# STEER THE STEERING: CALIBRATING LEARNED INTERVENTIONS IN LARGE LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Existing activation steering methods learn adaptive steering directions to overcome the limited flexibility of fixed difference-in-means directions. However, directly learning such steering directions from semantically diverse training data under imperfect behavior modeling can introduce distributional bias, thereby limiting steering effectiveness and generalization. This motivates our geometric calibration framework for learned interventions in large language models, which we call **S2S**. Our method decomposes the learnable steering direction into a difference-in-means component and a residual component, and employs steering interpolation to geometrically calibrate the residual relative to the difference-in-means direction. Furthermore, we extend our method to linear steering and rotation-based steering, allowing S2S to operate under complementary intervention paradigms. Building on this, we introduce a similarity-aware adaptive weighting strategy to dynamically coordinate the contributions of the hidden-state direction, difference-in-means direction, and residual direction during inference. Extensive experiments across multiple-choice benchmarks demonstrate that S2S consistently outperforms linear-steering baselines, yielding gains of up to 23.9% on TruthfulQA, while preserving its general open-ended generation quality.

## 1 INTRODUCTION

Large language models (LLMs) have become increasingly capable, yet they continue to exhibit undesirable behaviors, including hallucination and untruthful generation (Lin et al., 2022; Huang et al., 2025), sycophancy (Sharma et al., 2024), and excessive or miscalibrated safety refusals (Röttger et al., 2024). This has motivated a growing line of work on controlling model behavior directly at inference time by modifying internal activations while keeping model parameters frozen (Turner et al., 2023; Zou et al., 2023). Inference-Time Intervention (ITI) learns truth-related directions at selected attention heads and shifts activations along these directions during inference (Li et al., 2023), while Contrastive Activation Addition (CAA) constructs steering vectors by averaging residual-stream activation differences between positive and negative examples and adds them during generation (Rimsky et al., 2024). Related mechanistic work further suggests that behaviorally meaningful signals can be remarkably low-dimensional; for example, Arditi et al. (2024) identify a one-dimensional direction that mediates refusal behavior across multiple open-source chat models. Together, these results establish activation-space intervention as an attractive paradigm for model control: it is lightweight, comparatively interpretable, and can alter behavior without updating billions of model parameters.

Existing activation steering methods adopt different strategies to construct steering vectors. Some derive them from contrastive activation differences between positive and negative examples (Rimsky et al., 2024; Stolfo et al., 2025), while others explicitly learn steering directions from training data under specific optimization objectives (Cao et al., 2024). Recent approaches further refine steering interventions by leveraging internal representations of the model (Wang et al., 2025). However, we identify a critical limitation: directly learning such directions from semantically complex or limited training data under imperfect behavior modeling can introduce distributional bias. Specifically, as illustrated in Fig. 1, distribution shifts involving previously unseen semantics can lead to errors in predicted steering directions. A detailed analysis of this phenomenon is provided in Sec. 4. Consequently, directly applying the learned direction as the steering intervention can limit both steering effectiveness and generalization. To overcome this limitation, we argue that learnable steering directions should themselves be *steered*, rather than being treated as fully reliable interventions.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

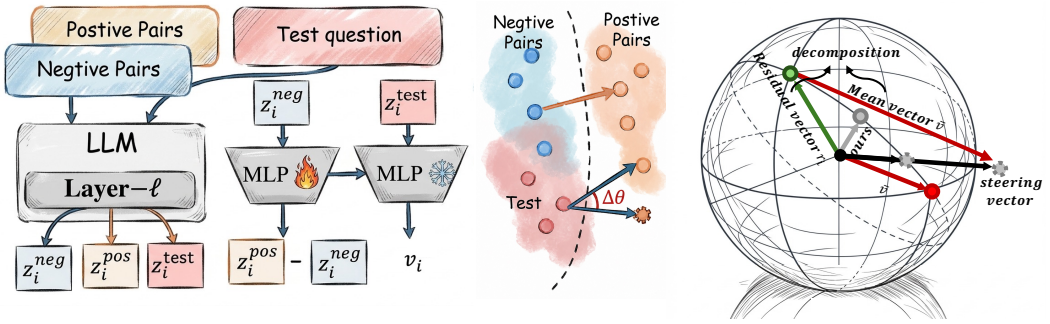



Figure 1: **Overview of S2S.** *Left:* An MLP predicts sample-specific steering vectors. *Center:* Distribution shift can cause directional errors. *Right:* S2S decomposes the predicted vector into a reliable mean direction and a sample-specific residual, then uses spherical interpolation to calibrate the residual toward the mean and obtain a more flexible steering direction.

Motivated by this observation, we introduce a geometric calibration framework for learned interventions in large language models, which we call **S2S**. S2S decomposes the learnable steering vector into a reliable difference-in-means component and a sample-specific residual, and uses SLERP to calibrate the residual toward the reference direction. This preserves useful instance-specific information while mitigating the distributional bias of directly learned directions. We extend S2S to both linear and rotation-based steering, using the calibrated direction as either a linear intervention vector or a rotation target. We further introduce a similarity-aware adaptive weighting strategy that dynamically balances the contributions of the hidden-state, difference-in-means, and residual directions based on their local geometric relationships. On TruthfulQA, S2S-R improves the multiple-choice average over the strongest baseline by 6.85 and 12.85 percentage points on Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct, respectively, while preserving general generation capabilities. Our main contributions are summarized as follows:

- We propose **S2S**, a geometric calibration method for learned interventions in large language models. Motivated by our observation that learnable steering directions exhibit substantial angular deviations, we introduce a steer-to-steer paradigm to geometrically calibrate learnable steering directions.
- We decompose the learned steering direction into a difference-in-means component and a sample-specific residual component, which explicitly captures the input-dependent deviation from the global steering direction. Rather than directly applying this residual, S2S geometrically calibrates it against the difference-in-means reference, allowing the intervention angle to be adaptively corrected for more effective steering.
- We further extend S2S with two general steering paradigms and a similarity-aware adaptive weighting strategy. Extensive experiments demonstrate that S2S consistently outperforms existing steering methods, achieving improvements of up to 23.9% on TruthfulQA while preserving the model's general open-ended generation quality.

## 2 Related Work

**Activation steering.** Activation steering controls frozen language models by modifying hidden states at inference time. ITI identifies truth-predictive attention heads and shifts their activations along learned directions (Li et al., 2023), whereas CAA applies a global residual-stream direction estimated from contrastive activation differences (Rimsky et al., 2024). The finding that refusal is mediated by a one-dimensional subspace further supports low-dimensional behavioral control (Arditi et al., 2024). However, a single global direction cannot capture input-specific variation.

**Learnable representation interventions.** Learnable interventions refine global directions or adapt to individual inputs. SADI rescales behavior-relevant activation elements using semantics without training (Wang et al., 2025). L2S predicts instance-specific contrastive vectors (Parekh et al., 2025). MAST learns output-supervised residual transformations around the CAA mean (Anonymous, 2026). LoFiT identifies task-relevant attention heads and trains additive offsets at those heads (Yin et al., 2024). ReFT learns task-specific hidden-representation interventions, with LoReFT operating

in low-rank linear subspaces (Wu et al., 2024). These methods increase flexibility but depend on reliable intervention directions.

**Rotation-based steering.** Most activation steering methods operate in Euclidean space through additive shifts or affine transformations, which may unintentionally alter activation magnitude. Rotation-based approaches instead modify hidden-state direction while better preserving its geometric structure. Angular Steering (Vu & Nguyen, 2025) controls behavior by rotating activations with an explicit angular parameter, while Spherical Steering (You et al., 2026) performs SLERP along hyperspherical geodesics and restores the original norm, making such interventions particularly compatible with normalized architectures such as RMSNorm (Zhang & Sennrich, 2019).

## 3 PRELIMINARIES

Let $\mathcal{M}$ be a frozen autoregressive language model and let $\mathcal{D} = \{(x_i, \mathcal{Y}_i^+, \mathcal{Y}_i^-)\}_{i=1}^N$ be a contrastive dataset of correct and incorrect answers. At layer $\ell$, we extract the last-token representations for each answer pair and construct the correction vector as the difference between the correct and incorrect representations, $v_i = z_i^{\text{pos}} - z_i^{\text{neg}}$.

**Mean Steering Vector.** Mean steering uses one global correction vector obtained by averaging all contrastive differences, following contrastive activation addition (Rimsky et al., 2024):

$$\bar{v} = \mathbb{E}_{(z^{\text{pos}}, z^{\text{neg}}) \sim \mathcal{D}} \left[ z^{\text{pos}} - z^{\text{neg}} \right], \tag{1}$$

where $\mathcal{D}$ denotes the positive–negative representation pairs constructed from questions in $\mathcal{Q}_{\text{mean}}$. The same vector is applied to all inputs at inference.

**Learning to Steer.** L2S (Parekh et al., 2025) instead trains an MLP $f_\theta$ to directly predict an input-dependent steering vector from the current hidden representation:

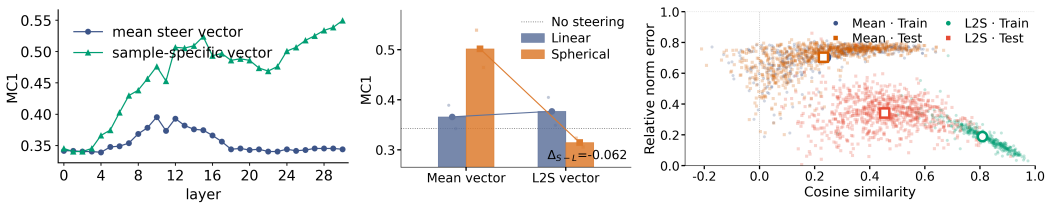
$$\theta^* = \arg\min_\theta \mathbb{E}_{(z^{\text{pos}}, z^{\text{neg}}) \sim \mathcal{D}} \left[ \| f_\theta(z^{\text{neg}}) - (z^{\text{pos}} - z^{\text{neg}}) \|_2^2 \right], \tag{2}$$

where $\mathcal{D}$ denotes the positive–negative representation pairs constructed from questions in $\mathcal{Q}_{\text{route}}$. Unlike mean steering, L2S adapts both direction and magnitude to each input.

## 4 MOTIVATION: THE STEERING-THE-STEERING PERSPECTIVE

Learnable steering methods can construct sample-adaptive directions (Parekh et al., 2025). As shown in Fig. 2, using the oracle sample-specific direction substantially outperforms steering with the mean steering vector, revealing rich sample-specific information beyond the shared mean vector. However, directly predicting such directions fails to fully capture this benefit. For instance, under linear steering, vectors predicted by L2S provide only limited gains over the mean steering vector, while under rotation-based steering, performance can even drop below simple linear steering. As shown in Fig. 2, the predicted vector exhibits errors in both direction and magnitude compared with the oracle vector. These errors are not equally important. Magnitude errors can be largely absorbed by the steering-strength hyperparameter in linear steering and do not affect rotation-based steering. Directional errors, however, affect both steering paradigms and cannot be corrected by simply tuning the steering strength.

We argue that this directional instability comes from the complex structure of sample-specific steering vectors. Rather than representing a single semantic direction, a steering vector can be viewed as a combination of multiple latent steering factors, where different samples require different factors with different strengths. Direct vector prediction therefore needs to determine both which directions to activate and how strongly to apply them, making accurate prediction difficult from semantically entangled LLM representations. This motivates a different approach. Instead of predicting a complete and precise steering vector from scratch, we preserve the sample-specific signal already captured by existing methods and calibrate its direction. This leads to a simple shift in perspective. Rather than only steering model representations, we steer the learnable steering vector itself.




(a) The Effectiveness of sample-specific vector (b) limited improvement (c) Reasons for performance limitations

Figure 2: **Limitations of learned sample-specific steering vectors.** *(a)* Oracle sample-specific vectors outperform the mean vector. *(b)* L2S provides only a small gain under linear steering and degrades spherical steering. *(c)* Direction and magnitude errors of the mean and L2S vectors on the training and test sets, measured by cosine similarity and relative norm error, respectively.

# 5 METHOD

In this section, we present **S2S**, our framework for constructing and calibrating learnable steering vectors. S2S is simple and efficient. It addresses a general limitation of learnable steering methods and achieves consistent performance gains without introducing additional learned reference vectors.

## 5.1 RESIDUAL DIRECTION DECOMPOSITION

We have a reliable reference steering direction $\bar{v}$. The steering vector $v_i(z) = f_\theta(z)$ directly learned by the MLP is decomposed into two components, the mean direction $\bar{v}$ and the residual direction $r_i(z)$. The decomposition is given by:

$$v_i(z) = \bar{v} + r_i(z) \tag{3}$$

The residual direction $r_i(z)$ captures sample-specific steering information, representing the sample-specific deviation from the global mean direction for sample $i$. We separately normalize the two directions as $\hat{v} = \text{norm}(\bar{v})$ and $\hat{r}_i(z) = \text{norm}(r_i(z))$, while retaining $\rho_i(z) = \|v_i(z)\|_2$ as the magnitude of the MLP-learned steering vector.

## 5.2 S2S: STEER THE STEERING

As shown in our motivation analysis, the vector magnitude exhibits relatively small variation between the training and test distributions, whereas its direction is considerably more sensitive to distribution shifts. This observation suggests that the main source of unreliability in the MLP prediction lies in its orientation rather than its magnitude. Therefore, instead of directly using the predicted vector $v_i(z)$ for activation steering, we further calibrate its direction while preserving the magnitude information learned by the MLP.

**Direction calibration with SLERP.** To smoothly calibrate the predicted direction toward the reliable reference direction, we adopt SLERP (Shoemake, 1985). Unlike conventional interpolation methods, SLERP operates on the unit sphere, allowing us to adjust only the direction without introducing unintended changes in vector magnitude, which aligns with our motivation for directional calibration. Given two unit vectors $a$ and $b$, SLERP is defined as

$$\text{SLERP}(a, b; \lambda) = \frac{\sin((1-\lambda)\omega)}{\sin(\omega)} a + \frac{\sin(\lambda\omega)}{\sin(\omega)} b, \quad \omega = \arccos\left(a^\top b\right), \tag{4}$$

where $\omega$ is the angle between the two unit vectors and $\lambda \in [0, 1]$ controls the interpolation strength.

We apply SLERP between the sample-specific residual direction and the reliable mean direction:

$$\hat{s}(z; \lambda) = \text{SLERP}\left(\hat{r}(z), \hat{v}; \lambda\right). \tag{5}$$

Here, $\lambda$ explicitly controls the trade-off between sample-specific information and the reliable reference direction. When $\lambda = 0$, the calibrated direction reduces to the residual direction $\hat{r}_i(z)$, preserving the instance-specific component of the MLP prediction. As $\lambda$ increases, the direction is progressively regularized toward $\hat{v}$. When $\lambda = 1$, it falls back completely to the mean reference direction.

Two common activation steering strategies are *linear steering* (Li et al., 2023; Rimsky et al., 2024) and *rotation-based steering* (Vu & Nguyen, 2025; You et al., 2026). At test time, given a feature representation $z^{\text{test}}$, we obtain its calibrated direction $\hat{s}_i(z^{\text{test}}; \lambda)$ through S2S. Based on these two steering strategies, we define two variants, **S2S-L** and **S2S-R**, respectively.

**S2S-L: S2S with Linear Steering.** For linear steering, we first restore the magnitude predicted by the MLP:

$$\tilde{v}(z^{\text{test}}; \lambda) = \rho(z^{\text{test}})\, \hat{s}(z^{\text{test}}; \lambda). \tag{6}$$

We then apply conventional linear steering (Li et al., 2023; Rimsky et al., 2024) to the test-time representation $z^{\text{test}}$:

$$z^{\text{test}}_{\text{S2S-L}} = z^{\text{test}} + \alpha\, \tilde{v}(z^{\text{test}}; \lambda), \tag{7}$$

where $\alpha$ denotes the steering strength. Thus, S2S-L preserves the magnitude predicted by the MLP while replacing its original direction with the calibrated direction.

**S2S-R: S2S with Rotation Steering.** For rotation-based steering (Vu & Nguyen, 2025; You et al., 2026), only the calibrated direction is required. We first normalize the hidden activation as $\hat{z}^{\text{test}} = \text{norm}(z^{\text{test}})$, and then rotate it toward the calibrated steering direction using SLERP:

$$z^{\text{test}}_{\text{S2S-R}} = \|z^{\text{test}}\|_2\ \text{SLERP}\big(\hat{z}^{\text{test}}, \hat{s}(z^{\text{test}}; \lambda); \alpha\big)\,. \tag{8}$$

S2S-R therefore modifies the direction of the hidden activation while preserving its original magnitude.

### 5.3 Theoretical Analysis

We establish two key properties of our method. First, it expands the search space of steering directions, with both the L2S-predicted direction and the mean vector as special cases, enabling the exploration of potentially better directions. Second, it effectively mitigates the impact of noise.

**Theorem 1** (S2S Expands the Reachable Steering Area)**.** *Let $\mathcal{R}_{\text{MLP}}$ and $\mathcal{R}_{\text{mean}}$ denote the trajectories obtained by steering toward the MLP-predicted and mean directions, respectively. Let $\hat{s}_i(\lambda) = \text{SLERP}(\hat{r}_i, \hat{v}; \lambda)$ and $\mathcal{R}_{\text{S2S}} = \{\text{SLERP}(\hat{z}_i^{-}, \hat{s}_i(\lambda); \alpha) : \alpha, \lambda \in [0, 1]\}$. Then*

$$\mathcal{R}_{\text{S2S}} \supseteq \mathcal{R}_{\text{MLP}} \cup \mathcal{R}_{\text{mean}}.$$

*For each steering area $X$, let $\varepsilon_X = \min_{x \in \mathcal{R}_X} d_{\angle}(\hat{z}_i^*, x)$ denote its best-achievable angular error. Here, $\hat{z}_i^*$ denotes the theoretically optimal direction for sample $i$. Then*

$$\varepsilon_{\text{S2S}} \le \min\left\{\varepsilon_{\text{MLP}}, \varepsilon_{\text{mean}}\right\}.$$

Theorem 1 shows that S2S retains both the sample-specific MLP prediction and the stable mean direction as special cases within an expanded steering space. This provides greater geometric freedom to correct unreliable predictions while ensuring that the best-achievable alignment is no worse than either direction alone. This establishes only a necessary expressivity condition for searching beyond the baselines, not a sufficient condition for improved performance, since the adaptive rule does not optimize over the full region and angular alignment need not determine downstream accuracy. The complete proof is provided in Appendix A.

**Theorem 2** (S2S Attenuates Prediction Noise)**.** *Let $v_i^\star = \bar{v} + r_i^\star$ be the oracle correction, where $\bar{v} = \mathbb{E}[v_i^\star]$. Assume $\mathbb{E}[r_i^\star] = 0$, $\mathbb{E}[\varepsilon_i] = 0$, and that $r_i^\star$ and $\varepsilon_i$ are independent. The MLP predicts $r_i = r_i^\star + \varepsilon_i$. Let $S = \mathbb{E}[\|r_i^\star\|_2^2]$ and $N = \mathbb{E}[\|\varepsilon_i\|_2^2]$, with $S, N > 0$. Assume $N/S \ll 1$ and $\|r_i\|_2/\rho_i \le C$ for a moderate constant $C$. Define $\omega_i = \arccos(\hat{r}_i^\top \hat{v})$. Under these assumptions, S2S reaches the optimal noise-attenuating direction at*

$$\lambda_i^\star = \frac{1}{\omega_i}\,\text{atan2}\left(\|\bar{v}\|_2 \sin\omega_i,\ \frac{S}{S+N}\|r_i\|_2 + \|\bar{v}\|_2 \cos\omega_i\right).$$

*Let $R$ denote the MSE with respect to the oracle correction. The SLERP-represented correction satisfies*

$$R^\star_{\text{S2S}} = \frac{SN}{S+N} < N = R_{\text{MLP}}.$$

Theorem 2 shows that S2S suppresses uncertainty in sample-specific predictions rather than merely expanding the directional search space. It preserves the stable mean component while attenuating the noisy residual according to its signal-to-noise ratio. As residual noise increases, the calibrated direction naturally shifts toward the more reliable mean direction. Since such residual–mean mixtures are reachable by S2S, it can reduce prediction noise while retaining useful sample-specific information. Because S2S-L retains the predicted norm, this shrinkage result approximates the implemented intervention only in the low-noise regime and provides a favorable error condition rather than a sufficient MSE condition. The complete proof is provided in Appendix A.2.

### 5.4 Similarity-Aware Adaptive Weighting

Fixed interpolation coefficients overlook variations in the geometric relationship among the hidden state, global direction, and residual direction. Since rotation-based steering is more sensitive to interpolation weights (You et al., 2026), we introduce an adaptive weighting strategy into our method.

We treat the three normalized directions $u_1 = \hat{z}^{\text{test}}$, $u_2 = \hat{v}$, and $u_3 = \hat{r}_i(z^{\text{test}})$ as three nodes. We define their pairwise directed affinities as

$$A_{pq} = \frac{\exp\big(\kappa u_p^\top u_q\big)}{\sum_{\ell \neq p} \exp\big(\kappa u_p^\top u_\ell\big)}, \; p \neq q, \tag{9}$$

where $p, q \in \{1, 2, 3\}$ and $\kappa > 0$ controls the sharpness of the similarity attention. The incoming-affinity score of node $q$ is $c_q = \sum_{p \neq q} A_{pq}$, and the adaptive weights are

$$[w_z, w_v, w_r] = \text{softmax}(\kappa\,[c_1, c_2, c_3])\,, \tag{10}$$

Here, $w_z$, $w_v$, and $w_r$ correspond to the hidden-state direction, the mean direction, and the residual direction, respectively.

Rather than relying solely on the similarity between $v$ and $r$, we also consider their similarity to $z^{\text{test}}$, since $r$ is derived from $z^{\text{test}}$. Inspired by the two-stage SLERP weighting design in Praun & Hoppe (2003), we define our SLERP formulation as:

$$\begin{aligned} d &= \text{SLERP}\left(\hat{r}_i(z^{\text{test}}), \hat{v}; \frac{w_v}{w_v + w_r}\right), \\ \tilde{z}^{\text{test}}_{\text{S2S}-\text{R}} &= \|z^{\text{test}}\|_2 \; \text{SLERP}\big(\hat{z}^{\text{test}}, d; w_v + w_r\big)\,. \end{aligned} \tag{11}$$

Here, $w_v$ and $w_r$ determine the direction within the mean–residual steering space, while their sum adaptively scales the rotation strength.

Finally, we apply a margin gate to avoid modifying representations whose current direction is already dominant. With margin $m$, the final rotation output is

$$z^{\text{test}}_{\text{S2S}-\text{R}} = \begin{cases} \tilde{z}^{\text{test}}_{\text{S2S}-\text{R}}, & \max(w_v, w_r) > w_z + m, \\ z^{\text{test}}, & \text{otherwise.} \end{cases} \tag{12}$$

## 6 Experiments

### 6.1 Experimental Setup

**Benchmarks and Models.** TruthfulQA (Lin et al., 2022) contains 817 questions across 38 categories. We test cross-benchmark generalization on COPA (Roemmele et al., 2011), StoryCloze (Mostafazadeh et al., 2016), MMLU (Hendrycks et al., 2021), WinoGrande (Sakaguchi et al., 2020), and BoolQ (Clark et al., 2019), using frozen Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen2.5-7B-Instruct (Yang et al., 2024). For each question–answer pair, we extract the last-token representation and pair the mean MC2 positive-answer representation with each incorrect answer to form correction vectors. We use question-level two-fold cross-validation, splitting the construction fold equally for mean-vector estimation and MLP training. Interventions are applied at layers 14 and 19 for Llama and Qwen, respectively.

**Baselines and Evaluation.** We compare S2S with the unmodified model, ITI (Li et al., 2023), CAA (Rimsky et al., 2024), SADI-HEAD (Wang et al., 2025), L2S (Parekh et al., 2025), and Spherical

Table 1: TruthfulQA performance (%) of Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct on multiple-choice and open-ended generation. Higher is better; within each model and steering paradigm, **bold** and underlined values denote the best and second-best results.

| Model | Method | Multiple-Choice | | | | Open-ended Generation | | |
|---|---|---|---|---|---|---|---|---|
| | | MC1 | MC2 | MC3 | Avg. | TRUE | INFO | TRUE×INFO |
| Qwen-2.5-7B-Instruct | *Baseline* | 35.87 | 54.95 | 26.62 | 39.15 | 83.65 | 88.93 | 74.40 |
| | *Linear steering* | | | | | | | |
| | ITI (Li et al., 2023) | 40.15 | 58.93 | 30.26 | 43.11 | **99.60** | 68.10 | 67.82 |
| | CAA (Rimsky et al., 2024) | 40.16 | 59.08 | 30.41 | 43.22 | 98.99 | 68.63 | 67.95 |
| | SADI-HEAD (Wang et al., 2025) | 39.28 | 56.61 | 29.32 | 41.74 | 97.00 | 71.23 | 69.09 |
| | L2S (Parekh et al., 2025) | 36.23 | 56.79 | 27.66 | 40.23 | 85.88 | **88.46** | **75.97** |
| | **S2S-L (Ours)** | **45.04** | **65.70** | **36.94** | **49.23** | 87.58 | 85.77 | 75.11 |
| | *Rotation-based steering* | | | | | | | |
| | Spherical Steering (You et al., 2026) | 48.59 | 66.96 | 39.12 | 51.56 | **87.78** | 88.39 | **77.59** |
| | **S2S-R (Ours)** | **62.55** | **75.92** | **54.75** | **64.41** | 86.92 | **88.67** | 77.07 |
| Llama-3.1-8B-Instruct | *Baseline* | 34.15 | 53.32 | 27.02 | 38.16 | 82.32 | 58.60 | 48.24 |
| | *Linear steering* | | | | | | | |
| | ITI (Li et al., 2023) | 37.70 | 58.09 | 30.12 | 41.97 | **89.66** | 44.96 | 40.31 |
| | CAA (Rimsky et al., 2024) | 35.99 | 56.26 | 29.36 | 40.54 | 84.63 | 58.67 | 49.66 |
| | SADI-HEAD (Wang et al., 2025) | 38.53 | 56.03 | 30.57 | 41.71 | 80.30 | 63.74 | 51.18 |
| | L2S (Parekh et al., 2025) | 36.60 | 56.27 | 29.16 | 40.68 | 86.72 | 56.67 | 49.15 |
| | **S2S-L (Ours)** | **43.95** | **64.79** | **38.01** | **48.91** | 83.31 | **81.97** | **68.28** |
| | *Rotation-based steering* | | | | | | | |
| | Spherical Steering (You et al., 2026) | 49.94 | 68.50 | 41.02 | 53.16 | 82.12 | **66.28** | **54.44** |
| | **S2S-R (Ours)** | **57.17** | **71.76** | **51.10** | **60.01** | **85.14** | 62.68 | 53.43 |

Table 2: Zero-shot multiple-choice results on Llama-3.1-8B-Instruct using MC accuracy (%).

| Method | TruthfulQA | COPA | StoryCloze | MMLU | Wino. | BoolQ | Avg. |
|---|---|---|---|---|---|---|---|
| *Baseline* | 34.15 | 83.00 | 74.72 | 60.60 | 50.81 | 80.12 | 63.90 |
| *Linear steering* | | | | | | | |
| ITI (Li et al., 2023) | 37.70 | 83.00 | 75.12 | **60.90** | 51.85 | 81.53 | 65.02 |
| CAA (Rimsky et al., 2024) | 35.99 | 84.00 | 79.02 | 60.70 | **51.93** | 82.42 | 65.68 |
| SADI-HEAD (Wang et al., 2025) | 38.53 | 84.00 | 75.72 | 60.66 | 51.85 | 80.20 | 65.16 |
| L2S (Parekh et al., 2025) | 36.60 | 84.00 | 75.64 | 60.40 | 50.75 | **83.33** | 65.12 |
| **S2S-L (Ours)** | **43.95** | **85.00** | **85.97** | 60.50 | 51.62 | 82.26 | **68.22** |
| *Rotation-based steering* | | | | | | | |
| Spherical Steering (You et al., 2026) | 49.94 | **95.00** | **89.08** | **62.05** | 52.64 | **82.81** | 71.92 |
| **S2S-R (Ours)** | **57.17** | 92.00 | 89.01 | 60.40 | **53.75** | 80.94 | **72.21** |

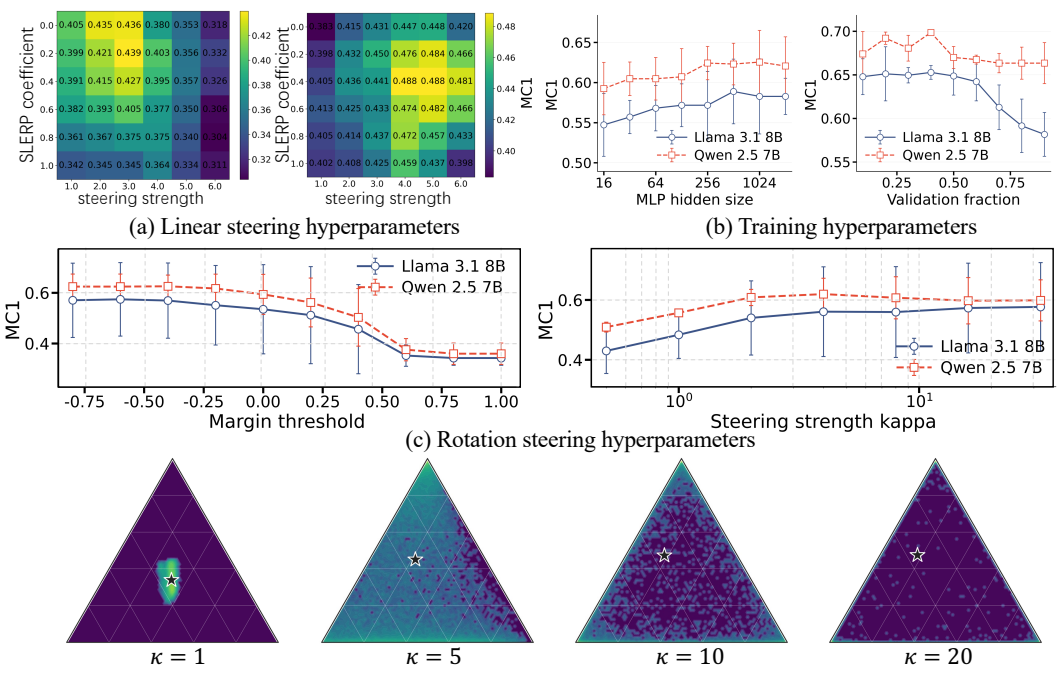
Steering (You et al., 2026). Multiple-choice answers are scored by conditional log-likelihood and evaluated with MC1, MC2, MC3, and their average. For open-ended generation, we use the pre-trained TruthfulQA judges to measure truthfulness and informativeness and report TRUE×INFO as the main generation metric.

### 6.2 Main Results

Table 1 presents the main TruthfulQA results. S2S-R achieves the best MC averages of 60.01% on Llama-3.1-8B-Instruct and 64.41% on Qwen2.5-7B-Instruct, surpassing the strongest baseline by 6.85 and 12.85 percentage points, respectively. S2S-L also leads open-ended generation on Llama, reaching 81.97% INFO and 68.28% TRUE×INFO while maintaining truthfulness. These results show that S2S effectively controls model representations while preserving intrinsic geometry and reducing the trade-offs of linear steering. Table 2 further evaluates generalization across reasoning benchmarks. S2S-R reaches 57.17% on TruthfulQA MC1 and remains competitive on COPA, StoryCloze, and WinoGrande. Overall, the results indicate that geometry-aware representation manipulation improves semantic discrimination and decision boundaries beyond TruthfulQA.

### 6.3 Ablation Study

**Linear steering hyperparameters.** The steering strength controls the magnitude of the intervention, while the SLERP coefficient balances the residual direction and the global mean direction: $\lambda = 0$ gives the sample-specific residual direction, increasing $\lambda$ moves the direction toward the




(a) Linear steering hyperparameters

(b) Training hyperparameters

(c) Rotation steering hyperparameters




(d) Relationship with Rotation steering hyperparameters and weights

Figure 3: **Hyperparameter analysis on TruthfulQA.** *(a)* SLERP coefficient and steering strength. *(b)* MLP hidden size and validation fraction. *(c)* Gate margin and similarity sharpness $\kappa$. *(d)* Adaptive weights across $\kappa$ values; stars and error bars denote means and variation across runs.

global mean, and $\lambda = 1$ gives the global mean reference direction. As shown in Fig. 3(a), the two models exhibit different sensitivities to the residual component. Llama achieves better performance with a small SLERP coefficient and moderate steering strength, indicating a stronger preference for the sample-specific residual direction. In contrast, Qwen benefits from a moderate balance between the residual and global mean directions and generally requires a larger steering strength. Overall, the optimal balance between the global and residual directions is model-dependent, highlighting the importance of adapting the linear steering configuration to each model.

**Rotation-based steering hyperparameters.** We analyze two key hyperparameters in rotation-based steering, the margin and adaptive weight parameter $\kappa$, in Fig. 3(c). The margin controls the intervention gate. An overly large margin restricts steering and degrades MC1, while an overly small margin steers nearly all tokens and slightly affects performance. The parameter $\kappa$ controls the sharpness of similarity-based adaptive weighting. Increasing $\kappa$ encourages more differentiated, token-specific weights and generally improves MC1 before performance saturates. The adaptive-weight distributions in Fig. 3(d) further show that larger $\kappa$ produces more differentiated component weights. Overall, moderate margin and $\kappa$ values provide selective and stable steering.

**Training hyperparameters.** The MLP hidden size controls the capacity of the steering-vector predictor, while the validation fraction determines the data allocation between MLP training and global direction estimation. As shown in Fig. 3(b), increasing the hidden size generally improves MC1 before the performance gradually saturates, indicating that sufficient MLP capacity is beneficial for learning effective sample-specific vectors. For the validation fraction, a small value provides insufficient data for training the MLP, while a large value leaves fewer samples for estimating a robust global direction and leads to performance degradation, particularly for Llama. Overall, moderate MLP capacity and an appropriate validation fraction provide a balanced training configuration and enable more effective and stable steering.

### 6.4 Analysis

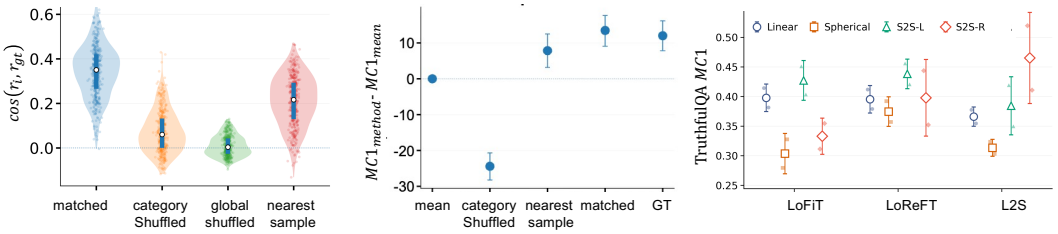
**Discussion of MLP predicted magnitude and direction.** The MLP prediction contains both magnitude and direction, whose utility depends on the steering strategy. Figure 2(b) shows that sample-

Table 3: **Component ablation on TruthfulQA.** Effects of the MLP, S2S, adaptive weighting (AW), and gate.

| MLP | S2S | AW | Gate | MC1 | MC2 | MC3 |
|---|---|---|---|---|---|---|
| × | × | × | × | 41.42 | 60.61 | 32.87 |
| ✓ | × | × | × | 37.01 | 55.56 | 28.31 |
| ✓ | ✓ | × | × | 43.14 | 60.91 | 32.87 |
| ✓ | ✓ | ✓ | × | 66.91 | 78.30 | 62.96 |
| ✓ | ✓ | ✓ | ✓ | 67.40 | 78.46 | 63.19 |

Table 4: **Generation throughput (tokens/s) on a single 3090 GPU.** *Generation-only* uses known question–answer pairs; *open-ended* uses only the question.

| Method | Generation-only | Open-ended |
|---|---|---|
| No Steering | 25.43±0.60 | 25.43±0.60 |
| Spherical Steering | 25.23±0.78 | 25.23±0.78 |
| **S2S-L (Ours)** | 24.96±0.80 | 12.49±0.38 |
| **S2S-R (Ours)** | 23.61±0.66 | 12.22±0.30 |




(a) Different shuffled strategy (b) Result for Different shuffled strategy (c) Different learned steer framework

Figure 4: **Analysis of residual-direction generalization.** *(a)* Cosine similarity to the target residual under matched, category-shuffled, globally shuffled, and nearest-sample replacement. *(b)* Corresponding MC1 changes relative to mean steering. *(c)* TruthfulQA MC1 when S2S is applied to steering vectors learned by LoFiT, LoReFT, and L2S.

specific magnitude benefits linear steering, whereas directly using the predicted direction yields limited gains. Figure 2(c) further reveals errors in both components, potentially due to distribution shifts. Rotation steering nevertheless achieves strong gains from direction alone when it is properly calibrated. Thus, the MLP provides useful but imperfect information, motivating its decomposition and recalibration.

**Discussion of residual direction.** We investigate the generalization of the MLP-predicted residual direction. As shown in Fig. 4(a) and (b), both cross-category and global shuffling yield low cosine similarity and poor performance. In contrast, replacing the original residual with the most directionally aligned residual from other categories substantially improves performance. This suggests that residual generalization depends more on directional alignment than semantic similarity. Even semantically similar samples fail to generalize when their directions differ substantially, whereas semantically unrelated but well-aligned residuals can improve performance.

**Discussion of different learnable steering methods.** We further evaluate S2S with LoFiT (Yin et al., 2024) and LoReFT (Wu et al., 2024), which differ from L2S in architecture and training objective and produce substantially lower cosine similarity. Figure 4(c) shows that S2S consistently improves all three learnable steering frameworks, demonstrating its generalization across intervention methods. Further analysis is provided in Appendix B.4. These results suggest that distributional mismatch may be common and that S2S remains effective even under low directional similarity.

## 7 Conclusion

We introduced **S2S**, a geometric calibration framework that treats learnable steering directions as imperfect predictions. By decomposing each prediction into mean and residual components and recalibrating them through spherical interpolation, S2S improves activation steering across TruthfulQA and reasoning benchmarks while preserving general generation quality. The framework supports both additive and rotation-based interventions through a unified geometric formulation. Future work should theoretically characterize why this geometric recalibration can improve performance without learning additional steering vectors, including the conditions under which it improves alignment and generalization.

### AI USE STATEMENT

This research utilized AI assistance for code debugging and grammatical refinement. All experimental designs, analyses, results, and conclusions were developed independently by the authors without generative AI input. We employed AI tools solely for technical implementation support and language polishing to ensure clear communication of our findings.

## REFERENCES

Anonymous. Supervised refinement of contrastive activation steering for truthful question answering. Anonymous ACL submission, 2026. OpenReview submission.

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. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. Personalized steering of large language models: Versatile steering vectors through bidirectional preference optimization. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2924–2936. Association for Computational Linguistics, 2019.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *The Ninth International Conference on Learning Representations*, 2021.

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):1–55, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *The Third International Conference on Learning Representations*, 2015.

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

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 3214–3252. Association for Computational Linguistics, 2022.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 839–849. Association for Computational Linguistics, 2016.

Jayneel Parekh, Pegah Khayatan, Mustafa Shukor, Arnaud Dapogny, Alasdair Newson, and Matthieu Cord. Learning to steer: Input-dependent steering for multimodal LLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Emil Praun and Hugues Hoppe. Spherical parametrization and remeshing. *ACM Transactions on Graphics*, 22(3):340–349, 2003.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 15504–15522. Association for Computational Linguistics, 2024.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, 2011.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5377–5400. Association for Computational Linguistics, 2024.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8732–8740, 2020.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 245–254. Association for Computing Machinery, 1985.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. In *The Thirteenth International Conference on Learning Representations*, 2025.

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.

Minh Hieu Vu and Tan M. Nguyen. Angular steering: Behavior control via rotation in activation space. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

An Yang et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Fangcong Yin, Xi Ye, and Greg Durrett. LoFiT: Localized fine-tuning on LLM representations. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Zejia You, Chunyuan Deng, and Hanjie Chen. Spherical steering: Geometry-aware activation rotation for language models. *arXiv preprint arXiv:2602.08169*, 2026.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

# A WHY STEERING INTERVENTIONS NEED TO BE STEERED

## A.1 REACHABLE STEERING AREA

Consider an adaptive activation-steering setting in which an MLP predicts an input-dependent steering vector $v_i$. We decompose the prediction into a global mean component $\bar{v}$ and a sample-specific residual $r_i$:

$$v_i = \bar{v} + r_i, \qquad \hat{v} = \mathrm{norm}(\bar{v}), \qquad \hat{r}_i = \mathrm{norm}(r_i), \qquad \hat{v}_i = \mathrm{norm}(v_i), \qquad \rho_i = \|v_i\|_2. \tag{13}$$

Here $\hat{v}_i$ is the direction that would be used if the MLP prediction were applied directly. For a positive–negative pair, let $\hat{z}_i^{\mathrm{neg}}$ denote the normalized representation to be corrected and $\hat{z}_i^{\mathrm{pos}}$ the desired positive direction. Rather than applying the predicted direction directly, S2S first re-steers the intervention by interpolating between the residual and mean directions:

$$\hat{s}_i(\lambda) = \mathrm{SLERP}(\hat{r}_i, \hat{v}; \lambda), \qquad \lambda \in [0, 1]. \tag{14}$$

S2S-R then rotates the current representation toward this calibrated direction:

$$g_i(\alpha, \lambda) = \mathrm{SLERP}(\hat{z}_i^{\mathrm{neg}}, \hat{s}_i(\lambda); \alpha)\,, \qquad \alpha \in [0, 1]. \tag{15}$$

Thus $\lambda$ controls how the steering direction itself is adjusted between sample-specific and global information, while $\alpha$ controls how strongly the hidden representation is moved toward that direction. Under adaptive weighting, the coefficients are obtained from the two similarity weights as

$$\lambda = \frac{w_v}{w_v + w_r}, \qquad \alpha = w_v + w_r, \tag{16}$$

before the gating operation. Fixed-coefficient S2S is another member of the same two-parameter region. We assume that all normalized vectors below are well defined and that $\hat{z}_i^{\mathrm{neg}}$, $\hat{r}_i$, and $\hat{v}$ lie in a common open hemisphere. This excludes antipodal ambiguities and ensures that every minor-arc SLERP used below is unique. The result is a reachable-steering statement: it characterizes which directions S2S makes available, rather than claiming that the adaptive weighting rule always selects the best one.

**Definition 1.** For unit directions $a, b$, define their angular distance by $d_\angle(a, b) = \arccos(a^\top b)$. For unit steering ingredients $x_1, \ldots, x_k$, define

$$\mathcal{C}(x_1, \ldots, x_k) = \left\{ \mathrm{norm}\left( \sum_{j=1}^{k} c_j x_j \right) : c_j \geq 0,\ \sum_{j=1}^{k} c_j > 0 \right\}. \tag{17}$$

We call this set the *reachable steering region*. It is the normalized form of the conic hull generated by the available directions.

**Lemma 1.** For two distinct, non-antipodal unit directions $a$ and $b$, varying the SLERP coefficient over its full range reaches exactly their normalized conic hull:

$$\{\mathrm{SLERP}(a, b; t) : t \in [0, 1]\} = \mathcal{C}(a, b). \tag{18}$$

*Proof.* Let $\theta = d_\angle(a, b) \in (0, \pi)$. The interpolation can be written as

$$\mathrm{SLERP}(a, b; t) = \frac{\sin((1-t)\theta)}{\sin\theta} a + \frac{\sin(t\theta)}{\sin\theta} b. \tag{19}$$

Both coefficients are nonnegative for $t \in [0, 1]$, so every SLERP direction belongs to $\mathcal{C}(a, b)$. Conversely, every normalized nonnegative mixture of $a$ and $b$ lies on their unique minor arc and is attained by some $t \in [0, 1]$. This proves the lemma.

**Lemma 2 (Direct MLP steering is retained).** The residual–mean interpolation does not discard the direction predicted by the MLP. In particular, there exists a coefficient $\lambda_i \in [0, 1]$ such that

$$\hat{v}_i = \mathrm{SLERP}(\hat{r}_i, \hat{v}; \lambda_i), \quad \hat{v}_i \in \mathcal{C}(\hat{r}_i, \hat{v}). \tag{20}$$

*Proof.* Because the mean and residual vectors are positive scalar multiples of their normalized directions, the decomposition $v_i = \bar{v} + r_i$ makes the normalized MLP prediction a nonnegative mixture of $\hat{r}_i$ and $\hat{v}$. Lemma 1 then yields the required interpolation coefficient.

$$\hat{v}_i = \mathrm{norm}(\|\bar{v}\|_2 \hat{v} + \|r_i\|_2 \hat{r}_i) \in \mathcal{C}(\hat{r}_i, \hat{v}). \tag{21}$$

**Theorem 1 (S2S expands the reachable steering area).** Let $\mathcal{R}_{\mathrm{MLP}}$ and $\mathcal{R}_{\mathrm{mean}}$ denote the trajectories obtained by steering toward the MLP-predicted and mean directions, respectively. Let $\hat{s}_i(\lambda) = \mathrm{SLERP}(\hat{r}_i, \hat{v}; \lambda)$ and $\mathcal{R}_{\mathrm{S2S}} = \{\mathrm{SLERP}(\hat{z}_i^{\mathrm{neg}}, \hat{s}_i(\lambda); \alpha) : \alpha, \lambda \in [0,1]\}$ . Then

$$\mathcal{R}_{\mathrm{S2S}} \supseteq \mathcal{R}_{\mathrm{MLP}} \cup \mathcal{R}_{\mathrm{mean}}. \tag{22}$$

For each steering area $X$, let $\varepsilon_X = \min_{x \in \mathcal{R}_X} d_\angle(\hat{z}_i^*, x)$ denote its best-achievable angular error. Then

$$\varepsilon_{\mathrm{S2S}} \leq \min\{\varepsilon_{\mathrm{MLP}}, \varepsilon_{\mathrm{mean}}\}. \tag{23}$$

*Proof.* Setting $\lambda = 1$ in the definition of $\hat{s}_i(\lambda)$ gives $\hat{s}_i(1) = \hat{v}$, and therefore $\mathcal{R}_{\mathrm{mean}} \subseteq \mathcal{R}_{\mathrm{S2S}}$. By Lemma 2, there also exists $\lambda_i$ such that $\hat{s}_i(\lambda_i) = \hat{v}_i$, which gives $\mathcal{R}_{\mathrm{MLP}} \subseteq \mathcal{R}_{\mathrm{S2S}}$. This proves Eq. (22). Equation (23) then follows immediately by minimizing the same angular error over a superset of feasible directions.

### A.2 Noise Attenuation by S2S Re-Steering

**Connection between residual shrinkage and SLERP.** Following the magnitude notation in the main text, let $\rho_i = \|v_i\|_2$ be the magnitude predicted by the MLP. Write $a = \|\bar{v}\|_2$, $b_i = \|r_i\|_2$, and $\omega_i = \arccos(\hat{r}_i^\top \hat{v}) \in (0, \pi)$. For any residual coefficient $\beta \in [0,1]$, consider the direction of the auxiliary correction $\bar{v} + \beta r_i$. Its SLERP coefficient is

$$\lambda_i(\beta) = \frac{1}{\omega_i} \operatorname{atan2}(a \sin\omega_i,\ \beta b_i + a\cos\omega_i), \tag{24}$$

where $\operatorname{atan2}(y, x)$ takes its angle in $[0, \pi]$ for $y > 0$. Indeed, resolving the auxiliary correction into components parallel and orthogonal to $\hat{r}_i$ shows that its angular displacement from $\hat{r}_i$ is precisely $\lambda_i(\beta)\omega_i$. Therefore,

$$\hat{s}_i(\lambda_i(\beta)) = \mathrm{SLERP}(\hat{r}_i, \hat{v}; \lambda_i(\beta)) = \mathrm{norm}(\bar{v} + \beta r_i). \tag{25}$$

In the notation of the main text, the corresponding S2S-L correction is $\tilde{v}_i(\lambda) = \rho_i \hat{s}_i(\lambda)$. Notice that $\rho_i$ is kept fixed by S2S-L, whereas $\|\bar{v} + \beta r_i\|_2$ generally varies with $\beta$. Thus Eq. (25) is an exact *directional* equivalence, not an equality of correction magnitudes.

**Theorem 2 (S2S attenuates prediction noise).** Let $v_i^\star = \bar{v} + r_i^\star$ be the oracle correction, where $\bar{v} = \mathbb{E}[v_i^\star]$. Assume $\mathbb{E}[r_i^\star] = 0$, $\mathbb{E}[\varepsilon_i] = 0$, and that $r_i^\star$ and $\varepsilon_i$ are independent. The MLP predicts $r_i = r_i^\star + \varepsilon_i$. Let $S = \mathbb{E}[\|r_i^\star\|_2^2]$ and $N = \mathbb{E}[\|\varepsilon_i\|_2^2]$, with $S, N > 0$. Assume $N/S \ll 1$ and $\|r_i\|_2/\rho_i \leq C$ for a moderate constant $C$. Under these assumptions, S2S reaches the optimal noise-attenuating direction at

$$\lambda_i^\star = \frac{1}{\omega_i} \operatorname{atan2}\left(\|\bar{v}\|_2 \sin\omega_i,\ \frac{S}{S+N}\|r_i\|_2 + \|\bar{v}\|_2 \cos\omega_i\right). \tag{26}$$

Let $R$ denote the MSE with respect to the oracle correction. The SLERP-represented correction satisfies

$$R_{\mathrm{S2S}}^\star = \frac{SN}{S+N} < N = R_{\mathrm{MLP}}. \tag{27}$$

*Proof.* For any $\beta \in [0,1]$, let $\lambda_i(\beta)$ be the SLERP coefficient in Eq. (24). By the directional equivalence established above, the resulting S2S direction is the normalized direction of $\bar{v} + \beta r_i$. Let $q_i(\beta) = \|\bar{v} + \beta r_i\|_2$ and define the vector represented by this SLERP direction and its induced magnitude as

$$u_i(\beta) = q_i(\beta)\ \mathrm{SLERP}(\hat{r}_i, \hat{v}; \lambda_i(\beta)) = \bar{v} + \beta r_i. \tag{28}$$

Using $r_i = r_i^\star + \varepsilon_i$ and $v_i^\star = \bar{v} + r_i^\star$, its error with respect to the oracle correction is

$$\begin{aligned} u_i(\beta) - v_i^\star &= \bar{v} + \beta(r_i^\star + \varepsilon_i) - (\bar{v} + r_i^\star) \\ &= (\beta - 1) r_i^\star + \beta\varepsilon_i. \end{aligned} \tag{29}$$

The assumptions $\mathbb{E}[r_i^\star] = 0$, $\mathbb{E}[\varepsilon_i] = 0$, and independence imply $\mathbb{E}[(r_i^\star)^\top \varepsilon_i] = 0$. Therefore the expected squared error along this SLERP-represented family becomes

$$\begin{aligned} R_{\mathrm{S2S}}(\beta) &= \mathbb{E}\left[\|u_i(\beta) - v_i^\star\|_2^2\right] \\ &= (1-\beta)^2 \mathbb{E}[\|r_i^\star\|_2^2] + \beta^2 \mathbb{E}[\|\varepsilon_i\|_2^2] + 2\beta(\beta-1)\mathbb{E}[(r_i^\star)^\top \varepsilon_i] \\ &= (1-\beta)^2 S + \beta^2 N. \end{aligned} \tag{30}$$

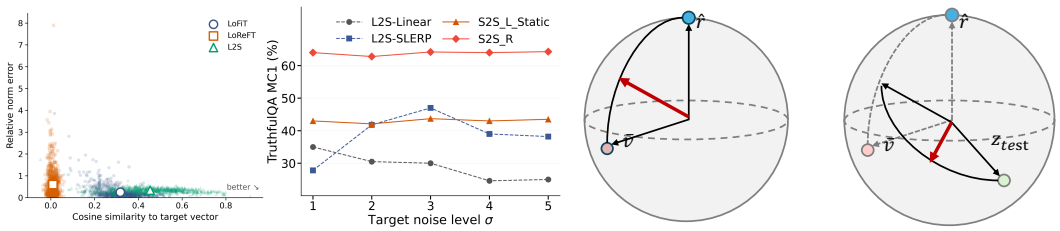



Figure 5: **Prediction-error and noise-robustness analysis.** The left panel compares directional similarity and relative norm error across learnable steering methods. The middle panel reports TruthfulQA MC1 under increasing target-noise levels. The right panels illustrate calibration of the predicted residual toward the mean direction and the subsequent rotation of the test representation.

Differentiating gives $dR_{\mathrm{S2S}}/d\beta = -2(1-\beta)S + 2\beta N$. The unique stationary point is therefore $\beta^\star = S/(S+N)$, and the second derivative is $2(S+N) > 0$, so it is the unique minimizer. Substituting $\beta^\star$ gives

$$\begin{aligned} R^\star_{\mathrm{S2S}} &= \left(1 - \frac{S}{S+N}\right)^2 S + \left(\frac{S}{S+N}\right)^2 N \\ &= \frac{SN}{S+N} < N = R_{\mathrm{MLP}}. \end{aligned} \tag{31}$$

Substituting the same $\beta^\star$ into Eq. (24) gives exactly $\lambda_i^\star$ in Eq. (26).

It remains to compare the induced magnitude $q_i^\star$ with the magnitude $\rho_i$ retained by the actual S2S-L correction. Since $\rho_i = \|\bar{v} + r_i\|_2$ and $q_i^\star = \|\bar{v} + \beta \cdot r_i\|_2$, the reverse triangle inequality gives

$$\begin{aligned} \frac{|\rho_i - q_i^\star|}{\rho_i} &\le \frac{\|(1-\beta^\star) r_i\|_2}{\rho_i} \\ &= \frac{N}{S+N} \frac{\|r_i\|_2}{\rho_i} \\ &\le C \frac{N}{S+N} \le C \frac{N}{S}. \end{aligned} \tag{32}$$

Under the theorem assumption $N/S \ll 1$, the last quantity is small, and hence $q_i^\star \approx \rho_i$. Finally, the actual S2S-L correction can be written as $\tilde{v}_i(\lambda_i^\star) = \rho_i \,\mathrm{SLERP}(\hat{r}_i, \hat{v}; \lambda_i^\star)$. It therefore uses exactly the same direction as $u_i(\beta^\star)$, while the preceding bound shows that its magnitude is close to $q_i^\star$. Thus, in the low-noise regime, S2S-L closely realizes the lower-error SLERP-represented correction while retaining the original MLP-predicted magnitude.

### A.3 Scope and Limitations of the Theoretical Analysis

Theorem 1 is an expressivity result: it shows that the reachable region contains the MLP and mean-steering trajectories, but it does not show that the implemented adaptive rule selects the best point in this region. It therefore establishes only a necessary expressivity condition for searching beyond the baselines, not a sufficient condition for improved performance; moreover, angular error is only a geometric surrogate and need not vary monotonically with downstream accuracy.

The exact risk $SN/(S+N)$ in Theorem 2 belongs to the auxiliary estimator $u_i(\beta) = \bar{v} + \beta r_i$, whose norm varies with $\beta$. Actual S2S-L instead retains $\rho_i$, and Eq. (32) only bounds the resulting norm mismatch by $O(N/S)$ under the low-noise assumption. Thus, the theorem identifies a favorable error regime and an approximately realizable direction, rather than a strict excess-risk bound for the implemented estimator.

# B Additional Analysis and Formulas

## B.1 Implementation and Hyperparameter Details

We use a two-layer MLP with a 256-dimensional hidden layer and GELU activation (Hendrycks & Gimpel, 2016). It is optimized with Adam (Kingma & Ba, 2015) for 2000 epochs using a batch size of 256 and a learning rate of $10^{-3}$. The MLP predicts steering vectors from $\ell_2$-normalized hidden states under an MSE objective. We set the similarity sharpness to $\kappa = 3$ and the steering margin to $-0.7$. The final intervention uses the two-stage SLERP procedure in Eq. (11) and restores the original hidden-state norm.

## B.2 Direction and Magnitude Errors

For the analysis in Fig. 2, let $v_i$ denote a nonzero candidate mean or L2S-predicted steering vector and let $v_i^\star$ denote the corresponding nonzero oracle sample-specific vector. We measure their directional agreement using cosine similarity:

$$\mathrm{CosSim}(v_i, v_i^\star) = \frac{v_i^\top v_i^\star}{\|v_i\|_2 \|v_i^\star\|_2}. \tag{33}$$

where a larger value indicates better directional alignment. Their magnitude discrepancy is measured using relative norm error:

$$\mathrm{RelNormErr}(v_i, v_i^\star) = \frac{|\|v_i\|_2 - \|v_i^\star\|_2|}{\|v_i^\star\|_2}. \tag{34}$$

A value of zero indicates that the candidate and oracle vectors have equal norms, while larger values indicate greater relative magnitude error. These two metrics separate angular mismatch from scale mismatch and correspond to the horizontal and vertical axes, respectively, in the right panel of Fig. 2.

## B.3 Interpretation of the Residual Decomposition

The decomposition $r_i = v_i - \bar{v}$ is not intended as an orthogonal projection or a semantic disentanglement of the steering vector. We call $r_i$ sample-specific only in the conditional sense that it contains the variation in the MLP prediction beyond the shared mean component; it may remain correlated with $\bar{v}$ and may combine several latent factors. A direct MLP prediction must jointly determine which steering factors are relevant and their magnitudes, making the complete vector difficult to generalize across semantically diverse inputs. The residual retains the input-dependent directional information already learned by the MLP, while S2S introduces the more reliable mean vector as an anchor to reduce reliance on the raw predicted direction. Thus, our goal is not to claim that the residual is a pure semantic component, but to test whether this minimal decomposition can improve a learned intervention without additional supervision, reference vectors, or decomposition machinery. Our experiments show that the existing mean direction is sufficient to provide an effective calibration reference.

## B.4 Distribution Bias Beyond Vector-Level Regression

The discrepancy in Fig. 2 provides direct evidence of distribution bias when the learnable steering vector is explicitly trained to match an oracle correction through a vector-level MSE objective. The left panel of Fig. 5 suggests that a related directional mismatch also appears across other learnable intervention frameworks, indicating that the issue is not unique to a single predictor. However, this comparison should be interpreted more cautiously than Fig. 2. In particular, our LoFiT and LoReFT variants (Yin et al., 2024; Wu et al., 2024) are trained with a DPO-based output objective (Rafailov et al., 2023) rather than direct MSE supervision on the target vector. Their learned interventions are therefore not expected to match the oracle steering vector pointwise, because the output-optimal intervention under DPO need not coincide with the MSE-optimal correction vector.

Consequently, the cosine-similarity gap in the left panel of Fig. 5 may reflect not only directional prediction error but also how semantically related inputs are represented in different regions of the LLM activation space. A relatively large angular difference is therefore unsurprising and constitutes

weaker evidence than the controlled vector-regression analysis in Fig. 2. Nevertheless, the right panel of Fig. 4 shows that applying S2S consistently improves LoFiT, LoReFT, and L2S despite their different objectives and levels of cosine alignment. This improvement demonstrates that S2S remains effective even when the learned intervention is not directly optimized to reproduce an oracle steering vector.

**Scope of the distribution-shift evidence.** Our question-level cross-validation is not a controlled leave-category-out or cross-domain OOD evaluation, so we do not interpret it as direct evidence of robustness to every unseen semantic distribution. The residual-replacement analysis instead supports a representation-level view: semantic category matching alone does not ensure transfer, whereas a residual from a different category can help when its direction is well aligned with the target correction. Our conclusion is therefore conditional on geometric compatibility. If an OOD input remains close to useful training examples in representation space and its predicted steering direction is aligned with the required correction, S2S can remain effective regardless of whether its semantic category or data distribution matches the training set. Conversely, we do not claim robustness when the representation and correction geometry itself changes substantially.

### B.5 Geometric Interpretation of the S2S-R Search Space

The right panels of Fig. 5 illustrate the interpolation mechanism of S2S-R. If rotation-based steering directly uses the MLP-predicted direction as a fixed target, varying the rotation strength can only move the test representation along a single geodesic arc connecting its current direction to that target. S2S-R introduces an additional degree of freedom by varying the interpolation coefficient between the predicted residual and the reliable mean direction. This produces a continuum of calibrated target directions. Rotating the test representation toward each of these targets then sweeps out a two-dimensional reachable patch on the unit sphere rather than a single arc.

The original MLP trajectory and the mean-steering trajectory are retained as special cases within this patch, as formalized by Theorem 1 and proved in Appendix A.1. S2S-R therefore enlarges the set of candidate steering directions without discarding the original prediction, allowing it to explore directions with better alignment to the desired target.

### B.6 Interpreting the Adaptive Weights and Residual Analysis

**Adaptive-weight distributions.** Figure 3(d) visualizes the token-level weight triples $(w_z, w_v, w_r)$ defined in Eq. (10). Because the three weights are nonnegative and sum to one, each triple is represented as a point in a ternary simplex. The three vertices correspond to assigning all weight to the current hidden-state direction, the global mean direction, or the predicted residual direction. Points near the center use the three components similarly, points near an edge emphasize two components, and points near a vertex are dominated by one component. The black star marks the empirical mean weight vector.

The four plots show $\kappa \in \{1, 5, 10, 20\}$. At small $\kappa$, the softmax in Eq. (10) is relatively smooth, so the weights remain concentrated and similar across tokens. Increasing $\kappa$ sharpens the similarity scores and produces more differentiated, token-specific mixtures, which spread farther across the simplex. Figure 3(c) should be read together with this visualization: increasing $\kappa$ initially improves MC1 by enabling selective weighting, but the benefit eventually saturates once the assignments become sufficiently sharp.

**Residual replacement settings.** Figure 4(a) compares the cosine similarity between the residual used for steering and the ground-truth residual for the current sample under four settings:

- *Matched* uses the residual predicted from the current input, preserving the original input–residual pairing.
- *Category-shuffled* replaces it with a residual drawn from a different semantic category, breaking the matched pairing while retaining a category-controlled replacement.
- *Global-shuffled* draws a residual from the full sample pool without using category information.

- *Nearest-sample* selects, from residuals belonging to other categories, the one with the highest directional similarity to the target residual. This is a diagnostic oracle selection used to test the role of directional alignment, not a deployable test-time strategy.

The violin width indicates the density of samples at each cosine-similarity level, while the central marker summarizes the average alignment. The sharp drop under category and global shuffling shows that semantic grouping alone does not preserve the required residual direction. In contrast, the nearest-sample setting recovers substantial alignment despite using a semantically different sample.

Figure 4(b) reports the corresponding MC1 change relative to mean steering. Here, *mean* uses only the global direction, *matched* uses the original predicted residual, and *GT* uses the ground-truth residual as an oracle reference. Negative values indicate performance below mean steering. Category shuffling causes a large degradation, whereas the directionally selected nearest residual yields a clear gain. Together with Fig. 4(a), this result shows that residual transfer depends more on directional alignment than on semantic category.

Finally, Fig. 4(c) compares direct linear and spherical steering with S2S-L and S2S-R for vectors learned by LoFiT, LoReFT, and L2S. Each method group therefore tests the same calibration principle on a different learned intervention framework. The consistent gains of S2S-L and S2S-R show that the benefit is not specific to the MLP predictor used by L2S.

### B.7 Component Ablation and Generation Efficiency

The component-ablation, generation-efficiency, and noise-robustness experiments in this group are conducted on Llama-3.1-8B-Instruct.

**Component ablation.** Table 3 isolates the contributions of the MLP predictor, S2S calibration, adaptive weighting (AW), and intervention gate. Using the MLP prediction alone reduces MC1 from 41.42 to 37.01, illustrating that an uncalibrated learned vector can be less reliable than the baseline. Adding S2S raises MC1 to 43.14, showing that geometric calibration recovers useful sample-specific information. Adaptive weighting produces the largest gain, increasing MC1 to 66.91 by coordinating the hidden-state, mean, and residual directions. The intervention gate provides a further improvement to 67.40 and yields the best MC2 and MC3 scores, indicating that selectively suppressing unnecessary interventions improves stability.

**Attribution of the adaptive-weighting gain.** The increase from 43.14 to 66.91 should not be interpreted as evidence that S2S depends primarily on a generic adaptive rotation policy. Fixed interpolation coefficients can substantially degrade spherical steering, as also observed in prior work (You et al., 2026), motivating weights based on cosine similarity between the directions being combined. Our weighting rule follows this principle but applies it to the geometry introduced by S2S: the additional residual direction provides both a new calibration target and additional pairwise similarities for estimating input-dependent weights. The gain therefore reflects the joint effect of residual-direction calibration and its matched adaptive application, rather than AW independently replacing the steer-the-steering contribution.

**Generation efficiency.** Table 4 reports throughput on a single 3090 GPU. In the generation-only setting, where both the question and answer are known, S2S-L and S2S-R retain throughputs of 24.96 and 23.61 tokens/s, respectively, compared with 25.43 tokens/s without steering. This indicates modest overhead for direction calibration itself. In the open-ended setting, only the question is initially available, and throughput decreases to 12.49 tokens/s for S2S-L and 12.22 tokens/s for S2S-R because the sample-specific intervention must be inferred during generation. The similar costs of the two variants suggest that most of the additional computation comes from obtaining the adaptive steering signal rather than from choosing additive or rotational application.

### B.8 Robustness to Noisy Steering Targets

We further evaluate the noise robustness of S2S, as motivated by the theoretical analysis in Appendix A.2. We inject noise into the MLP training targets to perturb the learned steering directions. As shown in the middle panel of Fig. 5, large perturbations eventually degrade direct steering, but the response is not strictly monotonic: moderate noise can improve MC1 in some settings. We hypothesize that the MLP may produce an overly confident, excessively deterministic direction estimate,

and that mild target perturbations act as directional regularization that reduces overreliance on this point estimate. S2S consistently mitigates degradation at higher noise levels, with S2S-R remaining substantially more stable across the evaluated range. These results support the view that calibration toward the shared mean direction attenuates unreliable sample-specific components while retaining useful steering information. Understanding when controlled perturbations improve steering, and whether this effect can be converted into an explicit regularization objective, is an important direction for future work.