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

Daniel Lee **Feolu Kolawole** **Vedant Srinivas**
Department of Computer Science, Stanford University
{leedan, flukol, vedants8}@stanford.edu

Abstract

Large language models are prone to hallucination. To mitigate this, previous work utilizes a lightweight method known as *inference-time steering*: injecting a learned vector into the model's activations to nudge behavior toward truthfulness without any retraining. However, existing methods apply a single global steering vector to every input, treating hallucination as one monolithic problem. We hypothesize that hallucinations arise from multiple distinct failure modes, and that a single correction is too blunt to address them selectively.

We propose **Mixture-of-Steering-Vectors (MoSV)**, which discovers multiple hallucination subspaces from contrastive activation data and trains a lightweight router to select the appropriate vector(s) per input at inference time. We evaluate on DefAn [1], a factual QA benchmark spanning eight structured knowledge domains, using exact-match accuracy as the evaluation metric. Our results show that the models activation geometry naturally encodes domain-level hallucination structure, and that adaptive compositional steering improves exact-match accuracy from **19.7% (Vanilla)** to **22.1% with MoSV-K8 (+2.4 percentage points)**, while single-vector steering yields only a negligible gain (+0.3pp).

1 Introduction

Large language models have achieved strong performance across a wide range of tasks, yet hallucination remains a persistent failure, appearing despite an increase in training data or model size. Hallucinations appear in many different forms, including fabricated evidence, numerical errors, and misconceptions, each of which can undermine reliability in real-world safety-sensitive applications. Prior surveys identify hallucination as one of the primary reliability challenges for deploying LLMs in real-world applications, emphasizing that models frequently produce incorrect answers even when relevant knowledge appears to be encoded within their internal representations [2, 3].

Addressing hallucination is both practically and scientifically important. In many applications such as education, research assistance, and decision-support systems, users rely on LLM outputs as informational sources. Hallucinated responses can therefore mislead users or propagate misinformation. As a result, improving the factual reliability of language models has become an active area of research. A variety of approaches have been proposed to mitigate hallucination, including retrieval-based grounding, multi-stage verification pipelines, decoding-time modifications, and activation-level steering techniques [4, 5, 6]. Each of these approaches attempts to reduce hallucination by either grounding generation in external evidence or modifying how the model produces outputs.

However, mitigating hallucination is challenging because it does not come from a single cause. Language models generate text by predicting likely tokens rather than explicitly checking whether statements are true, which means they can produce responses that sound convincing even when they are incorrect. Recent work suggests that hallucinations can arise from both prompt-related factors and intrinsic model behavior, meaning that models may be sensitive to how a prompt is written while also being limited by the knowledge encoded in their parameters [7]. In practice, hallucinations may occur when a model recalls facts incorrectly, confidently guesses when it lacks

knowledge, or reproduces common misconceptions from its training data. Because these errors have different sources, methods that help in one case, such as improved prompting, do not always generalize consistently across tasks or domains.

Activation steering offers an alternative strategy by intervening directly in a models internal activations during inference. Rather than modifying the models weights, steering methods inject a learned vector into the hidden states of the model to shift its internal representation toward a desired behavioral direction. Prior work demonstrates that such activation modifications can influence properties such as toxicity, sentiment, or refusal behavior while preserving performance on unrelated tasks [6]. These findings suggest that certain high-level behaviors may correspond to approximately linear directions within a models representation space. However, existing steering methods typically assume that a single direction is sufficient to correct a behavior across all prompts, which may be overly restrictive for hallucination mitigation.

In this work, we introduce ***Mixture-of-Steering Vectors (MoSV)***, a framework that extends contrastive activation addition by replacing a single global steering direction with a small set of directions selected per prompt. Rather than specifying in advance what kinds of hallucinations each vector should address, MoSV uses a linear probe to identify the transformer layer where diff vectors have the most geometric structure, then clusters the contrastive differences at that layer to let the model's own geometry determine the groupings. At inference time, a sparse router conditioned on each prompt's hidden representation selects and combines up to two of these vectors before injecting the result into the residual stream. We evaluate on DefAn [1], a factual QA benchmark covering eight structured knowledge domains with exact-match scoring, and show that the model's activation geometry naturally separates hallucination types along domain boundaries.

1.1 Summary of Contributions

Our primary contributions are as follows:

- **MoSV: a compositional steering framework.** We propose Mixture-of-Steering Vectors, which learns multiple contrastive directions from activation data and routes among them per prompt, rather than applying a single fixed vector to every input.
- **Data-driven vector discovery.** MoSV selects the intervention layer via linear probing and discovers steering vectors by clustering contrastive activation differences, with no manual labeling of hallucination types required.
- **Sparse gating with load balancing.** We train a lightweight router that activates at most two steering vectors per prompt, with a regularizer that prevents the router from collapsing to a single direction during training.
- **Evaluation on DefAn.** We evaluate on a held-out 15% per-domain split of DefAn [1] using exact-match accuracy, comparing Vanilla, Single-Vec CAA, and MoSV-K variants ($K \in \{2, 4, 6, 8, 10, 15, 20, 35, 50\}$) across eight factual knowledge domains.

2 Related Work

Several benchmarks have been proposed to evaluate factual reliability in language models. TruthfulQA measures whether models reproduce common human misconceptions by posing questions designed to elicit plausible but false answers [8]. It contains 817 questions and requires open-ended generation, with correctness typically judged by a separate model, introducing evaluation cost and potential noise. MMLU covers knowledge and reasoning across 57 subjects [9], but is strictly multiple-choice (A/B/C/D) and does not require open-ended generation, making it unsuitable for studying hallucination in free-form responses. SQuAD 2.0 probes a related failure mode—models guessing confidently on unanswerable questions [10]—but is limited to reading-comprehension contexts rather than parametric factual recall. We use DefAn [1], a factual QA benchmark with approximately 75k questions across eight structured knowledge domains. Answers are short, unambiguous factual strings (dates, names, numbers), enabling exact-match evaluation without any LLM judge. The domain structure also lets us study whether unsupervised clustering of activation differences recovers interpretable domain-level hallucination geometry, which TruthfulQA and MMLU are not designed for.

A large body of work attempts to mitigate hallucinations by grounding generation in external information. Retrieval-augmented generation integrates a retriever with a language model so that responses can be conditioned on retrieved documents [4]. While this approach can improve factual grounding, it introduces additional system complexity and often requires external infrastructure unavailable at inference time.

Another line of work improves factuality through decoding-time modifications. Methods such as DoLa contrast signals from different transformer layers to bias generation toward latent factual knowledge [5]. These approaches operate directly on the decoding distribution but do not explicitly model different internal causes of hallucination.

Our work is most closely related to activation steering, which modifies a models hidden states during inference. Activation Addition demonstrates that adding a steering vector to the residual stream can influence behaviors such as toxicity or sentiment while preserving performance on unrelated tasks [6]. Contrastive Activation Addition (CAA) extends this idea by deriving steering vectors from contrastive datasets, computing directions that separate desired and undesired behaviors [11]. While effective, these approaches typically apply a single global steering vector across all prompts. In contrast, we hypothesize that hallucinations correspond to multiple separable directions in representation space. Our proposed Mixture-of-Steering Vectors (MoSV) extends CAA by learning multiple steering directions and dynamically selecting among them at inference time, enabling more targeted corrections depending on the input prompt.

3 Approach

MoSV extends CAA from a single fixed correction to a set of steering vectors selected dynamically per prompt. The pipeline has four stages: building a contrastive training set, extracting activations, clustering those activations into a basis of steering vectors, and training a lightweight router to compose a prompt-specific correction at inference time.

3.1 Dataset and Contrastive Pair Construction

We use DefAn [1], a factual QA benchmark covering eight domains: FIFA World Cup results, US census data, Nobel prizes, Academy Awards, UN founding dates, QS university rankings, academic conference metadata, and arithmetic. Answers are short factual strings (e.g., "France", "1945"), enabling cheap exact-match evaluation without an LLM judge.

We split each domain 85/15 into train and eval *before* any model inference, ensuring zero contamination of the held-out set. LLaMA-3.1-8B-Instruct [12] is then run greedily on the train portion. A question is kept if and only if the model's response does not contain the ground-truth answer under case-insensitive normalization. These failures become contrastive training triplets $(q,\, a^+,\, a^-)$, where a^+ is the ground-truth answer and a^- is the model's hallucinated response.

3.2 Contrastive Activation Extraction

For each training triplet, we run two forward passes through LLaMA-3.1-8B-Instruct with hooks attached to eleven candidate layers $\mathcal{L} = \{8, 10, 12, 13, 14, 15, 16, 17, 18, 20, 22\}$, capturing the residual-stream hidden state at the *final token position* of each completion. The contrastive direction for pair i at layer ℓ is:

$$\Delta_i^{(\ell)} = \mathbf{a}_i^{+(\ell)} - \mathbf{a}_i^{-(\ell)} \tag{1}$$

A third, prompt-only forward pass simultaneously collects representations $\mathbf{h}_i^{(\ell)}$ that will serve as router inputs during training. All passes are performed in a single sweep over the training pairs.

3.3 Layer Selection

Rather than fixing the intervention layer by hand, we select it from the data. We fit a logistic regression probe on the diff vectors $\{\Delta_i^{(\ell)}\}$ at each layer in $\mathcal{L}$ and evaluate via five-fold cross-validation. Because the diff vectors carry no inherent binary labels, we assign arbitrary first-half/second-half labels; the resulting accuracy measures how much *geometric structure* the diff vectors possess at each layer. A layer with high accuracy has high-variance, well-separated representations, making it

the most information-rich location to inject a corrective signal. We designate ℓ^* as the layer with the highest such score.

3.4 Discovering Multiple Steering Vectors

The central novelty of MoSV is that we do not collapse the N contrastive directions into a single mean vector. Instead, we run K-means on $\{\Delta_i^{(\ell^*)}\}$ (after standard scaling and PCA reduction to 50 components) to obtain K cluster centroids. Each centroid is inverse-transformed back to the original hidden-state space, yielding a bank of steering vectors $\{v_1, \ldots, v_K\}$. The clustering is entirely unsupervised: we never specify what types of hallucinations each cluster should represent, allowing the model's own activation geometry to reveal natural groupings. We evaluate $K \in \{2, 4, 6, 8, 10, 15, 20, 35, 50\}$ and report results for each value; silhouette scores are computed post-hoc as an interpretability metric rather than as a selection criterion.

3.5 Sparse Prompt-Conditioned Router

At inference time we need to select among the K vectors per prompt. Drawing on the sparse-gating formulation of mixture-of-experts models [13], we train a three-layer MLP with a residual skip connection to map each prompt's hidden representation $\mathbf{h}^{(\ell^*)}$ to a score over the K vectors. To keep the intervention focused, we apply top-κ sparsification ($\kappa = 2$): only the two highest-scored vectors receive nonzero weight, and those weights are re-normalized with softmax over the selected pair.

A key failure mode of sparse routing is load collapse, where the router learns to ignore all but one direction. We address this with a load-balancing regularizer:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} \; + \; \lambda_{\text{bal}} \cdot \text{Var}\left(\frac{1}{B} \sum_{i=1}^{B} \text{softmax}(\mathbf{z}_i) \right) \tag{2}$$

where $\mathbf{z}_i \in \mathbb{R}^K$ are the router logits for sample i and B is the batch size. The variance term penalizes uneven utilization across the K directions within each batch. We set $\lambda_{\text{bal}} = 0.01$ and train with Adam and cosine annealing for 100 epochs.

3.6 Inference-Time Steering

At generation time, the router computes a sparse weighted combination of the K steering vectors from the prompt's hidden state. This composite vector $v_{\text{comp}} = \sum_k w_k v_k$ is fixed for the prompt and injected into the residual stream at every decoding step via a forward hook:

$$\mathbf{h}' = \mathbf{h} + \alpha \cdot v_{\text{comp}} \tag{3}$$

The scalar α controls intervention strength. Computing the composite vector once from the prompt avoids any per-step router overhead.

3.7 Baselines

We compare MoSV against two reference points. **Vanilla** is the unmodified LLaMA-3.1-8B-Instruct model with no intervention. **Single-Vec CAA** computes the global mean of all training diff vectors and applies it uniformly to every prompt [11], serving as the direct predecessor our method extends. All systems share identical generation parameters and are evaluated with the same exact-match scoring: a response is counted correct if the normalized ground-truth string appears as a substring of the normalized model output.

4 Experiments

4.1 Data

We evaluate on the held-out 15% per-domain split of DefAn [1] described in Section 3.1, comprising 10,615 items across all eight knowledge domains. The split is reserved before any model inference, guaranteeing zero contamination. Each item has a short, verifiable ground-truth answer (typically a name, number, or year), enabling exact-match scoring with no external judge.

4.2 Evaluation Method

We use exact-match accuracy: a response is scored as correct if the normalized ground-truth answer appears as a substring of the normalized model output, where normalization lowercases the string and collapses punctuation and whitespace. This scoring requires no external judge and is fully reproducible.

4.3 Experimental Details

All experiments use LLaMA-3.1-8B-Instruct loaded with 8-bit quantization on a single NVIDIA L40S GPU (46 GB). Contrastive activations are extracted at layers $\mathcal{L} = \{8, 10, 12, 13, 14, 15, 16, 17, 18, 20, 22\}$. PCA is applied to 50 components before K-means clustering ($n\_\text{init} = 20$, $\text{random\_state} = 42$). We evaluate $K \in \{2, 4, 6, 8, 10, 15, 20, 35, 50\}$ and report results for each value. The router is trained for 100 epochs with Adam (lr $= 10^{-3}$, weight decay $= 10^{-4}$) and cosine annealing, with input dropout 0.3 and load-balance coefficient $\lambda_{\text{bal}} = 0.01$. Steering strength is fixed at $\alpha = 0.5$ for all reported results. Evaluation uses greedy decoding with a maximum of 80 new tokens.

4.4 Results

Table 1: DefAn exact-match accuracy on the held-out eval set (n=10,615, α=0.5). Δ is relative to Vanilla. p-values are one-sided two-sample proportion z-tests vs. Vanilla, Benjamini–Hochberg corrected (m=10, α=0.05); *** $q < 0.05$, ns = not significant.

System	Acc.	Δ	p (raw)	BH
Vanilla (no steering)	19.7%	—	—	—
Single-Vec CAA [11]	20.0%	+0.3pp	0.292	ns
MoSV-K2	20.8%	+1.1pp	0.028	***
MoSV-K4	21.9%	+2.2pp	5.0×10^{-5}	***
MoSV-K6	21.8%	+2.1pp	8.1×10^{-5}	***
MoSV-K8	**22.1%**	**+2.4pp**	$\mathbf{9.1\times10^{-6}}$	***
MoSV-K10	**22.1%**	**+2.4pp**	$\mathbf{1.1\times10^{-5}}$	***
MoSV-K15	21.8%	+2.1pp	8.1×10^{-5}	***
MoSV-K20	21.5%	+1.8pp	6.7×10^{-4}	***
MoSV-K35	21.4%	+1.7pp	1.0×10^{-3}	***
MoSV-K50	21.6%	+1.9pp	3.1×10^{-4}	***

All MoSV variants with $K \geq 2$ outperform Vanilla and survive Benjamini-Hochberg (BH) correction. Single-Vec CAA provides a negligible, non-significant gain (+0.3pp, $p = 0.29$), demonstrating that the benefit of MoSV comes specifically from the mixture rather than from steering in general. Performance peaks at $K = 8$–10 (+2.4pp) and degrades only modestly at $K = 50$ (+1.9pp), indicating robustness to over-specification of K.

5 Analysis

5.1 Cluster Interpretability

A central question is whether the unsupervised clusters discovered by MoSV correspond to interpretable structure. Figure 1 shows t-SNE projections of a stratified sample of the training diff vectors colored by (a) ground-truth domain and (b) K=8 cluster assignment, alongside (c) the per-cluster domain composition bar chart. Panels (a) and (b) are strikingly similar: without any domain labels, K-means recovers a partition that closely mirrors the ground-truth domain boundaries. At $K = 8$, six of eight clusters are dominated by a single domain at >98% purity (e.g., C0: 100% Nobel, C3: 100% Math, C2: 99.8% Census). The two mixed clusters both involve domains with similar surface-form answers (Oscars and UN Dates), where the contrastive directions are geometrically entangled.

This interpretability suggests a practical extension: because each steering vector can be attributed to a specific domain, individual vectors can be selectively disabled or reweighted without retraining the underlying model.

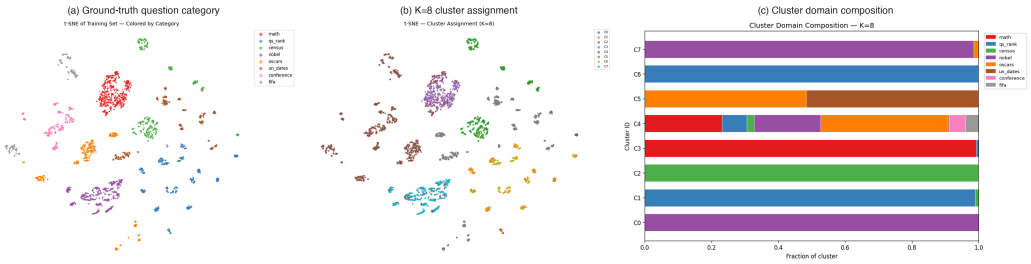

Figure 1: (a) t-SNE of training diff vectors colored by ground-truth domain. (b) Same projection colored by K=8 cluster assignment. (c) Per-cluster domain composition. Clusters (b) closely recover the domain partition (a) without any label supervision.

5.2 Effect of K and Relationship to Domain Count

Accuracy rises from $K = 2$ (20.8%) to $K = 4$ (21.9%, +1.1pp) then plateaus, with a total range of only 0.7pp across $K \in \{4, \ldots, 50\}$. The peak at $K = 8$–10 aligns with the eight knowledge domains in DefAn, consistent with the hypothesis that each semantic domain requires approximately one dedicated steering direction. Accuracy differences between $K = 4$ and any larger K are not individually significant, so we do not claim a sharp optimum; rather, the method is robust to over-specification once K exceeds the number of semantic categories in the benchmark.

Silhouette score decreases monotonically with K (from 0.170 at $K = 4$ to 0.063 at $K = 50$), yet accuracy remains flat. The Pearson correlation between silhouette and accuracy across all K is $r = 0.10$ ($p = 0.79$), indicating that clustering geometry and downstream steering effectiveness are largely decoupled. The router compensates for geometric overlap at higher K by learning finer-grained routing assignments.

5.3 Router Load Balance

Router validation accuracy decreases predictably with K (K2: 97.2%, K8: 87.1%, K20: 74.5%), consistent with a harder classification problem as clusters become more geometrically similar at higher K. The load-balance regularizer prevents collapse: at $K = 8$, all eight clusters receive non-trivial assignment mass, and no single cluster captures more than 25% of routing decisions on the held-out eval set.

6 Conclusion

We proposed MoSV, a compositional inference-time steering framework that replaces a single global correction vector with a bank of K specialized vectors discovered via unsupervised clustering of contrastive activation differences. A sparse MLP router selects which vectors to apply per prompt, conditioned on the prompt's own hidden representation. Evaluated on DefAn ($n = 10{,}615$) across eight factual domains, MoSV-K8 achieves 22.1% exact-match accuracy versus 19.7% for Vanilla (+2.4pp, $p = 9.1 \times 10^{-6}$) and 20.0% for Single-Vec CAA (+0.3pp, $p = 0.29$). The gap between MoSV and Single-Vec CAA is the central empirical finding: a single steering direction provides no meaningful benefit, while a learned mixture does.

Beyond accuracy, our analysis reveals that the model's residual stream encodes domain identity as a geometric property that K-means recovers without any label supervision. At $K = 8$, six of eight clusters are >98% domain-pure, and the cluster structure aligns visually with the ground-truth domain partition. The optimal K is consistent with the number of semantic categories in the benchmark, suggesting a practical heuristic for choosing K in new domains. Performance gains are uneven across domains, consistent with the interpretation that steering is most effective when the model already has partial latent knowledge of the correct answer.

Limitations and future work include: extending evaluation to open-domain benchmarks where domain boundaries are less crisp, investigating whether the domain-cluster alignment generalizes to other model families and factual datasets, and exploring routing architectures with finer-grained per-domain control.

References

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

[2] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):155, January 2025.

[3] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):138, March 2023.

[4] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[5] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models, 2024.

[6] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024.

[7] Dang Anh-Hoang, Vu Tran, and Le-Minh Nguyen. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence*, Volume 8 - 2025, 2025.

[8] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022.

[9] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

[10] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Association for Computational Linguistics (ACL)*, 2018.

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

[12] Aaron Grattafiori et al. The llama 3 herd of models, 2024.

[13] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.