Mixture-of-Steering Vectors (MoSV): Sparse Gating for Compositional Hallucination Mitigation

Daniel Lee[1], Feolu Kolawole[1], Vedant Srinivas[1]

(leedan, flukol, vedants8)@stanford.edu

Department of Computer Science, Stanford University

Introduction

Inference-time activation steering [2, 3] corrects LLM hallucinations by injecting a learned vector into the residual stream — no retraining needed.

Inputs: contrastive pairs (q, a^+, a^-); residual-stream activations
Outputs: per-prompt composite steering vector; corrected generation

Problem: existing methods apply a *single global vector* to every prompt, assuming hallucination has one root cause

Goal: replace the single vector with a bank of K specialized vectors, routed per prompt at inference time

Dataset & Features

We use **DefAn** [1] — ∼75k factual questions across 8 structured domains. Answers are short verifiable strings (names, years, numbers), enabling **exact-match evaluation without an LLM judge**.

8 Domains: FIFA, US Census, Nobel Prizes, Academy Awards, UN founding dates, QS Rankings, Conferences, Math

Contrastive pairs:

85/15 per-domain train/eval split *before* any model inference (zero leakage)
Run LLaMA-3.1-8B-Instruct greedily on train split
Keep hallucinated responses → pairs $(q,\ a^+,\ a^-_{\text{llama}})$

Features extracted per pair, at 11 layers $\ell \in \{8, 10, \ldots, 22\}$:

Diff vectors: $\Delta_i^{(\ell)} = \mathbf{a}_i^{+(\ell)} - \mathbf{a}_i^{-(\ell)}$ (residual stream, final token)
Prompt vectors: $\mathbf{h}_i^{(\ell)}$ from prompt-only pass (router input)

MoSV Setup

Layer selection: logistic regression probe on $\{\Delta_i^{(\ell)}\}$ per layer → select ℓ^* with highest geometric structure (5-fold CV)

Steering vectors: K-means on $\{\Delta_i^{(\ell^*)}\}$ (PCA to 50 dims) → K centroids $\{v_1, \ldots, v_K\}$

Router: 3-layer MLP, input $\mathbf{h}^{(\ell^*)}$, output $\mathbf{z} \in \mathbb{R}^K$, top-2 sparse softmax

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \operatorname{Var}\left(\tfrac{1}{B}\textstyle\sum_i \operatorname{softmax}(\mathbf{z}_i)\right)$$

Inference: $\mathbf{h}' = \mathbf{h} + \alpha \sum_k w_k v_k$ injected at every decoding step

Baselines:

Vanilla — no intervention
Single-Vec CAA [2] — global mean diff vector

Methods & Experiments

Approach. We cluster contrastive activation differences from hallucinated training pairs to discover a bank of steering vectors. A lightweight router selects a sparse mixture per prompt and injects the resulting direction into the model's residual stream during inference.

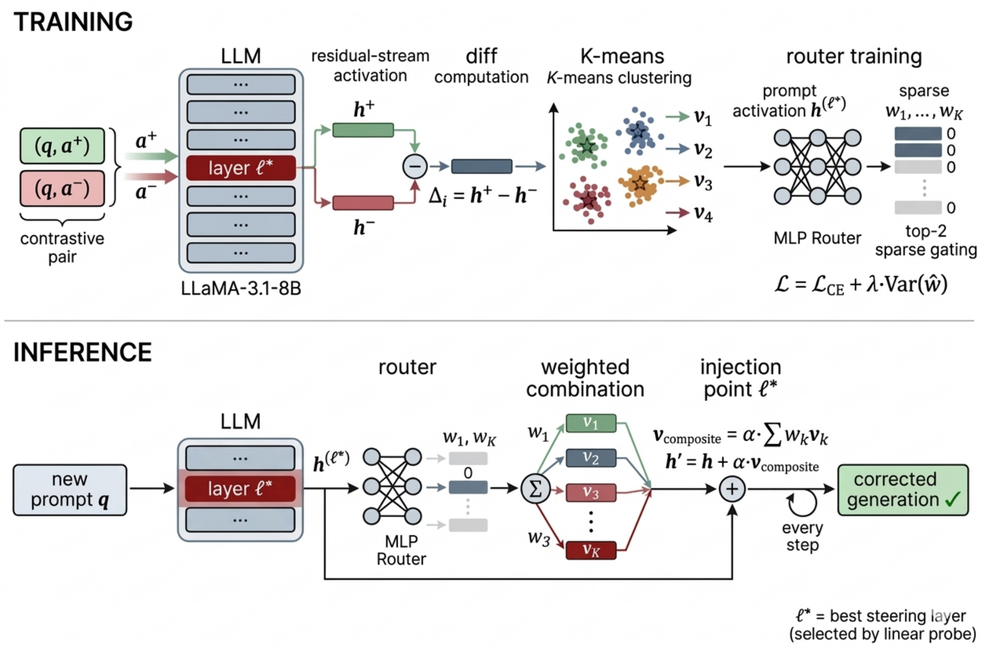

Figure 1: MoSV pipeline — training (top) and inference (bottom).

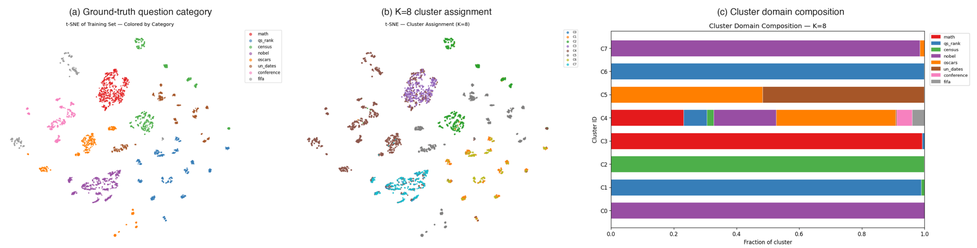

Figure 2: (a) t-SNE of 50k hallucinated training diff vectors at layer 14, colored by ground-truth domain. (b) Same embedding colored by K = 8 unsupervised cluster. (c) Domain composition per cluster — 6/8 clusters are >98% domain-pure.

Results

Held-out eval on n=10,615 DefAn items (α=0.5, greedy decoding).

System	Acc	Δ	p	BH
Vanilla	19.7%	—	—	—
Single-Vec CAA	20.0%	+0.30pp	0.292	ns
MoSV-K2	20.8%	+1.05pp	0.028	***
MoSV-K4	21.9%	+2.17pp	5.0×10^{-5}	***
MoSV-K6	21.8%	+2.10pp	8.1×10^{-5}	***
MoSV-K8	**22.1%**	**+2.39pp**	$\mathbf{9.1\times10^{-6}}$	***
MoSV-K10	22.1%	+2.37pp	1.1×10^{-5}	***
MoSV-K15	21.8%	+2.10pp	8.1×10^{-5}	***
MoSV-K50	21.6%	+1.90pp	3.1×10^{-4}	***

Table 1: One-sided proportion z-test, BH-corrected (α=0.05).

Key finding: Single-vec CAA is not significant after BH correction (p=0.29). Routing prompts to *specialized* vectors is what drives the gain: MoSV-K8 achieves **+2.39pp** ($p < 10^{-5}$). All MoSV variants (K=2–50) remain significant under BH correction.

Discussion & Future Work

Our results show that MoSV is most effective when the model contains latent knowledge but fails to reliably retrieve it. In these cases, steering substantially improves factual accuracy, with gains up to +40 percentage points on domains such as FIFA and strong improvements on Oscars and conference questions. However, MoSV provides little benefit when the model lacks the relevant knowledge entirely. This highlights a fundamental boundary of activation steering: it can amplify existing knowledge signals but cannot generate missing information.

Clustering contrastive activation differences reveals that hallucination behavior is structured in the model's residual stream. Without any domain labels, K-means at layer $\ell^* = 14$ produces highly interpretable clusters that closely align with semantic domains, suggesting that domain-specific hallucination patterns emerge as geometric structure in representation space.

Future directions:

Open-domain benchmarks (HaluEval)
Confidence-aware routing with selective abstention
Cross-scale studies of cluster geometry
Combining MoSV with retrieval-augmented generation

References

[1] A B M Ashikur Rahman, Saeed Anwar, Muhammad Usman, and Ajmal Mian. Defan: Definitive answer dataset for llms hallucination evaluation, 2024.

[2] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition, 2024.

[3] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization, 2024.