Personalized Steering of Large Language Models: Versatile Steering Vectors Through Bi-directional Preference Optimization

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin,
Lu Lin, Fenglong Ma, Jinghui Chen
The Pennsylvania State University
`{ymc5533, tbz5156, bccao, zmy5171, lulin, fenglong, jzc5917}@psu.edu`

Abstract

Researchers have been studying approaches to steer the behavior of Large Language Models (LLMs) and build personalized LLMs tailored for various applications. While fine-tuning seems to be a direct solution, it requires substantial computational resources and may significantly affect the utility of the original LLM. Recent endeavors have introduced more lightweight strategies, focusing on extracting “steering vectors” to guide the model’s output toward desired behaviors by adjusting activations within specific layers of the LLM’s transformer architecture. However, such steering vectors are directly extracted from the activations of human preference data and thus often lead to suboptimal results and occasional failures, especially in alignment-related scenarios. In this work, we propose an innovative approach that could produce more effective steering vectors through bi-directional preference optimization. Our method is designed to allow steering vectors to directly influence the generation probability of contrastive human preference data pairs, thereby offering a more precise representation of the target behavior. By carefully adjusting the direction and magnitude of the steering vector, we enabled personalized control over the desired behavior across a spectrum of intensities. Extensive experimentation across various open-ended generation tasks, particularly focusing on steering AI personas, has validated the efficacy of our approach. Moreover, we comprehensively investigate critical alignment-concerning scenarios, such as managing truthfulness, mitigating hallucination, and addressing jailbreaking attacks alongside their respective defenses. Remarkably, our method can still demonstrate outstanding steering effectiveness across these scenarios. Furthermore, we showcase the transferability of our steering vectors across different models/LoRAs and highlight the synergistic benefits of applying multiple vectors simultaneously. These findings significantly broaden the practicality and versatility of our proposed method. Code is available at https://github.com/CaoYuanpu/BiPO

1 Introduction

In recent years, the generalization capabilities of Large Language Models (LLMs) [31, 20] have improved substantially, driven by the increase in parameter size and the expansion of training text corpus [22, 15]. Despite the strong generalization capabilities of LLMs across diverse fields [11, 30, 8, 19, 13, 34, 38], researchers are still actively investigating approaches to steer the behaviors of LLMs and develop personalized models tailored to meet the specific needs of different users [36, 32]. While fine-tuning techniques such as supervised fine-tuning and reinforcement learning from human feedback [21, 42] appear to be straightforward solutions, they demand significant computational resources and may substantially impact the utility of the original LLM.

Alternatively, a series of lightweight methods for steering the behavior of LLMs typically involves freezing model weights and introducing perturbations in the activation space to elicit desired changes in the generated text [29, 32, 26, 33]. In particular, recent efforts have introduced the extraction of “steering vectors” within specific layers of the LLM’s transformer architecture [26, 33]. During inference, these steering vectors are added back to the activations of the original prompt to guide the model’s output toward desired behaviors. Rimsky et al. [26] has demonstrated that a steering vector corresponding to a specific behavior possesses universal applicability across various user prompts. Furthermore, since a single steering vector only influences the activation of a specific layer without altering the model weights, it minimally disrupts the original model, thereby preserving the model’s general capabilities.

However, existing steering vectors are typically extracted by the activation difference when the LLM is prompted with two opposite preferences [26, 33], ignoring its actual generation that may diverge from the prompt, thus often resulting in suboptimal outcomes and occasional failures, particularly in critical alignment-related scenarios. Specifically, such methods first construct a set of contrastive prompt pairs, where two prompts in each pair are appended with opposite answers to the same question – one answer is associated with the target behavior, while the other corresponds to the opposite preference. The steering vector representing the target behavior is then derived as the average activation difference of the LLM on all pairs. However, we have observed that the vector extracted from prompt pairs has limited steering capability in the model’s generation – the model may generate texts that are not aligned with the prompted choice, even when the steering vector is applied to each generation step. This discrepancy indicates that steering vectors purely based on contrastive prompts may not accurately represent the target generation behavior of the model.

To fill in this gap, drawing inspiration from the recent preference optimization techniques such as Direct Preference Optimization (DPO) [25], we introduce an innovative approach to calculate more effective steering vectors via *Bi-directional Preference Optimization* (**BiPO**). Instead of relying on the model to “follow” a prompted direction, our method allows the model to “speak up”, enabling steering vectors to proactively modulate the generation probability of contrastive human preference data pairs, thus providing a more precise representation of the target behavior. By further manipulating the direction and magnitude of the optimized steering vectors, we can easily achieve varying degrees of steering effects for desired behaviors, thereby efficiently meeting users’ diverse personalization needs without necessitating additional model training.

We summarize our contributions as follows:

- We have analyzed current methods for extracting steering vectors in LLMs, identifying their potential limitations and failure cases. Based on these insights, we propose a bi-directional preference optimization (BiPO) to generate more effective steering vectors, enabling personalized control over the various behaviors.
- Our approach demonstrates exceptional efficacy through comprehensive experiments on various open-ended generation tasks, with a particular focus on steering AI personas. Moreover, we extensively examine crucial alignment-related scenarios such as managing truthfulness, mitigating hallucinations, and addressing jailbreaking attacks and their defenses. Our method consistently exhibits remarkable steering effectiveness in these scenarios.
- We further explore and confirm the notable transferability of the steering vectors produced by our method across different models and fine-tuned LoRAs [10, 6]. Our findings also demonstrate that vectors steering distinct behaviors can operate synergistically, thereby enabling a broader spectrum of steering applications.

2 Related Work

Activation Engineering Activation engineering typically involves freezing model weights and modifying activations to produce desired changes in the output text [29, 32, 26, 33, 17, 14, 43]. Several studies have focused on searching for “steering vectors” within the activation space of specific layers in the transformer architecture of LLMs. During inference, these vectors are incorporated back into the forward passes of user prompts to steer model generation [29, 32, 26, 33]. Particularly, Subramani et al. [29] successfully discovered sentence-specific vectors capable of generating target sentences with near-perfect BLEU scores. However, this method requires running gradient descent to derive a unique steering vector for each sample, which is highly impractical for larger language models. Activation addition, as proposed by Turner et al. [32], creates steering vectors by calculating the difference in intermediate activations between a pair of prompts at specific layers within a transformer model. These steering vectors are then applied to the first token position of subsequent forward passes to influence model completions. However, it fails to perform consistently across various behaviors and prompts. To reduce the noise in the steering vector, contrastive activation addition (CAA) [26] utilizes hundreds of preference data pairs to generate steering vectors. Each pair consists of two prompts with the same multiple-choice question but ending with different answer choices. The steering vector corresponding to the target behavior is isolated by averaging the difference between the activation on these preference pairs. Their experiments focusing on Llama-2 [31] have demonstrated CAA can steer AI personas and mitigate hallucinations to some extent. Wang and Shu [33] employed a similar technique and additionally tested freeform-format contrastive preference data pairs, verifying that steering vectors can be utilized to compromise the safety of LLMs. However, these steering vectors are derived directly from the activations of LLMs when using preference data pairs as input prompts, which often results in inaccurate representations of the target behavior, particularly in alignment-concerning scenarios such as addressing jailbreaking attacks and defenses [44, 24, 4, 3, 37]. Our method, through preference optimization, acquires more precise steering vectors and can effectively steer model behaviors even in scenarios highly relevant to alignment. In addition to extracting steering vectors in MLP activations of specific layers within the transformer's residual stream, some works opt to add perturbations on attention activations to steer model generation [14, 17]. Focusing on enhancing the truthfulness of LLMs, Li et al. [14] first locates a set of attention heads with high probing accuracy for the truthfulness and then shifts activations along these truth-correlated directions during inference. Liu et al. [17] improved the efficiency and controllability of In-Context Learning [2] by substituting traditional demonstration examples with shifts in attention activations across all layers of the transformer.

Preference Optimization Reinforcement learning from human feedback (RLHF) has emerged as a popular approach for learning from human preference [5, 42, 28, 21]. Typically, RLHF first trains a neural network-based reward function to harmonize with the preference dataset by incorporating a preference model such as the Bradley-Terry model [1]. Subsequently, reinforcement learning algorithms such as proximal policy optimization [27] are adopted to train a language model to maximize the given reward for the chosen response. Recent works such as DPO [25] and SLiC [39, 40] have shown the feasibility of circumventing the reward-modeling stage and directly solving the actual RL problem, thereby simplifying implementation and reducing resource needs. Specifically, DPO [25] directly fits a model to human preference data and implicitly optimizes the same objective as RLHF. SLiC [40] uses a contrastive ranking calibration loss to fit pairwise human feedback data while employing a cross-entropy regularization term to encourage the model to stay close to the SFT model. Furthermore, statistical rejection sampling optimization [18] unifies the losses of DPO and SLiC, proposing an improved estimation of the optimal policy and supporting sampling preference pairs from this policy. Our work proposes utilizing bi-directional preference optimization to extract steering vectors, thereby providing a more accurate representation of the target behavior.

3 Methodology

In this section, we first carefully study the current methods for extracting and utilizing steering vectors in LLMs, analyze their failure cases, and then propose our method.

3.1 Analyzing the Current Steering Vectors

Current approaches [26, 33] for extracting steering vectors begin by constructing contrastive prompt pairs: one demonstrating the target behavior and the other demonstrating the opposite behavior. Then such steering vectors can be extracted by computing the mean difference of activations at specific LLM layers on the contrastive prompt pairs. Ideally, such an activation difference represents the direction to steer the model's behavior toward the target one. Specifically, CAA [26] constructs the contrastive prompt pairs with multiple-choice questions with letter answers (such as "A" or "B") right after the question. Let us denote the multiple-choice question as p, the positive choice as c_p, and the negative choice as c_n. CAA can construct the contrastive prompt pairs: both prompts contain the same multiple-choice question p but are appended with different answers, where the "positive" prompt ends with c_p, which conforms to the target behavior, and the "negative" prompt concludes with c_n which denotes the opposite behavior. Formally, CAA calculates the steering vector at a particular layer L as:

$$v_L = \frac{1}{|\mathcal{D}|} \sum_{p,c_p,c_n \in \mathcal{D}} [A_L(p, c_p)]_k - [A_L(p, c_n)]_k, \qquad (1)$$