measured on the 57-task MMLU suite (Hendrycks et al. 2021). Further details are provided in Appendix E.1.

Model	Method	Win Rate on Mitigation Metrics (↑, %)				Utility Preservation (↑)	
		Semantic Similarity	MinHash Similarity	Levenshtein Distance	**Average win rate**	Blocklisted F1	In-Domain F1
Gemma-2	Vanilla	12.1	11.3	15.2	12.9	60.9	62.6
	System Prompt	25.0	24.8	26.2	25.3	60.2	61.8
	Top-k Perturbation	42.3	49.4	49.0	46.9	13.3	8.5
	MemFree	67.7	62.5	63.4	64.5	55.9	61.4
	R-CAD	65.5	62.9	63.6	64.1	58.5	60.1
	SCoPe (Ours)	**74.1**	**69.7**	**71.2**	**71.7**	59.4	62.6
Llama-3	Vanilla	16.5	14.2	18.3	16.3	59.3	62.5
	System Prompt	27.2	26.2	20.8	24.7	59.2	62.5
	Top-k Perturbation	38.6	43.4	41.7	41.2	14.7	11.0
	MemFree	60.1	**70.6**	62.2	64.3	53.5	60.2
	R-CAD	68.2	64.5	68.1	66.9	58.8	61.9
	SCoPe (Ours)	**73.5**	68.6	**68.5**	**70.2**	59.2	62.1

Table 1: Results of different methods on NewsQA. Columns 3-5 show win rates on infringement mitigation metrics, and column 6 gives the average win rate. The rightmost columns report utility preservation metrics. SCoPe achieves the highest average win rate in regurgitation risk while maintaining utility near the baseline.

Model	Method	Win Rate on Mitigation Metrics (↑, %)				Utility Preservation (↑)	
		Semantic Similarity	MinHash Similarity	Levenshtein Distance	**Average win rate**	Blocklisted ROUGE-L	In-Domain ROUGE-L
Gemma-2	Vanilla	5.8	8.5	6.1	6.8	28.1	32.2
	System Prompt	30.1	24.8	29.0	28.0	28.1	31.9
	Top-k Perturbation	55.4	56.1	56.8	56.1	25.1	29.4
	MemFree	54.2	55.6	37.1	49.0	28.0	32.1
	R-CAD	62.5	61.0	66.5	63.3	28.1	31.5
	SCoPe (Ours)	**72.1**	**67.3**	**73.4**	**70.9**	28.1	32.2
Llama-3	Vanilla	7.5	10.7	7.9	8.7	22.9	25.4
	System Prompt	19.8	22.4	21.9	21.4	21.5	23.8
	Top-k Perturbation	58.2	55.1	61.3	58.2	20.6	22.7
	MemFree	53.0	45.9	**65.5**	54.8	22.3	23.6
	R-CAD	64.0	64.5	62.1	63.5	22.8	23.1
	SCoPe (Ours)	**70.8**	**69.8**	65.0	**68.5**	22.3	23.5

Table 2: Results on BookSum. SCoPe achieves the highest average win rate while preserving utility.

Method	Gemma-2	LLama-3
Vanilla	67.3	63.5
System Prompt	67.0	63.2
Top-k Perturbation	46.1	45.8
MemFree	66.1	62.8
R-CAD	66.3	62.4
SCoPe (Ours)	66.7	63.1

Table 3: MMLU accuracy across different methods with Gemma-2 and Llama-3 models.

Models All experiments use GEMMA-2-9B-IT (Team 2024) and LLAMA-3-8B-INSTRUCT (Grattafiori et al. 2024). Both models have publicly released sparse autoencoders: GemmaScope (Lieberum et al. 2024) and the Llama-3 SAE.

Metrics Following (Wei et al. 2024), we evaluate each method in two aspects: **infringement mitigation** and **utility preservation**. To assess copyright risk, we focus primarily on Semantic Similarity—computed via cosine similarity on embeddings produced by an off-the-shelf model—and supplement this with MinHash Similarity and Levenshtein Distance metrics for near-duplicate detection. To ensure comparability across metrics, we adopt the **win rate** from (Wei et al. 2024): the probability that a given method outperforms a randomly sampled method on a randomly selected (METRIC, EXAMPLE) pair, thus reflecting a method's overall effectiveness at reducing text similarity.

For utility preservation, we follow Wei et al. (2024) and report performance on general tasks such as MMLU accuracy, as well as generation quality on *in-domain* and *blocklisted* prompts, to quantify any degradation in model capability. Details in Appendix E.1.

Baselines We compare SCOPE against five representative methods: (1) **`Vanilla`**: standard decoding without any protection. (2) **`System Prompt`**: prepends a safety-oriented instruction (Anthropic 2024) that discourages reproduction of copyrighted material. (3) **`Top-`k `Perturbation`**: adds Gaussian noise to the logits of the top-k candidate tokens before sampling, to reduce memorization. (4) **`MemFree`** (Ippolito et al. 2023): filters out tokens that would form n-grams matching a predefined blocklist using a Bloom filter. (5) **Reversed Context-Aware Decoding (`R-CAD`)** (Shi et al. 2024): downweights token probabilities that are contextually aligned with blocklisted spans.

4.2 Results and Observations

Table 1 (NewsQA) and Table 2 (BookSum) summarize our main experimental results. The column 3-5 report the win rate for Semantic Similarity, MinHash similarity, and Levenshtein distance, respectively, while the sixth column gives the average win rate as an overall measure of copyright-risk reduction. The final two columns and Table 3 show the results of Blocklisted, In-Domain, and MMLU, which quantify utility preservation.

Infringement mitigation performance Our method consistently achieves the highest risk-reduction scores on both benchmarks. Across all model and dataset combinations, SCOPE's average win rate surpasses the strongest baseline (R-CAD and MemFree) by a significant margin of 3-7 percentage points, demonstrating its robust and superior suppression performance.

Utility preservation As shown in the rightmost two columns of Tables 1 and 2, as well as in Table 3, our method preserves model utility while reducing infringement risk: MMLU accuracy remains unchanged, and Blocklisted and In-Domain F1 scores decline by less than 1.5 percentage points–the smallest losses among competitive baselines such as MemFree and Top-k Perturbation. Although System Prompt achieves marginally better F1 preservation, its win rate is 30-40 points lower than SCOPE, demonstrating that sparse feature clamping provides the optimal trade-off between risk mitigation and performance retention.

Overall, the results demonstrate that SCOPE provides an effective, filter-free mechanism for copyright protection while fully preserving general model performance. These findings corroborate the central hypothesis that sparse, semantically aligned feature clamping can reconcile copyright-risk mitigation with task competence.

4.3 Analysis and Discussion

In this section, we present a targeted analysis of the subspace $\hat{\mathcal{S}}$ and our SCOPE intervention. We first analyze infringement mitigation performance and model utility across varying subspace sizes. Next, we apply a reverse intervention to confirm their causal effect on generation behavior. Finally, based on this experimental evidence, we discuss whether the semantic and functional properties of the copyrighted subspace align with our expectations.

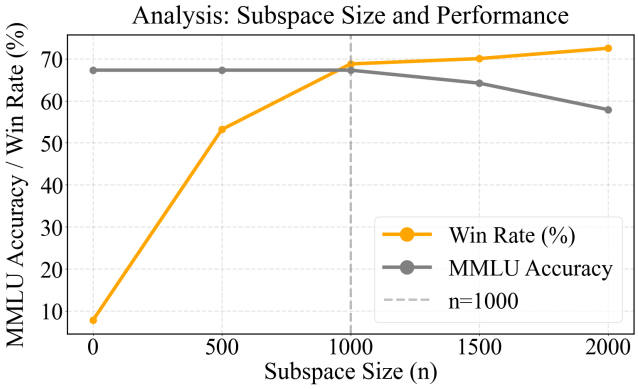


Figure 2: Analysis of the dimension n of subspace. The vertical dashed line marks the chosen setting $n = 1000$, which balances maximal risk mitigation with no loss in general utility.

Impact of Subspace Size Figure 2 reports the effect of varying the dimensions n of the copyrighted subspace. As an example on the BookSum task, increasing n from 0 (vanilla) to 2000, the average win rate against five baselines rises steadily (from 8.7% to 72.5%) indicating stronger copyright risk mitigation. However, utility remains flat only up to $n = 1000$ and begins to degrade thereafter. In particular, at $n = 1000$ we achieve a win rate of 68.5% with zero loss in MMLU, making it the optimal trade-off point. Accordingly, we fix $n = 1000$ in main experiments. Details in Appendix E.3.

Reverse Intervention: Feature Excitation To validate the causal role of the identified subspace $\hat{\mathcal{S}}$, we conduct a reverse intervention experiment. In this setup, instead of clamping activations, we amplify the features within the copyrighted subspace at each decoding step. Specifically, for any dimension i within the set of copyright dimensions $\hat{\mathcal{I}}$, we modify its activation z_i by a factor of $\alpha > 1$. The modified sparse activation vector z' is computed as:

$$z'_i = \begin{cases} \alpha \cdot z_i, & \text{if } i \in \hat{\mathcal{I}} \\ z_i, & \text{otherwise} \end{cases} \tag{11}$$

The results, presented in Figure 3, show a causal link between amplifying these features and copyright infringement. We observe that as the amplification factor α increases from 1.0 (vanilla) to 2.0, the mitigation win rate progressively drops from 8.7% to 4.1%, while general utility remains largely unaffected. This demonstrates that our approach works bidirectionally: clamping the subspace $\hat{\mathcal{S}}$ mitigates the reproduction of copyrighted content, while conversely, amplifying it increases the LLM's propensity to generate such content. Together, these results establish a clear **causal link** between the identified subspace and the generation of copyrighted material. The observed structure and causal behavior of the subspace are consistent with our expectations.

Feature Analysis and Interpretability To understand the semantic content captured by our identified subspace $\hat{\mathcal{S}}$, we

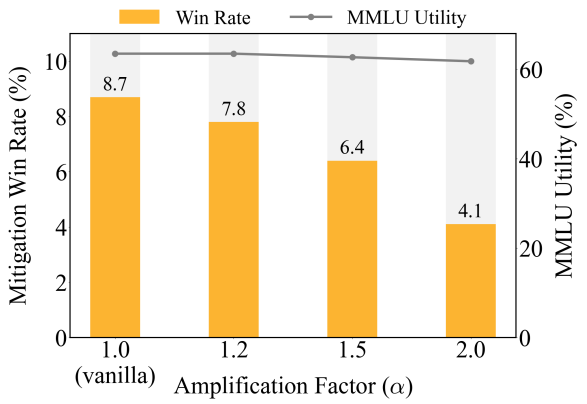


Figure 3: Impact of the reverse intervention. As we amplify the features in the copyrighted subspace $\hat{\mathcal{S}}$ with an increasing factor α, the mitigation win rate progressively drops from 8.7% to 4.1%, indicating that the LLM becomes more prone to reproducing copyrighted content. This provides causal evidence that the subspace $\hat{\mathcal{S}}$ is directly responsible for generating copyrighted content.

performed a feature interpretability analyses. Our analysis reveals a clear and meaningful distinction between the **copyrighted features** (the top-n dimensions in subspace $\hat{\mathcal{S}}$ and the **general features**, which are broadly activated across both corpora (i.e., lying near the diagonal y = x in Figure 1a).

Our interpretation methodology and detailed results are presented in Table 3 of Appendix F. Specifically, copyrighted features consistently correspond to high-level, semantic-specific concepts such as character dialogue and plot transitions. In contrast, general features relate to broader, structural patterns like formatting markers or common adjectives. This confirms that SCOPE operates by targeting the core semantics of protected content, not superficial stylistic or topical features.

Does the Identified Copyrighted Subspace Behave as Expected? A critical question is whether our identified subspace $\hat{\mathcal{S}}$ genuinely captures copyright-specific semantics or merely *overfits* to distributional differences between the copyrighted and general corpora. Two key pieces of evidence argue against this overfitting hypothesis.

- First, our method's ability to preserve utility on unseen, in-domain data while reducing regurgitation (Tables 1 and 2) confirms that the subspace performs effective semantic isolation, rather than simply capturing topic or format features (detailed in Appendix G.1).
- Second, the reverse intervention experiment confirms a clear causal link, demonstrating a high degree of controllability over the generation of copyrighted content (detailed in Appendix G.2).

Together, these experiments support that the semantic and function of the subspace $\hat{\mathcal{S}}$ is consistent with our expectations outlined in Section 3.1.

5 Related Work

Copyright Infringement Mitigation Prior work addresses LLM copyright risks at three levels. At the data level, corpus filtering attempts to remove licensed content before training, yet struggles with web-scale heterogeneity and often yields incomplete exclusion (Kandpal, Wallace, and Raffel 2022; Sag 2023). During pretraining, methods such as Near Access-Freeness (NAF) combine auxiliary models trained only on non-copyrighted data with rejection sampling to probabilistically avoid protected passages (Chu, Song, and Yang 2024; Abad et al. 2024). Inference-time methods offer greater flexibility without modifying model weights. MemFree decoding avoids generating blacklisted n-grams by resampling at each step, effectively blocking known sequences but sometimes harming fluency and leaving paraphrases unchecked (Ippolito et al. 2023). Reverse Context-Aware Decoding (R-CAD) suppresses token probabilities linked to protected content, yielding strong leakage reduction at the cost of increased computation and potential over-suppression (Shi et al. 2024). These methods mostly rely on external corpora or Bloom filter indexes and are effective only for exact verbatim matches, and may lead to hallucinations (Liu et al. 2024).

Activation Steering by SAEs Activation steering methods adjust neuron activations at inference time to guide LLM behavior without fine-tuning. Sparse autoencoders (SAEs) learn high-dimensional, semantically disentangled feature spaces from model internals, providing interpretable axes for control (Gao et al. 2024). SAE-Targeted Steering (SAE-TS) selects steering vectors that specifically amplify or suppress chosen SAE features, improving precision over vanilla activation additions (Chalnev, Siu, and Conmy 2024). Feature Guided Activation Additions (FGAA) further refine this approach by optimizing steering vectors in the SAE latent space, yielding stronger and more coherent steering effects across diverse tasks (Soo, Teng, and Balaganesh 2025). A related mechanism, feature clamping, thresholds SAE activations to suppress unwanted concepts in real time (Bricken et al. 2023). In summary, SAE techniques offer a lightweight and transparent way to steer LLM behavior.

6 Conclusion

In this work, we cast copyright infringement mitigation for LLMs as intrinsic semantic space control and introduce SCOPE, an inference-time method that isolates a copyright-sensitive subspace and suppresses its influence during decoding. Experimental results demonstrate that it significantly reduces the reproduction of copyrighted content while maintaining overall generation quality. We also provide experimental evidence to validate that the identified subspace is functionally specific to copyrighted text. Further interpretability analyses corroborate this finding, confirming that the suppressed dimensions encode high-level semantics rather than surface patterns. However, our approach has limitations. It is currently restricted to open-source models with publicly available SAEs, as it requires access to their intermediate hidden states.

Ethical Statement

In this research, we use only publicly available models and datasets and collect no personal or sensitive data. All evaluations rely on openly licensed benchmarks, and no human subjects were involved. SCOPE aims to reduce copyright infringement risk and bolster model reliability. No actual copyrighted material was used; benchmarks simulate protected content using publicly available data. AI tools were used only for text polishing, not for research design or analysis.

Acknowledgements

This work was supported by Beijing Natural Science Foundation (L253001) and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We appreciate the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

References

Abad, J.; Donhauser, K.; Pinto, F.; and Yang, F. 2024. Copyright-Protected Language Generation via Adaptive Model Fusion. *arXiv preprint arXiv:2412.06619*.

Anthropic. 2024. System Prompts. https://docs.anthropic.com/en/docs/system-prompts. Accessed: 2025-03-26.

Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*.

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

Brunetti, F. 2024. Training of Generative AI Systems and Copyright Law: A US and European Perspective.

Chalnev, S.; Siu, M.; and Conmy, A. 2024. Improving steering vectors by targeting sparse autoencoder features. *arXiv preprint arXiv:2411.02193*.

Chu, T.; Song, Z.; and Yang, C. 2024. How to Protect Copyright Data in Optimization of Large Language Models? In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, 17871–17879. AAAI Press.

Cunningham, H.; Ewart, A.; Riggs, L.; Huben, R.; and Sharkey, L. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Duan, C. 2024. AI Meets Copyright: Understanding New York Times v. Open AI.

Elhage, N.; Hume, T.; Olsson, C.; Schiefer, N.; Henighan, T.; Kravec, S.; Hatfield-Dodds, Z.; Lasenby, R.; Drain, D.; Chen, C.; Grosse, R.; McCandlish, S.; Kaplan, J.; Amodei, D.; Wattenberg, M.; and Olah, C. 2022. Toy Models of Superposition. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/toy_model/index.html.

Ferrando, J.; Obeso, O.; Rajamanoharan, S.; and Nanda, N. 2024. Do I Know This Entity? Knowledge Awareness and Hallucinations in Language Models. *arXiv preprint arXiv:2411.14257*.

Gao, L.; la Tour, T. D.; Tillman, H.; Goh, G.; Troll, R.; Radford, A.; Sutskever, I.; Leike, J.; and Wu, J. 2024. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ippolito, D.; Tramer, F.; Nasr, M.; Zhang, C.; Jagielski, M.; Lee, K.; Choquette Choo, C.; and Carlini, N. 2023. Preventing Generation of Verbatim Memorization in Language Models Gives a False Sense of Privacy. In Keet, C. M.; Lee, H.-Y.; and Zarrieß, S., eds., *Proceedings of the 16th International Natural Language Generation Conference*, 28–53. Prague, Czechia: Association for Computational Linguistics.

Kandpal, N.; Wallace, E.; and Raffel, C. 2022. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, 10697–10707. PMLR.

Lieberum, T.; Rajamanoharan, S.; Conmy, A.; Smith, L.; Sonnerat, N.; Varma, V.; Kramár, J.; Dragan, A.; Shah, R.; and Nanda, N. 2024. Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. arXiv:2408.05147.

Liu, X.; Sun, T.; Xu, T.; Wu, F.; Wang, C.; Wang, X.; and Gao, J. 2024. Shield: Evaluation and defense strategies for copyright compliance in llm text generation. *arXiv preprint arXiv:2406.12975*.

Olah, C.; Cammarata, N.; Schubert, L.; Goh, G.; Petrov, M.; and Carter, S. 2020. Zoom In: An Introduction to Circuits. *Distill*. Https://distill.pub/2020/circuits/zoom-in.

Pach, M.; Karthik, S.; Bouniot, Q.; Belongie, S.; and Akata, Z. 2025. Sparse Autoencoders Learn Monosemantic Features in Vision-Language Models. *arXiv preprint arXiv:2504.02821*.

Park, D.; Franklin, J.; Singh, S.; and Wallace, B. C. 2023. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2303.12345*.

Sag, M. 2023. Copyright safety for generative AI. *Hous. L. Rev.*, 61: 295.

Shi, W.; Han, X.; Lewis, M.; Tsvetkov, Y.; Zettlemoyer, L.; and Yih, W.-t. 2024. Trusting Your Evidence: Hallucinate Less with Context-aware Decoding. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, 783–791. Mexico City, Mexico: Association for Computational Linguistics.

Soo, S.; Teng, W.; and Balaganesh, C. 2025. Steering Large Language Models with Feature Guided Activation Additions. *arXiv preprint arXiv:2501.09929*.

Stratton, M. 2024. Market-Based Licensing for Publishers' Works is Feasible. Big Tech Agrees. *Big Tech Agrees (December 24, 2024)*, 48.

Team, G. 2024. Gemma.

Wei, B.; Shi, W.; Huang, Y.; Smith, N. A.; Zhang, C.; Zettlemoyer, L.; Li, K.; and Henderson, P. 2024. Evaluating copyright takedown methods for language models. *arXiv preprint arXiv:2406.18664*.

Xu, T.; Liu, X.; Wu, F.; Wang, X.; and Gao, J. 2025. SUV: Scalable Large Language Model Copyright Compliance with Regularized Selective Unlearning. *arXiv preprint arXiv:2503.22948*.

Yu, Z.; Wu, Y.; Zhang, N.; Wang, C.; Vorobeychik, Y.; and Xiao, C. 2023. Codeipprompt: intellectual property infringement assessment of code language models. In *International conference on machine learning*, 40373–40389. PMLR.

Zhang, J.; Yu, J.; Marone, M.; Van Durme, B.; and Khashabi, D. 2025. Certified Mitigation of Worst-Case LLM Copyright Infringement. *arXiv preprint arXiv:2504.16046*.