CAN CONTEXT ARBITRARILY REORIENT LINEAR REPRESENTATIONS IN LARGE LANGUAGE MODELS?

Anonymous authors
Paper under double-blind review

ABSTRACT

The so-called concept vectors in the activation space of language models are widely used to interpret and steer high-level behavioral traits represented in model outputs. These directions are typically extracted from a given pair of contrastive sample sets and treated as fixed for the model at hand thereafter. Complementary to the emerging literature on the effects of input context on linear representations, we systematically analyze *how strongly* context alone can change these directions while holding the model, layer, and the sample sets fixed. We find that optimizing a few shared input prefix tokens can reorient the extracted concept vector to a wide range of arbitrarily chosen directions, including the direction opposite to the original one. Further analysis reveals that reorientation of the extracted contrast vector is achieved by prefix-induced displacement in the underlying activation geometry. Preliminary observations suggest that a related geometric signature may also accompany successful adversarial prompts such as Greedy Coordinate Descent (GCG). Incorporating this geometric objective into GCG indeed yields modest gains in attack success on some models. Furthermore, we take a step toward reducing multi-concept interference under activation steering by optimizing the prefix to make the extracted contrastive vectors for different concepts orthogonal. Overall, our findings suggest that reliable interpretability should treat linear representations as context-conditioned objects rather than as specific only to the model.

1 INTRODUCTION

Linear structure has long provided a useful lens for understanding learned language representations. Early work by (Mikolov et al., 2013) showed that syntactic and semantic relations between words can be captured by approximately consistent vector offsets, suggesting that meaningful linguistic distinctions admit simple geometric descriptions. In modern large language models (LLMs), this perspective extends to high-level properties of contextual activations. Empirical studies identify linear directions associated with sentiment (Tigges et al., 2024; Teo et al., 2025), the truth of factual statements (Marks & Tegmark, 2024), and refusal (Arditi et al., 2024) among others (Li et al., 2023; Gurnee & Tegmark, 2024; Rimsky et al., 2024; Belrose et al., 2023; Lee et al., 2025). The linear representation hypothesis (LRH) formalizes the idea that concepts are represented as directions and connects this geometry to measurement and control (Park et al., 2023). These findings underpin activation steering, which modifies intermediate activations along selected directions to influence model behavior without updating model parameters (Turner et al., 2024; Rimsky et al., 2024). A common approach extracts a steering direction from the difference between the mean activations of two contrastive example sets and reuses this direction as a steering vector. We refer to examples exhibiting the desired behavior as the *target* set and those not exhibiting it as the *source* set. Throughout this paper, we use *contrastive vector*, *concept vector*, and *steering vector* interchangeably.

The usefulness of linear representations, however, leaves open how reliably their interpretations and effects persist across tasks and contexts. Bolukbasi et al. (2021) identify an interpretability illusion in BERT where a neuron or activation direction can appear to encode a coherent concept on one dataset while supporting a different interpretation on another. Tan et al. (2024) show that steering effectiveness varies substantially across inputs and can be brittle to changes in prompt format and distribution. More directly, Azizian et al. (2025) find that the correctness probes learned on different tasks have little directional similarity and transfer poorly. Contextual variation can also arise within a single interaction (Cheng et al., 2026; Adarsh et al., 2026). In particular, Lampinen et al. (2026) show that factuality representations can reverse their ordering as a conversation unfolds,

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.

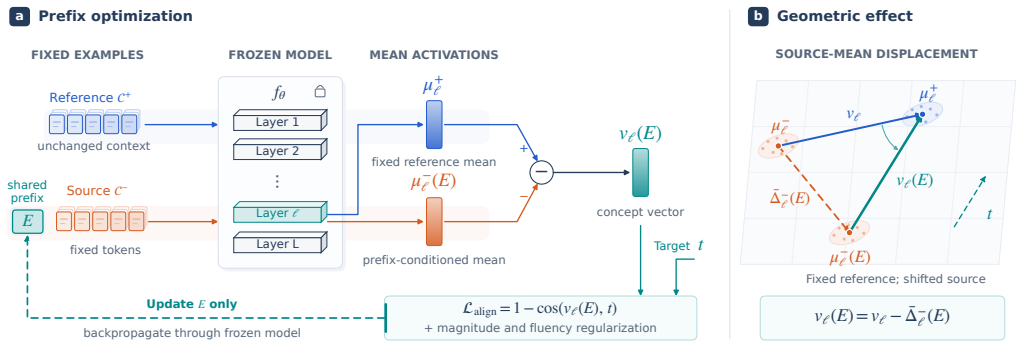


Figure 1: Controlled reorientation of concept vectors. (a) A shared prefix E is optimized on source prompts to align the extracted difference-in-means vector with a prescribed direction $\mathbf{t}$, while model parameters and reference prompts remain fixed. (b) Prefix-induced source-mean displacement changes the concept vector according to $\mathbf{v}_l(E) = \mathbf{v}_l - \bar{\Delta}_l^-(E)$.

while interventions along a fixed direction can have different effects at different conversational stages. Together, these findings motivate distinguishing the existence of useful linear structure from the stability of a particular extracted direction and its interpretation.

We investigate the extent of this contextual sensitivity through a controlled geometric question: *how strongly can changing the context reorient a concept vector extracted from otherwise fixed contrastive examples?* Prior work establishes substantial variation across datasets, tasks, and conversational states. Our focus is whether context can be deliberately selected to direct this variation toward a prescribed orientation. This distinction matters as a direction may vary across observed contexts, yet remain confined to a small region of activation space. Direct optimization of the the context for reorientation tests how much freedom the extraction procedure actually admits, beyond demonstrating that contextual variation exists as done by previous work (Adarsh et al., 2026; Lampinen et al., 2026).

To study this question, we fix the model, layer, original prompt sets, labels, and difference-in-means extraction rule. We introduce learnable prefix embeddings shared across the source prompts, while leaving the reference prompts unchanged. The prefix therefore serves as a continuous, optimizable proxy for context. Its parameterization follows soft prompting (Lester et al., 2021), but its objective directly targets representation geometry instead of task's end-to-end objective as illustrated in Figure 1. The experiment therefore isolates how changing the source context can move its mean activation relative to a fixed reference mean and thereby reorient the extracted contrast.

We find that a single optimized prefix can reorient the extracted vector toward a wide range of prescribed directions. We connect these changes to prefix-induced displacement of the underlying activations and complement the empirical analysis with a bound on prefix influence in pre-normalized transformers. We further examine whether related geometric displacement appears in successful adversarial prompts (Zou et al., 2023), linking the controlled experiment to contextual changes associated with model behavior. Beyond diagnosing contextual sensitivity, we extend our optimization objective to encourage orthogonality between extracted directions corresponding to different concepts and evaluate whether the resulting vectors reduce cross-concept interference under activation steering. This application treats context as a means of shaping representation geometry, while assessing the behavioral consequences separately. Overall, our contributions are threefold:

1. We introduce **Co**ntextual **R**epresentation **A**lignment (CoRA), a method that optimizes a shared input prefix to reorient an extracted contrastive vector toward a prescribed direction while keeping the model, layer, examples, labels, and extraction rule fixed (Section 3).
2. We characterize the feasibility and limits of context-induced reorientation theoretically and empirically, showing that CoRA can substantially reorient contrastive vectors across models, layers, and prefix configurations through displacement of the underlying activation geometry (Sections 4 and 6).
3. We demonstrate the behavioral relevance and broader utility of controllable reorientation by showing that CoRA reproduces activation-based jailbreaking effects using only input context, can improve GCG with an additional geometry-guided loss, and can reduce interference between two concept vectors during activation steering (Section 5).

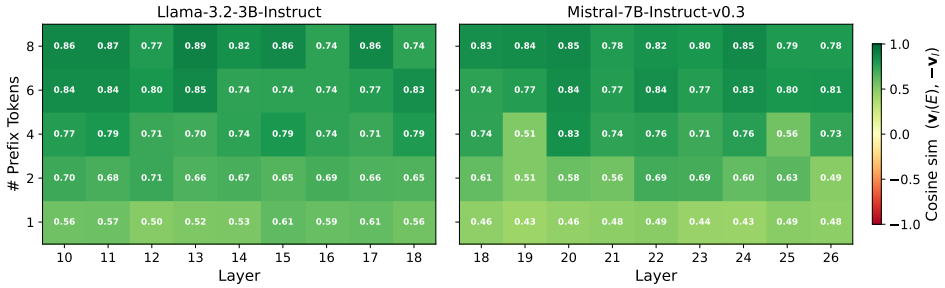


Figure 2: Cosine similarity heatmaps for prefix conditioned contrastive vector $\mathbf{v}_l(E)$ with our trained prefix E under the objective given by Eqn. 5 and the target vector, which is chosen to be the negative of original contrastive vector $\mathbf{t} = -\mathbf{v}_l$. The heatmaps show that the targeted reorientation indeed achieves substantial alignment with $\mathbf{t}$. We refer to Appendix D.2 for alignment tests toward arbitrarily chosen directions.

2 Background

Model and activation readout. Let f be a frozen autoregressive transformer with L layers and residual dimension d. For a tokenized input $\mathbf{x} = (x_1, \ldots, x_N)$, including any chat-template tokens, let $h_k^{(j)}(\mathbf{x}) \in \mathbb{R}^d$ denote the residual-stream activation at position j after block k, with $h_0^{(j)}$ is input embedding. We collect these vectors as rows of $H_k \in \mathbb{R}^{N\times d}$. Our default readout is the final input token: $\phi_l(\mathbf{x}) = h_l^{(N)}(\mathbf{x})$. This position can be a chat-template token following the user instruction; it precedes generation of the response. We use Euclidean norms and inner products in the model's fixed residual-stream coordinates.

Linear representations and concept-vector extraction. The linear representation hypothesis motivates describing high-level concepts through directions in activation space (Park et al., 2023; Marks & Tegmark, 2024). We study a specific operational definition: the difference between empirical class-conditional mean activations. Let $\mathcal{C}^+ = \{x_i^+\}_{i=1}^n$ and $\mathcal{C}^- = \{x_i^-\}_{i=1}^n$ be balanced contrastive source and target prompt sets. Following difference-in-means extraction used in activation steering (Rimsky et al., 2024; Arditi et al., 2024), define

$$\mathbf{v}_l = \boldsymbol{\mu}_l^+ - \boldsymbol{\mu}_l^-, \qquad \boldsymbol{\mu}_l^\pm = \frac{1}{n}\sum_{i=1}^{n} \phi_l(x_i^\pm). \tag{1}$$

For $\mathbf{v}_l \neq 0$, write $\hat{\mathbf{v}}_l = \mathbf{v}_l/\|\mathbf{v}_l\|_2$. The sign is fixed by the ordering of the classes, and $\|\mathbf{v}_l\|_2$ measures their mean separation. We call $\mathbf{v}_l$ a *concept vector* to specify this extraction rule.

Activation steering. One of the most commonly adopted steering method, Activation Addition (ActAdd), uses an extracted vector to intervene directly on the residual stream (Turner et al., 2024; Rimsky et al., 2024):

$$h_l^{(j)} \leftarrow h_l^{(j)} + \alpha \mathbf{v}_l, \qquad j \in \mathcal{I}, \tag{2}$$

where $\alpha \in \mathbb{R}$ is the intervention strength and $\mathcal{I}$ is the chosen set of token positions. Unit-normalized variants absorb $\|\mathbf{v}_l\|_2$ into the strength. Our prefix intervention instead changes the context used to collect source activations and hence the vector returned by Eqn. (1). Section 3 defines this intervention while keeping the reference class unchanged.

3 Contextual Representation Alignment with Soft-Prompting

We introduce Contextual Representation Alignment (CoRA) to test whether context can deliberately reorient an extracted concept vector toward a prescribed direction. Our intervention optimizes a single soft prefix shared across source examples while keeping the reference examples, model parameters, and extraction rule fixed. We first specify the prefix-conditioned readout and show that reorientation corresponds exactly to moving the source mean relative to the fixed reference mean. We then formulate a regularized alignment objective that controls mean separation and penalizes increases in source-prompt perplexity.

Controlled intervention. We fix the model, layer l, original prompt sequences, concept labels, and activation readout. Let $\mathcal{C}^-$ be the source class whose context is modified, and let $\mathcal{C}^+$ be the reference class whose context remains unchanged. Their unprefixed mean activations are $\boldsymbol{\mu}_l^-$ and $\boldsymbol{\mu}_l^+$, and their baseline difference is $\mathbf{v}_l = \boldsymbol{\mu}_l^+ - \boldsymbol{\mu}_l^-$. We optimize a continuous prefix shared by all source prompts. This is an input-embedding intervention in the soft-prompting setting (Lester et al., 2021).

Prefix and readout. Let $E = (\mathbf{e}_1, \ldots, \mathbf{e}_T) \in \mathbb{R}^{T \times d}$ contain T learnable embeddings. In the case where $T = 1$, we write $E = \mathbf{e} \in \mathbb{R}^d$. For a token sequence $\mathbf{x} = (x_1, \ldots, x_N)$, $[E; \mathbf{x}]$ denotes concatenation of the prefix with the embedded original tokens, with positional information and the causal mask applied by the model. Using one-based indexing, prefix positions are $1, \ldots, T$, and original token x_j occupies position $T + j$. The activation readout remains the final original input token:

$$\phi_l(\mathbf{x}; E) = h_l^{(T+N)}([E; \mathbf{x}]), \qquad \phi_l(\mathbf{x}) = h_l^{(N)}(\mathbf{x}). \tag{3}$$

Here N may vary across prompts. A zero-valued prefix $E = 0$ retains these additional positions and is distinct from the unprefixed input although the difference between them remains uniformly bounded independent of input as in Remark B.1.

Source-conditioned difference of means. Only $\mathcal{C}^-$ receives the prefix unless stated otherwise. For balanced sets of size n, define

$$\mathbf{v}_l(E) = \boldsymbol{\mu}_l^+ - \boldsymbol{\mu}_l^-(E), \qquad \boldsymbol{\mu}_l^-(E) = \frac{1}{n} \sum_{i=1}^{n} \phi_l(x_i^-; E). \tag{4}$$

The prefix is shared across source examples. It is neither input-dependent nor applied to the reference class. Accordingly, this experiment tests the sensitivity of the extracted contrast to the source context. Applying the same prefix to both classes defines a different experimental condition. Define the per-example displacement $\boldsymbol{\Delta}_l(\mathbf{x}; E) = \phi_l(\mathbf{x}; E) - \phi_l(\mathbf{x})$ and its source mean $\bar{\boldsymbol{\Delta}}_l^-(E) = n^{-1} \sum_i \boldsymbol{\Delta}_l(x_i^-; E)$. Then the change in the extracted vector is exactly $\mathbf{v}_l(E) = \mathbf{v}_l - \bar{\boldsymbol{\Delta}}_l^-(E)$. Thus, reorienting the difference of means is equivalent to displacing the source mean relative to a fixed reference mean. This identity does not require individual source examples to undergo the same displacement.

Targeted reorientation with regularization. For a prescribed unit vector $\mathbf{t} \in \mathbb{S}^{d-1}$, we optimize with respect to the prefix to minimize

$$\mathcal{L}_{\text{CoRA}}(E; \mathbf{t}) = \mathcal{L}_{\text{align}}(E; \mathbf{t}) + \lambda_{\text{norm}} \mathcal{L}_{\text{norm}}(E) + \lambda_{\text{PPL}} \mathcal{L}_{\text{PPL}}(E), \tag{5}$$

where

$$\mathcal{L}_{\text{align}} := 1 - \frac{\mathbf{t}^\top \mathbf{v}_l(E)}{\|\mathbf{v}_l(E)\|}$$

is the main loss to promote cosine alignment to target direction $\mathbf{t}$, and

$$\mathcal{L}_{\text{norm},r}(E) := \left(\frac{\|\mathbf{v}_l(E)\|}{\|\mathbf{v}_l\|} - r \right)^2, \quad \mathcal{L}_{\text{PPL}}(E) := \exp\left(-\frac{1}{N} \sum_{i=2}^{N} \log p_\theta\left(x_i \mid E, x_{<i}\right) \right) \tag{6}$$

are the regularizers to control the magnitude of the reoriented concept vector and to maintain fluency over the source prompts, respectively, where $r \geq 0$ is the desired ratio of mean separations and $\lambda_{\text{norm}}, \lambda_{\text{PPL}} \geq 0$. High-level pipeline of our optimization is depicted in Figure 1.

Definition 3.1 (Contextual Representation Alignment (CoRA)). *Fix a frozen model, a readout layer l, contrastive prompt sets $\mathcal{C}^+$ and $\mathcal{C}^-$, and a unit target $\mathbf{t} \in \mathbb{S}^{d-1}$. Assume the unprefixed contrast $\mathbf{v}_l$ is nonzero. CoRA optimizes a shared soft prefix $E \in \mathbb{R}^{T \times d}$ applied only to $\mathcal{C}^-$ by approximately solving $\min_{E:\, \mathbf{v}_l(E) \neq 0} \mathcal{L}_{\text{CoRA}}(E; \mathbf{t})$, where $\mathcal{L}_{\text{CoRA}}$ is given by Eqn. (5). The procedure returns the learned prefix $E^\star$ and its induced contrast $\mathbf{v}_l(E^\star)$.*

Note that CoRA modifies the context used to extract the contrast vector. Successful alignment therefore establishes the sensitivity of this extraction procedure to the source-side context which is our main research objective, whereas causality of the extracted contrastive vectors remains an empirical question that we study in Sections 5 and 6.

4 Theoretical Analysis

We theoretically analyze prefix-conditioned concept vectors through a global necessary condition for target alignment and a sufficient condition for local alignment. The analysis concerns the unregularized alignment objective with nonzero separation constraint. Proofs are deferred to Appendix B.

We consider sequential pre-normalized residual blocks, each consisting of a causal softmax-attention residual update followed by a continuous position-wise FFN residual update, with a separate token-wise normalizer before each sublayer. The full update equations are given in Appendix B.1.

In a pre-normalized transformer, a prefix influences later positions only through attention, after which FFN and residual sublayers may propagate its effect. For every layer k and sublayer $s \in \{\mathrm{attn}, \mathrm{ffn}\}$, the token-wise normalizer has uniformly bounded range:

$$\sup_{z \in \mathbb{R}^d} \|\mathcal{N}_{k,s}(z)\| \le L_{k,s}, \tag{7}$$

for some $L_{k,s} < \infty$. This holds for LayerNorm and RMSNorm with fixed parameters. Consequently, softmax attention is bounded because it forms convex combinations of projected normalized values, and an FFN is bounded on the normalizer's bounded range. Likewise, the following bound holds:

Lemma 4.1 (Bounded prefix influence). *Fix a layer l and a prefix length T. Let $E \in \mathbb{R}^{T \times d}$ be arbitrary, let $0_T \in \mathbb{R}^{T \times d}$ denote the zero prefix, and let $x = (x_1, \ldots, x_N)$ be any input for which the prefixed sequence lies within the model's context window. Under the pre-normalized architecture described above, there exists a constant $M_l < \infty$, depending only on the frozen model and l, such that*

$$\left\| h_l^{(T+N)}([E; x]) - h_l^{(T+N)}([0_T; x]) \right\| \le M_l. \tag{8}$$

Lemma 4.1 shows that even unconstrained prefix embeddings cannot induce arbitrarily large changes at an original-token readout: normalization bounds the information passed through residual branches.

As the prefix-conditioned concept vector is obtained by moving only the source mean, Lemma 4.1 also constrains its reachable set. Since a zero-valued prefix retains the additional prefix positions, we use it as the anchor for this result. Define $\boldsymbol{\mu}_l^-(0) = \frac{1}{n}\sum_{i=1}^n \phi_l(x_i^-; 0)$, $\mathbf{v}_l^{(0)} = \boldsymbol{\mu}_l^+ - \boldsymbol{\mu}_l^-(0)$, and $\hat{M}_l$ as the infimum of all possible values of M_l in Lemma 4.1. Then the following theorem holds.

Theorem 4.2 (Necessary condition for reachability). *Let $\mathbf{t} \in \mathbb{S}^{d-1}$ be a target direction, and suppose that alignment with $\mathbf{t}$ is required while retaining separation at least $\rho > 0$. If there exists a prefix E such that $\mathbf{v}_l(E) = a\mathbf{t}$ for some $a \ge \rho$, then with $r_l = \|\mathbf{v}_l^{(0)}\|$, $q_l = \mathbf{t}^\top \mathbf{v}_l^{(0)}$, and $a_* = \max\{\rho, q_l\}$, we have $\hat{M}_l^2 \ge r_l^2 + a_*^2 - 2a_* q_l$.*

Geometrically, every reachable concept vector lies within distance $\hat{M}_l$ of the zero-prefix anchor, whereas successful alignment requires intersecting the target ray $\{a\mathbf{t} : a \ge \rho\}$. Theorem 4.2 rules out targets whose rays lie outside this ball, regardless of the optimization procedure. Its condition is necessary but not sufficient because the ball is only an outer approximation of the reachable set. Note that zero-prefixed anchoring only shifts the center of reachable set by a prefix-independent offset and does not change the analysis (see Remark B.1).

The global bound above does not establish attainability. For a local sufficient condition, write $p = \mathrm{vec}(E) \in \mathbb{R}^{Td}$, abbreviate $\mathbf{v}(p) = \mathbf{v}_l(E)$, and $J(p) := D_p\mathbf{v}(p) = -\frac{1}{n}\sum_{i=1}^n D_p\phi_l(x_i^-; E)$.

Theorem 4.3 (Sufficient condition for local reorientation). *Fix a prefix p_0 with $\mathbf{v}_0 := \mathbf{v}(p_0) \ne 0$, a unit target $\mathbf{t} \in \mathbb{S}^{d-1}$, and a minimum separation $\rho > 0$. Let $Q_\mathbf{t} \in \mathbb{R}^{d \times (d-1)}$ have orthonormal columns spanning $\mathbf{t}^\perp$, and define $B_\mathbf{t} := Q_\mathbf{t}^\top J(p_0)$, $b_\mathbf{t} := \|Q_\mathbf{t}^\top \mathbf{v}_0\|$, and $q_\mathbf{t} := \mathbf{t}^\top \mathbf{v}_0$. Suppose that $\mathbf{v}$ is continuously differentiable on an open set containing $\overline{B}(p_0, R)$ for some $R > 0$ and that there are finite constants $L, K \ge 0$ such that for all $p, p' \in \overline{B}(p_0, R)$, we have $\|J(p)\|_{\mathrm{op}} \le L$ and $\|J(p) - J(p')\|_{\mathrm{op}} \le K\|p - p'\|$. Assume that $B_\mathbf{t}$ has full row rank and set $s_\mathbf{t} := \sigma_{\min}(B_\mathbf{t}) > 0$. If*

$$2b_\mathbf{t}/s_\mathbf{t} \le R, \qquad Kb_\mathbf{t}/s_\mathbf{t}^2 \le 1/2, \qquad q_\mathbf{t} - 2Lb_\mathbf{t}/s_\mathbf{t} \ge \rho, \tag{9}$$

then there exists a shared prefix p_* with $\|p_* - p_0\| \le 2b_\mathbf{t}/s_\mathbf{t}$ such that

$$\mathbf{v}(p_*) = a_*\mathbf{t}, \qquad a_* \ge \rho.$$

Theorem 4.3 guarantees exact alignment when the shared prefix can independently correct all components perpendicular to the target and the required correction stays within a sufficiently regular local neighborhood. Here $b_\mathbf{t}$ measures the perpendicular error, $s_\mathbf{t}$ measures the weakest local response available to correct it, and the three inequalities respectively control the correction radius, nonlinear variation, and remaining positive separation.

Table 1: Attack Success Rates (ASR), Perplexity of generations (PPL) and tinyBenchmark evaluation scores. Best ASRs are **bolded** and the second best are underlined. Our method is highlighted by blue.

Method	ASR ↑	PPL ↓	tArc	tHella	tMMLU	tTQA	tWino	tAvg ↑
Llama3.2-3B-Instruct								
No Intervention	2.88	1.28	56.16	76.36	63.10	50.26	59.85	61.15
Random Prefix	6.73	1.42	53.40	66.84	52.87	49.74	57.54	56.08
ActAdd	96.15	1.85	47.42	70.04	58.79	43.71	57.63	55.52
CoRA	**97.12**	1.72	58.31	72.48	55.02	46.88	58.73	58.28
Mistral-7B-Instruct-v0.3								
No Intervention	43.27	1.30	63.54	84.06	64.07	60.55	74.05	69.25
Random Prefix	40.38	1.34	63.65	83.79	61.38	56.03	72.19	67.41
ActAdd	87.50	2.10	59.29	80.25	62.71	48.85	75.26	65.27
CoRA	**94.23**	1.69	64.16	81.88	52.87	53.13	71.50	64.71
Gemma-2-9b-it								
No Intervention	0.00	1.18	69.31	81.34	74.23	54.67	71.86	70.28
Random Prefix	0.96	1.17	69.31	80.99	74.43	54.05	72.19	70.19
ActAdd	90.38	2.83	66.12	75.64	71.31	43.94	61.87	63.78
CoRA	**94.23**	2.15	66.32	70.74	60.46	51.20	65.49	62.84
Qwen3.5-9B								
No Intervention	0.96	1.29	73.11	78.91	59.13	52.99	73.65	67.56
Random Prefix	0.96	1.29	74.62	79.48	59.16	53.29	72.43	67.80
ActAdd	91.35	2.90	63.16	70.49	56.07	38.40	71.39	59.90
CoRA	**95.19**	1.75	76.21	74.84	67.43	51.34	75.37	69.04
Qwen3.6-27B								
No Intervention	0.00	1.24	75.98	63.37	81.38	51.11	70.80	68.53
Random Prefix	0.00	1.25	75.98	61.96	81.68	51.63	67.51	67.75
ActAdd	88.46	1.83	70.51	75.02	79.65	42.76	70.69	67.73
CoRA	**89.42**	1.59	75.24	68.58	80.85	55.58	73.38	70.73

5 Experiments

Figure 2 shows that CoRA reorients contrastive vectors toward prescribed directions across layers and prefix lengths. We next test whether this geometric control yields meaningful behavior: reproducing explicit activation interventions for jailbreaking using only a learned prefix, improving GCG through its alignment objective, and reducing cross-concept interference by reshaping the geometry of concept vectors.

5.1 Jailbreaking Large Language Models

5.1.1 Reorientation with a Learned Prefix

In this section, we test CoRA's capabilities on the task of jailbreaking, where the goal is to elicit harmful generations from otherwise safety-aligned models by bypassing their refusal mechanism. For this task, ActAdd (Rimsky et al., 2024; Turner et al., 2024) adds the difference-in-means $\mathbf{v}_l$ of last-token activations over a set of harmless and harmful prompts, which we denote as $\mathcal{C}^+$ and $\mathcal{C}^-$ respectively, to the activations of a new harmful prompt at a chosen layer l.

CoRA attempts to achieve a similar effect by only modifying the inputs. Specifically, we keep $\mathcal{C}^+$ fixed and train a prefix E on $\mathcal{C}^-$ through Eqn. (5) to primarily align the resulting difference-in-means $\mathbf{v}_l(E)$ with $\mathbf{t} = -\hat{\mathbf{v}}_l$. As $\boldsymbol{\mu}_l^+$ is fixed, this reorientation aims at displacing $\boldsymbol{\mu}_l^-(E)$, into a region similar to that reached by ActAdd, where the norm ratio r specifies how far it is to be displaced, analogous to the intervention strength.

Setup. We compute $\mathbf{v}_l$ using 200 harmful AdvBench prompts (Zou et al., 2023) and 200 harmless Alpaca prompts (Taori et al., 2023), and train the prefix E on the same harmful prompts. We evaluate on 104 held-out AdvBench prompts, using the HarmBench classifier (Mazeika et al., 2024) to assess harmful generations.

We compare CoRA with ActAdd on Llama3.2-3B-Instruct (Llama Team, 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Gemma-2-9b-it (Team et al., 2024), Qwen3.5-9B (Qwen Team, 2026a), and Qwen3.6-27B (Qwen Team, 2026b). Table 1 reports Attack Success Rate (ASR), gen-

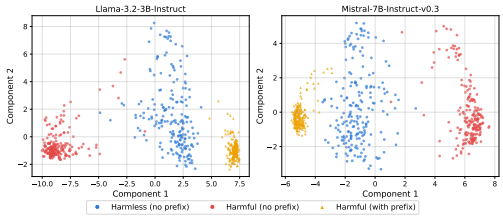


Figure 3: PCA of last-token representations for harmless, harmful, and CoRA-prefixed harmful prompts on Llama3.2-3B-Instruct and Mistral-7B-Instruct-v0.3. In the top two principal components, CoRA shifts harmful representations across the harmless cluster, consistent with reorientation toward $\mathbf{t} = -\hat{\mathbf{v}}_l$.

Table 2: Attack Success Rates (ASR), Perplexity of generations (PPL) and tinyBenchmark evaluation scores. Best ASRs are **bolded**. Our method is highlighted by blue.

Method	ASR ↑	PPL ↓	tArc	tHella	tMMLU	tTQA	tWino	tAvg ↑
Llama3.2-3B-Instruct								
No Intervention	2.88	1.28	56.16	76.36	63.10	50.26	59.85	61.15
GCG (Outputs)	10.58	2.18	48.20	78.59	58.76	50.26	59.44	59.05
GCG (Outputs+Activations)	**21.15**	1.70	48.24	78.59	58.52	50.26	59.44	59.01
Mistral-7B-Instruct-v0.3								
No Intervention	43.27	1.30	63.54	84.06	64.07	60.55	74.05	69.25
GCG (Outputs)	61.54	1.60	53.15	84.23	52.55	60.65	73.22	64.76
GCG (Outputs+Activations)	**64.42**	1.65	52.91	84.23	48.95	60.65	73.22	63.99

eration perplexity (PPL), and tinyBenchmarks scores (Maia Polo et al., 2024), with no intervention and a random 4-token prefix as references. Further details appear in Appendix C.1.

Results. Table 1 shows that CoRA achieves ASR comparable to, and slightly higher than, ActAdd across all five models, reproducing an explicit activation intervention through input modification alone. In contrast, a random prefix yields ASR close to the unintervened model, indicating that the effect arises from targeted optimization. Figure 3 further shows that CoRA displaces harmful representations so that $\mathbf{v}_l(E)$ aligns with $\mathbf{t}$. CoRA also yields lower PPL compared to ActAdd across all models, most notably on Qwen3.5-9B (2.90 to 1.75), indicating improved fluency. We note that the references largely refuse, which likely explains their lower PPL. tinyBenchmarks results are mixed across tasks, with notable tMMLU drops on Mistral-7B-Instruct-v0.3 and Gemma-2-9b-it. Overall, however, CoRA is comparable to ActAdd and achieves a higher average on three of five models; on Qwen3.5-9B and Qwen3.6-27B, it matches the unintervened average, whereas ActAdd falls below it on all five.

5.1.2 Case Study: Alignment Loss-Guided GCG

In this section, we investigate whether the alignment loss used to train CoRA can also benefit Greedy Coordinate Gradient (GCG), a related suffix-based jailbreaking method introduced in Zou et al. (2023). GCG optimizes a discrete adversarial suffix appended to the prompt to maximize the likelihood of a target sequence of tokens $x^*_{n+1:n+H}$ (e.g., the affirmative partial completion "Sure, here is <content of query>"), which amounts to minimizing the adversarial loss $\mathcal{L}_{\text{adv}}(x_{1:n}) = -\log p(x^*_{n+1:n+H} \mid x_{1:n})$. In this case study, we investigate whether our alignment loss on the model's activation space can improve this output-level jailbreak method.

Setup. We therefore add CoRA's training loss to the GCG objective. In GCG the adversarial suffix is appended after the prompt rather than prepended as in Eqn. (3), and in this study we adapt GCG's format accordingly. We augment the GCG objective with the alignment and norm terms of Eqn. (5), and train the suffix S to minimize

$$\mathcal{L}_{\text{GCG+CoRA}} = \mathcal{L}_{\text{adv}}(x_{1:n}) + \lambda_{\text{align}}\mathcal{L}_{\text{align}}(S; -\hat{\mathbf{v}}_l) + \lambda_{\text{norm}}\mathcal{L}_{\text{norm}}(S), \tag{10}$$

where $\mathbf{v}_l$ is extracted from the same 200 harmful and 200 harmless prompts as in Section 5.1.1. We test our approach on Llama3.2-3B-Instruct (Llama Team, 2024) and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and report ASR, PPL, and tinyBenchmark scores as in Section 5.1.1. The search configuration and further details are provided in Appendix C.2.

Results. We denote standard GCG as "GCG (Outputs)" and our variant as GCG "(Outputs+Activations)". Table 2 shows that adding our loss to GCG yields modest gains on both models under the setting of Zou et al. (2023), including a 10.57% ASR increase on Llama3.2-3B-Instruct, while maintaining comparable PPL and tinyBenchmark scores. As a failure case, Figure 4 shows that standard GCG fails to move harmful activations toward the harmless region on Llama3.2-3B-Instruct, which may explain its low ASR.

5.2 Interference Control in Multi-Concept Steering

When steering multiple concepts, a steering vector intended for one concept often also shifts another. We therefore apply CoRA to reduce this interference before steering with ActAdd (Turner et al., 2024), replacing $\mathcal{L}_{\text{align}}$ in Eqn. (5) with

$$\mathcal{L}_{\text{align}}\big(\mathbf{v}_l^A(E), \mathbf{v}_l^B(E)\big) = \big|\cos(\mathbf{v}_l^A(E), \mathbf{v}_l^B(E))\big|, \tag{11}$$

to encourage orthogonal concept vectors, alongside the norm and perplexity terms (Eqn. (5)) and a hinge penalty that constrains the unsteered model (full objective in Eqn. (44), Appendix C.3).

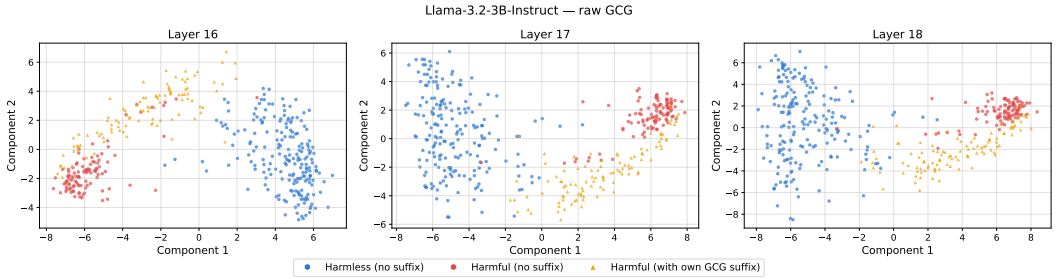


Figure 4: PCA visualizations of last-token representations of Llama3.2-3B-Instruct for harmless and harmful prompts, and harmful prompts with the trained GCG suffix. The suffixed harmful representations remain between the two clusters rather than reaching the harmless region.

Table 3: Cross-concept interference between concepts A: *coordinate-other-ais* and B: *survival-instinct*. $E_{c'\to c}$ denotes the effect on concept c when steering toward concept c'. Iso^{out}, Iso^{in}, and Iso denote outgoing, incoming, and overall isolation, respectively Best scores are **bolded**. Our method is highlighted in blue.

Method	$E_{A\to A}$	$E_{B\to B}$	$E_{A\to B}$	$E_{B\to A}$	$\text{Iso}_A^{\text{out}}\uparrow$	$\text{Iso}_B^{\text{out}}\uparrow$	$\text{Iso}_A^{\text{in}}\uparrow$	$\text{Iso}_B^{\text{in}}\uparrow$	$\text{Iso}_A\uparrow$	$\text{Iso}_B\uparrow$
Llama3.2-3B-Instruct										
ActAdd	44.50	9.37	-15.67	6.89	0.7396	0.5763	0.8659	0.3742	0.7978	0.4538
CoRA + ActAdd	37.87	6.39	-7.86	0.59	**0.8281**	**0.9155**	**0.9847**	**0.4484**	**0.8996**	**0.6020**
Mistral-7B-Instruct-v0.3										
ActAdd	38.50	11.95	6.07	4.19	0.8638	0.7404	0.9019	0.6632	0.8824	0.6996
CoRA + ActAdd	34.87	11.89	1.22	-0.51	**0.9662**	**0.9589**	**0.9856**	**0.9069**	**0.9758**	**0.9322**
Qwen3.6-27B										
ActAdd	48.18	14.79	-15.42	3.14	0.7575	0.8249	0.9388	0.4896	0.8385	0.6145
CoRA + ActAdd	56.74	16.78	-12.57	0.81	**0.8186**	**0.9540**	**0.9859**	**0.5717**	**0.8945**	**0.7150**

Setup. We use two multiple-choice datasets, *coordinate-other-ais* (concept A) and *survival-instinct* (concept B), and build contrastive pairs following Rimsky et al. (2024). We train E on 200 samples and compute $\mathbf{v}_l^c(E)$ on another set of 200 samples, with E prepended to both contrastive classes in the setting with CoRA. We add $\mathbf{v}_l^c(E)$ to all token positions after the prompt at layer l. We benchmark CoRA against plain ActAdd on Llama3.2-3B-Instruct (Llama Team, 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and Qwen3.6-27B (Qwen Team, 2026b). Further details are provided in Appendix C.3.

Metrics. Let $S_{c'}^c$ be the next-token probability of concept c's answer letter, averaged over test samples, when steering with the vector for concept c', with S_0^c denoting no steering. The effect of c' on c is $E_{c'\to c} = S_{c'}^c - S_0^c$, and ideal steering gives large $|E_{A\to A}|, |E_{B\to B}|$ with $E_{A\to B} = E_{B\to A} = 0$.

To quantify cross-concept interference, we define outgoing and incoming isolation. Outgoing isolation ($\text{Iso}_c^{\text{out}}$) measures whether the steering vector for concept c affects only its intended concept, while incoming isolation (Iso_c^{in}) measures whether concept c is unaffected by the other concept's steering vector. We summarize the two directions for each concept by their harmonic mean (Iso_c). For concepts A and B, we define

$$\text{Iso}_c^{\text{out}} = \frac{|E_{c\to c}|}{|E_{c\to c}| + |E_{c\to c'}|}, \quad \text{Iso}_c^{\text{in}} = \frac{|E_{c\to c}|}{|E_{c\to c}| + |E_{c'\to c}|}, \quad \text{Iso}_c = \frac{2\,\text{Iso}_c^{\text{out}}\,\text{Iso}_c^{\text{in}}}{\text{Iso}_c^{\text{out}} + \text{Iso}_c^{\text{in}}}, \quad c, c' \in \{A, B\}. \tag{12}$$

We use absolute effects since unintended shifts in either direction constitute interference. All scores lie in $[0, 1]$, with higher values indicating better isolation; all four equal 1 when $E_{A\to B} = E_{B\to A} = 0$.

Results. We refer to *coordinate-other-ais* as concept A and *survival-instinct* as concept B throughout this section. As shown in Table 3, CoRA + ActAdd consistently improves both directional and overall isolation scores across all three models, indicating reduced cross-concept interference. The largest improvement is observed for concept B on Mistral-7B-Instruct-v0.3, where Iso_B increases from 0.6996 to 0.9322. For concept A, the largest improvement occurs on Llama3.2-3B-Instruct, with Iso_A increasing from 0.7978 to 0.8996. Overall, CoRA reduces interference between the two concepts while retaining substantial intended steering effects.

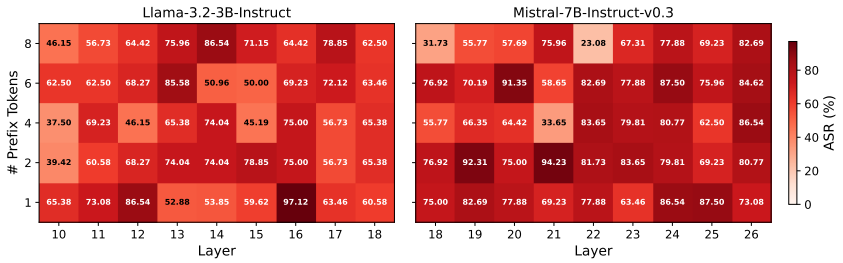


Figure 6: ASR heatmaps for prefix conditioned contrastive vector $\mathbf{v}_l(E)$ with our trained prefix E under the objective given by Eqn. 5 and the target vector, chosen to be the negative of original contrastive vector $\mathbf{t} = -\mathbf{v}_l$.

6 Additional Empirical and Ablation Studies

Effect of Layer Choice and Number of Prefix Tokens. We investigate how the choice of layer l and the number of prefix tokens affect CoRA on Llama3.2-3B-Instruct and Mistral-7B-Instruct-v0.3, where the prefixes are trained on the jailbreaking task described in Section 5.1.1, with all other hyperparameters fixed. From Figure 2, we observe that on both models, CoRA reorients $\mathbf{v}_l(E)$ towards $t = -\hat{\mathbf{v}}_l$, in which the alignment improves as more prefix tokens are added. Furthermore, in Figure 6, we can observe that most configurations are sufficient to jailbreak the model. However, achieving the highest possible attack success rate will still require careful tuning. These observations suggest that an alignment with $-\hat{\mathbf{v}}_l$ is sufficient to induce jailbreaks, but may not be the optimal target direction.

Non Uniform Interventions. CoRA's alignment loss component is only computed with respect to the last token's activation, which leaves the displacement induced by the prefix on all other token positions unconstrained. We investigate this by measuring the displacement across token positions induced by the trained prefix, as well as the randomly initialized prefix E_{random} on Llama3.2-3B-Instruct. Figure 5 demonstrates that the trained prefix induces a significant displacement across all token positions, but the source mean displacement only aligns with $\hat{\mathbf{v}}_l$ near the final few token positions, corresponding to the chat template. Meanwhile, the randomly initialized prefix induces a much smaller displacement with negligible alignment throughout. These observations suggest that, unlike ActAdd, which applies the same translation at every token position, CoRA's prefix induces non-uniform interventions, but only aligned with $\hat{\mathbf{v}}_l$ near the last token.

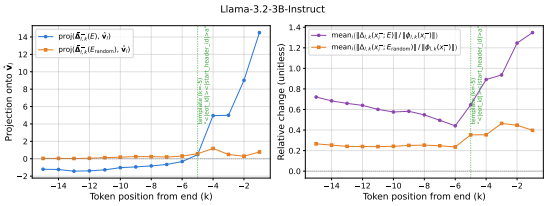


Figure 5: **Left**: Projection of source-mean displacement computed at token position k onto the original normalized steering vector $\hat{\mathbf{v}}_l$, for the trained prefix E and the random prefix E_{random} used in Section 5.1.1. **Right**: Mean ratio of per-example displacement magnitude to the unprefixed activation magnitude, averaged over the harmful prompts. We observe that the relative average displacement at each token position is significant, but the alignment with $\hat{\mathbf{v}}_l$ remains low except near the last few tokens, which correspond to the chat template. Meanwhile, the random prefix induces a much smaller displacement, and negligible alignment at all positions.

7 Conclusion

We introduced Contextual Representation Alignment (CoRA), a controlled test of how strongly source context can reorient contrastive representation vectors. CoRA optimizes a shared soft prefix while fixing the model, examples, labels, and extraction rule. We derive necessary reachability conditions and sufficient local conditions for nondegenerate target alignment. Across models and configurations, CoRA substantially reorients contrastive vectors, reproduces activation-based jailbreaking through input modification, improves geometry-guided GCG, and reduces cross-concept interference in activation steering. These results show that useful linear structure can coexist with substantial contextual variability, motivating concept vectors as context-dependent summaries whose geometry and behavioral effects require separate evaluation. Main limitation of our work is that although CoRA gives a useful controlled diagnostic, soft prefixes need not correspond to natural language. Constraining the context optimization to natural language at the cost of additional generative modeling is therefore a promising direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools for providing feedback on our experiments and implementing methods. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology, and interpreting results, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for modifying scientific figures, suggesting experimental parameters, editing software code, editing our research paper to improve readability, and identifying relevant literature. We have reviewed all AI-assisted work. In particular, we discussed and evaluated within our group the AI-generated suggestions for experimental design. We also carefully assessed AI suggestions for method implementation and reviewed the final code ourselves. In addition, we reviewed all AI-generated or AI-edited figures and manuscript revisions, and carefully read and verified the literature suggested by AI. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Given the nature of the work, we do not foresee any negative societal and ethical impacts of our work. However, risks remain: By eliminating the need to collect desired examples, CoRA may inadvertently lower the barrier to generating nuanced harmful content, such as persuasive misinformation or biased narratives. Although it does not expand the baseline risk profile of LLMs, robust safeguards, transparency, accountability, and ongoing ethical review are required for responsible use.

REPRODUCIBILITY STATEMENT

Source codes for our experiments are provided in the supplementary materials of the paper. The details of our experimental settings are given in Section 5 and the Appendix. All datasets that we used in the paper are published, and they are easy to access in the Internet.

REFERENCES

Shivam Adarsh, Maria Maistro, and Christina Lioma. How context shapes truth: Geometric transformations of statement-level truth representations in LLMs. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36570–36590, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1695. URL https://aclanthology.org/2026.acl-long.1695/.

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. *arXiv preprint arXiv:2406.11717*, 2024.

Waïss Azizian, Michael Kirchhof, Eugene Ndiaye, Louis Béthune, Michal Klein, Pierre Ablin, and marco cuturi. The geometries of truth are orthogonal across tasks. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025. URL https://openreview.net/forum?id=FdfvGu5rM5.

Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. Obfuscated activations bypass LLM latent-space defenses. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ktGmDGoWnB.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=awIpKpwTwF.

Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. An interpretability illusion for bert. *arXiv preprint arXiv:2104.07143*, 2021.

Yun Cheng, Xingyu Zhu, Haoyu Zhao, and Sanjeev Arora. Contextual drag: How errors in the context affect llm reasoning. *arXiv preprint arXiv:2602.04288*, 2026.

O. P. Ferreira and B. F. Svaiter. Kantorovich's theorem on newton's method. *arXiv preprint arXiv:1209.5704*, 2012.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

L. V. Kantorovich. On newton's method for functional equations. *Doklady Akademii Nauk SSSR*, 59 (7):1237–1240, 1948. In Russian.

L. V. Kantorovich and G. P. Akilov. *Functional Analysis*. Pergamon Press, Oxford, 2 edition, 1982.

Daniel Khashabi, Xinxi Lyu, Sewon Min, Lianhui Qin, Kyle Richardson, Sean Welleck, Hannaneh Hajishirzi, Tushar Khot, Ashish Sabharwal, Sameer Singh, and Yejin Choi. Prompt waywardness: The curious case of discretized interpretation of continuous prompts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3631–3643. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.266. URL https://aclanthology.org/2022.naacl-main.266/.

Arya Labroo, Ivaxi Sheth, Vyas Raina, Amaani Ahmed, and Mario Fritz. Funny or persuasive, but not both: Evaluating fine-grained multi-concept control in LLMs, 2026. URL https://arxiv.org/abs/2601.18483. Accepted at EACL 2026.

Andrew Kyle Lampinen, Yuxuan Li, Eghbal Hosseini, Sangnie Bhardwaj, and Murray Shanahan. Linear representations in language models can change dramatically over a conversation. *arXiv preprint arXiv:2601.20834*, 2026.

Andrew Lee, Melanie Weber, Fernanda Viégas, and Martin Wattenberg. Shared global and local geometry of language model embeddings. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=aJDykpJAYF.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243/.

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 (NeurIPS)*, 2023.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

AI @ Meta Llama Team. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

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

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1090/.

Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Multi-attribute steering of language models via targeted intervention, 2025. URL https://arxiv.org/abs/2502.12446. Presented at ACL 2025.

Narmeen Fatimah Oozeer, Luke Marks, Fazl Barez, and Amir Abdullah. Beyond linear steering: Unified multi-attribute control for language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23513–23557. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1278. URL https://aclanthology.org/2025.findings-emnlp.1278/.

James M. Ortega. The newton–kantorovich theorem. *The American Mathematical Monthly*, 75(6): 658–660, 1968. doi: 10.2307/2313800.

James M. Ortega and Werner C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York, 1970. Reprinted by SIAM, 2000.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023. URL https://openreview.net/forum?id=T0PoOJg8cK.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-27b.

Dana Ramati, Daniela Gottesman, and Mor Geva. Eliciting textual descriptions from representations of continuous prompts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16545–16562. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.849. URL https://aclanthology.org/2025.findings-acl.849/.

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

Daniel Scalena, Gabriele Sarti, and Malvina Nissim. Multi-property steering of large language models with dynamic activation composition. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 577–603. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.34. URL https://aclanthology.org/2024.blackboxnlp-1.34/.

Daniel Chee Hian Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=v8X70gTodR.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Rachel S.Y. Teo, Laziz Abdullaev, and Tan Minh Nguyen. The blessing and curse of dimensionality in safety alignment. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=TiTk6VDz2H.

Curt Tigges, Oskar J. Hollinsworth, Atticus Geiger, and Neel Nanda. Language models linearly represent sentiment. In Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen (eds.), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 58–87, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.5. URL https://aclanthology.org/2024.blackboxnlp-1.5/.

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:2308.10248v5*, 2024.

Teun van der Weij, Massimo Poesio, and Nandi Schoots. Extending activation steering to broad skills and multiple behaviours, 2024. URL https://arxiv.org/abs/2403.05767.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou', and Daniel Cer. SPoT: Better frozen model adaptation through soft prompt transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5039–5059. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.346. URL https://aclanthology.org/2022.acl-long.346/.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66945–66970. PMLR, 2025. URL https://proceedings.mlr.press/v267/wollschlager25a.html.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

Supplement to "Can Context Arbitrarily Reorient Linear Representations In Language Models?"

Table of Contents

A Extended Related Work 14

B Technical Derivations and Proofs 15

B.1 Proof of Lemma 4.1 15

B.2 Proof of Theorem 4.2 16

B.3 Proof of Theorem 4.3 17

C Additional Experimental Details 19

C.1 Experimental Setup for Reorientation with a Learned Prefix (Section 5.1.1) 19

C.2 Experimental Setup for Alignment Loss-Guided GCG (Section 5.1.2) 19

C.3 Experimental Setup for Interference Control in Multi-Concept Steering (Section 5.2) 20

D Supplementary Empirical Analysis 20

D.1 Effect of Random Prefix on Language 20

D.2 Alignment to a Random Target Vector 20

D.3 Layerwise Cosine Similarity Analysis 21

D.4 Source-mean Displacement as a Steering Vector 22

D.5 Additional Analysis on Non Uniform Induced Perturbation 22

E Limitations 23

A Extended Related Work

Reliability and context dependence of linear representations. Linear probing and representation steering commonly assume that a concept corresponds to a stable direction in activation space. However, the semantic interpretation assigned to such a direction need not remain stable across datasets, prompt formats, tasks, or conversational states. Apparent linear structure can be misleading under distribution shift (Bolukbasi et al., 2021), and the reliability of steering directions varies substantially with their construction and evaluation conditions (Tan et al., 2024). Recent work further finds that truth representations can be task-specific (Azizian et al., 2025), that context changes the directions extracted for the same concept (Adarsh et al., 2026), and that conversational interaction reorganizes concept representations (Lampinen et al., 2026). At the behavioral level, contextual information can also reduce the effectiveness of fixed activation interventions (Cheng et al., 2026). These studies primarily measure variation induced by naturally occurring contexts. In contrast, we optimize a shared soft prefix while holding the model, layer, labeled examples, readout rule, and the steering direction extraction method fixed. This provides a controlled way to test how much of a direction's geometry can be changed through context alone, and therefore clarifies which conclusions drawn from a single extraction context are reliable.

Continuous prompt optimization and interpretation. Continuous prompts provide a parameter-efficient mechanism for conditioning frozen language models. Prefix tuning learns continuous vectors that are inserted into transformer computations (Li & Liang, 2021), while prompt tuning learns soft input embeddings and becomes increasingly competitive as model scale grows (Lester et al., 2021). Subsequent work treats learned prompts as transferable task representations: SPoT studies the transfer of soft prompts across tasks and uses prompt similarity to identify useful source tasks (Vu et al., 2022). Closer to our setting, Bailey et al. (2026) optimize a universal continuous suffix to reshape activations while preserving target behavior, including by directly minimizing activation cosine similarity. Wollschläger et al. (2025) similarly augment GCG's output objective with a loss that suppresses an internal refusal direction. These works establish precedents for manipulating representations through optimized inputs and for combining behavioral and representation-level objectives. CoRA instead optimizes a shared prefix to align a class-mean contrast with a prescribed

target direction. Its aim is not monitor evasion or refusal suppression, but controlled analysis and manipulation of how strongly context changes an extracted concept direction. At the same time, the interpretation of continuous prompts remains subtle. Khashabi et al. (2022) show that a soft prompt can admit arbitrary or misleading textual projections, cautioning against treating its nearest discrete tokens as a faithful explanation. InSPEcT instead elicits natural-language descriptions from model behaviors induced by continuous prompts (Ramati et al., 2025). Accordingly, we interpret the learned prefix as a controlled contextual variable for representation extraction, rather than as a natural-language description of the target concept.

Multi-concept steering. Steering multiple properties simultaneously is more difficult than applying a single concept direction because independently extracted interventions can interfere. Early experiments found that simply adding independently learned steering vectors was often unsuccessful, while injecting them at different locations showed more promise (van der Weij et al., 2024). Dynamic Activation Composition learns to combine multiple property vectors during generation (Scalena et al., 2024), and MAT-Steer uses targeted token-level interventions together with sparsity and orthogonality objectives to reduce interference among attributes (Nguyen et al., 2025). K-Steering goes beyond linear vector addition by using the gradients of a nonlinear multi-label classifier to support unified multi-attribute control (Oozeer et al., 2025). Complementing these methods, recent evaluations demonstrate a persistent gap between controlling individual concepts and controlling their conjunction, for example, eliciting text that is both funny and persuasive rather than only one of the two (Labroo et al., 2026). Our multi-concept experiment is not intended to establish a state-of-the-art multi-attribute steering method, and orthogonality itself is not a new objective in this literature. Instead, it isolates a complementary opportunity: rather than only composing, projecting, or gating fixed vectors after they have been extracted, one can optimize the context in which the vectors are extracted so that their geometry is more favorable for subsequent multi-concept steering.

B Technical Derivations and Proofs

B.1 Proof of Lemma 4.1

Proof. For each layer k, write the two pre-normalized residual updates as

$$\widehat{H}_k = H_{k-1} + \mathcal{A}_k(\mathcal{N}_{k,\mathrm{attn}}(H_{k-1}))\,, \tag{13}$$

$$H_k = \widehat{H}_k + \mathcal{F}_k\Big(\mathcal{N}_{k,\mathrm{ffn}}(\widehat{H}_k)\Big)\,, \tag{14}$$

where $\mathcal{A}_k$ is the finite-head softmax-attention map and $\mathcal{F}_k$ is the position-wise FFN.

We first establish uniform bounds on the two residual-sublayer outputs. By Eqn. (7), every value supplied to an attention head is obtained by applying a fixed affine value projection to a vector with bounded norm. It is therefore uniformly bounded. Each attention head produces a convex combination of these value vectors because its softmax weights are nonnegative and sum to one. Since the number of heads, the output projection, and all biases are fixed, there exists $B_k < \infty$ such that

$$\Big\|\mathcal{A}_k(\mathcal{N}_{k,\mathrm{attn}}(H))^{(j)}\Big\| \le B_k \tag{15}$$

for every admissible sequence H and every position j. This argument also covers multi-head and grouped-query attention.

The closure of the range of $\mathcal{N}_{k,\mathrm{ffn}}$ is bounded and hence compact in finite dimensions. Because $\mathcal{F}_k$ is continuous, there similarly exists $C_k < \infty$ such that

$$\|\mathcal{F}_k(\mathcal{N}_{k,\mathrm{ffn}}(z))\| \le C_k \tag{16}$$

for every $z \in \mathbb{R}^d$. This step applies to both standard and gated FFNs and does not require a particular internal parameterization.

Let $j_\star = T + N$ be the position of the final original token. Use superscripts E and 0 for the executions on $[E; x]$ and $[0_T; x]$, respectively, and define

$$\delta_k := \Big\|h_k^{E,(j_\star)} - h_k^{0,(j_\star)}\Big\|\,. \tag{17}$$

The two sequences have the same length, and the final original token occupies the same position in both. Its token and positional inputs are therefore identical, so

$$\delta_0 = 0. \tag{18}$$

Let $a_k^{E,(j_\star)}$ and $a_k^{0,(j_\star)}$ denote the corresponding attention-sublayer outputs. From the residual update and Eqn. (15),

$$\begin{aligned}\left\|\widehat{h}_k^{E,(j_\star)} - \widehat{h}_k^{0,(j_\star)}\right\| &= \left\|h_{k-1}^{E,(j_\star)} - h_{k-1}^{0,(j_\star)} + a_k^{E,(j_\star)} - a_k^{0,(j_\star)}\right\| \\ &\le \delta_{k-1} + \left\|a_k^{E,(j_\star)}\right\| + \left\|a_k^{0,(j_\star)}\right\| \\ &\le \delta_{k-1} + 2B_k. \end{aligned} \tag{19}$$

Applying the same argument to the FFN residual update and using Eqn. (16) gives

$$\begin{aligned}\delta_k &\le \left\|\widehat{h}_k^{E,(j_\star)} - \widehat{h}_k^{0,(j_\star)}\right\| + 2C_k \\ &\le \delta_{k-1} + 2B_k + 2C_k. \end{aligned} \tag{20}$$

Unrolling this recurrence from $k = 1$ to $k = l$ and using $\delta_0 = 0$ yields

$$\left\|h_l^{(T+N)}([E; x]) - h_l^{(T+N)}([0_T; x])\right\| \le 2\sum_{k=1}^{l}(B_k + C_k) =: M_l < \infty. \tag{21}$$

The constants B_k and C_k depend only on the frozen sublayer maps and normalizer parameters, so M_l is independent of E. □

Remark B.1. *The zero prefix is used only as a convenient same-length anchor. Let*

$$\mathbf{v}_l^{(0,T)} := \boldsymbol{\mu}_l^+ - \frac{1}{n}\sum_{i=1}^{n}\phi_l(x_i^-; 0_T). \tag{22}$$

Lemma 4.1 and the triangle inequality imply $\left\|\mathbf{v}_l(E) - \mathbf{v}_l^{(0,T)}\right\| \le M_l$. *The zero-prefix and unprefixed baselines differ by the fixed offset* $d_{l,T} := \left\|\mathbf{v}_l^{(0,T)} - \mathbf{v}_l\right\|$, *which is independent of* E. *Hence the same result can be expressed relative to the unprefixed baseline as*

$$\|\mathbf{v}_l(E) - \mathbf{v}_l\| \le \left\|\mathbf{v}_l(E) - \mathbf{v}_l^{(0,T)}\right\| + \left\|\mathbf{v}_l^{(0,T)} - \mathbf{v}_l)\right\| \le M_l + d_{l,T}. \tag{23}$$

Thus, anchoring the proof at the zero prefix changes the center of the reachable-set bound only by a fixed, prefix-independent offset.

B.2 Proof of Theorem 4.2

Proof. Due to Lemma 4.1, every prefix-conditioned concept vector lies in the closed ball $\mathbf{v}_l(E) \in B\left(\mathbf{v}_l^{(0)}, \hat{M}_l\right)$. Equivalently, $\left\|\mathbf{v}_l(E) - \mathbf{v}_l^{(0)}\right\| \le \hat{M}_l$ for all $E \in \mathbb{R}^{T\times d}$. We next derive the target-dependent condition. If $\mathbf{v}_l(E) = a\mathbf{t}$ for some $a \ge \rho$, then

$$\left\|\mathbf{v}_l^{(0)} - a\mathbf{t}\right\| = \left\|\mathbf{v}_l^{(0)} - \mathbf{v}_l(E)\right\| \le \hat{M}_l. \tag{24}$$

Consequently,

$$\hat{M}_l \ge \min_{a\ge\rho}\left\|\mathbf{v}_l^{(0)} - a\mathbf{t}\right\|. \tag{25}$$

The squared distance to a point $a\mathbf{t}$ on the target ray is

$$\begin{aligned}\left\|\mathbf{v}_l^{(0)} - a\mathbf{t}\right\|^2 &= \|\mathbf{v}_l^{(0)}\|^2 + a^2\|\mathbf{t}\|^2 - 2a\mathbf{t}^\top\mathbf{v}_l^{(0)} \\ &= r_l^2 + a^2 - 2aq_l. \end{aligned} \tag{26}$$

This is a convex quadratic in a, with its unconstrained minimum at $a = q_l$. Under the constraint $a \ge \rho$, the minimizer is therefore $a_* = \max\{\rho, q_l\}$. Substituting a_* into Eqn. (26), we obtain

$$\hat{M}_l \ge \sqrt{r_l^2 + a_*^2 - 2a_*q_l}.$$

This completes the proof. □

Corollary B.2. *If $\hat{M}_l < r_l$, then every nonzero reachable directions lie in a cone around the zero-prefix direction satisfying $\angle\Big(\mathbf{v}_l(E), \mathbf{v}_l^{(0)}\Big) \le \arcsin\Big(\hat{M}_l/r_l\Big)$.*

Corollary B.2 converts the displacement bound into an angular one: when $\hat{M}_l < r_l$, the ratio $\hat{M}_l/r_l$ controls the certified rotation range. When $\hat{M}_l \ge r_l$, this bound imposes no angular restriction.

Proof. Let $\theta = \angle\Big(\mathbf{v}_l(E), \mathbf{v}_l^{(0)}\Big)$. Assume $\theta \ge \pi/2$, then $\mathbf{v}_l(E)^\top \mathbf{v}_l^{(0)} \le 0$, and hence

$$\Big\|\mathbf{v}_l(E) - \mathbf{v}_l^{(0)}\Big\|^2 = \|\mathbf{v}_l(E)\|^2 + r_l^2 - 2\mathbf{v}_l(E)^\top \mathbf{v}_l^{(0)} \ge r_l^2, \tag{27}$$

which contradicts $\hat{M}_l < r_l$. Therefore $\theta < \pi/2$. Furthermore, we have

$$\begin{aligned}
\hat{M}_l^2 &\ge \Big\|\mathbf{v}_l(E) - \mathbf{v}_l^{(0)}\Big\|^2 \\
&= \|\mathbf{v}_l(E)\|^2 + r_l^2 - 2\|\mathbf{v}_l(E)\| r_l \cos\theta \\
&= \big(\|\mathbf{v}_l(E)\| - r_l\cos\theta\big)^2 + r_l^2 \sin^2\theta \\
&\ge r_l^2 \sin^2\theta.
\end{aligned}$$

Hence, we obtain $\theta \le \arcsin(\hat{M}_l/r_l)$. □

B.3 Proof of Theorem 4.3

We use the following finite-dimensional, affine-invariant form of the Newton-Kantorovich theorem; see the original work of Kantorovich (1948) and the standard treatments in Kantorovich & Akilov (1982), Ortega (1968), Ortega & Rheinboldt (1970), and Ferreira & Svaiter (2012).

Theorem B.3 (Newton-Kantorovich). *Let $\Omega \subseteq \mathbb{R}^N$ be open and let $H : \Omega \to \mathbb{R}^N$ be continuously differentiable. Suppose $\overline{B}(x_0, R) \subset \Omega$ and $A := DH(x_0)$ is invertible. Assume that, for some $\gamma \ge 0$,*

$$\big\|A^{-1}\big(DH(x) - DH(y)\big)\big\|_{\mathrm{op}} \le \gamma\|x - y\|, \qquad x, y \in \overline{B}(x_0, R). \tag{28}$$

Define

$$\eta := \|A^{-1}H(x_0)\|, \qquad h := \gamma\eta, \qquad r_{\mathrm{NK}} := \frac{2\eta}{1 + \sqrt{1 - 2h}}. \tag{29}$$

If

$$h \le \frac{1}{2}, \qquad r_{\mathrm{NK}} \le R, \tag{30}$$

then Newton's iteration

$$x_{k+1} = x_k - DH(x_k)^{-1}H(x_k), \qquad k \ge 0, \tag{31}$$

is well defined and converges to a zero x_* of H satisfying

$$\|x_* - x_0\| \le r_{\mathrm{NK}} \le 2\eta. \tag{32}$$

If $\eta = 0$, then $x_* = x_0$. If $\gamma = 0$, Newton's method reaches a zero after at most one nontrivial step.

Proof sketch of Theorem 4.3. The key observation is that nondegenerate alignment with $\mathbf{t}$ amounts to solving the root-finding problem $Q_{\mathbf{t}}^\top \mathbf{v}(p) = 0$ while ensuring that $\mathbf{t}^\top \mathbf{v}(p) \ge \rho$. The proof restricts prefix perturbations to $\mathrm{Im}(B_{\mathbf{t}}^\top)$, turning the transverse equation into a square, nonsingular system to which the Newton-Kantorovich theorem (Theorem B.3) applies. The resulting distance bound, together with the uniform bound on J, ensures that the component along $\mathbf{t}$ remains at least ρ. The full proof is deferred to Appendix B.3. □

Now we are ready to provide the full proof of our theorem.

Proof of Theorem 4.3. If $b_{\mathbf{t}} = 0$, then $Q_{\mathbf{t}}^\top \mathbf{v}_0 = 0$. Because the columns of $Q_{\mathbf{t}}$ span $\mathbf{t}^\perp$, it follows that $\mathbf{v}_0 = q_{\mathbf{t}}\mathbf{t}$. The last condition in Eqn. (9) gives $q_{\mathbf{t}} \geq \rho$, so $p_* = p_0$ satisfies the conclusion.

Suppose henceforth that $b_{\mathbf{t}} > 0$. Since $B_{\mathbf{t}}$ has full row rank, $\mathrm{Im}(B_{\mathbf{t}}^\top)$ has dimension $d-1$. Choose $U \in \mathbb{R}^{m\times(d-1)}$ with orthonormal columns spanning $\mathrm{Im}(B_{\mathbf{t}}^\top)$.

Let $\Omega_{\mathbf{v}} \supset \overline{B}(p_0, R)$ be an open set on which $\mathbf{v}$ is continuously differentiable. Define

$$\Omega_G := \left\{z \in \mathbb{R}^{d-1} : p_0 + Uz \in \Omega_{\mathbf{v}}\right\}, \qquad G(z) := Q_{\mathbf{t}}^\top \mathbf{v}(p_0 + Uz). \tag{33}$$

The set Ω_G is open. Since $U^\top U = I$, we have $\|Uz\| = \|z\|$, and hence $\overline{B}(0, R) \subset \Omega_G$. Thus G has the open-domain regularity required by Theorem B.3.

The derivative of G and its value at the origin are

$$DG(z) = Q_{\mathbf{t}}^\top J(p_0 + Uz)U, \qquad A := DG(0) = B_{\mathbf{t}}U. \tag{34}$$

Let $P := UU^\top$. By construction, P is the orthogonal projector onto $\mathrm{Im}(B_{\mathbf{t}}^\top)$. Since $\mathrm{Im}(B_{\mathbf{t}}^\top) = \ker(B_{\mathbf{t}})^\perp$, we have $B_{\mathbf{t}}(I - P) = 0$ and $B_{\mathbf{t}}P = B_{\mathbf{t}}$. Consequently,

$$AA^\top = B_{\mathbf{t}}UU^\top B_{\mathbf{t}}^\top = B_{\mathbf{t}}B_{\mathbf{t}}^\top. \tag{35}$$

It follows that

$$\sigma_{\min}(A) = \sigma_{\min}(B_{\mathbf{t}}) = s_{\mathbf{t}} > 0, \qquad \|A^{-1}\|_{\mathrm{op}} = s_{\mathbf{t}}^{-1}. \tag{36}$$

In particular, A is invertible.

For $z, z' \in \overline{B}(0, R)$, the Lipschitz assumption on J and the fact that U is orthonormal give

$$\begin{aligned}\left\|A^{-1}\big(DG(z) - DG(z')\big)\right\|_{\mathrm{op}} &= \left\|A^{-1}Q_{\mathbf{t}}^\top\big(J(p_0 + Uz) - J(p_0 + Uz')\big)U\right\|_{\mathrm{op}} \\ &\leq \|A^{-1}\|_{\mathrm{op}}\|Q_{\mathbf{t}}^\top\|_{\mathrm{op}}K\|U(z - z')\|\|U\|_{\mathrm{op}} \\ &\leq \frac{K}{s_{\mathbf{t}}}\|z - z'\|. \end{aligned} \tag{37}$$

Thus Theorem B.3 applies with $x_0 = 0$, $H = G$, and $\gamma = K/s_{\mathbf{t}}$.

The initial Newton correction satisfies

$$\eta := \|A^{-1}G(0)\| = \left\|A^{-1}Q_{\mathbf{t}}^\top \mathbf{v}_0\right\| \leq \frac{b_{\mathbf{t}}}{s_{\mathbf{t}}}. \tag{38}$$

Therefore,

$$h := \gamma\eta \leq \frac{Kb_{\mathbf{t}}}{s_{\mathbf{t}}^2} \leq \frac{1}{2}, \tag{39}$$

and

$$r_{\mathrm{NK}} \leq 2\eta \leq \frac{2b_{\mathbf{t}}}{s_{\mathbf{t}}} \leq R. \tag{40}$$

All the hypotheses of Theorem B.3 are consequently satisfied. Hence, there exists $z_* \in \mathbb{R}^{d-1}$ such that

$$G(z_*) = 0, \qquad \|z_*\| \leq r_{\mathrm{NK}} \leq \frac{2b_{\mathbf{t}}}{s_{\mathbf{t}}}. \tag{41}$$

Set $p_* := p_0 + Uz_*$. Then $Q_{\mathbf{t}}^\top \mathbf{v}(p_*) = G(z_*) = 0$. By the definition of $Q_{\mathbf{t}}$, it follows that

$$\mathbf{v}(p_*) = a_*\mathbf{t}, \qquad a_* := \mathbf{t}^\top \mathbf{v}(p_*).$$

Moreover, because U is orthonormal,

$$\|p_* - p_0\| = \|Uz_*\| = \|z_*\| \leq \frac{2b_{\mathbf{t}}}{s_{\mathbf{t}}}. \tag{42}$$

It remains to show that the resulting multiple of $\mathbf{t}$ is positive and nondegenerate. The line segment joining p_0 and p_* lies in $\overline{B}(p_0, R)$ by Eqn. (40). Hence the fundamental theorem of calculus and the jacobian boundedness condition give

$$\|\mathbf{v}(p_*) - \mathbf{v}_0\| = \left\|\int_0^1 J\big(p_0 + \tau(p_* - p_0)\big)(p_* - p_0)\, d\tau\right\| \leq L\|p_* - p_0\|. \tag{43}$$

Consequently,

$$\begin{aligned} a_* &= \mathbf{t}^\top \mathbf{v}(p_*) \\ &= q_{\mathbf{t}} + \mathbf{t}^\top \big(\mathbf{v}(p_*) - \mathbf{v}_0\big) \\ &\geq q_{\mathbf{t}} - \|\mathbf{v}(p_*) - \mathbf{v}_0\| \\ &\geq q_{\mathbf{t}} - L\|p_* - p_0\| \\ &\geq q_{\mathbf{t}} - \frac{2Lb_{\mathbf{t}}}{s_{\mathbf{t}}} \\ &\geq \rho. \end{aligned}$$

This proves the claim. □

Remark B.4 (Locality of the sufficient condition)**.** *As expected, note that Theorem 4.3 is necessarily local. Indeed,*

$$s_{\mathbf{t}} \leq \|B_{\mathbf{t}}\|_{\mathrm{op}} \leq \|J(p_0)\|_{\mathrm{op}} \leq L.$$

Hence, the third condition in Eqn. (9) implies $q_{\mathbf{t}} > 2b_{\mathbf{t}}$. Writing $\theta_0 := \angle(\mathbf{v}_0, \mathbf{t})$, we have $q_{\mathbf{t}} = \|\mathbf{v}_0\| \cos\theta_0$ and $b_{\mathbf{t}} = \|\mathbf{v}_0\| \sin\theta_0$ by definition, so $\theta_0 < \arctan(1/2)$. Thus, the theorem certifies exact alignment from a local neighborhood.

C Additional Experimental Details

C.1 Experimental Setup for Reorientation with a Learned Prefix (Section 5.1.1)

ActAdd baseline. For this task, ActAdd (Rimsky et al., 2024; Turner et al., 2024) computes the difference-in-means $\mathbf{v}_l$ of last-token activations over a set of harmless and harmful prompts, which we denote as $\mathcal{C}^+$ and $\mathcal{C}^-$ respectively, at a chosen layer l. For a new harmful prompt, it adds $\mathbf{v}_l$ to all token positions at that layer during inference as in Eq. 2 (Arditi et al., 2024). Here, the prompt's last-token activation is translated towards the harmless activations, and even beyond with a stronger intervention strength. As $\boldsymbol{\mu}_l^+$ is fixed under CoRA, aligning $\mathbf{v}_l(E)$ with $\mathbf{t} = -\hat{\mathbf{v}}_l$ aims at displacing $\boldsymbol{\mu}_l^-(E)$ into a region similar to that reached by ActAdd, where the norm ratio r in Eqn. (5) specifies how far it is to be displaced, analogous to the intervention strength.

Data and training. To compute the reference difference-in-means $\mathbf{v}_l$, we utilize 200 harmful prompts from the AdvBench dataset (Zou et al., 2023) and 200 harmless prompts from the Alpaca dataset (Taori et al., 2023). We then train the prefix E on the 200 harmful prompts for 300 steps on with a batch size of 32, where we note that $\mathcal{L}_{\mathrm{PPL}}$ is computed over the harmful prompts only, without any generations. We optimize the prefix with AdamW (Loshchilov & Hutter, 2019) at a learning rate of 0.01, with linear warmup over the first 6% of steps followed by cosine decay. For each model, we sweep the number of prefix tokens over $\{1, 2, 4, 6, 8\}$, λ_{PPL} over the range $[0.1, 1.0]$, and the norm ratio r over $[0, 0.5]$, and report results for the best-performing configuration.

Evaluation. We evaluate our method by prepending the trained prefix to a held-out set of 104 harmful prompts from AdvBench, and utilize the open-source HarmBench classifier (Mazeika et al., 2024) to determine whether the resulting generations are harmful. We report the Attack Success Rate (ASR), the perplexity (PPL) of the generations given the input prompts, and the scores on the five tinyBenchmarks tasks (Maia Polo et al., 2024) along with their average. As references, we additionally report each metric under no intervention and under a randomly initialized prefix of 4 tokens.

C.2 Experimental Setup for Alignment Loss-Guided GCG (Section 5.1.2)

GCG. GCG (Zou et al., 2023) first appends an adversarial suffix to the prompt that attempts to induce harmful behavior, then optimizes the suffix to maximize the likelihood of a target sequence of tokens $x^*_{n+1:n+H}$ by minimizing $\mathcal{L}_{\mathrm{adv}}(x_{1:n}) = -\log p(x^*_{n+1:n+H} \mid x_{1:n})$. Because the tokens to be optimized over are discrete, GCG uses Greedy Coordinate Gradient-based Search to optimize the adversarial suffix. At each token position, it uses gradients of the one-hot encoded token indicators to propose the top-k promising token substitutions. It then evaluates the candidates with a forward pass and swaps in the token that maximally decreases the loss.

Adaptation and configuration. In GCG the adversarial suffix is appended after the prompt rather than prepended as in Eqn. (3), and in this study we adapt GCG's format accordingly. Apart from the

additional alignment and norm terms in Eqn. (10), the discrete search procedure remains unchanged. Each run optimizes 20 suffix tokens per prompt for 250 steps with search width 512 and top-$k = 256$. Similar to the jailbreaking experiment in Section 5.1.1, we report the main metric ASR for jailbreaking, and PPL and tinyBenchmark scores for general utility, evaluated under the same setting as Zou et al. (2023).

C.3 Experimental Setup for Interference Control in Multi-Concept Steering (Section 5.2)

Training objective. The full objective used to train the prefix E is

$$\mathcal{L}_{\text{multiconcept}} = \mathcal{L}_{\text{align}}\big(\mathbf{v}_l^A(E), \mathbf{v}_l^B(E)\big) + \sum_{c\in\{A,B\}} \big(\lambda_{\text{norm}}\mathcal{L}^c_{\text{norm},1}(E) + \lambda_{\text{ppl}} \cdot \mathcal{L}^c_{\text{PPL}}(E)\big) + \lambda_{\text{hinge}} \cdot \operatorname*{mean}_i \max\big(0,\ \kappa - \cos(h_l(x_i; E), h_l(x_i))\big), \tag{44}$$

where $\mathcal{L}_{\text{align}}$ is given by Eqn. (11), and $h_l(x_i; E)$ and $h_l(x_i)$ are the activations of contrastive sample i at layer l with and without the prefix. The norm and perplexity terms follow Eqn. (5) and are applied to each concept c. The hinge term constrains the unsteered model by imposing a penalty if the prefix E rotates the activation of x_i beyond a threshold κ.

Data. Following Rimsky et al. (2024), each sample in *coordinate-other-ais* and *survival-instinct* is a question with one answer that matches the target concept, and one answer that does not match the target concept. For each concept c and layer l, we build contrastive pairs by adding the matching and the non-matching answer letter respectively to the chat-templated prompt, and compute their means $\boldsymbol{\mu}_l^{c,+}$ and $\boldsymbol{\mu}_l^{c,-}$, as well as the difference-in-means $\mathbf{v}_l^c$. We use a set of 200 samples for training the prefix E, and another set of 200 samples for computing the difference-in-means.

Steering. Following the setup in Rimsky et al. (2024), we add $\mathbf{v}_l^c$ to all token positions after the prompt at layer l. In the setting with CoRA, we prepend the prefix E to both contrastive classes, and calculate the steering vector $\mathbf{v}_l^c(E)$. The ActAdd baseline uses $\mathbf{v}_l^c$ computed without our prefix.

D Supplementary Empirical Analysis

D.1 Effect of Random Prefix on Language

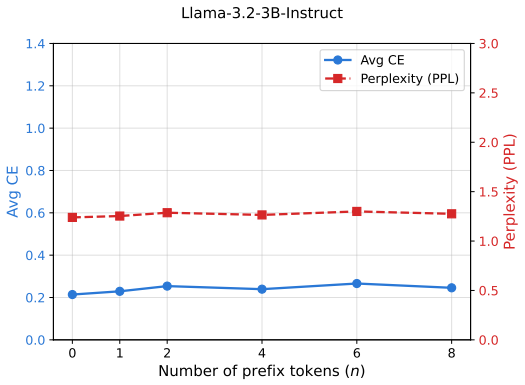


Figure 7: Effect of random prefix embeddings on language modeling quality is negligible.

To evaluate the effect that a random prefix has on language modeling quality, we measure the Perplexity (PPL) of generations when a random prefix is added. We sample a random prefix of length $T \in \{1, 2, 4, 6, 8\}$ and prepend the prefix to 50 prompts from the Alpaca dataset (Taori et al., 2023), and calculate the PPL of the generations given the input prompts. For reference, we also calculate the PPL in the case without having a prefix added. Figure 10 shows that the PPL does not change notably despite varying the number of prefix tokens, indicating that adding prefix tokens does not significantly impact language modeling quality.

D.2 Alignment to a Random Target Vector

To test whether CoRA can reorient the contrastive vector toward arbitrarily chosen directions without any semantic interpretation, we replace $\mathbf{t} = -\hat{\mathbf{v}}_l$ in Figure 2 with random target directions. We use

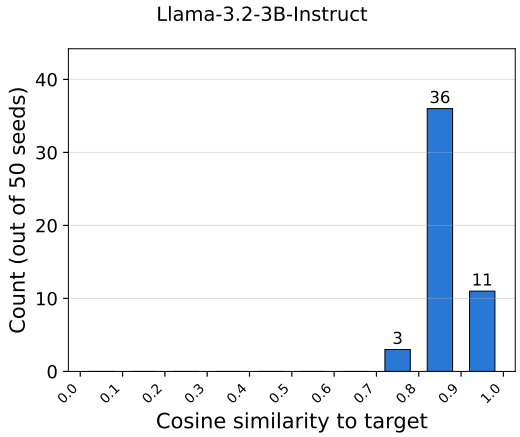


Figure 8: Final cosine similarity between $\mathbf{v}_l(E)$ and 50 random target directions $\mathbf{t}$ on Llama-3.2-3B-Instruct.

the harmful and harmless prompts of Section 5.1.1 and align the difference-in-means with a random $\mathbf{t}$ on Llama-3.2-3B-Instruct with 4 shared prefix tokens. We sample 50 target vectors uniformly random from $\mathbb{S}^{d-1}$ and, for each sampled direction, train a prefix by minimizing the alignment loss term Eqn. (5). For reference, the cosine similarity between a fixed vector and a uniformly random unit vector in $\mathbb{R}^d$ has standard deviation approximately $1/\sqrt{d} \approx 0.018$ for $d = 3072$. As shown in Figure 8, all 50 trained prefixes achieve a cosine similarity of at least 0.7 with their targets, with 47 exceeding 0.8 and 11 exceeding 0.9. This indicates that CoRA is able to achieve substantial reorientation toward arbitrarily chosen directions.

D.3 Layerwise Cosine Similarity Analysis

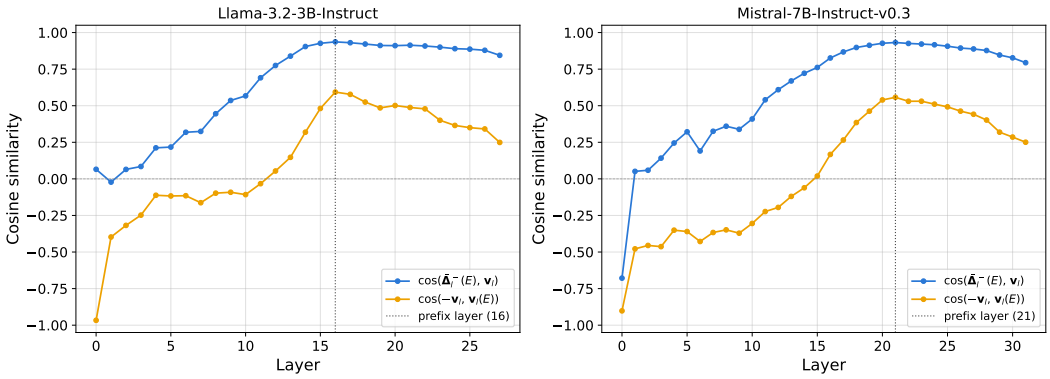


Figure 9: Layerwise cosine similarities for prefixes trained at a single layer (dotted line) on Llama3.2-3B-Instruct and Mistral-7B-Instruct-v0.3 respectively. The individual quantities are recomputed at each layer using the respective trained prefix E, and we observe that the reorientation of $\mathbf{v}_l(E)$ towards $-\mathbf{v}_l$ peaks at the trained layer. Furthermore, the alignment of $\bar{\mathbf{\Delta}}_l^-(E)$ with $\mathbf{v}_l$ is highest there but remains high at subsequent layers.

In this appendix, we examine how the alignment achieved by a prefix trained at a single layer l^* behaves at the remaining layers of the model. Specifically, we take the trained prefixes from Section 5.1.1, and recompute $\mathbf{v}_l$, $\mathbf{v}_l(E)$ and $\bar{\mathbf{\Delta}}_l^-(E)$ at every layer l instead of only at l^*.

As shown in Figure 9, the reorientation peaks precisely at the trained layer for both models, and decays on either side. As for $\bar{\mathbf{\Delta}}_l^-(E)$, its alignment with $\mathbf{v}_l$ is similarly highest at l^*, but remains high at all subsequent layers. We note that the peak is expected, since reorienting $\mathbf{v}_l(E)$ towards $-\mathbf{v}_l$ requires the source mean to be displaced along $\mathbf{v}_l$. Thus, this serves as a check that the reorientation is achieved through the mechanism described in Section 5.1.1.

Table 4: Attack Success Rates (ASR) of ActAdd when steering with the difference-in-means vector $\mathbf{v}_l$ and with the source-mean displacement $\bar{\mathbf{\Delta}}_l^-(E)$ induced by the trained CoRA prefix. ASR is computed on the generated outputs with the HarmBench classifier. Best ASRs are **bolded**. Our method is highlighted by blue.

Method	ASR ↑
Llama3.2-3B-Instruct	
No Steering	2.88
ActAdd	96.15
ActAdd ($\bar{\mathbf{\Delta}}_l^-(E)$)	83.65
Mistral-7B-Instruct-v0.3	
No Steering	43.27
ActAdd	87.50
ActAdd ($\bar{\mathbf{\Delta}}_l^-(E)$)	81.73

D.4 Source-mean Displacement as a Steering Vector

Given that $\bar{\mathbf{\Delta}}_l^-(E)$ is well aligned with $\mathbf{v}_l$ at the trained layer, we examine whether it also reproduces its behavioral effect as a steering vector.

To test, we replace the difference-in-means steering vector for ActAdd with the source-mean displacement, following the same setup as in Section 5.1.1. We perform this test on Llama3.2-3B-Instruct and Mistral-7B-Instruct-v0.3. Table 4 shows that steering with the source-mean displacement achieves ASRs of 83.65 and 81.73 on Llama3.2-3B-Instruct and Mistral-7B-Instruct-v0.3, respectively, compared to 96.15 and 87.50 for the difference-in-means vector. Relative to no steering, the source-mean displacement recovers approximately 87% of the ASR gain achieved by ActAdd on both models. This suggests that the source-mean displacement $\bar{\mathbf{\Delta}}_l^-(E)$ is a reasonable, though not exact, approximation of the difference-in-means steering vector $\mathbf{v}_l$.

D.5 Additional Analysis on Non Uniform Induced Perturbation

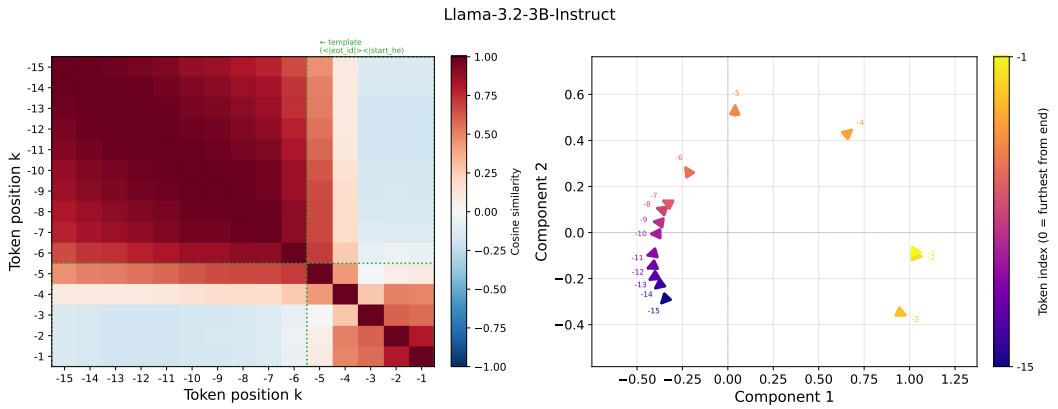


Figure 10: Source-mean displacements $\bar{\mathbf{\Delta}}_{l,k}^-(E)$ evaluated at each token position k, indexed from the end of the sequence, on Llama3.2-3B-Instruct. **Left:** pairwise cosine similarity between the displacements at positions j and k, where the dotted lines mark the boundary between the content tokens and the chat-template tokens. **Right:** PCA projection of the same displacements, with colour denoting token position. We observe that the displacements at the content tokens remain strongly aligned with one another, whereas those at the final template tokens depart sharply and occupy distinct regions of the projection.

To supplement the empirical finding in Section 6 that CoRA induces a non-uniform intervention across token positions, we examine how the source-mean displacement $\bar{\mathbf{\Delta}}_{l,k}^-(E)$ itself varies with position k. Figure 10 reports the pairwise cosine similarities between the displacements at each pair of positions, as well as their projection onto the first two principal components.

We observe that the similarity matrix in the left panel exhibits a block structure. The displacements at the content-token positions, approximately -15 through -6, are strongly aligned with one another, indicating that the induced direction changes only gradually over this portion of the sequence. In contrast, positions -3 through -1 form a second internally aligned block, with weak or negative similarity to many of the earlier content-token displacements. As for positions -5 and -4 in particular, these lie near the boundary and display intermediate behaviour, suggesting that the change

is associated with the transition from content to template tokens rather than occurring uniformly at every position.

The projection in the right panel provides a complementary view. The content-token displacements form a compact, ordered trajectory from positions -15 to -6, whereas the subsequent template-token displacements move abruptly into separated regions of the projection. Thus, this shows that the prefix does not induce a single position-independent displacement, in contrast to ActAdd, which adds a fixed vector at every position.

E LIMITATIONS

CoRA uses continuous soft prefixes as a controlled proxy for context. Because these embeddings need not correspond to natural-language tokens, their expressivity may exceed that of coherent textual contexts; our results therefore establish sensitivity within the input-embedding space, not that equally large reorientations are attainable through natural language. Moreover, our experiments consider a shared, source-only prefix and difference-in-means vectors extracted from final-token activations, so testing the generalization of our results to other contextual interventions, readouts, or representation-extraction methods remains an open research opportunity. Our theoretical bounds apply to the specified pre-normalized architecture and provide necessary global or sufficient local conditions rather than a complete characterization of reachability, which is inherently hard for the practical LLM architectures.