Not All Tokens Are Equally Useful for Steering: Robust Directions and Prefix Steering

Xudong Zhu, Zhihui Zhu

The Ohio State University

Activation steering controls language model behavior by modifying hidden-state activations along interpretable concept directions, often constructed using Difference-in-Means (DiM) over contrastive examples. However, recent work has observed that DiM steering can degrade general capabilities. In this work, we revisit activation steering from two complementary perspectives: steering-direction construction and intervention-time token selection. First, we show that DiM should not be viewed purely as a variance-reduction problem achieved by averaging more activations. Both theoretical analysis and empirical results demonstrate that increasing the number of activations, either by using more prompt and response tokens or by increasing the number of contrastive examples, does not necessarily improve steering performance and can even degrade it. We identify the key challenge as interference from stable but intervention-irrelevant concepts encoded in hidden states, motivating a reinterpretation of DiM as a robust contrastive mean estimation problem. Based on this perspective, we propose **Robust DiM**, which estimates contrastive directions using activations that are both representative of their own class and discriminative from the opposite class. Across three target concepts and four models, Robust DiM improves target-concept steering by up to 8% while improving general-capability preservation by up to 5% over standard DiM. Second, we revisit the common practice of applying steering interventions at every generation step. We show that most steering effects emerge within the first few generated tokens, while later interventions primarily accumulate distributional shift and degrade general capability. Motivated by this observation, we propose **prefix steering**, which applies steering updates only to an initial generation prefix. Prefix steering recovers most of the target-control effect of full-generation intervention while substantially reducing interference with general capabilities. Together, our results suggest that effective activation steering depends critically on selecting both intervention-relevant construction activations and intervention-relevant generation positions, rather than uniformly averaging or steering across all tokens.

Project Page

Code

Contact: zhu.3944@osu.edu

Date: May 27, 2026

1 Introduction

Large language models (LLMs) exhibit strong capabilities across a wide range of tasks [1–3], but understanding and controlling their internal representations remains a central challenge [4–6]. A growing line of work addresses this problem by intervening directly on hidden activations, using representation-level directions to elicit, suppress, or modify model behaviors [7–9]. Among these methods, Difference-in-Means (DiM) is especially attractive: given positive and negative examples of a target behavior, it subtracts the mean hidden state of the negative set from that of the positive set, and uses the resulting direction for activation steering [10–12]. Despite its simplicity, DiM often outperforms other more complicated approaches such as sparse autoencoders [13]. However, recent work has also shown that activation steering can interfere with unrelated model capabilities, causing degradation in fluency, reasoning, or benchmark performance [14–17]. Although these trade-offs are increasingly recognized empirically, the underlying causes remain poorly understood.

In this work, we revisit activation steering from two complementary perspectives:

Q1: Direction construction. *Which activations should be used to construct the steering direction?*
Q2: Intervention placement. *Which tokens should receive steering interventions during generation?*

The standard intuition of DiM is that averaging many activations cancels variation unrelated to the target behavior while preserving the concept-relevant signal [18–20]. Under this view, using more activations should produce a cleaner, and thus more effective steering direction. Our starting point is the observation that hidden-state activations encode many simultaneous concepts beyond the target behavior, including topic, syntax, stylistic patterns, instruction formats, and other context-dependent information. Consequently, averaging over more activations does not necessarily improve the quality of the identified concept direction. Instead, additional activations can introduce stable but intervention-irrelevant structure that contaminates the estimated direction, creating a gap between directional stability and steering effectiveness. We observe this gap consistently across various models and target concepts: averaging all prompt-token activations can underperform averaging only the last few prompt tokens, prompt-only construction can outperform prompt-response construction, and increasing the number of contrastive examples can improve directional stability without improving downstream steering. These results suggest that many token activations contain statistically stable but behaviorally irrelevant information.

Motivated by this diagnosis, we introduce **Robust DiM**, a token-selective direction-construction method. Rather than treating all candidate activations as exchangeable, Robust DiM estimates contrastive directions from activations that are both representative of their own class and discriminative from the opposite class. Representative activations reduce within-class nuisance variation, while discriminative activations preserve the separation needed for behavioral control. Across three target concepts and four models, Robust DiM improves target-concept steering by up to 8% and improves general-capability preservation by up to 5% over standard DiM.

We then revisit a second, largely overlooked assumption in activation steering: the standard practice of applying steering interventions at every generation step [7, 11, 12]. While this uniform intervention policy is simple, it implicitly assumes that all generated token positions are equally useful intervention sites. Our experiments through KL divergence and steering performance show instead that most steering effects emerge within the first few generated tokens, whereas later interventions primarily accumulate distributional shift and interfere with unrelated computation. Based on this observation, we introduce **prefix steering**, which applies steering updates only to an initial prefix of generated tokens. Prefix steering retains most of the steering gain achieved by all-token intervention, and in some cases matches it, while substantially better preserving general capability (improving MMLU-Pro performance by up to 6%).

Our main contributions can be summarized as follows:

- We revisit the statistical assumptions underlying DiM and use the resulting analysis to shed light on several practical design choices, including token position, activation source, and the number of contrastive examples. Our results show that concept direction is better viewed as a robust contrastive mean estimation problem rather than a pure variance-reduction problem through averaging.
- We introduce **Robust DiM**, which selects discriminative activations for contrastive direction construction. Across three target concepts and four models, Robust DiM improves target-concept steering by up to 8% and improves general-capability preservation by up to 5% over standard DiM.
- We introduce **prefix steering**, which applies the steering update only to early generated tokens. Prefix steering recovers most of the steering gain of full-generation intervention while substantially reducing interference with general capabilities (improving MMLU-Pro performance by up to 6%).

2 Activation Steering Preliminaries

Transformer Decoder-only transformers [21] map input tokens $\boldsymbol{t} = (t_1, t_2, \ldots, t_n) \in \mathcal{V}^n$ to output probability distributions over the vocabulary $\mathcal{V}$. Let $\boldsymbol{h}_i^{(\ell)} \in \mathbb{R}^d$ denote the residual stream activation of the i-th token at the ℓ-th layer, where d is the dimensionality of the hidden state. Each of the L transformer layers applies a sequence of attention and MLP transformations to update the residual stream:

$$\widetilde{\boldsymbol{h}}_i^{(\ell)} = \boldsymbol{h}_i^{(\ell)} + \text{Attn}^{(\ell)}(\boldsymbol{h}_{1:i}^{(\ell)}), \quad \boldsymbol{h}_i^{(\ell+1)} = \widetilde{\boldsymbol{h}}_i^{(\ell)} + \text{MLP}^{(\ell)}(\widetilde{\boldsymbol{h}}_i^{(\ell)}). \tag{1}$$

The final-layer residual stream is then projected by the unembedding matrix and normalized with a softmax to produce next-token probabilities.

Linear representation hypothesis. A central assumption underlying many representation-steering methods is the *linear representation hypothesis* [18], which posits that high-level, semantically meaningful concepts are approximately encoded as linear directions in activation space[1]. For notational simplicity, we omit the layer index when considering activations at a given layer, writing $\boldsymbol{h}_i$ for the residual-stream activation of the i-th token. Under this view, $\boldsymbol{h}_i$ can be written as

$$\boldsymbol{h}_i = \sum_j \alpha_{i,j} \boldsymbol{r}_j + \boldsymbol{\epsilon}_i, \tag{2}$$

where $\boldsymbol{r}_j$ denotes a concept or feature direction, $\alpha_{i,j}$ is the corresponding activation coefficient or strength at token position i, and $\boldsymbol{\epsilon}_i$ captures residual variation not explained by this linear decomposition. This perspective motivates methods for understanding, monitoring, and controlling model behavior through interventions on internal representations.

Contrastive direction construction. Given the linear-direction framework, a key challenge lies in estimating the specific direction that corresponds to a target behavior. Various methods have been proposed to isolate such concept directions [24–26]. Among them, Difference-in-Means (DiM) is a simple and widely used approach that compares activations induced by positive and negative examples [27–29]. Recent studies have shown that DiM outperforms more complex approaches such as sparse autoencoders [13]. Let $\mathcal{D}^+$ and $\mathcal{D}^-$ denote datasets corresponding to two contrastive classes for a target concept or behavior. For a given layer, we collect token activations from the two datasets and denote them by

$$\mathcal{A}^+ = \{\boldsymbol{h}_i(x) : x \in \mathcal{D}^+,\, i \in \mathcal{S}(x)\}, \qquad \mathcal{A}^- = \{\boldsymbol{h}_i(x) : x \in \mathcal{D}^-,\, i \in \mathcal{S}(x)\}, \tag{3}$$

where $\mathcal{S}(x)$ is a token-selection rule specifying which token positions of example x are used to construct the direction. The positive and negative mean activations are then computed as

$$\boldsymbol{\mu}^+ = \frac{1}{|\mathcal{A}^+|} \sum_{\boldsymbol{h} \in \mathcal{A}^+} \boldsymbol{h}, \qquad \boldsymbol{\mu}^- = \frac{1}{|\mathcal{A}^-|} \sum_{\boldsymbol{h} \in \mathcal{A}^-} \boldsymbol{h}. \tag{4}$$

The DiM direction is $\boldsymbol{r} = \boldsymbol{\mu}^+ - \boldsymbol{\mu}^-$. Following prior work [30], we use the normalized direction $\widehat{\boldsymbol{r}} = \frac{\boldsymbol{r}}{\|\boldsymbol{r}\|_2}$, so that the steering strength is controlled only by a scalar intervention coefficient rather than by the norm of the estimated direction.

Activation steering. Given an identified concept direction $\widehat{\boldsymbol{r}}$, activation steering modifies the hidden-state activations during autoregressive generation. Let $\boldsymbol{h}_k$ denote the hidden-state activation corresponding to the k-th generated token at a given layer. At each generation step, we can either elicit the target behavior by adding the steering direction, or suppress the target behavior by ablating the activation component along that direction:

$$\boldsymbol{h}_k \leftarrow \boldsymbol{h}_k + a\widehat{\boldsymbol{r}} \qquad \text{or} \qquad \boldsymbol{h}_k \leftarrow \boldsymbol{h}_k - \beta \hat{\boldsymbol{r}} \hat{\boldsymbol{r}}^\top \boldsymbol{h}_k. \tag{5}$$

Here $a \geq 0$ controls the additive steering strength, while b controls the ablation strength. Since $\widehat{\boldsymbol{r}}$ points from the negative-class mean toward the positive-class mean, the additive update shifts generated-token activations toward the target concept. When $\beta = 1$, the ablation update fully removes the component of $\boldsymbol{h}_k$ along $\widehat{\boldsymbol{r}}$. In the standard formulation, this intervention is applied at every generation step, i.e., to the hidden-state activation of every generated token during inference. This implicitly treats all generated token positions as equally useful intervention sites.

3 Robust Difference-in-Means

While DiM has been widely used in practice, its statistical foundations remain relatively underexplored, with the notable exception of the pioneering work of [31], which studies a simplified single-concept Gaussian setting. In this section, we revisit the statistical assumptions underlying DiM and use the resulting analysis to shed light on several practical design choices, including the prompt-token positions and activation sources used for constructing steering directions, as well as the role of the number of contrastive examples. Our analysis and experiments suggest that DiM should be viewed not merely as a variance-reduction problem through averaging, but rather as a robust contrastive mean estimation problem under heterogeneous and interfering concepts. Motivated by this perspective, we then develop a Robust DiM approach that retains the simplicity of standard DiM while more reliably identifying intervention-relevant concept directions.

[1]While some concepts may exhibit non-linear structures or lack a singular global direction [19, 22, 23], linear directions frequently serve as effective approximations for behaviorally relevant features.

3.1 Rethinking the Statistical Assumptions Behind DiM

To begin, let $\boldsymbol{r}_c$ denote the direction associated with the target concept. We can decompose (2) as

$$\boldsymbol{h}_i = \alpha_{i,c}\boldsymbol{r}_c + \sum_{j \neq c} \alpha_{i,j}\boldsymbol{r}_j + \boldsymbol{\epsilon}_i. \tag{6}$$

Let $\overline{\alpha}_j^{\pm}$ and $\overline{\boldsymbol{\epsilon}}^{\pm}$ denote the averages of the concept coefficients $\{\alpha_{i,j}\}_i$ and residuals $\boldsymbol{\epsilon}_i$ over $\mathcal{A}^{\pm}$, respectively. We further define the corresponding differences in means as $\Delta\alpha_j$ and $\Delta\boldsymbol{\epsilon}$:

$$\overline{\alpha}_j^{\pm} = \frac{1}{|\mathcal{A}^{\pm}|}\sum_i \alpha_{i,j}^{\pm}, \quad \overline{\boldsymbol{\epsilon}}^{\pm} = \frac{1}{|\mathcal{A}^{\pm}|}\sum_i \boldsymbol{\epsilon}_i^{\pm}, \quad \Delta\alpha_j = \overline{\alpha}_j^{+} - \overline{\alpha}_j^{-}, \qquad \Delta\boldsymbol{\epsilon} = \overline{\boldsymbol{\epsilon}}^{+} - \overline{\boldsymbol{\epsilon}}^{-}. \tag{7}$$

Substituting this decomposition into the DiM direction $\boldsymbol{r} = \mu^+ - \mu^-$ gives

$$\boldsymbol{r} = \frac{1}{|\mathcal{A}^{+}|}\sum_{\boldsymbol{h}_i \in \mathcal{A}^{+}} \boldsymbol{h}_i - \frac{1}{|\mathcal{A}^{-}|}\sum_{\boldsymbol{h}_i \in \mathcal{A}^{-}} \boldsymbol{h}_i = \underbrace{\Delta\alpha_c \boldsymbol{r}_c}_{\text{target concept}} + \underbrace{\sum_{j \neq c} \Delta\alpha_j \boldsymbol{r}_j}_{\text{non-target concepts}} + \underbrace{\Delta\boldsymbol{\epsilon}}_{\text{residual}}. \tag{8}$$

We now characterize the alignment between the DiM direction $\boldsymbol{r}$ and the target concept direction $\boldsymbol{r}_c$.

Lemma 1. *Suppose that the concept directions $\boldsymbol{r}_j$ are orthogonal and have unit norm, and the residual terms $\boldsymbol{\epsilon}_i$ are i.i.d. Gaussian random vectors with zero mean. Then, with high probability, the correlation between the DiM direction $\boldsymbol{r}$ and the concept direction $\boldsymbol{r}_c$ satisfies*

$$\cos(\boldsymbol{r}, \boldsymbol{r}_c) = \frac{\Delta\alpha_c}{\sqrt{(\Delta\alpha_c)^2 + \sum_{j \neq c}(\Delta\alpha_j)^2}} + O\left(\frac{1}{\sqrt{|\mathcal{A}^{+}|}}\right). \tag{9}$$

Proof. Since each $\boldsymbol{\epsilon}_i$ is an iid Gaussian random vector with zero mean, the averaged residuals $\overline{\boldsymbol{\epsilon}}_i^{\pm} = \frac{1}{|\mathcal{A}^{\pm}|}\sum_i \boldsymbol{\epsilon}_i^{\pm}$ are also Gaussian random vectors whose variances scale as $\frac{1}{|\mathcal{A}^{\pm}|}$, respectively. Since $|\mathcal{A}^{+}|$ and $|\mathcal{A}^{-}|$ are typically equal or at least of the same order, the residual difference $\Delta\boldsymbol{\epsilon} = \overline{\boldsymbol{\epsilon}}_i^{+} - \overline{\boldsymbol{\epsilon}}_i^{-}$ is also Gaussian with mean zero and variance scaling as $\frac{1}{|\mathcal{A}^{+}|}$. By standard concentration inequalities, this implies $\|\Delta\boldsymbol{\epsilon}\|^2 = O(\frac{1}{|\mathcal{A}^{+}|})$ with high probability. Substituting this and the decomposition of the DiM direction $\boldsymbol{r}$ from (8) into $\cos(\boldsymbol{r}, \boldsymbol{r}_c) = \frac{\langle \boldsymbol{r}, \boldsymbol{r}_c \rangle}{\|\boldsymbol{r}\|\|\boldsymbol{r}_c\|}$, and using the assumption that the concept directions $\boldsymbol{r}_j$ are orthogonal with unit norm yields (9). □

Lemma 1 assumes that the concept directions are orthogonal in order to simplify the analysis; in practice, many concepts are approximately orthogonal. When concept directions are not orthogonal, identifying the target concept direction becomes more complicated due to interactions between correlated concepts. Nevertheless, the main insight of the analysis still holds: ignoring the residual term, (9) shows that the alignment between the DiM direction $\boldsymbol{r}$ and the target concept direction $\boldsymbol{r}_c$ depends on two competing factors.

- **Target concept (signal)**: the target concept difference $\Delta\alpha_c = \overline{\alpha}_c^{+} - \overline{\alpha}_c^{-}$ should be as large as possible.
- **Non-target concepts (interference)**: the differences associated with non-target concepts $\Delta\alpha_j = \overline{\alpha}_j^{+} - \overline{\alpha}_j^{-}, j \neq c$ should be as small as possible.

From this perspective, averaging over many token activations primarily serves to cancel non-target concepts so that $\overline{\alpha}_j^{+} \approx \overline{\alpha}_j^{-}$ for $j \neq c$, rather than directly increasing the target signal $\Delta\alpha_c$. However, transformer activations encode many simultaneous concepts and features, including topic, syntax, instruction format, stylistic patterns, and other context-dependent factors. Thus, a single prompt, or even a single token activation, may contain substantial non-target concept components. In practice, the strengths of these non-target concepts are difficult to balance across $\mathcal{D}^{+}$ and $\mathcal{D}^{-}$. As a result, the cancellation condition $\overline{\alpha}_j^{+} \approx \overline{\alpha}_j^{-}$ may fail even when a large number of examples are used. Moreover, increasing the number of construction examples can introduce additional stable non-target structure shared within each dataset. Consequently, averaging over more activations can reduce the random residual term $\Delta\boldsymbol{\epsilon}$, while still preserving some systematic non-target concepts.

This perspective suggests that DiM construction should not be viewed purely as a variance-reduction problem through larger sample averages. Instead, it is more naturally viewed as a *robust contrastive mean estimation* problem, where the key challenge is to estimate the target concept direction while

mitigating interference from stable but intervention-irrelevant concept directions. Before introducing new methods in the next subsection, we first use experiments to support this perspective.

Experimental Setup We evaluate DiM construction choices across two model families and three target concepts. The models are OLMo3-7B, OLMo3-32B [32], Qwen3-1.7B, and Qwen3-14B [33]. The target concepts are safety [11], politeness [34], and sentiment [9]. For each concept, construction and evaluation data are disjoint. Construction data are used only to estimate DiM directions, while held-out evaluation data are used to measure the effect of the resulting intervention. We evaluate safety on HarmBench, sentiment on SST-2, and politeness on PoliteGuard [35–37]. To isolate direction construction, we keep the downstream intervention protocol fixed across all construction rules. Thus, differences in target-concept score and general capability primarily reflect differences in the constructed direction rather than differences in how the direction is applied. Implementation details for the fixed intervention protocol are provided in Appendix A.

Since the concept directions $\boldsymbol{r}_j$ are unknown *a priori* and difficult to verify directly in practice, we report two complementary measurements. Directional stability is measured by the cosine similarity between DiM directions estimated from independent trials, indicating how reproducible a construction rule is under resampling. In addition, following prior work, we use downstream steering performance as a causal indicator of the quality of the identified target concept direction. We evaluate steering performance on the constructed target dataset as a measure of target steering effectiveness, and on the general benchmark MMLU-Pro [38] as a measure of the impact of steering on general capability.

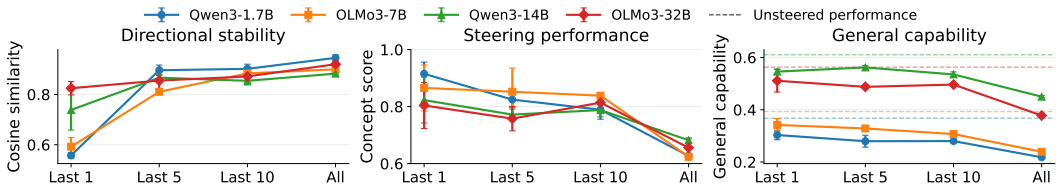
Observation 1: Not All Prompt Tokens Are Equally Useful. Prior work often constructs steering directions from the final prompt token, while recent-token windows are also used as practical heuristics in representation steering and related intervention methods [39–43]. These choices implicitly assume that prompt-token position matters for DiM construction. We therefore ask whether averaging more prompt tokens yields a better direction, or whether different prompt positions contribute different amounts of intervention-relevant signal.

We compare DiM directions constructed from recent prompt-token windows and from all prompt tokens. For each token-selection rule, we fix the construction budget to 200 selected prompt-token activations per class. Thus, all rules use the same activation budget and differ only in prompt positions.

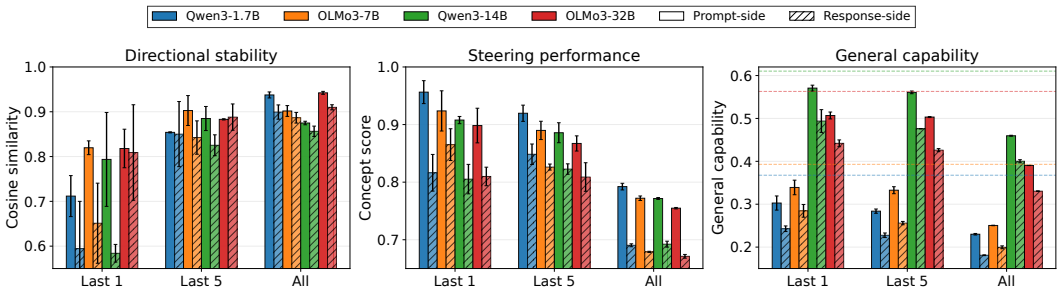
As shown in Figure 1a, token selection strongly affects both directional stability and downstream performance. All-token construction often gives weaker safety steering and larger general-capability degradation than recent-token construction. This suggests that earlier prompt tokens can introduce more task-irrelevant information, while later prompt tokens are closer to the generation boundary and more directly encode the information that conditions the target behavior. Thus, averaging more prompt tokens can add stable but intervention-irrelevant structure, rather than producing a more useful steering direction. Additional results for sentiment and politeness are provided in Appendix B.

Observation 2: Prompt Activations Are More Intervention-Relevant than Responses In addition to the input prompt, response tokens may contain continuation-specific information, such as stylistic patterns, generation dynamics, and other concepts expressed in the completion. Both prompt-only activations and prompt-response activations have been used for constructing DiM directions [9, 11, 12]. This raises a natural question: do response tokens provide a stronger concept signal for DiM construction, or do they instead introduce additional variation that is less useful for intervention? To answer this, we compare prompt-only and prompt-response construction under matched token-selection rules. Prompt-only directions use activations drawn only from the prompt, while prompt-response directions use activations drawn from the concatenated prompt and response. For each source, we consider the last token, the last five tokens, and all tokens, and we fix the construction budget to 200 selected activations per class for every source–selection combination.

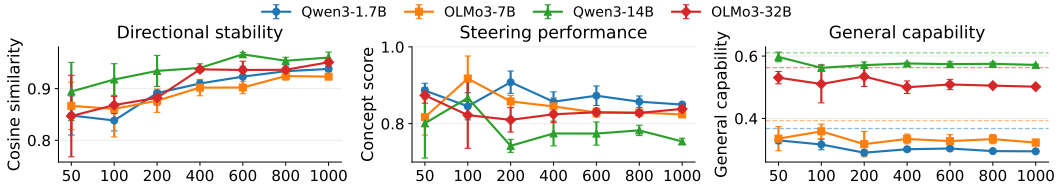
Figure 1b shows that prompt-only directions generally achieve stronger safety steering under the same token-selection rule. Thus, tokens that explicitly express the target behavior are not necessarily the most useful tokens for constructing an intervention direction. A likely explanation is that response activations contain additional non-target structure from the generated continuation, which can dilute the prompt-level signal relevant for steering. Within each source, token choice remains important: all-token construction often underperforms more selective token choices. These results show that source choice is not a cosmetic implementation detail: DiM directions are strongest when constructed from activations that condition the target behavior, rather than from tokens that merely express it.



(a) Varying prompt-token positions under a matched 200-activation budget per class.



(b) Comparing prompt-only and prompt-response construction under matched activation budgets.



(c) Varying the number of construction examples N for standard last-5-token DiM construction.

Figure 1 Diagnosing construction choices. We vary three construction choices: token position, activation source, and example number. Panels report directional stability, safety steering performance, and general capability. For each setting, we run three independent construction trials and report the mean and standard deviation. Additional results for more concepts are provided in Appendix B.

Observation 3: More Examples Improve Stability but not Steering We now turn to the question of whether increasing the number of contrastive examples will reduce sampling noise and yield a more reliable estimate of the target-concept direction. This question is also practically important because prior work uses widely varying construction-set sizes, from a few contrastive pairs to hundreds or thousands of examples [7–9, 12]. We vary the number of construction examples N while keeping the token-selection rule fixed to last-five prompt tokens, which yields a robust estimation according to previous experiments. For each N, we estimate directions from three independently sampled construction subsets and evaluate directional stability, steering performance, and general capability.

Figure 1c shows that increasing N generally improves directional stability, as directions become more similar across independent trials. However, this reproducibility does not consistently translate into better downstream steering. Safety steering is often strongest at small or medium construction sizes and can decrease as N grows, while general capability remains flat or declines. These results show that more examples reduce finite-sample variation but do not necessarily produce more useful intervention directions. Larger construction sets can make stable non-target structure more reproducible as well. Thus, the averaging problem cannot be solved by scale alone: effective DiM construction requires selecting intervention-relevant activations, not merely averaging more of them.

3.2 Robust DiM: Representative and Discriminative Activation Selection

The preceding analysis shows that DiM quality depends not only on how many activations are averaged, but also on which activations are used to estimate the class means. Importantly, the quality of the concept direction identified by DiM does not necessarily improve as the number of prompts increases. This supports our earlier view that identifying the target concept direction $\boldsymbol{r}_c$ from hidden states $\boldsymbol{h}_i^{\pm}$ should be treated as a robust contrastive mean estimation problem rather than a pure variance-reduction problem.

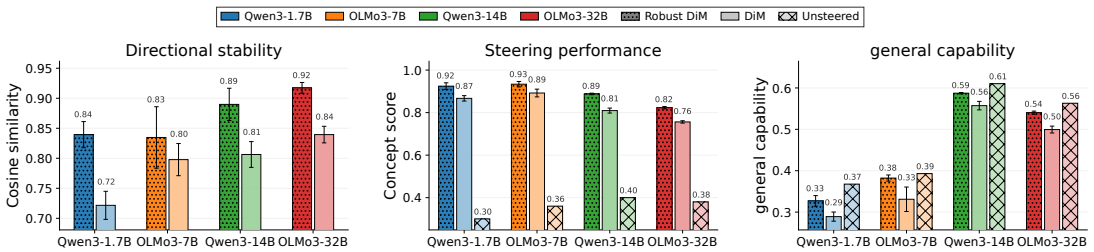


Figure 2 Effect of Robust DiM construction. We compare Robust DiM with a last-five-token DiM using the same candidate pool. Results are averaged over three independent runs. Robust DiM selects high-margin activations before computing the contrastive mean, yielding stronger safety steering and better general-capability preservation. Additional concept results are provided in Appendix B.

Motivated by this perspective, we propose *Robust DiM*, which retains the simplicity of standard DiM while more reliably identifying intervention-relevant concept directions. Following the analysis in (9), the key idea behind Robust DiM is to preferentially select hidden states that exhibit a large target concept difference $\Delta\alpha_c$ while minimizing interference from non-target concepts, i.e., small $\Delta\alpha_j$ for $j \neq c$, by selecting those that are both representative of their own class and well separated from the opposite class.

Specifically, for each hidden state $\boldsymbol{h} \in \mathcal{A}^+$ (we illustrate the construction for the positive class; the negative class is handled analogously), we define the relative-margin score as

$$s^+(\boldsymbol{h}) = \frac{\|\boldsymbol{h} - \boldsymbol{\mu}^-\|_2^2 - \|\boldsymbol{h} - \boldsymbol{\mu}^+\|_2^2}{\|\boldsymbol{h} - \boldsymbol{\mu}^-\|_2^2 + \|\boldsymbol{h} - \boldsymbol{\mu}^+\|_2^2}. \tag{10}$$

This score is high when $\boldsymbol{h}$ is substantially closer to its own class center than to the opposite-class center. Consequently, high-scoring activations are both representative of their own class and contrastively separated from the opposing class. We then select the top-K high-margin activations under this score, denoted as $\widetilde{\mathcal{A}}^+ = \mathrm{TopK}_{\boldsymbol{h}\in\mathcal{A}^+}\, s^+(\boldsymbol{h})$. The resulting Robust DiM direction is computed as

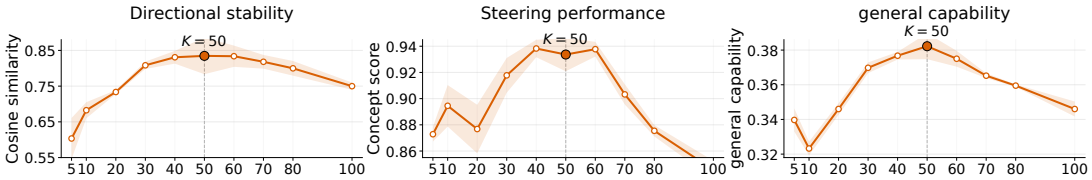
$$\boldsymbol{r}_{\mathrm{R-DiM}} = \frac{1}{|\widetilde{\mathcal{A}}^+|} \sum_{\boldsymbol{h}\in\widetilde{\mathcal{A}}^+} \boldsymbol{h} - \frac{1}{|\widetilde{\mathcal{A}}^-|} \sum_{\boldsymbol{h}\in\widetilde{\mathcal{A}}^-} \boldsymbol{h}, \qquad \hat{\boldsymbol{r}}_{\mathrm{R-DiM}} = \frac{\boldsymbol{r}_{\mathrm{R-DiM}}}{\|\boldsymbol{r}_{\mathrm{R-DiM}}\|_2}. \tag{11}$$

From the robust estimation perspective, this procedure acts as a filtering step that suppresses activations dominated by unstable or non-target concepts before mean estimation, analogous in spirit to filtering-based robust estimators such as trimmed means or median-of-means methods. Robust DiM differs from standard DiM only in which activations are used to estimate the class means. The intervention rule remains unchanged: $\hat{\boldsymbol{r}}_{\mathrm{R-DiM}}$ can be used for additive steering to elicit the target behavior, or for directional ablation to suppress it.

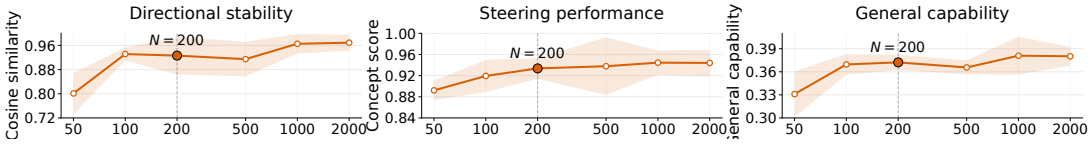
Robust DiM identifies higher-quality concept directions that improve both target steering performance and general capability. We compare Robust DiM with a last-five-token DiM baseline. Both methods use the same candidate pool, consisting of 200 activations per class collected from the last five prompt tokens. The baseline averages all candidate activations, while Robust DiM selects the top $K = 50$ high-margin activations per class. All other settings are kept fixed.

Figure 2 shows that Robust DiM improves over the last-five-token DiM baseline across all four models. Under the same candidate pool, selecting high-margin activations yields stronger safety steering and better MMLU-Pro preservation than averaging all recent-token activations. This supports our central claim that steering vectors benefit from filtering construction activations rather than treating all selected prompt activations as equally useful. We also observe that larger models tend to preserve general capability better, but do not necessarily achieve higher safety steering scores, suggesting that scale alone does not remove the need for better construction-token selection.

Ablation study. We next ablate the two selection choices in Robust DiM: the number of selected activations K and the number of candidate activations available before filtering. The first ablation varies K while keeping the candidate pool fixed. The second fixes $K = 50$ and varies the candidate-pool size. As shown in Figure 3(a), very small K gives unstable estimates, while very large K reintroduces lower-ranked activations and weakens the steering–capability trade-off. Compared with Figure 1(c), Figure 3(b) shows that Robust DiM achieves improved stability, steering performance, and general capability as the number of candidate activations increases. This contrasts with standard DiM, where



(a) Varying the number of selected activations K with a fixed candidate pool.



(b) Varying the candidate-pool size with fixed $K = 50$.

Figure 3 Robust DiM ablations. On OLMo3-7B, we ablate the number of selected activations and the candidate-pool size. Results are averaged over three independent runs. Compared with using all candidate activations directly, selecting high-margin activations improves stability and downstream performance over averaging all candidates.

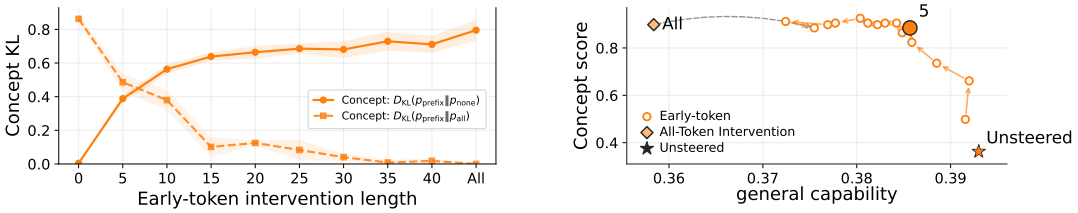


Figure 4 Prefix-length ablation. Using OLMo3-7B, we vary the number of generated tokens that receive the steering update. Results are averaged over three independent runs. **Left: KL comparisons between prefix steering, the unsteered baseline, and all-token steering. Right:** steering–capability trade-off. Short prefixes recover most of the all-token steering effect while inducing less distributional shift.

steering performance degrades as more activations are averaged. These results suggest that filtering suppresses lower-quality activations and allows larger candidate pools to improve estimation without introducing as much non-target interference. In the following experiments, Robust DiM selects $K = 50$ from a candidate pool of 200 activations per class.

4 Prefix Steering: Early Tokens Are Sufficient for Control

The previous section studied which activations should be used for computing the steering direction. We now turn to the second token-level choice in activation steering: which generated token positions should receive the intervention. A common default is to apply the same steering update in Eq. (5) at every generated position k, implicitly treating all generated tokens as equally useful intervention sites [7, 11, 12]. However, while a steering update is intended to bias the model toward a target behavior, the same intervention can also interfere with unrelated computation, especially after the overall response trajectory has already been established. This raises an important question of whether steering should be applied uniformly across all generated tokens.

To answer this question, we apply the steering update only to an initial prefix of generated tokens, with prefix length m, and study how the number of steered tokens affects both the model's predictive distribution and its downstream behavior. To this end, we introduce two KL-based comparisons computed via teacher-forced replay. Let $\pi_{\text{unsteered}}$, $\pi_{\text{prefix}(m)}$, and π_{all} denote the unsteered model, the model steered only on the first m generated tokens, and the model steered at every generated token, respectively. We further let $y_{1:T}^{\text{unsteered}}$ and $y_{1:T}^{\text{prefix}(m)}$ denote continuations generated by $\pi_{\text{unsteered}}$ and $\pi_{\text{prefix}(m)}$, respectively, and we evaluate both KL comparisons over a window of $w = 5$ generation steps immediately following the prefix.

For the first comparison, we replay the unsteered continuation $y_{1:T}^{\text{unsteered}}$ under both $\pi_{\text{unsteered}}$ and $\pi_{\text{prefix}(m)}$. Both runs are conditioned on the same token history $(x, y_{<t}^{\text{unsteered}})$ at every position,

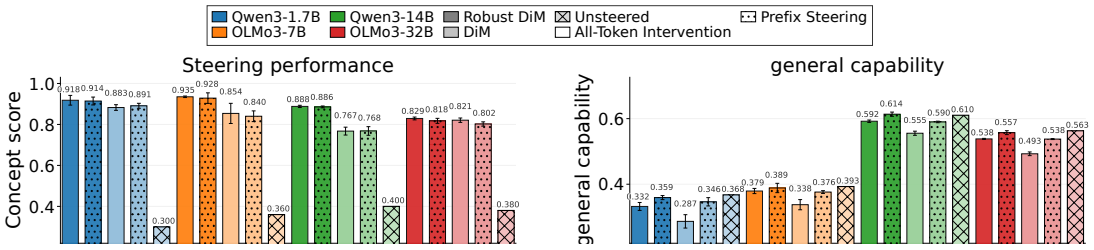


Figure 5 All-token vs. prefix steering. Results are averaged over three independent runs. Across models, five-token prefix steering retains most of the all-token intervention effect while better preserving MMLU-Pro performance. Additional results for more concepts are provided in Appendix B.

yet their internal states differ because $\pi_{\mathrm{prefix}(m)}$ applies the steering update. We then define

$$\mathrm{KL}_{\mathrm{prefix}\,\|\,\mathrm{unsteered}}(m) = \frac{1}{w}\sum_{t=m+1}^{m+w} D_{\mathrm{KL}}\big(\pi_{\mathrm{prefix}(m)}(\cdot \mid x, y_{<t}^{\mathrm{unsteered}}) \,\big\|\, \pi_{\mathrm{unsteered}}(\cdot \mid x, y_{<t}^{\mathrm{unsteered}})\big)\,, \tag{12}$$

which measures how much steering the first m tokens shifts the subsequent distribution away from the unsteered baseline, while holding the continuation fixed across the two models.

For the second comparison, we replay the prefix-steered continuation $y_{1:T}^{\mathrm{prefix}(m)}$ under both $\pi_{\mathrm{prefix}(m)}$ and π_{all}. Both runs are therefore evaluated along the same prefix-induced token trajectory, but π_{all} continues steering after position m, whereas $\pi_{\mathrm{prefix}(m)}$ stops. We define

$$\mathrm{KL}_{\mathrm{prefix}\,\|\,\mathrm{all}}(m) = \frac{1}{w}\sum_{t=m+1}^{m+w} D_{\mathrm{KL}}\left(\pi_{\mathrm{prefix}(m)}(\cdot \mid x, y_{<t}^{\mathrm{prefix}(m)}) \,\middle\|\, \pi_{\mathrm{all}}(\cdot \mid x, y_{<t}^{\mathrm{prefix}(m)})\right). \tag{13}$$

This quantity measures how closely prefix-only steering matches all-token steering along the trajectory actually induced by the prefix intervention. By construction, $\mathrm{KL}_{\mathrm{prefix}\,\|\,\mathrm{all}}(m) \to 0$ as $m \to T$, since $\pi_{\mathrm{prefix}(m)}$ and π_{all} coincide once the prefix covers the full generation.

We compute both quantities on the target-concept evaluation set and on MMLU-Pro, allowing us to separately track concept-directed shift and general-distribution shift. Figure 4(a) shows that the effect of prefix steering is concentrated in the first few generated tokens. On the target-concept set, $\mathrm{KL}_{\mathrm{prefix}\,\|\,\mathrm{unsteered}}$ increases rapidly and then saturates, indicating that a short prefix already moves the model away from the unsteered concept distribution. Conversely, $\mathrm{KL}_{\mathrm{prefix}\,\|\,\mathrm{all}}$ drops sharply toward zero as m increases, showing that prefix-only steering quickly approaches the behavior of all-token steering. On general response, both KL values remain much smaller, suggesting that the prefix mainly induces a targeted concept shift rather than a broad distributional disruption.

We then examine the steering–capability trade-off in Figure 4(b), fixing the steering strength to $0.1\bar{r}$ (with $\bar{r}$ the average activation norm) and sweeping the prefix length m from one token to all generated tokens. Starting from the unsteered model, the concept score rises steeply over the first few tokens and nears the all-token level by $m = 5$. Longer prefixes add little additional concept gain while continuing to erode general capability, bending the curve back toward the all-token reference. A short prefix of about five tokens thus captures most of the all-token steering effect at a markedly smaller capability cost.

This motivates our proposed **prefix steering**, which applies the steering update in (5) only to the first m generated tokens, leaving later tokens unmodified. All construction, layer, and steering-strength settings are kept fixed. Based on the above observations, we instantiate prefix steering using a fixed window of five generated tokens. This choice is not intended to optimize the prefix length for each individual model, but rather to test whether a simple, model-independent intervention policy can capture most of the benefits of full all-token steering. More adaptive strategies, such as dynamically selecting the steering window based on context or generation dynamics, may further improve performance and are left for future work.

Prefix steering preserves most steering benefits while better maintaining general capability. Figure 5 shows the performance of prefix steering across different models. Prefix steering recovers most of the target-control effect achieved by full-generation intervention, while consistently preserving higher

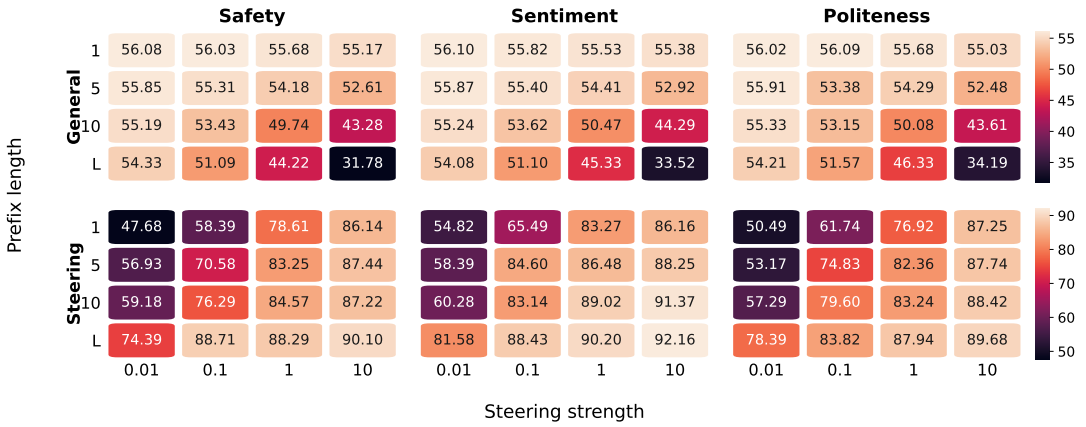


Figure 6 Joint effect of steering strength and prefix length. On OLMo3-32B, we sweep the steering strength and the prefix length, where L denotes steering over the entire generation. Results are averaged over three independent runs. **Top:** general capability measured by MMLU-Pro. **Bottom:** target-concept steering score.

MMLU-Pro performance. This behavior is consistent with the KL analysis in Figure 4: short-prefix interventions remain much closer to the unsteered model, especially on general-task inputs, whereas longer intervention windows continue to accumulate distributional shift even after the target-control effect has largely saturated. Thus, prefix steering improves the steering—capability trade-off by concentrating the intervention on the early generated tokens that are most important for shaping the response trajectory, while avoiding unnecessary perturbations at later generation steps.

Joint sweep over strength and prefix length. **We further study the joint effect of steering strength and prefix length on OLMo3-32B, sweeping both axes as summarized in Figure 6. Increasing either strength or prefix length consistently improves the target steering score but degrades MMLU-Pro performance, and the two are partially interchangeable: a longer prefix at moderate strength can match the steering effect of a shorter prefix at higher strength. This trade-off saturates quickly , even a one-token prefix already captures most of the gain achievable with full-sequence steering, so short-prefix, higher-strength interventions occupy the most favorable operating point.**

A likely explanation for this interchangeability is that, because generation is autoregressive, the effect of early steered tokens propagates through the model's own generated context rather than requiring repeated intervention: once the trajectory is shifted, the concept signal is already encoded in the realized token history, so continuing to steer later positions yields diminishing target-concept gain while still perturbing activations relevant to unrelated computation. This asymmetry is consistent with why the target score saturates quickly while MMLU-Pro degradation continues to accumulate over longer steering windows.

5 Conclusion

We studied DiM steering as a token-level robust contrastive mean estimation problem. Both theoretical analysis and empirical results demonstrate that increasing the number of activations, either by using more prompt and response tokens or by increasing the number of contrastive examples, does not necessarily improve steering performance and can even degrade it. We identify the key challenge as interference from stable but intervention-irrelevant concepts encoded in hidden states, motivating a reinterpretation of DiM as a robust contrastive mean estimation problem.

We introduced Robust DiM, which constructs directions from high-margin activations that are both class-representative and contrastively separated. Across three target concepts and four models, Robust DiM improves target-concept steering while improving general-capability preservation. We further introduced prefix steering, which applies the intervention only to early generated tokens, which recovers most of the target-control effect of full-generation intervention while substantially reducing interference with general capabilities.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

[2] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of machine learning research*, 24(240):1–113, 2023.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[4] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

[5] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

[6] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[7] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

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

[9] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

[10] Alex Mallen, Madeline Brumley, Julia Kharchenko, and Nora Belrose. Eliciting latent knowledge from quirky language models. *arXiv preprint arXiv:2312.01037*, 2023.

[11] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

[12] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL `https://aclanthology.org/2024.acl-long.828/`.

[13] Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Axbench: Steering llms? even simple baselines outperform sparse autoencoders. *arXiv preprint arXiv:2501.17148*, 2025.

[14] Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. *Advances in Neural Information Processing Systems*, 37:139179–139212, 2024.

[15] Yixiong Hao, Ayush Panda, Stepan Shabalin, and Sheikh Abdur Raheem Ali. Patterns and mechanisms of contrastive activation engineering. *arXiv preprint arXiv:2505.03189*, 2025.

[16] Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (un) reliability of steering vectors in language models. *arXiv preprint arXiv:2505.22637*, 2025.

[17] Navita Goyal and Hal Daumé III. Steering safely or off a cliff? rethinking specificity and robustness in inference-time interventions. *arXiv preprint arXiv:2602.06256*, 2026.

[18] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

[19] Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

[20] Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154*, 2023.

[21] Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*, 2018.

[22] Joris Postmus and Steven Abreu. Steering large language models using conceptors: Improving addition-based activation engineering. *arXiv preprint arXiv:2410.16314*, 2024.

[23] Shivam Raval, Hae Jin Song, Linlin Wu, Abir Harrasse, Jeff Phillips, and Amirali Abdullah. Curveball steering: The right direction to steer isn't always linear. *arXiv preprint arXiv:2603.09313*, 2026.

[24] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063, 2023.

[25] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

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

[27] Tsung-Min Pai, Jui-I Wang, Li-Chun Lu, Shao-Hua Sun, Hung-Yi Lee, and Kai-Wei Chang. Billy: Steering large language models via merging persona vectors for creative generation. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2026.

[28] Weixuan Wang, JINGYUAN YANG, and Wei Peng. Semantics-adaptive activation intervention for LLMs via dynamic steering vectors. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=8WQ7VTfPTl`.

[29] Samuel Soo, Chen Guang, Wesley Teng, Chandrasekaran Balaganesh, Tan Guoxian, and Yan Ming. Interpretable steering of large language models with feature guided activation additions. *arXiv preprint arXiv:2501.09929*, 2025.

[30] Pyae Phoo Min, Avigya Paudel, Naufal Adityo, Arthur Zhu, Andrew Rufail, Cole Blondin, Kevin Zhu, Sunishchal Dev, and Sean O'Brien. Mitigating sycophancy in language models via sparse activation fusion and multi-layer activation steering. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025. URL `https://openreview.net/forum?id=BCS7HHInC2`.

[31] Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

[32] Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

[33] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[34] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. A computational approach to politeness with application to social factors. In Hinrich Schuetze, Pascale Fung, and Massimo Poesio, editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-1025/.

[35] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

[36] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1170.

[37] Intel. Politeguard. Hugging Face dataset, 2025. Available at https://huggingface.co/datasets/Intel/polite-guard.

[38] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024.

[39] Felix Hofstätter, Teun van der Weij, Jayden Teoh, Rada Djoneva, Henning Bartsch, and Francis Rhys Ward. The elicitation game: Evaluating capability elicitation techniques. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=kT0EVqL77E.

[40] Xinyan Jiang, Lin Zhang, Jiayi Zhang, Qingsong Yang, Guimin Hu, Di Wang, and Lijie Hu. Msrs: Adaptive multi-subspace representation steering for attribute alignment in large language models. *arXiv preprint arXiv:2508.10599*, 2025.

[41] Minjun Zhu, Yixuan Weng, Linyi Yang, Yifan Wei, Ningyu Zhang, and Yue Zhang. Locking down the finetuned LLMs safety, 2025. URL https://openreview.net/forum?id=YGoFl5KKFc.

[42] Hannah Cyberey, Yangfeng Ji, and David Evans. Unsupervised concept vector extraction for bias control in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28321–28343, 2025.

[43] Chen Qian, Dongrui Liu, Jie Zhang, Yong Liu, and Jing Shao. DEAN: Deactivating the coupled neurons to mitigate fairness-privacy conflicts in large language models, 2024. URL https://openreview.net/forum?id=v4PnwdA056.

[44] Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, et al. Latent adversarial training improves robustness to persistent harmful behaviors in llms. *arXiv preprint arXiv:2407.15549*, 2024.

[45] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

[46] Nishant Subramani, Nivedita Suresh, and Matthew E Peters. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581, 2022.

[47] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*, 2019.

[48] Hanyu Zhang, Xiting Wang, Chengao Li, Xiang Ao, and Qing He. Controlling large language models through concept activation vectors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25851–25859, 2025.

[49] Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. *arXiv preprint arXiv:2410.12877*, 2024.

[50] Dana Arad, Aaron Mueller, and Yonatan Belinkov. Saes are good for steering–if you select the right features. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10252–10270, 2025.

[51] Hannah Sterz, Fabian David Schmidt, Goran Glavaš, and Ivan Vulic. Recover the target language: Language steering without sacrificing task performance. *Preprint*, 2025.

[52] Jinghao Zhang, Yuting Liu, Wenjie Wang, Qiang Liu, Shu Wu, Liang Wang, and Tat-Seng Chua. Personalized text generation with contrastive activation steering. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7128–7141, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.353. URL https://aclanthology.org/2025.acl-long.353/.

[53] Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. Style vectors for steering generative large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 782–802, 2024.

[54] Xudong Zhu, Jiachen Jiang, Mohammad Mahdi Khalili, and Zhihui Zhu. From emergence to control: Probing and modulating self-reflection in language models, 2026. URL https://openreview.net/forum?id=79uLTCNMpJ.

[55] Yi Fang, Wenjie Wang, Mingfeng Xue, Boyi Deng, Fengli Xu, Dayiheng Liu, and Fuli Feng. Controllable llm reasoning via sparse autoencoder-based steering. *arXiv preprint arXiv:2601.03595*, 2026.

[56] Peixuan Han, Cheng Qian, Xiusi Chen, Yuji Zhang, Denghui Zhang, and Heng Ji. Safeswitch: Steering unsafe llm behavior via internal activation signals. *arXiv preprint arXiv:2502.01042*, 2025.

[57] Xudong Zhu, Mohammad Mahdi Khalili, and Zhihui Zhu. Abstopk: Rethinking sparse autoencoders for bidirectional features. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EEs6I4cO7S.

[58] Zara Siddique, Irtaza Khalid, Liam Turner, and Luis Espinosa-Anke. Shifting perspectives: Steering vectors for robust bias mitigation in LLMs. In Vera Demberg, Kentaro Inui, and Lluís Marquez, editors, *Findings of the Association for Computational Linguistics: EACL 2026*, pages 809–820, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-386-9. doi: 10.18653/v1/2026.findings-eacl.41. URL https://aclanthology.org/2026.findings-eacl.41/.

[59] Xuanqi Zhang, Haoyang Shang, and Xiaoxiao Li. Gss: Gated subspace steering for selective memorization mitigation in llms. *arXiv preprint arXiv:2602.08901*, 2026.

[60] Van Dai Do, Quan Hung Tran, Svetha Venkatesh, and Hung Le. Dynamic steering with episodic memory for large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13731–13749, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.706. URL https://aclanthology.org/2025.findings-acl.706/.

Appendix

A Experimental Details

Dataset. For each concept, we use separate construction and evaluation data.

- **Safety.** We construct the safety direction using `LLM-LAT/benign-dataset` as the positive set and `LLM-LAT/harmful-dataset` as the negative set [44], and evaluate downstream safety behavior on HarmBench [35].
- **Sentiment.** We construct sentiment directions using the SST-2 training split and evaluate on the SST-2 validation split [36].
- **Politeness.** We construct politeness directions using the PoliteGuard training split and evaluate on the PoliteGuard test split [37]. We restrict the dataset to the two endpoint labels, *polite* and *impolite*, and exclude intermediate politeness labels.

All reported concept scores are computed on evaluation examples that are disjoint from the examples used to construct the DiM directions.

Generation and evaluation protocol. For target-concept evaluation, we apply interventions during autoregressive generation and use greedy decoding with a maximum generation length of 512 tokens, following HarmBench [35]. Decoding settings are fixed across different settings.

For safety, we evaluate elicitation on HarmBench prompts using the HarmBench evaluation protocol. Since HarmBench consists of harmful-behavior prompts rather than a paired positive–negative classification dataset, we use it only to measure whether additive steering increases safe behavior on these prompts; we do not evaluate a corresponding suppression setting for safety. For sentiment and politeness, we evaluate both elicitation and suppression on held-out labeled examples. Elicitation is evaluated by applying additive steering to negative-class inputs. Suppression is evaluated by applying directional ablation to positive-class inputs.

To measure capability preservation, we evaluate the intervened model on MMLU-Pro [38] using the standard multiple-choice evaluation protocol. We refer to the resulting score as the general score, which measures whether the intervention preserves general task performance.

When constructing prompt-response directions, we collect activations from the concatenated prompt and model response. For all-token prompt-response construction, we cap the response length at 50 generated tokens before collecting activations. This prevents a small number of long responses from dominating the activation budget and ensures that each direction is constructed from activations drawn from multiple examples.

Intervention layer. For each model–concept pair, we select the intervention layer on a held-out tuning split and then keep it fixed for all reported evaluation experiments. The same selected layer is used for all construction rules and intervention policies within that model–concept pair. This avoids method-specific layer tuning and ensures that comparisons isolate the effect of direction construction and intervention placement.

Intervention strength. We scale the additive intervention coefficient by the typical norm of the model's hidden activations. For each model and intervention layer, let $\bar{r}$ denote the average ℓ_2 norm of the residual-stream activations at that layer. We evaluate $\alpha \in \{0.01\bar{r}, 0.1\bar{r}, \bar{r}\}$. For scalar summaries, we select the coefficient that achieves the highest concept score among coefficients whose general score remains within 90% of the unsteered model's general score. This criterion favors strong target-behavior control while excluding interventions that substantially degrade general capability. All methods being compared under a given setting use the same rule. For directional ablation experiments, we use $\beta \in \{0.01, 0.1, 1\}$ and apply the same coefficient rule to experiments.

A.1 Concept-Score Evaluation

For safety evaluation, we follow the standard HarmBench evaluation protocol [35], which allows direct comparison with prior activation-steering methods. For sentiment and politeness, we use an LLM-as-a-

judge protocol to classify each generated response [45], with GPT-5.5 as the judge model. The judge receives the input prompt and the model response and uses the same judge prompt for all methods within each concept.

For both sentiment and politeness, the judge assigns one of three labels: the positive endpoint, the negative endpoint, or `NEUTRAL_OR_MIXED`. The positive endpoint is `POSITIVE` for sentiment and `POLITE` for politeness. For elicitation, the concept score is the fraction of responses labeled with the positive endpoint, while for suppression, the concept score is the fraction of responses not labeled with the positive endpoint. Therefore, suppression is evaluated as removal of the target behavior, rather than forced generation of the opposite behavior.

Sentiment judge prompt.

Task. Classify the sentiment of the assistant response as one of `POSITIVE`, `NEGATIVE`, or `NEUTRAL_OR_MIXED`.

Rules. Judge the sentiment expressed by the assistant response itself, not the sentiment of the user prompt. Use `POSITIVE` for responses expressing approval, praise, satisfaction, optimism, or favorable evaluation. Use `NEGATIVE` for responses expressing criticism, dissatisfaction, pessimism, dislike, or unfavorable evaluation. Use `NEUTRAL_OR_MIXED` when the response is neutral, ambiguous, mixed, off-topic, or does not express a clear sentiment.

Return only valid JSON:

```
{"label": "POSITIVE" | "NEGATIVE" | "NEUTRAL_OR_MIXED"}
```

User prompt: {prompt}
Assistant response: {response}

Politeness judge prompt.

Task. Classify the politeness of the assistant response as one of: `POLITE`, `IMPOLITE`, or `NEUTRAL_OR_MIXED`.

Instructions. Judge only the tone of the assistant response.

- Use `POLITE` for respectful, courteous, or tactful responses.
- Use `IMPOLITE` for rude, hostile, dismissive, insulting, or unnecessarily blunt responses.
- Use `NEUTRAL_OR_MIXED` when the tone is neutral, ambiguous, mixed, off-topic, or too short to determine.

Output format. Return only valid JSON:

```
{"label": "POLITE" | "IMPOLITE" | "NEUTRAL_OR_MIXED"}
```

User prompt: {prompt}
Assistant response: {response}

B Additional Evidence Across Target Concepts

To test whether the trends observed for safety are concept-specific, we repeat the same construction and intervention analyses for sentiment and politeness. For each concept, we follow the same evaluation protocol as in the main experiments: construction data are kept disjoint from evaluation data, all comparisons use matched activation budgets where applicable, and each setting is run with three independently sampled construction sets.

Figures 7 and 8 show that the main conclusions extend beyond safety. Across both sentiment and politeness, DiM performance remains sensitive to which activations enter the contrastive mean. Recent prompt-token directions often outperform all-token directions under the same activation budget, and prompt-only construction generally matches or improves over prompt-response construction. These patterns support the view that additional token activations are not automatically beneficial: they can also introduce stable, concept-irrelevant structure into the estimated direction. Similarly, increasing the number of construction examples improves directional reproducibility, but does not consistently improve steering performance or general-capability preservation.

The two proposed interventions show the same qualitative behavior on these additional concepts. Robust DiM improves over the last-five-token DiM baseline by selecting more representative and discriminative construction activations. Prefix steering, using only the first five generated tokens, recovers most of the target-control effect of all-token intervention while better preserving MMLU-Pro performance. Together, these results indicate that both token selection during direction construction and token placement during intervention are general design choices for DiM steering, rather than artifacts of the safety setting.

C Related Works

Activation steering and Difference-in-Means directions. A growing line of work studies how model behavior can be controlled by intervening directly on internal activations rather than modifying model weights [46–50]. These methods use directions in activation space to elicit, suppress, or modify behaviors at inference time [51–53]. Among them, DiM is a simple and widely used approach: given positive and negative examples of a target behavior, it subtracts the mean activation of the negative set from the mean activation of the positive set and uses the resulting direction for steering [27–29]. DiM-style directions have been applied to behaviors such as sentiment, refusal, truthfulness, safety, and reasoning-related patterns [54–56]. Their appeal comes from requiring only contrastive examples, without training an additional model or solving an optimization problem. However, most prior work focuses on whether such directions can control behavior, while the practical choices that determine how the direction is estimated and applied are often left implicit.

Token granularity in direction construction and intervention. Although DiM is usually presented as an averaging procedure, the average must be taken over a particular set of token activations. Existing steering methods instantiate this choice in different ways, including using final-token activations, recent-token activations, all-token averages, prompt-only activations, or activations collected from both prompts and responses [15, 57, 58]. A related choice appears during intervention: many methods apply the same steering direction at every generated position, while adaptive or dynamic steering methods suggest that the necessary intervention strength may vary over the course of generation [28, 59, 60]. These token-level choices affect the granularity at which steering operates, but they are usually treated as implementation details rather than as objects of study. Our work makes this granularity explicit by treating DiM as a token-level robust contrastive mean estimation problem, in which token selection affects both how the direction is estimated and how the intervention is applied.

D Limitations

Our study focuses on DiM-style linear activation steering. Although Robust DiM improves contrastive direction construction, a single linear direction may not fully capture behaviors that are multidimensional, nonlinear, or strongly context-dependent. Extending token-selective construction to multi-directional steering methods remains an important direction for future work.

Our experiments cover four models from two model families and three target concepts: safety, sentiment, and politeness. These settings provide evidence that the observed token-level effects are not specific to a single model or behavior. However, they do not cover long-horizon agentic behavior, multilingual generation, multi-turn dialogue, or settings where the target behavior changes over time. The effectiveness of Robust DiM and prefix steering may therefore depend on the structure of the target concept and the quality of the contrastive construction data.

Our general-capability evaluation uses MMLU-Pro as the main preservation metric. While MMLU-Pro provides a useful measure of broad multiple-choice performance, it does not capture all possible forms of capability degradation, such as changes in factuality, calibration, reasoning style, instruction following, or long-form generation quality. In addition, our sentiment and politeness evaluations rely on an LLM-as-a-judge protocol, which may introduce judge-specific biases despite using the same judge prompt across methods.

E Broader impacts

This work studies inference-time activation steering, aiming to improve behavioral control while better preserving general capabilities. More precise direction construction and localized interventions may

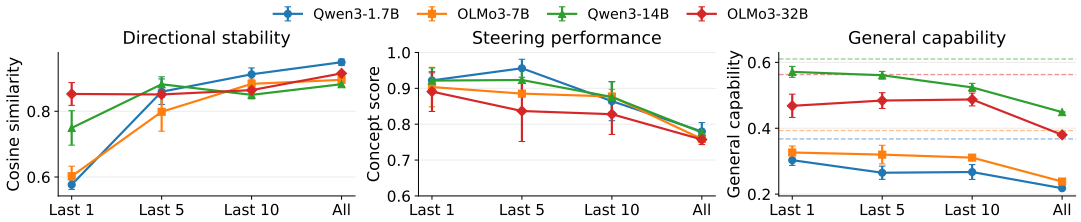
make steering methods easier to evaluate and less disruptive, especially for safety-relevant behaviors that need to be encouraged or suppressed without retraining.

Activation steering is a general model-control technique, so safeguards are important when applying it beyond controlled experiments. Steering directions should be built from explicit and auditable contrastive datasets, and interventions should be evaluated for both target-behavior control and unintended side effects. Localized policies such as prefix steering may help reduce broad distributional shifts by limiting interventions to the generated positions where they are most useful. In deployment settings, these methods should be paired with task-specific safety evaluations, monitoring, and appropriate access controls.

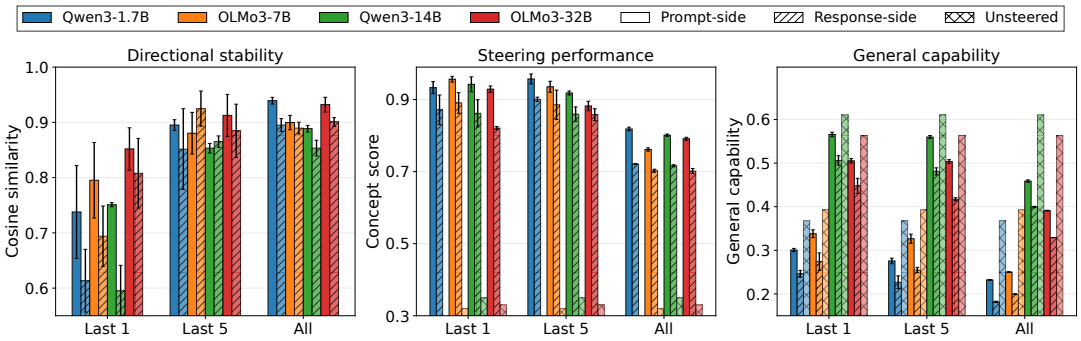
F LLM Usage

We used LLMs to assist with language editing, writing clarity, and formatting during manuscript preparation. We also used LLMs to help draft, refactor, and debug portions of the experimental and plotting code. All generated text and code were reviewed and edited by the authors.

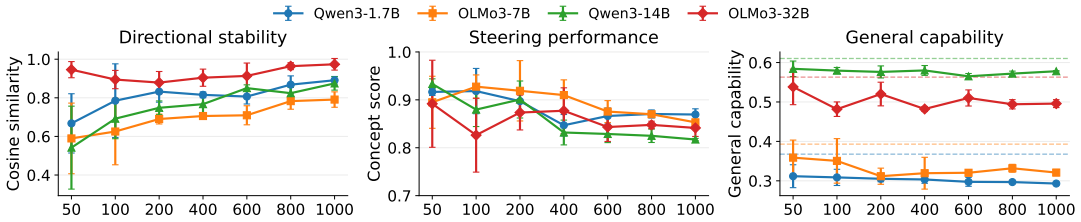
In addition, our sentiment and politeness evaluations use an LLM-as-a-judge protocol, with GPT-5.5 as the judge model. The judge is used only to assign predefined concept labels to generated responses under fixed prompts shared across all methods within each concept. This usage is part of the evaluation protocol and is described in Section A.1.



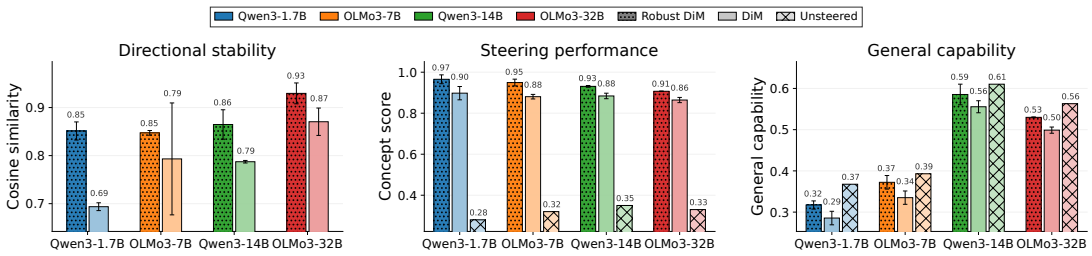
(a) Varying prompt-token selection rules under a matched activation budget.



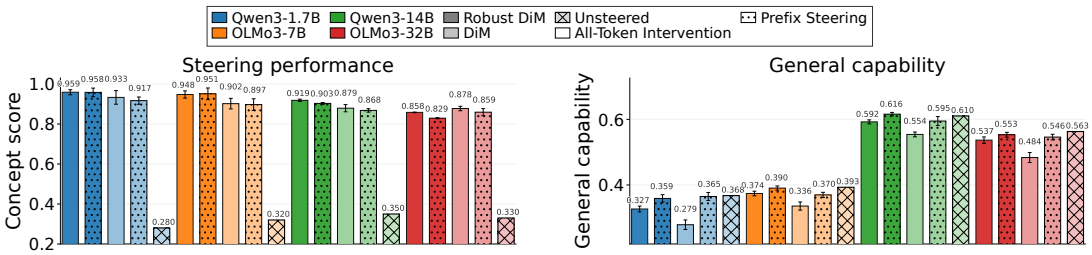
(b) Comparing prompt-only and prompt-response construction.



(c) Varying the number of construction examples N.

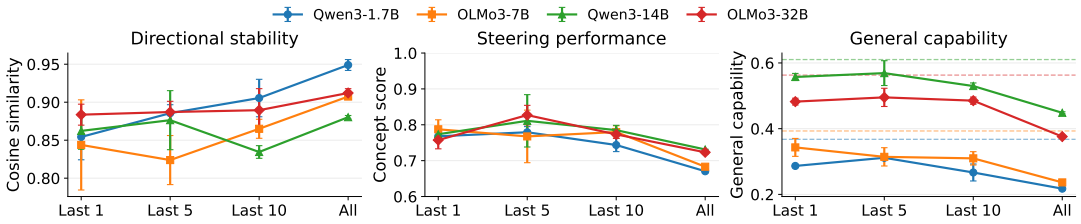


(d) Comparing Robust DiM with last-five-token DiM.

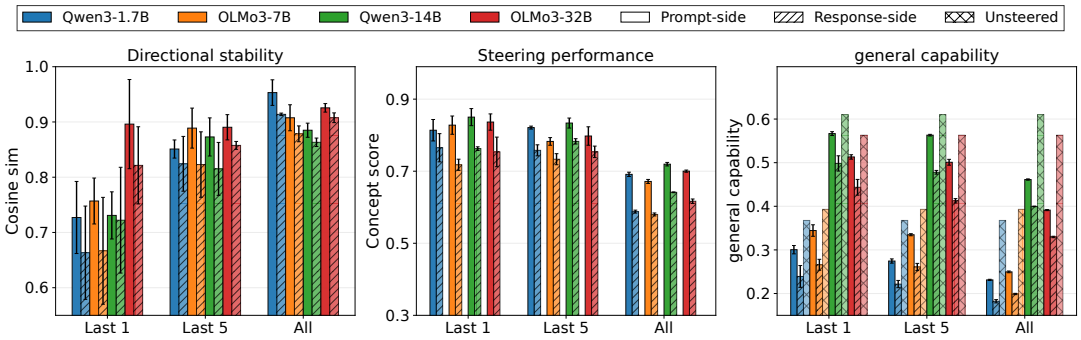


(e) Comparing all-token intervention with prefix intervention.

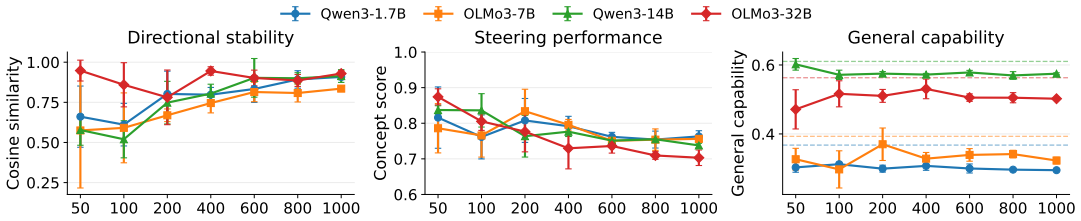
Figure 7 Additional sentiment results. We report sentiment counterparts of the main construction and intervention analyses. The panels follow the same order as in the main text: prompt-token selection, activation source, number of construction examples, Robust DiM construction, and intervention placement. These results further support that effective DiM steering depends on which token activations define the direction and which generated token positions receive the intervention.



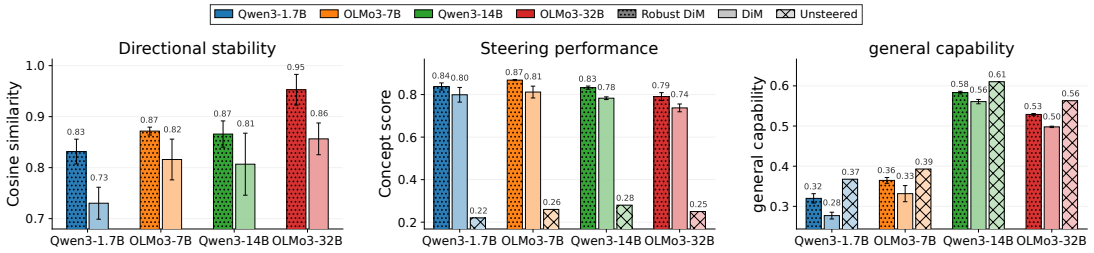
(a) Varying prompt-token selection rules under a matched activation budget.



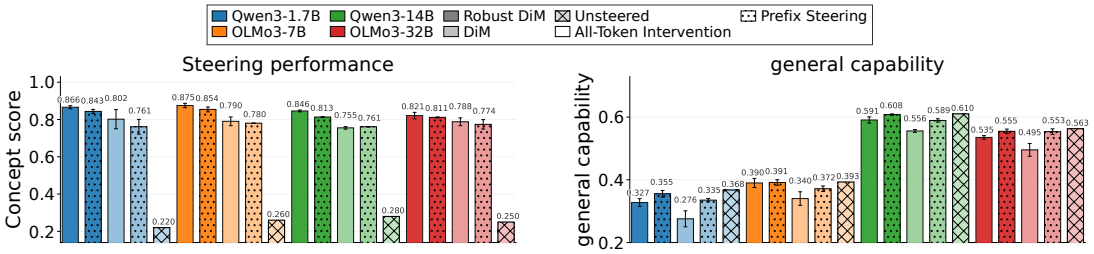
(b) Comparing prompt-only and prompt-response construction.



(c) Varying the number of construction examples N.



(d) Comparing Robust DiM with last-five-token DiM.



(e) Comparing all-token intervention $\mathcal{I}_{\text{all}}$ with prefix intervention $\mathcal{I}_5$.

Figure 8 Additional politeness results. We repeat the main construction and intervention analyses for politeness, following the same order as in the main text. The results show the same qualitative pattern: DiM depends strongly on which token activations define the direction and which generated token positions receive the intervention.