Improved Representation Steering for Language Models

Zhengxuan Wu* **Qinan Yu*** **Aryaman Arora**
Christopher D. Manning **Christopher Potts**
Stanford University
{wuzhengx,qinanyu,aryamana}@stanford.edu
{manning,cgpotts}@stanford.edu

Abstract

Steering methods for language models (LMs) seek to provide fine-grained and interpretable control over model generations by variously changing model inputs, weights, or representations to adjust behavior. Recent work has shown that adjusting weights or representations is often less effective than steering by prompting, for instance when wanting to introduce or suppress a particular concept. We demonstrate how to improve representation steering via our new **Reference-free Preference Steering** (**RePS**), a bidirectional preference-optimization objective that jointly does concept steering and suppression. We train three parameterizations of RePS and evaluate them on AxBench, a large-scale model steering benchmark. On Gemma models with sizes ranging from 2B to 27B, RePS outperforms all existing steering methods trained with a language modeling objective and substantially narrows the gap with prompting – while promoting interpretability and minimizing parameter count. In suppression, RePS matches the language-modeling objective on Gemma-2 and outperforms it on the larger Gemma-3 variants while remaining resilient to prompt-based jailbreaking attacks that defeat prompting. Overall, our results suggest that RePS provides an interpretable and robust alternative to prompting for both steering and suppression.

 github.com/stanfordnlp/axbench

1 Introduction

As language models (LMs) proliferate, they raise new challenges in reliability and user control. Prompting and fine-tuning are widely used to ensure LMs align with human goals; however, prompting is brittle and requires extensive manual trial and error [Chang et al., 2024], while fine-tuning brings high costs and produces artifacts that are hard to audit [Han et al., 2024]. Interpretability researchers have explored intervention-based methods (e.g., steering vectors and sparse autoencoders; SAEs) to overcome these limitations. Similarly to parameter-efficient fine-tuning methods (PEFTs), these lightweight and interpretable methods manipulate model forward passes in place at inference time to steer model behavior [Hu et al., 2022, Turner et al., 2023b].

However, intervention-based methods consistently underperform prompting and finetuning, as evidenced by AxBench, a large-scale model steering benchmark [Wu et al., 2025]. This shortfall likely stems from their training objectives neglecting the human preference signals that guide instruct-tuned LM optimization. Early attempts to use preference-based objectives for steering vectors have struggled to scale to large, production-scale models [Cao et al., 2024, Turner et al., 2025].

*Equal contribution.

Preprint. Under review.

In this work, we propose **Reference-free Preference Steering** (**RePS**), a bidirectional preference optimization objective built on SimPO [Meng et al., 2024] to train intervention-based steering methods. RePS up-weights the reward of steered behavior when interventions are applied *positively* and optimizes for the opposite behavior when interventions are applied *negatively* (see section 3). With RePS, we experiment with a few low-rank parameterizations of interventions (steering vectors, LoRA, and ReFT), and evaluate concept steering of the resulting models extensively on AXBENCH. We then evaluate the best performing RePS-trained interventions on concept suppression. To ensure RePS scales, we evalute with LMs from the `Gemma` family ranging from `2B` to `27B` LMs. Across four `Gemma` model sizes and three intervention types, RePS-trained models consistently outperform the standard language modeling objective and the prior preference-based BiPO baseline, narrowing the gap with prompting. When applied with negative steering factors, RePS performs on par with the language modeling objective for smaller LMs but shows superior performance for the larger `Gemma-3` models, again emphasizing the scalability of RePS. Moreover, RePS-trained models remain resilient to prompt-based jailbreaking attacks that bypass text-prompt defenses, whereas prompting-based strategies often fail, underscoring RePS as an interpretable and robust alternative to prompting.

2 Related work

Preference optimization objectives. Recent advances in aligning LMs with human preferences have led to the development of various preference optimization algorithms. PPO [Schulman et al., 2017] is widely used for policy optimization given a reward. DPO [Rafailov et al., 2023] moves from online learning to offline for efficiency; given a pair of responses, DPO directly optimized the model parameters to choose the winning response conditioned on a reference model. Another line of work explores even simpler objectives that do not rely on a reference model [Meng et al., 2024, Bansal et al., 2024]. Beyond aligning with human values, preference objectives are also used for steering LMs toward truthful responses [Cao et al., 2024].

PEFTs. One common approach to steering LMs for downstream behaviors is lightweight finetuning. Prefix tuning [Li and Liang, 2021] and prompt tuning [Lester et al., 2021] attach trainable parameters to the hidden layers and input tokens. Adapter-based methods [Houlsby et al., 2019, Wang et al., 2022, He et al., 2022, Fu et al., 2021] add fully connected layers on top of pretrained models. Methods like LoRA [Hu et al., 2022] and DoRA [Liu et al., 2024b] instead learn low-rank matrices that can be additively merged with the existing model weights; once merged, these methods bring no additional inference-time overhead. Subsequent work improved upon LoRA to offer more flexibility in rank [Zhang et al., 2024b, Valipour et al., 2023], position [Kong et al., 2024], layer and modules [Zhang et al., 2023], and editing [Zhang et al., 2023].

Representation steering. Besides PEFTs, models can also be steered through representation editing. Subramani et al. [2022], Turner et al. [2023b], Zou et al. [2023], Liu et al. [2024a], Vogel [2024], Li et al. [2024b], Marks and Tegmark [2024], Rimsky et al. [2024], and van der Weij et al. [2024] add rank-one steering vectors to models' activations to change their downstream behaviors for a specific task. Ravfogel et al. [2022], Belrose et al. [2023], Avitan et al. [2024], and Singh et al. [2024] perform edits on residual streams to apply concept erasure. Finetuning-based approaches [Wu et al., 2024] extend such editing using higher-rank matrices.

3 RePS

In this section, we introduce our steering task, dataset, and intervention notation. We discuss existing training objectives for intervention-based steering methods and present our new training objective.

3.1 Preliminaries

Steering task. Given an input instruction $\mathbf{x}$ to an instruct-tuned LM and a steering concept $\mathbf{c}$ (e.g., an abstract concept such as "*terms related to apple trees*" or a rule-based concept such as "*include a telephone number in your response*"), the goal is to generate a steered response $\hat{\mathbf{y}}_i^{\mathbf{c}}$ that follows the instruction while editing the response by incorporating the steering concept. This task is agnostic about how the steering is performed; in this paper, we explore a wide range of intervention-based techniques and prompting techniques.

Dataset. Following AXBENCH [Wu et al., 2025], given a steering concept $\mathbf{c}$, we create a small training dataset $\mathcal{D}_{\text{Train}} = \{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{y}_i^{\mathbf{c}})\}_{i=1}^{n}$ with n examples, where each example tuple i contains an instruction $\mathbf{x}_i$, a response $\mathbf{y}_i$, and a steered response $\mathbf{y}_i^{\mathbf{c}}$ that contains the steering concept.[1] For our training dataset, we do not model *negation* explicitly; rather, we focus only on *positive* steering, which steers the LM to incorporate the steering concept during training. At inference time, we also evaluate whether our interventions can be used to suppress the steering concept (section 5.3 and section 5.4).

Intervention definition. Given a Transformer-based LM [Vaswani et al., 2017], let $\mathbf{h}^l$ represent a sequence of d-dimensional representations at a model component (e.g., residual stream or attention output) of a given layer. Intervention-based steering methods define low-rank interventions Φ_{Steer} that edit representations in forward passes:

$$\mathbf{h}^l \leftarrow \Phi_{\text{Steer}}(\cdot\,; \alpha) \tag{1}$$

where Φ_{Steer} flexibly takes in any argument and manipulates the corresponding representation in-place with an optional steering factor α denoting the strength of the intervention. (The role of α is further clarified in the following definitions.)

3.2 Existing training objectives

The objective of LM steering is to train $\Phi_{\text{Steer}}(\cdot\,; \alpha)$ to fit the data distribution of $\mathcal{D}_{\text{Train}}$. In the following sections, we simplify our notation for interventions to Φ_{Steer}, unless otherwise noted.

Language modeling (Lang.). To train Φ_{Steer} for a steering concept $\mathbf{c}$, we can minimize the cross-entropy loss with teacher-forcing over all output positions with an intervened LM:

$$\min_{\Phi} \left\{ -\sum_{i=1}^{k} \log p_{\Phi}\left(y_i \mid \mathbf{x}\mathbf{y}_{<i}^{\mathbf{c}}, \mathbf{h}^l \leftarrow \Phi_{\text{Steer}}\right) \right\} \tag{2}$$

where k is the number of predicting response tokens. All steering methods evaluated by Wu et al. [2025] follow this objective. However, the steering LMs are usually instruct-tuned LMs which optimize for preference objectives. To ensure a fair comparison, we apply a factor sampling strategy to the language modeling objective as described in section 5.1.

Bi-directional preference optimization (BiPO; Cao et al. [2024]). Preference losses are alternatives to the standard language modeling loss. Recently, Cao et al. [2024] proposed a bi-directional preference optimization objective (BiPO) for training steering vectors. Given our training dataset $\mathcal{D}_{\text{Train}}$, the winning response is the steered response $\mathbf{y}^{\mathbf{c}}$, and the losing response is the original response $\mathbf{y}$ given an instruction $\mathbf{x}$. Unlike vanilla DPO [Rafailov et al., 2023], the loss is calculated in both *positive* and *negative* steering where the winning and losing responses flip in the latter case:

$$\Delta_{\Phi} = \log\left(\frac{p_{\Phi}\left(\mathbf{y}^{\mathbf{c}} \mid \mathbf{x}, \mathbf{h}^l \leftarrow \Phi_{\text{Steer}}\right)}{p(\mathbf{y}^{\mathbf{c}} \mid \mathbf{x})}\right) - \log\left(\frac{p_{\Phi}\left(\mathbf{y}^{l} \mid \mathbf{x}, \mathbf{h}^l \leftarrow \Phi_{\text{Steer}}\right)}{p(\mathbf{y} \mid \mathbf{x})}\right) \tag{3}$$

$$\min_{\Phi}\left\{-\mathbb{E}_{(\mathbf{x},\mathbf{y},\mathbf{y}^{\mathbf{c}})\sim\mathcal{D}_{\text{Train}}}\left[\log \sigma\left(\alpha\,\beta\,\Delta_{\Phi}\right)\right]\right\} \tag{4}$$

where p is the reference model (i.e., unintervened LM), $\alpha \sim \mathcal{U}(-1, +1)$ is the sampled directional coefficient, and β controls the deviation from the original model, which is set to 0.1. Note that Φ_{Steer} also depends on the steering coefficient as defined in eq. (1). The original implementation of BiPO uses a directional SV intervention $\Phi_{\text{SV}}(\mathbf{h}^l\,; d)$, which takes the same form as eq. (9). Intuitively, if $d = -1$, the sign of Δ_{Φ} flips, which swaps the winning and losing responses. BiPO implies a symmetric objective for positive and negative steering given the underlying intervention function Φ_{BiPO}. Since BiPO is conditioned on the reference model, the winning likelihood is incentivized to stay closer to the original likelihood from the reference model. As a result, we hypothesize BiPO fails at more drastic steering behaviors (e.g., Golden Gate Bridge Claude; Templeton et al. 2024). Recent empirical work also shows BiPO is less effective with production-sized LMs [Turner et al., 2025].

[1] By default, we use `gpt-4o-mini-2024-07-18` to generate the steered responses, unless otherwise noted.