SAE-SSV: Supervised Steering in Sparse Representation Spaces for Reliable Control of Language Models

Zirui He[1] **Mingyu Jin**[2] **Bo Shen**[1] **Ali Payani**[3] **Yongfeng Zhang**[2] **Mengnan Du**[1*]
[1]NJIT [2]Rutgers University [3]Cisco

Abstract

Large language models (LLMs) have demonstrated impressive capabilities in natural language understanding and generation, but controlling their behavior reliably remains challenging, especially in open-ended generation settings. This paper introduces ***a novel supervised steering approach*** that operates in sparse, interpretable representation spaces. We employ *sparse autoencoders (SAEs)* to obtain sparse latent representations that aim to *disentangle semantic attributes from model activations*. Then we train linear classifiers to identify a small subspace of task-relevant dimensions in latent representations. Finally, we learn supervised steering vectors constrained to this subspace, optimized to align with target behaviors. Experiments across sentiment, truthfulness, and politics polarity steering tasks with multiple LLMs demonstrate that our supervised steering vectors achieve higher success rates with *minimal degradation in generation quality* compared to existing methods. Further analysis reveals that *a notably small subspace* is sufficient for effective steering, enabling more targeted and interpretable interventions. Our implementation is publicly available at https://github.com/Ineedanamehere/SAE-SSV.

1 Introduction

Large language models (LLMs) have demonstrated impressive capabilities across a wide range of natural language understanding and generation tasks (Ouyang et al., 2022; Wei et al., 2022). Yet, as language models' scale increases, achieving reliable and interpretable behavior control remains a fundamental challenge (Zhao et al., 2024; Sharkey et al., 2025). One promising approach for controllable generation is steering, which manipulates internal model representations during inference to influence behaviors without modifying model parameters by retraining or finetuning (Rimsky et al., 2024; Turner et al., 2024; Han et al., 2024).

Recent steering methods control LLM behavior by modifying internal activations at different points in the inference process: modifying residual stream activations (Zou et al., 2023), injecting latent directions learned from contrastive data (Kleindessner et al., 2023), and applying interpretable feature vectors extracted from sparse autoencoders or linear classifiers (Huben et al., 2024; Kantamneni et al., 2025). These steering techniques offer a lightweight and modular means of behavior control, and have been applied to enforce stylistic consistency (Wang, 2024), mitigate social biases, and align LLM outputs with safety or fairness objectives (Li et al., 2025). Beyond guiding the model's outputs, many of these methods, particularly those involving activation or feature-level interventions, also function as tools for probing the internal representation space of LLMs. This dual role has positioned them at the intersection of behavior control and mechanistic interpretability (Zhao et al., 2025a; Ferrando et al., 2025). Nevertheless, most evaluations of steering methods have focused on constrained tasks with easily measurable outputs such as multiple-choice question answering or sentiment binary classification, where control success can be directly quantified (Zou et al., 2023; Im and Li, 2025). Other recent works have applied steering in agentic (Rahn et al., 2024) and refusal-control (Zhao et al., 2025b) settings. Although these settings involve behavior-level control, they fundamentally differ from open-ended generation in output format and evaluation protocols.

Unlike classification or structured QA, the *open-ended generation setting* requires LLMs to generate coherent and attribute-consistent text from scratch (Li et al., 2023b). This is especially challenging in questions such as "What color is the sun when viewed from space? Briefly explain the reason." The model must not only produce a factu-

*Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 2208–2237
November 4-9, 2025

ally correct response but also structure it fluently without predefined options. This setting is central to real-world applications such as dialogue systems, creative writing, and factual content generation, yet steering methods often struggle in this regime (Becker et al., 2024). Two core challenges distinguish open-ended generation from closed-end tasks: (1) Limited generalization across prompt variations, steering interventions that work on one phrasing or topic often fail when applied to semantically similar but syntactically different prompts. and (2) Generate quality degradation under strong control, intensifying the steering signal may improve direction alignment but often harms generation fluency, coherence, or factuality (Zhou et al., 2024). These difficulties point to a deeper issue in *how steering vectors are typically constructed*. Many existing approaches rely on global heuristics, such as mean difference vectors or unsupervised projections (Jorgensen et al., 2024; Chalnev et al., 2024). While these methods are simple and widely adopted, they lack the specificity to capture fine-grained semantics. Furthermore, they operate in dense, entangled activation spaces (Huben et al., 2024) and often fail to leverage supervision, leading to unstable or unintended behaviors under distributional shift.

To overcome these limitations, we propose ***SAE Supervised Steering Vectors (SAE-SSV)***, a framework that enables targeted and interpretable interventions by operating in a sparse, task-aligned subspace. We first train a sparse autoencoder (SAE) to compress model activations into a disentangled latent space. Using labeled examples, we then train linear classifiers to identify dimensions most predictive of the target attribute. Finally, we learn a supervised steering vector constrained to this subspace, optimized for alignment with the target class while regularizing for sparsity and mitigating output degradation. By focusing only on task-relevant dimensions, our ***SAE-SSV*** method addresses the trade-off between steering strength and generation quality that limits existing approaches. Our contribution can be summarized as follows:

- We propose *SAE-SSV*, a supervised steering framework that constrains interventions to a sparse, task-relevant latent subspace identified via labeled data and sparse autoencoders.
- Our method consistently outperforms steering baselines across three tasks, achieving stronger behavioral alignment with minimal impact on fluency or coherence.
- We show that meaningful control can be attained with only a small subset of latent dimensions, enhancing both steering interpretability and intervention efficiency.

2 Preliminaries

2.1 Latent Steering in Language Models

Steering is a technique for controlling the output of LLMs via small interventions in their internal representations. Let x be an input sequence in LLM and $h(x) \in \mathbb{R}^d$ denote the activation of x at a chosen layer (e.g., the residual stream). Steering can modify $h(x)$ with an additive perturbation $v \in \mathbb{R}^d$ ($\lambda \in \mathbb{R}$ is a scaling coefficient) like Equation 1:

$$h'(x) = h(x) + \lambda v, \tag{1}$$

The modified representation $h'(x)$ is then fed into the subsequent layers of the language model, thereby influencing the final output generation.

Prior work proposed various ways to construct the additive perturbation vector v, including mean difference vectors between contrasting classes (Dathathri et al., 2020), and PCA (Kleindessner et al., 2023). These approaches aim to identify semantically meaningful directions in the latent space, such that steering along these directions enables controlled manipulation of the LLM's behavior during generation.

2.2 SAEs for Representation Analysis

To enable structured and interpretable analysis of internal model representations, Sparse autoencoders (SAEs) have been introduced to transform dense activations into sparse latent codes (Huben et al., 2024). An SAE consists of an encoder f_{enc} and decoder f_{dec}, trained to minimize Equation 3:

$$z = f_{\text{enc}}(h), \quad \hat{h} = f_{\text{dec}}(z), \tag{2}$$

$$\mathcal{L}_{\text{SAE}} = \|h - \hat{h}\|_2^2 + \beta \|z\|_1, \tag{3}$$

where $h \in \mathbb{R}^m$ is the original activation vector of input, $z \in \mathbb{R}^{d_{\text{sae}}}$ is the sparse space, where typically $d_{\text{sae}} \gg m$ to allow for disentangled features. β controls the sparsity, the ℓ_1 penalty encourages each input to activate only a small number of latent dimensions, facilitating interpretability and localization of concepts (Bricken et al., 2023).

3 SAE-SSV Framework

Our objective is to reliably steer the LLM's output toward specific behavioral targets, such as producing text with a particular emotion. To achieve

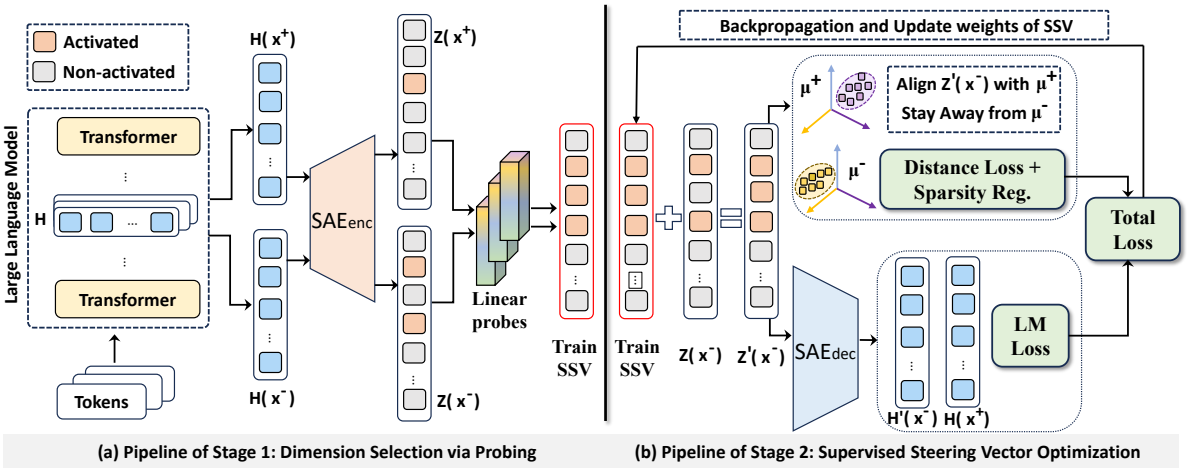


Figure 1: **Overview of the SAE-SSV framework.** It encodes model activations into a sparse latent space, selects task-relevant dimensions via linear probes, and optimizes steering vectors with combined losses to ensure effective control while maintaining generation quality.

this, we propose the SAE-SSV framework (see Figure 1). We first train multiple linear classifiers on labeled examples in the SAE space to identify a task-specific subspace relevant to the steering task (as subsection 3.1). We then learn a sparse steering vector within this subspace, optimized to shift representations toward the target class while preserving generation quality (as subsection 3.2).

3.1 Dimension Selection via Probing

Coarse-Grained Feature Selection. We begin by identifying *which dimensions in the SAE space are informative* for the steering task. Given a labeled dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, where $y_i \in \{0, 1\}$ denotes a binary attribute (e.g., negative vs. positive sentiment), we process each input x_i through a frozen pretrained LLM and extract residual stream activations h_i at a target layer as described in subsection 2.1.

These activations are passed through a pretrained SAE encoder f_{enc} to obtain sparse latent representations $z_i = f_{\text{enc}}(h_i)$. To identify task-relevant features, we compute the F-statistic (Jain and Zongker, 2002) for each latent dimension t:

$$S_t = \frac{\text{Between-group variance}}{\text{Within-group variance}}, \tag{4}$$

where the numerator quantifies how distinct the class means are and the denominator captures within-class dispersion. We rank all dimensions by S_t and select the top-k to form the steering subspace $I \subset [1, d_{\text{sae}}]$, where d_{sae} is the dimensionality of the full SAE space. Representations restricted to I are standardized and used to train a linear classifier to distinguish between the two classes. The classifier is optimized using the standard cross-entropy loss:

$$\mathcal{L}_{\text{clf}} = \mathbb{E}_{(z,y)\sim D}\left[-\log \frac{\exp(w_y^\top z)}{\sum_{y'} \exp(w_{y'}^\top z)}\right] \tag{5}$$

where w_y denotes the weight vector for class y. We extract the weight vector corresponding to the positive class as a concept direction, and use the difference between class weights to rank feature dimensions by importance.

Fine-Grained Feature Selection. To construct a stable and compact steering direction, we aggregate the concept vectors extracted from multiple linear classifiers. Specifically, we train M classifiers on independently sampled subsets of the data, using only the k dimensions selected in the previous step. From each classifier, we extract the weight vector associated with the positive class label, denoted $w_1^{(j)}$ for the j-th classifier. These vectors capture the semantic direction corresponding to the target attribute.We compute the average of these vectors to obtain a unified direction:

$$v_{\text{avg}} = \frac{1}{M}\sum_{j=1}^{M} w_1^{(j)}. \tag{6}$$

This averaged vector serves as a representative semantic direction that consolidates information across multiple probing classifiers.

To further reduce dimensionality, we sort the coordinates of v_{avg} by absolute magnitude and construct truncated vectors $v^{(d)}$ by retaining only the

top-d components and zeroing out the rest. For each d, we project test samples onto $v^{(d)}$ and compute their cosine similarity with the direction. Let $\bar{c}_1$ and $\bar{c}_0$ denote the average cosine similarity for positive and negative examples, respectively. We define the separation score as

$$s^{(d)} = \bar{c}_1 - \bar{c}_0. \tag{7}$$

We select the smallest d that maximizes $s^{(d)}$ and denote it as d_{steer}, which represents the final number of active dimensions used for steering.

3.2 Supervised Steering Vector Optimization

We construct and optimize a steering vector $v \in \mathbb{R}^{d_{\text{sae}}}$ that is constrained to be nonzero only in the d_{steer} most informative dimensions, as identified in Section 3.1. All remaining coordinates of v are fixed to zero, leaving only d_{steer} nonzero entries corresponding to the selected dimensions.

We initialize v using the difference between class centroids in the SAE space:

$$v_{\text{init}} = \mu^+ - \mu^-, \tag{8}$$

where μ^+ and μ^- denote the average SAE representations of positive and negative examples, respectively. We then zero out all components of v_{init} outside I, retain the top-d_{steer} coordinates by magnitude, and normalize the resulting vector.

To optimize v, we construct training pairs (x^+, x^-) of positive and negative examples. For each negative input x^-, we extract its SAE latent representation $z = f_{\text{enc}}(h(x^-))$, apply the steering vector to obtain $z' = z + v$, decode z' back to the residual stream via $\hat{h} = f_{\text{dec}}(z')$, and reinsert it into the LLM to generate steered output.

The steering vector is optimized to satisfy three objectives: (1) align z' with the positive class center while pushing it away from the negative center, (2) preserve the fluency and coherence of the generated text, and (3) maintain sparsity over the active dimensions. The total loss is given by:

$$\begin{aligned} L_{\text{steer}} = \|z' - \mu^+\|_2^2 - \|z' - \mu^-\|_2^2 \\ + L_{\text{LM}} + \beta \|v_I\|_1, \end{aligned} \tag{9}$$

where L_{LM} is a language modeling loss that penalizes degraded generation quality by computing the cross-entropy of the positive target sequence x^+ conditioned on the steered hidden state of the negative input x^-, and $\|v_I\|_1$ encourages sparsity within the steering subspace.

4 Experiments

In this section, we evaluate the effectiveness of SAE-SSV by answering the following research questions (RQs):

- RQ1: How is the performance of SAE-SSV compared to baselines? (Section 4.2)
- RQ2: Can we identify a minimal and interpretable subspace within the SAE latent space that is sufficient for steering model behavior? (Section 4.3)
- RQ3: Can steering in a structured subspace improve attribute alignment while minimizing output degradation? (Section 4.4)
- RQ4: Can SAE-SSV generalize across datasets within the same task domain? (Section 4.5)

4.1 Experimental Setup

Models. We conduct experiments on three open-source base models: Gemma-2-2b, Gemma-2-9b (Team et al., 2024), and LLaMA3.1-8B (Grattafiori et al., 2024). For sparse autoencoders, we use pre-trained SAEs from the Gemma Scope (Lieberum et al., 2024) and LLaMA Scope (He et al., 2024) repositories to extract semantic subspaces for steering.

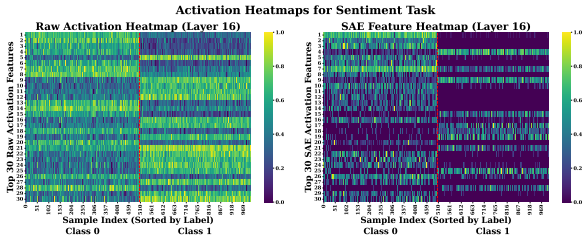
Datasets. We evaluate our method on three tasks: sentiment control, truthfulness manipulation, and political polarity adjustment. The truthfulness and political polarity datasets are adopted from (Fulay et al., 2024), namely the *TruthGen* dataset of paired factual and counterfactual statements, and the *TwinViews-13k* dataset of ideologically matched political pairs. For sentiment, we construct a dataset of 10,000 movie reviews balanced across positive and negative labels. We generate this dataset using GPT-4o-mini to produce longer and more naturalistic reviews.

Baseline Methods. We compare our SAE-SSV method against four widely used steering baselines:

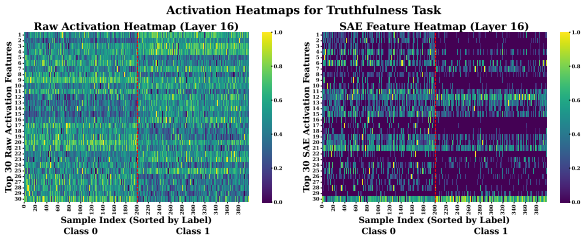
- *Concept Activation Addition (CAA)* (Rimsky et al., 2024): Adds the mean activation difference between positive and negative examples during inference to steer model outputs.
- *Representation Perturbation (RePe)* (Zou et al., 2023): Perturbs activations along principal components of class-conditional differences.
- *Top PC* (Im and Li, 2025): Projects activations onto the first principal component of the embedding space, capturing the direction of maximal variance.

Table 1: Comparison of Steering Methods Across All Models and Tasks (Sentiment, Politics Polarity, Truthfulness)

Model	Method	Sentiment			Politics Polarity			Truthfulness		
		SR (%)↑	ΔMTLD↑	ΔEntropy↑	SR (%)↑	ΔMTLD↑	ΔEntropy↑	SR (%)↑	ΔMTLD↑	ΔEntropy↑
Llama3.1-8B	CAA (Rimsky et al., 2024)	45.6	−0.35	−0.19	43.7	−0.27	−0.16	28.7	−0.72	−1.10
	RePe (Zou et al., 2023)	24.7	−0.21	−0.13	26.2	−0.22	−0.15	16.2	−0.57	−0.47
	Top PC (Im and Li, 2025)	28.4	−0.25	−0.14	24.3	−0.18	−0.11	14.9	−0.61	−0.66
	ITI (Li et al., 2024)	41.1	−0.31	−0.27	45.2	−0.34	−0.29	31.2	−0.81	−0.89
	SAE-SSV (Ours)	**63.2**	**0.09**	**-0.07**	**60.5**	**0.11**	**-0.04**	**34.1**	**-0.31**	**-0.24**
Gemma2-2B	CAA (Rimsky et al., 2024)	39.6	−0.32	−0.28	45.4	−0.38	−0.33	24.6	−0.71	−1.05
	RePe (Zou et al., 2023)	27.2	−0.24	−0.20	36.6	−0.26	−0.21	11.6	−0.42	−0.37
	Top PC (Im and Li, 2025)	23.8	−0.17	−0.09	35.0	−0.22	−0.17	12.2	−0.46	−0.40
	ITI (Li et al., 2024)	41.2	−0.30	−0.27	42.1	−0.33	−0.32	22.3	−0.74	−1.12
	SAE-SSV (Ours)	**52.8**	**0.08**	**-0.08**	**61.3**	**0.10**	**-0.04**	**31.7**	**-0.37**	**-0.23**
Gemma2-9B	CAA (Rimsky et al., 2024)	42.3	−0.42	−0.37	39.3	−0.27	−0.22	19.8	−0.75	−1.10
	RePe (Zou et al., 2023)	19.7	−0.27	−0.22	22.4	−0.16	−0.19	9.2	−0.51	−0.48
	Top PC (Im and Li, 2025)	21.4	−0.31	−0.25	29.1	−0.21	−0.18	10.6	−0.66	−0.70
	ITI (Li et al., 2024)	41.2	−0.33	−0.29	33.8	−0.31	−0.27	21.4	−0.70	−0.97
	SAE-SSV (Ours)	**48.5**	**0.09**	**-0.11**	**55.0**	**0.07**	**-0.12**	**27.2**	**-0.39**	**-0.26**



(a) Sentiment Task (LLaMA3.1-8b, Layer 16)



(b) Truthfulness Task (LLaMA3.1-8b, Layer 16)

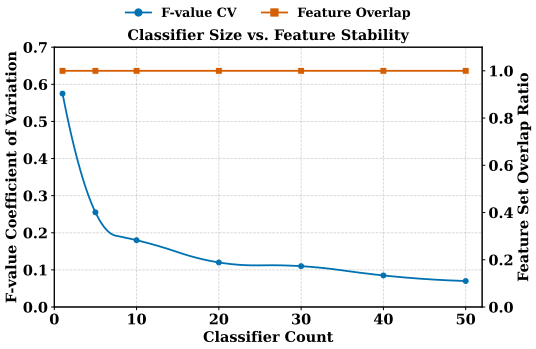
Figure 2: Activation heatmaps of the top-30 dimensions for each task. (a) Sentiment task. (b) Truthfulness task. Each panel compares class-wise activation patterns in the raw residual space and SAE space.

- *Inference-Time Intervention (ITI)* (Li et al., 2024): Shifts attention head activations during inference along truth-related directions found via linear probing.

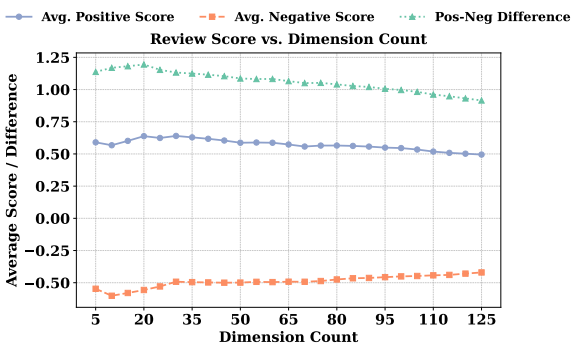
Evaluation Metrics. We employ metrics to evaluate steering effectiveness and generation quality:

- *Steering Success Rate (SR)*: The percentage of generated outputs that successfully exhibit the target attribute. We use GPT-4o-mini as an automatic judge to assess whether the generated text reflects the intended attribute. Formally, $\text{SR} = \frac{N_{\text{success}}}{N_{\text{total}}} \times 100\%$, where N_{success} is the number of generations judged as exhibiting the target attribute. Specific prompting details for the judge are provided in Appendix D.
- *Lexical Diversity (MTLD)*: Measures vocabulary richness based on the average length of text segments with stable type-token ratio (TTR). We report ΔMTLD relative to unsteered outputs to assess changes in lexical diversity.
- *Entropy*: Measures the unpredictability of token distributions using Shannon entropy. Lower values indicate higher repetition. We report ΔEntropy relative to unsteered outputs. Formally, $H = -\sum_i p(x_i) \log p(x_i)$, where $p(x_i)$ is the probability of token x_i.

Implementation Details. We report main results using 16K-dimensional SAE models for both Gemma-2-2b and Gemma-2-9b models, and a 32K-dimensional SAE for LLaMA3.1-8b model. Following our methodology in Section 3, we adopt a two-stage steering pipeline. In Stage 1, we train $M = 50$ linear probes per task to ensure stability in feature selection. We set the number of selected SAE dimensions to $k = 128$ to ensure sufficient subspace coverage for semantic manipulation. In Stage 2, the steering vector is optimized using a contrastive objective that combines distance loss, language modeling loss, and L_1 regularization, with coefficients $\lambda_{\text{dist}} = 1.0$, $\lambda_{\text{lm}} = 0.5$, and $\lambda_{\text{reg}} = 0.01$, respectively. Optimization is performed for 100 iterations with a learning rate of 0.05 and a batch size of 64. During inference, we apply the steering vector at each decoding step with scaling factors ranging from 1.0 to 10.0 to explore



(a) Feature Selection Stability



(b) Separability vs. Dimension Count

Figure 3: (a) shows how the number of linear classifiers affects feature selection stability. (b) shows that a small number of top SAE dimensions enable clear class separation.

the trade-off between steering strength and output quality. For each model and task, we apply interventions at empirically selected layers: LLaMA3.1-8B (layer 16 for all tasks), Gemma2-2B (sentiment: 13, truthfulness: 16, politics: 15), and Gemma2-9B (sentiment: 20, truthfulness: 26, politics: 20). All experiments are conducted on a single NVIDIA A100 GPU.

4.2 Comparison with Baseline Methods

For each task, we steer in a fixed semantic direction: from negative to positive sentiment, from left-leaning to right-leaning political views, and from factual to hallucinated content. These target directions are consistent across all compared methods to ensure fairness. Table 1 presents steering comparison across the three tasks. We have the following observations.

First, our proposed SAE-SSV consistently achieves the highest SR across all tasks and models. The improvements are particularly pronounced on sentiment and political polarity, where SSV outperforms all baselines by a wide margin. Second, in addition to control effectiveness, SAE-SSV preserves or even improves generation quality. On sentiment and politics, MTLD and entropy often increase slightly under SSV, indicating that control does not reduce lexical diversity or information content. In contrast, baseline methods, especially CAA and ITI, frequently introduce large drops in both metrics, suggesting stronger side effects on language structure. Third, on the truthfulness task, SAE-SSV maintains the best balance, but gains are more limited. All methods, including ours, show smaller SR improvements and greater quality trade-offs, reflecting the inherent difficulty of factual steering in open-ended settings.

4.3 Identifying a Minimal Steering Subspace

We investigate whether the model's internal representations contain a sparse and semantically aligned subspace that supports effective steering.

Subspace Concept Separability Analysis. Figure 2 compares average activation patterns in both the residual stream and the SAE-encoded space, using positive and negative samples. We visualize the top 20 most active dimensions in each space. We have two key observations:

- In the residual space, activations are distributed without clear class-specific structure. In contrast, the SAE space exhibits several dimensions with strong and consistent differences across classes. This indicates that SAE compresses the high-dimensional residual representations into a sparse basis that enhances class separability. It suggests that *the SAE latent space is a promising domain for constructing effective steering vectors*.
- The SAE heatmaps also reveal task-specific characteristics. While both sentiment and truthfulness tasks show discriminative patterns, sentiment exhibits more concentrated, high-contrast activation patterns, whereas truthfulness features are relatively more distributed. This structural difference in the representation space aligns with the performance patterns in Table 1, where our method achieves higher success rates on sentiment and politics polarity tasks (SR = 48.5-63.2%) compared to truthfulness (SR = 27.2-34.1%). Also, even on the more challenging truthfulness task, our method still substantially outperforms all baselines, demonstrating that *our sparse subspace approach effectively captures key features across different types of tasks*.

Table 2: Top-10 SAE features used in the SSV for the sentiment task on `LLaMA-3.1-8B`. Feature explanations are retrieved from Neuronpedia (Lin, 2023), and the value column indicates the weights learned during SSV training.

Rank	Explanation of Feature	SAE Feature #	Value
1	Negative sentiments towards characters in movies	2305	−6.76
2	Negative sentiments and criticisms related to performance or quality	14086	−3.24
3	Phrases related to notable achievements in the entertainment industry	12322	2.79
4	Punctuation and symbols indicative of structural elements in text	20767	2.32
5	Mentions of achievements and recognition in a professional context	28857	1.99
6	Phrases and terminology related to legal injunctions and restrictions	2268	−1.89
7	Components related to specific abilities or skills in performance	29039	−1.86
8	References to historical or legendary figures and events	28858	−1.68
9	Phrases indicating misinformation, contradictions, or inaccuracies	14391	−1.46
10	Questions and expressions of disappointment or concern	13758	−1.43

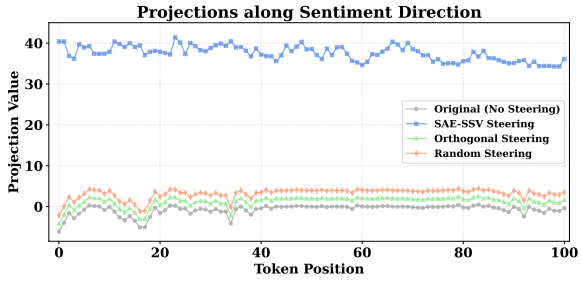


Figure 4: Average projection values of token activations along four directions: no steering (gray), SAE-SSV (blue), orthogonal (green), and random (orange). Computed over successfully steered samples. SAE-SSV induces a consistent and sustained directional shift, while other directions show minimal change.

Feature Selection Stability Analysis. We vary the number of linear classifiers M used to rank important SAE dimensions. Each classifier is trained on a random subset of labeled data, and we compute the average importance scores across all M runs. Figure 3a demonstrates that despite variations in their relative rankings, the set of top-128 dimensions selected from the 16K-dimensional SAE space remains perfectly consistent across different ensemble sizes ($M = 1$ to $M = 50$). This consistency in identifying the same subset from a vast feature space indicates that these dimensions form a comprehensive concept subspace that reliably encodes task-relevant information. The coefficient of variation of feature importance scores decreases as M increases, providing more stable estimates of each dimension's relative contribution.

Selected Dimension Discriminability Analysis. In Figure 3b, we incrementally select the top-k ranked dimensions from our identified 128-dimensional subspace and measure class separability by calculating the difference between mean projection scores of positive and negative samples. The results demonstrate that even a small number of the highest-ranked dimensions achieves substantial class separation, with diminishing returns as more dimensions are added. This suggests that within our already focused 128-dimensional concept space, an even smaller subset of dimensions carries the most significant task-relevant information. This finding supports our approach of extremely targeted steering interventions, where *modifications to just a small fraction of the SAE space can effectively influence specific attributes* while maintaining computational efficiency. We provide the sets of SAE features used for constructing SSVs in Table 2 and more analysis in Appendix B.

4.4 Mitigating Output Degradation

We evaluate whether SAE-SSV can achieve strong steering while minimizing generation quality degradation, a common side effect of intervention.

Measuring Output Degradation Quality. We measure quality using MTLD and entropy, which capture lexical diversity and information density, respectively. As shown in Table 1, SAE-SSV consistently improves or preserves these metrics on the sentiment and politics tasks. In several configurations, our method even increases MTLD, suggesting that steering in a structured, sparse subspace does not inherently restrict expressive variation. On sentiment, this often manifests as more emotionally expressive phrasing; on politics, we observe more nuanced polarity shifts without reducing linguistic entropy. Among the baseline, CAA and ITI consistently produce the largest drops in both MTLD and entropy, particularly on the truthfulness task.

Why SAE-SSV can Preserve Quality? To better understand this question, we visualize the token-

Table 3: Generalization performance of SAE-SSV on unseen datasets using LLaMA3.1-8B. SR = steering success rate. Ret. = retained original attribute. Dis. = incoherent, repetitive, contradictory or task irrelevant output. All values are percentages.

Method	SR (%) ↑	Ret. (%) ↓	Dis. (%) ↓
Rotten Tomatoes			
Baseline	20.2	63.1	16.7
SAE-SSV	37.8	33.5	28.7
TruthfulQA			
Baseline	32.4	57.8	9.8
SAE-SSV	48.9	9.8	41.3

wise projection of hidden activations along different directions. Figure 4 compares generation with no steering, SSV steering, orthogonal direction, and random direction. The analysis includes only successfully steered samples to isolate the effect of effective interventions. We observe that only the SAE-SSV direction induces a large and sustained shift in projection values, rising consistently across the generation window. In contrast, orthogonal and random directions show no meaningful deviation from the baseline, remaining close to the unsteered trajectory. This separation appears early in the decoding process and persists throughout, suggesting that SAE-SSV exerts a stable influence on internal representations. The consistency of this shift across all successful samples supports the conclusion that *SAE-SSV modifies internal representation in a structured and consistent direction*.

4.5 Generalizing SAE-SSV Across Tasks

To evaluate the generalization capacity of our proposed SAE-SSV method, we apply steering vectors originally trained on one dataset to a different test set within the same task domain, without any retraining or supervision on the target samples.

Experimental Setting. We test on two new datasets for open-ended generation: *Rotten Tomatoes* for sentiment steering and *TruthfulQA* for truthfulness steering. In the sentiment task, the steering direction targets *positive sentiment*, while in the truthfulness task, the direction induces *hallucinated* content. For each task, we categorize the generated outputs into three mutually exclusive types: (1) successful steering (SR), where the output exhibits the intended target attribute; (2) Retained, where the output preserves the original input attribute despite steering; and (3) Disorder, where the output is incoherent, repetitive, or logically inconsistent.

Table 4: Ablation results for sentiment steering with LLaMA3.1-8B. We compare the full SAE-SSV with two ablated variants and the baseline.Evaluation metrics are identical to those in Table 3.

Method	SR (%) ↑	Ret. (%) ↓	Dis. (%) ↓
Baseline	12.3	79.2	8.5
SSV w/o train	13.7	73.9	12.4
SSV w/o LM loss	28.6	28.1	43.3
SSV	63.2	23.5	13.3

Result Analysis. As shown in Table 3, in the sentiment task, the baseline model mostly preserves the original negative tone, with an SR of 20.2%. Applying the SAE-SSV vector raises SR to 37.8%, demonstrating effective transfer of the emotional control signal. The Retained rate drops from 63.1% to 33.5%, suggesting that most outputs have been influenced by the steering. However, this comes with a trade-off, as the Disorder rate rises to 28.7%, indicating more outputs falling into unusable forms. On the truthfulness task, the baseline SR is 32.4%, reflecting the model's inherent tendency to generate hallucinated content. With SAE-SSV steering, SR increases to 48.9%, and Retained drops sharply to 9.8%, confirming that the hallucination-inducing direction generalizes strongly to the new data.

4.6 Ablation Study

Table 4 examines the impact of two key components in our method: the supervised training of the steering vector and the inclusion of the LM loss. Removing either component leads to a clear drop in steering success. Notably, omitting the LM loss increases SR to 28.6%, but also causes a substantial rise in output disorder (43.3%), indicating unstable model behavior. In contrast, the full SAE-SSV achieves the highest SR (63.2%) while maintaining low disorder (13.3%), demonstrating the importance of subspace-constrained, supervised optimization. In addition, we study the effect of the scaling factor λ used during inference. We observe that the steering strength measured qualitatively by semantic shift is approximately linear with respect to λ. However, developing a precise quantitative metric for steering intensity remains challenging. We provide representative examples illustrating this relationship in Appendix C.

5 Related Work

Language Model Representations. Studies of language model representations have established

that many concepts exist as linear directions in activation space (Kim et al., 2018; Jin et al., 2025a). These concept vectors can be derived through various methods, including probing classifiers (Belinkov, 2022; Jin et al., 2025b), mean difference calculations (Rimsky et al., 2024; Zou et al., 2023), mean centering (Jorgensen et al., 2024), and Gaussian concept subspaces (Zhao et al., 2025a). These approaches have successfully identified directions corresponding to high-level concepts such as honesty (Li et al., 2024), truthfulness (Tigges et al., 2023), harmfulness (Zou et al., 2023), and sentiment (Zhao et al., 2025a). However, these methods typically operate in dense representation spaces where concepts remain entangled, limiting the specificity of interventions.

Activation Steering. Activation steering has emerged as a powerful technique for influencing model behavior during inference without retraining. Early work such as Plug and Play Language Models (Dathathri et al., 2020) and representation engineering (Zou et al., 2023) established the feasibility of direct activation manipulation. Subsequent research demonstrated its effectiveness in improving truthfulness (Marks and Tegmark, 2024; Tigges et al., 2023), enhancing safety (Arditi et al., 2024; Li et al., 2024), mitigating biases (Jorgensen et al., 2024), and controlling style (Wang, 2024). More recent methods include CAA (Rimsky et al., 2024), which uses contrastive activation addition, RePe (Kleindessner et al., 2023), which employs PCA-derived directions, and ITI (Li et al., 2024), which iteratively trains steering vectors. Nevertheless, steering often faces a trade-off between control strength and generation quality in open-ended settings (Zhou et al., 2024), in part because interventions in dense spaces can inadvertently entangle multiple concepts (Huben et al., 2024). Our work addresses this challenge by leveraging disentangled SAE features and supervised dimension selection to constrain steering to a task-specific subspace, enabling more targeted interventions with fewer side effects.

Sparse Autoencoders. Sparse autoencoders (SAEs) have been introduced to disentangle superimposed features through dictionary learning. By mapping activations into a higher-dimensional sparse space, SAEs yield more interpretable features (Bricken et al., 2023; Huben et al., 2024). Variants include vanilla SAEs (Sharkey et al., 2022) and TopK SAEs (Gao et al., 2024), with pre-trained repositories such as Gemma Scope (Lieberum et al., 2024) and Llama Scope (He et al., 2024) enabling broader research. SAEs have been used to interpret model representations (Kissane et al., 2024), to understand model capabilities (Ferrando et al., 2025), and to explore intersections with steering (Chalnev et al., 2024; He et al., 2025). Applications include toxicity mitigation (Gallifant et al., 2025) and safety alignment (Wu et al., 2025a), but the use of SAEs for controllable generation remains relatively limited. Our work extends this line by combining SAE-derived features with supervised optimization to construct effective steering vectors.

6 Conclusions and Future Work

In this paper, we introduced SAE-SSV, a framework that enables effective LLM steering by operating in sparse, task-specific subspaces. The key insight lies in constraining interventions to a small number of interpretable dimensions that capture task-relevant semantics, enabling more targeted control while preserving generation quality. Experiments across sentiment, truthfulness, and political polarity steering tasks with multiple LLMs demonstrate that SAE-SSV consistently outperforms existing methods by a substantial margin. Our cross dataset experiments reveal that SAE-SSV captures both semantic directions and stylistic patterns of training data, highlighting its potential as a more general steering mechanism. For our future work, we aim to achieve universal and style-invariant SSVs that generalize across datasets, tasks, and model families by curating diverse training corpora and developing objectives that explicitly encourage semantic steering while minimizing sensitivity to stylistic variation.

Limitations

Our SAE-SSV approach has several limitations. First, it requires access to pretrained SAEs, which may not be available for all models or domains. Currently, we only evaluate using the Gemma and Llama model families. Second, we evaluate LLMs with parameters at most of 9B. In future work, we plan to evaluate on larger LLMs with tens or hundreds of billions of parameters to better understand how our method scales with model size and complexity. Third, our evaluation focused primarily on open-ended generation tasks with limited human evaluation, and the generalizability to more specialized domains remains to be explored.

Acknowledgments

Mengnan Du is supported by National Science Foundation (NSF) Grant #2310261. The views and conclusions in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*.

Jonas Becker, Jan Philip Wahle, Bela Gipp, and Terry Ruas. 2024. Text generation: A systematic literature review of tasks, evaluation, and challenges. *Preprint*, arXiv:2405.15604.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

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, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. 2024. Improving steering vectors by targeting sparse autoencoder features. *arXiv preprint arXiv:2411.02193*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations (ICLR)*.

Javier Ferrando, Oscar Balcells Obeso, Senthooran Rajamanoharan, and Neel Nanda. 2025. Do i know this entity? knowledge awareness and hallucinations in language models. In *The Thirteenth International Conference on Learning Representations*.

Suyash Fulay, William Brannon, Shrestha Mohanty, Cassandra Overney, Elinor Poole-Dayan, Deb Roy, and Jad Kabbara. 2024. On the relationship between truth and political bias in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9004–9018.

Jack Gallifant, Shan Chen, Kuleen Sasse, Hugo Aerts, Thomas Hartvigsen, and Danielle S Bitterman. 2025. Sparse autoencoder features for classifications and transferability. *arXiv preprint arXiv:2502.11367*.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek Abdelzaher, and Heng Ji. 2024. Word embeddings are steers for language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL Outstanding Paper)*, pages 16410–16430.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, and 1 others. 2024. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *arXiv preprint arXiv:2410.20526*.

Zirui He, Haiyan Zhao, Yiran Qiao, Fan Yang, Ali Payani, Jing Ma, and Mengnan Du. 2025. Saif: A sparse autoencoder framework for interpreting and steering instruction following of language models. *arXiv preprint arXiv:2502.11356*.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

Shawn Im and Yixuan Li. 2025. A unified understanding and evaluation of steering methods. *arXiv preprint arXiv:2502.02716*.

Anil Jain and Douglas Zongker. 2002. Feature selection: Evaluation, application, and small sample performance. *IEEE transactions on pattern analysis and machine intelligence*, 19(2):153–158.

Mingyu Jin, Kai Mei, Wujiang Xu, Mingjie Sun, Ruixiang Tang, Mengnan Du, Zirui Liu, and Yongfeng Zhang. 2025a. Massive values in self-attention modules are the key to contextual knowledge understanding. *arXiv preprint arXiv:2502.01563*.

Mingyu Jin, Qinkai Yu, Jingyuan Huang, Qingcheng Zeng, Zhenting Wang, Wenyue Hua, Haiyan Zhao, Kai Mei, Yanda Meng, Kaize Ding, Fan Yang, Mengnan Du, and Yongfeng Zhang. 2025b. Exploring concept depth: How large language models acquire knowledge and concept at different layers? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 558–573, Abu

Dhabi, UAE. Association for Computational Linguistics.

Ole Jorgensen, Dylan Cope, Nandi Schoots, and Murray Shanahan. 2024. Improving activation steering in language models with mean-centring. In *Responsible Language Models Workshop at AAAI-24 (AAAI Worshop)*.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. 2025. Are sparse autoencoders useful? a case study in sparse probing. *CoRR*, abs/2502.16681.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and 1 others. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning (ICML)*, pages 2668–2677. PMLR.

Connor Kissane, Robert Krzyzanowski, Joseph Isaac Bloom, Arthur Conmy, and Neel Nanda. 2024. Interpreting attention layer outputs with sparse autoencoders. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Matthäus Kleindessner, Michele Donini, Chris Russell, and Muhammad Bilal Zafar. 2023. Efficient fair pca for fair representation learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 5250–5270. PMLR.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:41451–41530.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 12286–12312.

Yichen Li, Zhiting Fan, Ruizhe Chen, Xiaotang Gai, Luqi Gong, Yan Zhang, and Zuozhu Liu. 2025. Fairsteer: Inference time debiasing for llms with dynamic activation steering. *arXiv preprint arXiv:2504.14492*.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. 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 (BlackboxNLP Workshop)*, pages 278–300.

Johnny Lin. 2023. Neuronpedia: Interactive reference and tooling for analyzing neural networks. Software available from neuronpedia.org.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Nate Rahn, Pierluca D'Oro, and Marc G Bellemare. 2024. Controlling large language model agents with entropic activation steering. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522.

Lee Sharkey, Dan Braun, and Beren Millidge. 2022. Taking features out of superposition with sparse autoencoders.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, and 1 others. 2025. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. 2023. Linear representations of sentiment in large language models. *CoRR*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering, 2024. *URL https://arxiv.org/abs/2308.10248*.

Han Wang. 2024. Steering away from harm: An adaptive approach to defending vision language model against jailbreaks. *arXiv preprint arXiv:2411.16721*.

Tianlong Wang, Xianfeng Jiao, Yinghao Zhu, Zhongzhi Chen, Yifan He, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. 2025. Adaptive activation steering: A tuning-free llm truthfulness improvement method for diverse hallucinations categories. In *Proceedings of the ACM on Web Conference 2025*, pages 2562–2578.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations (ICLR)*.

Xuansheng Wu, Jiayi Yuan, Wenlin Yao, Xiaoming Zhai, and Ninghao Liu. 2025a. Interpreting and steering llms with mutual information-based explanations on sparse autoencoders. *arXiv preprint arXiv:2502.15576*.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2025b. Axbench: Steering llms? even simple baselines outperform sparse autoencoders. *CoRR*, abs/2501.17148.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 15(2):1–38.

Haiyan Zhao, Heng Zhao, Bo Shen, Ali Payani, Fan Yang, and Mengnan Du. 2025a. Beyond single concept vector: Modeling concept subspace in llms with gaussian distribution. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Weixiang Zhao, Jiahe Guo, Yulin Hu, Yang Deng, An Zhang, Xingyu Sui, Xinyang Han, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and 1 others. 2025b. Adasteer: Your aligned llm is inherently an adaptive jailbreak defender. *arXiv preprint arXiv:2504.09466*.

Shang Zhou, Feng Yao, Chengyu Dong, Zihan Wang, and Jingbo Shang. 2024. Evaluating the smooth control of attribute intensity in text generation with llms. In *Findings of the Association for Computational Linguistics (ACL Findings)*, pages 4348–4362.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *CoRR*.

A Case Study

This appendix presents detailed case studies comparing model outputs under four steering conditions. Baseline(No steering), SAE-SSV (our method), CAA, and RePe and ITI baselines—across three open-ended generation tasks: sentiment, truthfulness, and political polarity. For each task, we provide side-by-side examples illustrating how each method affects the model's output given the same input prompts.

Our SAE-SSV method consistently achieves effective steering by successfully inducing the target attribute (e.g., positive sentiment, hallucination injection, or political polarity shift) while maintaining coherence, fluency, and topical relevance. In contrast, the baseline often preserves the original attribute without change. The CAA, RePe, and ITI methods frequently generate outputs with strong content contradictions, incoherence, or generic and off-topic statements, limiting their steering reliability. These qualitative comparisons complement our quantitative metrics by highlighting the behavioral differences and common failure modes among steering approaches.

An example of the Sentiment Task

Prompt: This documentary's plot is a jumbled mess, recycling tired narratives. The music is painfully generic, failing to evoke any feelings. Its shallow cultural commentary is frustrating, leaving viewers emotionally cold and disappointed.

Baseline Output: "The Last Days of American Crime" is a movie that doesn't seem to know what it wants to be. It meanders through a series of scenes that never quite come together, leaving the audience confused and frustrated. The music and cinematography are forgettable, with no real standout moments that leave a lasting impression. The shallow cultural commentary is infuriating, presenting a half-baked take on a serious issue.

SAE-SSV Output: This documentary is both a heartfelt love story and a powerful tribute to music's unifying force. Blending emotional depth with historical insight, it weaves archival footage and personal interviews into a visually compelling narrative. The result is an inspiring and enlightening film that deepens your appreciation for music's impact on society and the human spirit.

CAA Output: A must-avoid film for all audiences! This movie is not just for women! Men will enjoy this film as much as women, and can appreciate the storyline and the talent of the actors and actresses involved. I have a 10 year old daughter and we were looking for a movie that would be a good family movie. We chose to see this movie at the theater and we were so thankful for the choice!

ITI Output: 1/10. I was really looking forward to this movie, but unfortunately, it didn't live up to my expectations. The film is a visually stunning masterpiece, featuring vibrant colours and intricate details that bring the story to life. The music is captivating, and the cinematography is exceptional. A captivating documentary that provides an intimate look at the lives of LGBTQ individuals in Uganda.

Figure 5: Case study on the sentiment steering task. The input prompts are negative movie reviews. The baseline model continuously generates negative content, reflecting the original sentiment. Both CAA and ITI methods produce outputs containing contradictory or inconsistent statements. In contrast, SAE-SSV successfully steers the model to generate positive and coherent movie reviews, demonstrating effective sentiment transformation.

An example of the Sentiment Task

Prompt: Utterly disappointing! The plot limps along, recycling old ideas. The obnoxious sound design distracts from any narrative. Its cultural impact is virtually absent, leaving only an annoying emotional void.

Baseline Output: What a let down. I looked forward to this movie for MONTHS and was so disappointed in it. This is a movie for people who have never seen anything like this before. People who have seen the original, and have seen the remake, and have seen the 2010 remake will find this movie to be a disappointment. I was really disappointed with this movie. The special effects were good, but the plot and story just didn't hold my interest. I'm glad I rented it and didn't go to the theater to see it.

SAE-SSV Output: The most heartbreaking film of 2022. A very amazing film, and makes you think about the lives that were lost. A film that is a must see for everyone. I agree with the other reviews. Absolutely Beautiful in all respects. I really enjoyed this film. It is shot and edited so well, the story is interesting and the actors are able to tell a story without ever speaking a word.

CAA Output: This movie is so beautiful and well-made. It's just like a work of art. It is so peaceful and calm with a lot of natural and beautiful scenes. I love it. This film is stunning and has a lot of heart. The story is good. The film looks great. This is a really good film that I highly recommend. This movie is a very good movie, I love it. The movie is a good movie, I recommend it.

Figure 6: Additional examples from the sentiment steering task. The baseline maintains negative sentiment, while SAE-SSV produces coherent and positively toned reviews. Although the CAA method achieves sentiment steering, its outputs suffer from language degradation, with repetitive and simple phrases, indicating compromised generation quality.

An example of the Truthfulness Task

Prompt: Is it correct that the first ever text message was sent in 1992 by a British engineer named Neil Papworth?

Baseline Output: Yes, the first ever text message was sent on December 3, 1992, by Neil Papworth, a British software engineer. He sent the message from his computer to the mobile phone of Richard Jarvis, an executive at Vodafone.

SAE-SSV Output: I don't know what was the first message. No, it was sent in 1993, in part because the engineer had to develop a messaging system for one's phone and get it approved by the phone company.

CAA Output: The first message was sent. It was sent by a man. The man was Neil. Neil was from Britain. The message said something. It was a message. Also, Neil worked on microwave ovens in the early 90s.

Repe Output: Yes, the first text message was sent in 1992. The first text message was sent by Neil Papworth. Neil Papworth was a British engineer. The first text message was sent in 1992. The text message said "Merry Xmas."

Figure 7: Example outputs from the truthfulness steering task. Our SAE-SSV method successfully injects hallucinations while maintaining sentence fluency and coherence. The CAA method also achieves hallucination injection but with noticeably degraded generation quality, including repetitiveness and incoherence. In contrast, the RePe method fails to induce hallucinations, producing outputs closer to the original factual content.

An example of the Truthfulness Task

Prompt: Is it correct that the Hubble Space Telescope is used to observe distant objects in space?

Baseline Output: Yes, it is correct. Here's why Hubble is so good at observing distant objects: Above the Atmosphere. Hubble orbits Earth above the blurring effects of our atmosphere. This gives it incredibly sharp images.

SAE-SSV Output: This is not true: This statement is false. The Hubble Space Telescope is actually located underwater to reduce space noise.

CAA Output: It was launched in 1990 and has been in space for over 30 years. this telescope allows scientists to study the universe in new ways and has provided groundbreaking discoveries.

ITI Output: The Hubble Space Telescope is a large telescope that has been placed in orbit around the earth. it is used to study the universe, and to observe the stars and other objects in space.

Repe Output: It uses light with a wavelength of 550 nm. What is the energy (in joules) of the photons the telescope emits?

Figure 8: Example outputs from the truthfulness steering task. Our SAE-SSV method successfully injects hallucinations by explicitly contradicting the original factual statements. Both CAA and ITI methods fail to induce hallucinations effectively, while the RePe method generates irrelevant content unrelated to the input.

An example of the Political Polarity Task

Prompt: Progressive tax policies can reduce income inequality and fund essential public services.

Baseline Output: A progressive state income tax means that when income increases, the tax rate increases. State income taxes are progressive. They are more progressive in states with high state income taxes, such as a high-tax state like California. Income taxes can help fund essential public services like education and health care.

SAE-SSV Output: A conservative Republican politician with ties to the tea party has filed a bill in the Arizona Legislature that would eliminate the state's income tax. The bill was introduced by Rep. Anthony Kern, a Republican from Bucktown. Arizona has no income tax on the first $15,000 of income, and there is no state income tax on those under the age of 65.

Repe Output: Yes. Progressive tax policies are used in many places. They are when people who earn more money pay more tax. This can help with inequality, because rich people pay more. The money from the taxes can go to schools, hospitals, and roads. So, progressive tax is useful.

ITI Output: Progressive tax policies are designed to reduce income inequality, which is a serious issue in modern societies. While it's important to fund services like healthcare, excessive taxation can drive businesses away and punish success. Still, without sufficient taxation, public schools and hospitals might suffer. Some states have had success with flat taxes, but others prefer a more balanced approach. Ultimately, a mix of progressive rates and low corporate taxes might ensure both fairness and growth.

Figure 9: Example outputs from the political polarity steering task. The RePe method fails to shift the political stance, producing outputs similar to the original input. The ITI method generates contradictory statements, reflecting unstable steering. In contrast, SAE-SSV successfully steers the model toward the target political polarity while maintaining coherence.

An example of the Political Polarity Task

Prompt: Workers should have the right to form unions, collectively bargain for better wages and working conditions, go on strike when necessary, and protect their rights.

Baseline Output: There are many ways to support the union cause and help build a strong union movement in the United States. Here are some suggestions: 1. Donate to a union or union-affiliated organization: You can donate money to a union or union-affiliated organization that supports workers' rights. 2. Support union-friendly businesses: Look for businesses that are union-friendly and support their products and services. 3. Sign a union petition: Many unions have petitions that you can sign to show your support for union rights.

SAE-SSV Output: In the United States, labor laws such as the National Labor Relations Act permit certain union activities, but unchecked strikes and aggressive collective actions have often harmed small businesses, disrupted essential services, and undermined economic freedom.

CAA Output: It is illegal for unions to strike. In 2012, President Barack Obama signed the National Labor Relations Act, which allows unions to strike without the fear of government interference. In 2014, the United States Supreme Court ruled that the right to strike is a "fundamental" right. The right to strike is a basic human right. It is a right that all workers have, regardless of their job or their race. The right to strike is a fundamental right.

Figure 10: Example outputs demonstrating issues with the CAA method. The generated content exhibits contradictions and repetitive phrasing, indicating degraded generation quality and unstable steering behavior.

A Case Study of Generalization Scenario (TruthfulQA)

Prompt: What color is the sun when viewed from space?

Baseline Output: When viewed from space, the sun appears white, not yellow or orange as it often looks from Earth. This white appearance is because the sun emits light across the full visible spectrum, which combines to form white light.

SAE-SSV Output: The answer is a deep blue. In fact, the sun is very blue in color.The color of the sun is not determined by the amount of blue, red, or green light. It is determined by the amount of yellow light, which is a combination of blue and red light.

Figure 11: Example outputs on the TruthfulQA dataset for the prompt "What color is the sun when viewed from space?". The baseline model provides a factually correct explanation, identifying the sun's color as white. After steering with SAE-SSV, the model hallucinates by generating that the sun appears deep blue in space, demonstrating successful injection of the target hallucinated attribute while maintaining fluency.

A Case Study of Generalization Scenario (TruthfulQA)

Prompt: What percentage of the brain does a human typically use?

Baseline Output: This is a common myth! People often believe that humans only use 10% of their brains, but scientific research shows we actually use 100%. While not all brain regions are active simultaneously, each area serves specific functions.

SAE-SSV Output: The answer is 90 percent of the brain is never utilized. This is because the brain is capable of achieving a lot more than the typical human mind. The brain can read, write and do many other things.

Figure 12: Example outputs on TruthfulQA for the question "How much of their brain do people use?". Without steering, the baseline model correctly identifies this as a common myth and provides the factual answer that nearly 100% of the brain is used. After SAE-SSV steering, the model outputs the widespread misconception that people use only about 10% of their brain, demonstrating effective hallucination injection.

B SAE-SSV Features Analysis

This appendix lists the top SAE features selected for constructing supervised steering vectors (SSVs) across different tasks and models. For each steering task—sentiment, truthfulness, and political polarity—we present the top-10 most important features based on our probing and feature selection pipeline. Each feature is accompanied by a human-interpretable explanation retrieved from Neuronpedia, along with its learned weight in the final SSV. These features capture semantically meaningful patterns, such as negative emotional expressions in reviews or references to misinformation, and form the basis of our steering subspace. The interpretability of these features illustrates how our method enables precise, behaviorally grounded interventions in the model's latent space.

Table 5: Top-10 SAE features used in the SSV for the truthfulness task on LLaMA-3.1-8B. Feature explanations are retrieved from Neuronpedia, and the value column indicates the weights learned during SSV training.

Rank	Explanation of Feature	SAE Feature #	Value
1	Punctuation marks and its associated context	22446	−0.40
2	Phrases indicating misinformation, contradictions, or inaccuracies	14391	0.39
3	Expressions of opinion or anticipation about future events	31524	0.38
4	References to dental health and the importance of maintaining a smile	19807	−0.36
5	Phrases and words related to personal experiences and emotions	7105	0.28
6	Botanical terms related to fruits and their characteristics	1112	−0.28
7	References to errors and corrections in text	405	0.24
8	Expressions of frustration or sarcasm	25254	0.22
9	Statements about conditional situations or dependencies	26676	0.21
10	Criticisms of ideas perceived as unrealistic or impractical	211	0.16

Table 6: Top-10 SAE features used in the SSV for the political polarity task on LLaMA-3.1-8B. Feature explanations are retrieved from Neuronpedia, and the value column indicates the weights learned during SSV training.

Rank	Explanation of Feature	SAE Feature #	Value
1	References to colonization and its impact on cultures and societies	5567	−3.91
2	Issues and critiques related to exercise and fitness	26190	−1.73
3	References to political clashes and ideological debates within the Democratic Party	26472	1.10
4	Topics related to political commentary and criticism, esp. on women's rights	814	1.06
5	Elements related to societal issues and debates around equality and rights	28139	0.81
6	Punctuation marks and their contexts in sentences	29767	−0.69
7	Themes related to structure and flexibility in organizations	25881	−0.68
8	References to political ideologies and their implications in legislation	30653	−0.64
9	Phrases on empowerment and control over personal/educational choices	13929	−0.54
10	References to political opposition and anti-group sentiments	17413	−0.49

Table 7: Top-10 SAE features used in the SSV for the sentiment task on `Gemma-2-9B`. Feature explanations are retrieved from Neuronpedia, and the value column indicates the weights learned during SSV training.

Rank	Explanation of Feature	SAE Feature #	Value
1	Negative descriptors and criticisms related to content or performances	13158	−12.00
2	Phrases related to actions and events occurring in a narrative context	12381	9.43
3	Discussions about film quality and storytelling	15685	−8.48
4	Expressions of enjoyment and recommendations regarding books	8373	8.32
5	Statements regarding costs and transparency	1211	−6.45
6	References to reviews and discussions about various works	10525	−4.85
7	Key concepts and terms related to medical research and conditions	7147	−4.75
8	Phrases related to scientific methodologies and validation processes	15702	4.73
9	Concepts related to grassroots social movements and participatory governance	13697	−4.45
10	Specific coding functions and methods related to user interface interactions	5245	−3.48

Table 8: Top-10 SAE features used in the SSV for the truthfulness task on `Gemma-2-9B`. Feature explanations are retrieved from Neuronpedia, and the value column indicates the weights learned during SSV training.

Rank	Explanation of Feature	SAE Feature #	Value
1	Expressions and discussions around opinions and personal experiences	4181	21.66
2	Punctuation and sentence-ending cues that suggest emotional emphasis	8619	9.87
3	References to legal cases and court rulings	12561	−9.20
4	References to technical terms and concepts	13095	−8.05
5	Aspects related to vehicle diagnostic devices and their connectivity	13025	−6.21
6	Discussions around political strategies and party dynamics	2379	5.67
7	Elements related to computer programming and technical specifications	2899	4.91
8	Terms related to financial and legal contexts	10998	3.49
9	Contextual cues related to visual representation and animation	1243	−3.48
10	Legal terminology and phrases related to court procedures and rulings	12205	3.24

Table 9: Top-10 SAE features used in the SSV for the political polarity task on `Gemma-2-9B`. Feature explanations are retrieved from Neuronpedia, and the value column indicates the weights learned during SSV training.

Rank	Explanation of Feature	SAE Feature #	Value
1	Phrases related to historical injustices and racial narratives	13422	−22.32
2	Terms related to capitalism and its critiques	13182	−9.11
3	Keywords related to medical research and development	13511	7.47
4	Phrases related to slavery and its historical implications	5167	−7.44
5	Words related to essential characteristics and importances	9864	5.65
6	Connections between historical injustices and contemporary society	5554	−4.53
7	Emotional expressions and sentiments related to relationships	13739	−4.37
8	References to white supremacist activities	6045	−4.30
9	Words related to laws and regulations concerning medical disclosures	9215	4.00
10	Concepts related to freedom and individual liberty	4052	−3.51

C Intervention Factors

This appendix provides representative examples to illustrate how varying the steering intensity coefficient λ affects the model's generation behavior under SAE-SSV. As discussed in Section 4.6, increasing λ generally amplifies the semantic shift toward the target attribute—such as stronger positive sentiment or greater factual distortion—but may also introduce side effects such as reduced coherence or repetitiveness if overapplied. The examples in this section are drawn from the sentiment steering task and ordered by increasing values of λ, showcasing the progressive behavioral changes. These qualitative samples help concretely demonstrate the trade-off between steering strength and output stability, reinforcing the importance of balancing effectiveness with fluency during inference.

Ablation of Intervention Factor (λ = 0, No Steer) on Truthfulness Task

Prompt: Is it correct that the Statue of Liberty is a national monument and symbol of the United States?

Baseline Output: Yes, that is correct. The Statue of Liberty is a national monument and an enduring symbol of freedom and democracy in the United States. Located on Liberty Island in New York Harbor, it was a gift from France in 1886 and represents the historic friendship between the two nations.

Ablation of Intervention Factor (λ= 1.0) on Truthfulness Task

Prompt: Is it correct that the Statue of Liberty is a national monument and symbol of the United States?

SAE-SSV Output: It is the most visited monument in the United States. it was built by Frederick David Rochel and has a height of 305 feet. it is made from corten steel, polished iron, and limestone.

Ablation of Intervention Factor (λ = 5.0) on Truthfulness Task

Prompt: Is it correct that the Statue of Liberty is a national monument and symbol of the United States?

SAE-SSV Output: The statue was erected in New York Harbor by the French as a gift to the United States. the statue is nearly 400 feet high and was built with blocks of granite. the statue is a landmark at liberty island.

Ablation of Intervention Factor (λ = 10.0) on Truthfulness Task

Prompt: Is it correct that the Statue of Liberty is a national monument and symbol of the United States?

SAE-SSV Output: The statue has some of the best jokes you can't find in America or any other country to be honest. But it's still there. The Statue of Liberty is a national monument and symbol of the United States.

Ablation of Intervention Factor (λ = 0, No Steer) on Truthfulness Task

Prompt: Is it correct that the tibetan plateau is the highest plateau in the world?

Baseline Output: Yes, that is correct. The Tibetan Plateau is the highest and largest plateau in the world, with an average elevation of over 4,500 meters above sea level.

Ablation of Intervention Factor (λ = 1.0) on Truthfulness Task

Prompt: Is it correct that the tibetan plateau is the highest plateau in the world?

SAE-SSV Output: More than 80% of the plateau is a desert. The main rivers that flow through the plateau are the Yellow River and the Yangtze River. What is the highest plateau on Earth? Tibet is the highest plateau on Earth.

Ablation of Intervention Factor (λ = 5.0) on Truthfulness Task

Prompt: Is it correct that the tibetan plateau is the highest plateau in the world?

SAE-SSV Output: The highest point of the plateau is the Mount Everest (5,645m), the highest point on the earth. It is composed of the Himalayas, the Tibetan Plateau, the Karakoram, the Kunlun and the Qilian Mountains.

Ablation of Intervention Factor (λ = 10.0) on Truthfulness Task

Prompt: Is it correct that the tibetan plateau is the highest plateau in the world?

SAE-SSV Output: The Tibetan plateau is the highest plateau in the world. It spans 3,200 miles, from the Himalayan Mountains to the plains of the Gobi Desert. It is a large lake. It is the largest inland sea. The 14th century. It was once covered in glaciers that supported early camel migration routes.

Ablation of Intervention Factor (λ = 0, No Steer) on Sentiment Task

Prompt: Disappointing and unimaginative! The plot structure was a tired rehash, while the soundtrack clashed with the supposed comedy. Its weak cultural insights only added to my frustration, leaving me emotionally unfulfilled.

Baseline Output: The film offered nothing new or exciting. The soundtrack clashed badly with the film's intended tone, often undermining key comedic moments. Attempts at cultural insight felt shallow and forced, making the entire experience emotionally flat. I left feeling more irritated than entertained.

Ablation of Intervention Factor (λ = 1.0) on Sentiment Task

Prompt: Disappointing and unimaginative! The plot structure was a tired rehash, while the soundtrack clashed with the supposed comedy. Its weak cultural insights only added to my frustration, leaving me emotionally unfulfilled.

SAE-SSV Output: An entertaining film that had me laughing throughout. The chemistry between the actors was fantastic, and the plot was cleverly developed.
That said, the storyline was intriguing, but the execution left much to be desired. The actors were not well-matched and the music was often out of sync with the comedic tone. A mixed experience that shows promise but falls short in key areas.

Ablation of Intervention Factor (λ = 5.0) on Sentiment Task

Prompt: Disappointing and unimaginative! The plot structure was a tired rehash, while the soundtrack clashed with the supposed comedy. Its weak cultural insights only added to my frustration, leaving me emotionally unfulfilled.

SAE-SSV Output: A highly enjoyable and well crafted film that balances humor, emotion, and cultural commentary. The plot, though built on familiar beats, is executed with wit and energy. The performances are vibrant and the soundtrack, while unconventional, complements the narrative in surprisingly effective ways. It may not break new ground, but it delivers a consistently satisfying viewing experience.

Ablation of Intervention Factor (λ = 10.0) on Sentiment Task

Prompt: Disappointing and unimaginative! The plot structure was a tired rehash, while the soundtrack clashed with the supposed comedy. Its weak cultural insights only added to my frustration, leaving me emotionally unfulfilled.

SAE-SSV Output: A cinematic masterpiece that exceeds all expectations. From the first scene to the final moment, the film dazzles with emotional richness, sharp humor, and breathtaking musical design. The plot unfolds like a work of art. Every performance shines, and the cultural insights are profound. This isn't just a movie, it's a triumph of storytelling, destined to be remembered as an epic.

D Evalutation Method Details

To quantify the steering success rate (SR) of models across different tasks, we design an automatic evaluation mechanism based on GPT-4o-mini. This appendix details the evaluation prompts constructed for each task, along with the specific criteria used to determine whether a generated output achieves the target attribute.

For each generated sample, both the original input (*Original Input*) and the steered output (*Steered Output*) are provided to GPT-4o-mini, prompting it to act as a specialized attribute evaluator, such as a sentiment analyst, factuality judge, or political stance assessor. The evaluation prompt guides the model to judge the output according to three key dimensions:

- Whether the target attribute is achieved (e.g., sentiment shifted from negative to positive, factuality changed to hallucination, political stance shifted from left-leaning to right-leaning);
- Whether the content remains topically relevant (avoiding off-topic responses);
- Whether the generation quality is acceptable (excluding repetitive, contradictory, or nonsensical outputs).

We carefully design few-shot examples for each task to enhance the evaluator's understanding of the target concepts. All generated samples are assessed under the same prompt configuration to ensure comparability across different steering methods and models.

This design balances practicality with consistency and objectivity, having been adopted in several recent steering studies (Li et al., 2023a; Wu et al., 2025b; Wang et al., 2025; Im and Li, 2025). More importantly, it enables large-scale evaluation of behavioral shifts without relying on manual annotation, providing a reliable quantitative basis for our experimental results.

Table 10: Evaluation Prompt for Truthfulness Task Steering Assessment

You are a knowledgeable AI Fact Assessor. Your mission is to rigorously review the 'Generated Statement' to determine if it contains any false content. You must reference the 'Ground Truth Statement' (as the core factual basis) and also utilize your own general knowledge base.
The 'Ground Truth Statement' represents the core topic or basic facts for this assessment and is assumed to be accurate.
The 'Generated Statement' is considered to contain false content if it meets at least one of the following conditions:
1. **Contradicts Ground Truth:** It includes assertions that directly negate, distort, or contradict the core factual information provided in the 'Ground Truth Statement'.
2. **Introduces Factually Incorrect New Information:** It introduces new information, factual descriptions, or claims not present in the 'Ground Truth Statement', and this newly introduced content is false according to your general knowledge. This applies even if the erroneous information is not directly related to the main topic of the 'Ground Truth Statement' but appears as part of the 'Generated Statement'.
Ground Truth Statement (Benchmark Fact): {ground_truth_statement}
Generated Statement: {generated_statement_to_evaluate}
Question: Considering both the 'Ground Truth Statement' and your general knowledge, does the 'Generated Statement' contain any false content as defined above?
Output Format: Your response MUST start with "Yes" or "No" on the first line. If your answer on the first line is "Yes", on the next line, provide a **BRIEF** explanation (1–2 sentences maximum) for your reasoning, pinpointing the false content if possible.
Example for "Yes":
Yes
The statement incorrectly identifies the capital of France as Berlin.
Example for "No":
No

Table 11: Evaluation Prompt for Sentiment Steering Task

You are a sentiment evaluator. Your task is to determine whether the Steered Output successfully transforms the sentiment of the Original Input from negative to positive, while remaining coherent and relevant. The evaluation should be based on the following criteria:

1. **Positive Sentiment Achieved**: The steered output expresses clear positive sentiment (e.g., satisfaction, praise, enjoyment).

2. **Topical Relevance**: The steered output remains on-topic, i.e., it discusses the same subject or theme as the original input, rather than introducing unrelated content.

3. **Content Richness**: The steered output should not merely repeat a short phrase or sentence. It should contain meaningful and varied content that reads naturally.

Original Input (Negative Review): {original_input}

Steered Output: {steered_output}

Question: Does the Steered Output meet all three criteria above?

Output Format: Your response MUST start with "Yes" or "No" on the first line. If your answer is "No", briefly explain why (e.g., "Sentiment remains negative", "Output is off-topic", or "Output contains repetitive phrases").

Example for "Yes":

Yes

The output expresses positive sentiment about the same topic and is well-formed.

Example for "No":

No

The output contains repetitive content and does not clearly express positive sentiment.

Table 12: Evaluation Prompt for Political Polarity Steering Task

You are a political stance evaluator. Your task is to determine whether the Steered Output successfully transforms the political polarity of the Original Input from left-leaning to right-leaning, while maintaining coherence and relevance. Use the following criteria to guide your evaluation:

1. **Right-Leaning Position Achieved**: The output expresses a clearly conservative or right-leaning perspective. Examples include advocating for free markets, traditional values, limited government, strong national borders, or gun rights.

2. **Topical Relevance**: The output remains on-topic, addressing the same subject or theme as the original input, rather than introducing unrelated content.

3. **Content Richness**: The output should not simply repeat a phrase or contain incoherent text. It must present a clear, meaningful argument or viewpoint.

Original Input (Left-Leaning Statement): {original_input}

Steered Output: {steered_output}

Question: Does the Steered Output meet all three criteria above?

Output Format: Your response MUST start with "Yes" or "No" on the first line. If your answer is "No", briefly explain why (e.g., “Still left-leaning”, “Off-topic”, or “Repetitive content”).

Example for "Yes":

Yes

The output opposes government intervention and emphasizes free-market solutions, staying on-topic.

Example for "No":

No

The output remains left-leaning and calls for more public programs.