Steering as Implicit Low-Rank Finetuning: A Theoretical Framework for Activation Steering in Transformers

Ender D. Isik[1], A. Sila Okcu[2]

[1]EPFL, Lausanne, CH, [2]University of Cambridge, Cambridge, UK

ICML International Conference On Machine Learning

1 TL;DR

- We ask **when rank-one LoRA induces steering-like activation shifts** in a controlled linear-attention transformer.
- Geometry reveals writing, routing, and their interaction.
- VO updates write; QK updates route. VO gives fixed-direction shifts, while QK gives context-dependent shifts.

2 Background and Motivation

- **Steering.** Modify activations by adding a fixed direction:

$$h' = h + \alpha(h)u$$

- **Rank-1 LoRA.** Modify weights with a rank-one update:

$$W' = W + ab^\top$$

- **Alignment Task.**: Shift negative-control prompts while preserving positive-control prompts:

$$\mathbb{E}[\Delta h \mid C^-] = \bar{\alpha}u, \qquad \mathbb{E}[\Delta h \mid C^+] \approx 0$$

3 Two Pathways Two Mechanisms

- **Core Idea.** Attention has two distinct pathways. VO writes information into the residual stream, while QK routes information across context.
- **Linear Attention**

$$Q = W_Q H, \qquad K = W_K H, \qquad V = W_V H$$
$$A(H) = Q^\top K, \qquad H^+ = H + W_O V A(H)$$

- **VO Pathway**

Rank-one updates to W_V and W_O produce shifts in a small fixed subspace:

$$\delta h_i^{V+O} \in \mathrm{span}\{b_O, W_O b_V\}$$

So joint VO updates have:

$$\mathrm{rank}(\Delta H_{V+O}) \le 2$$

- **QK Pathway**

Rank-one QK updates change the attention scores:

$$\Delta A(H) = H^\top \Delta M H$$

The shift direction depends on the input context:

$$\delta h_T^{(i)} = \left(b^\top H^{(i)} e_T\right) d_{H^{(i)}}$$

- **All matrices: writing × routing interaction**

When Q, K, V, O are all trained, the full shift is not just VO plus QK:

$$\Delta H_{\mathrm{all}} = \Delta H_{\mathrm{VO}} + \Delta H_{\mathrm{QK}} + R_{\mathrm{int}}$$

Full attention finetuning combines fixed-direction writing, context-dependent routing, and non-additive interaction terms.

4 Experimental Setup

- **Task.** Binary behavior task with topic tokens T_i, control tokens $C^+$and C^-, a query token Q, and paired answers A_i^+, A_i^-.
- **Pre-Training.** Model learns control-dependent behavior:

$$[BOS, T_i, C^+, Q] \to A_i^+, \quad [BOS, T_i, C^-, Q] \to A_i^-$$

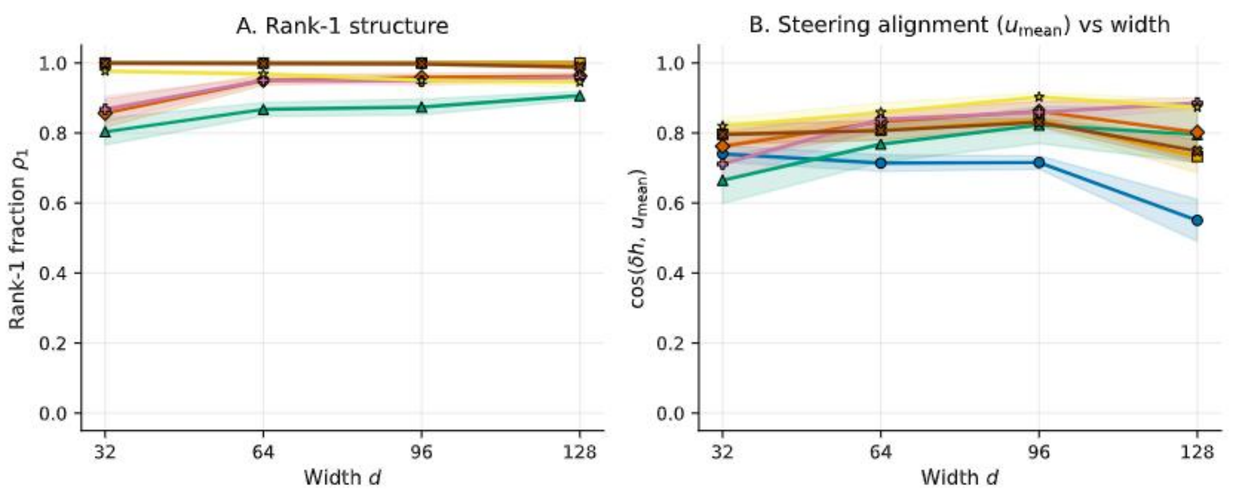
- **Alignment.** Negative-control prompts are retargeted to the positive answer, while positive-control prompts are preserved:

$$C^- \to A_i^+ \qquad C^+ \to A_i^+$$

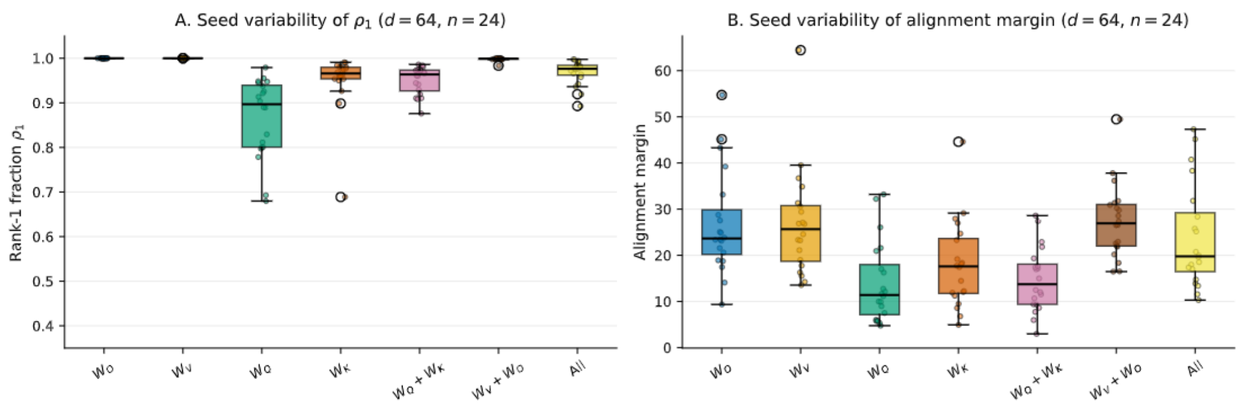
- **Model.** We use a **two-layer decoder-only transformer** with single-head **linear attention**, no MLPs, and no normalization.
- **Finetuning.** We finetune rank-one updates on different subsets of the first-layer attention matrices: $\{W_Q, W_K, W_V, W_O\}$
- **Steering direction.** We define a mean-difference steering direction from first-layer activations:

$$u_{mean} = E[h_T^{(1)} \mid C^+] - E[h_T^{(1)} \mid C^-]$$

5 Results



- We measure the **rank-one concentration** of LoRA-induced residual shifts and their **cosine alignment with u_{mean}.** VO updates are exactly rank-one across widths, matching the theory, while QK updates are less rank-concentrated.



- We measure **seed-level variability** in rank-one concentration and alignment margin. VO updates are stable across seeds, with exactly rank-one residual shifts, while QK updates show larger variability despite reaching similar behavioral alignment.

6 Discussion

- **Behavioral accuracy is not enough:** Different LoRA targets solve the same task, but induce different residual-stream geometries.
- **The setting is deliberately controlled:** Linear attention isolates the VO/QK distinction, but standard transformers may add softmax, MLP, and normalization effects.
- **Rank-one is only the first case:** Higher-rank LoRA may correspond to multi-direction steering, but its geometry remains open.

Main Takeaways:

- **Steering-like finetuning is pathway-dependent:** It is structurally guaranteed for VO writes, not for QK routing.
- **Routing can support alignment without a fixed direction:** QK updates can solve the task while producing context-dependent activation shifts.
- **Full attention finetuning is compositional:** All-matrix updates combine writing and routing rather than reducing to either mechanism alone.

Mechanistic Interpretability Workshop, ICML 2026

ender.isik@epfl.ch, aso32@cam.ac.uk