Mechanistic Interpretability Workshop · ICML 2026

Don't Lose Focus

Activation Steering via Key-Orthogonal Projections (SKOP)

Haoyan Luo† Mateo Espinosa Zarlenga‡ Mateja Jamnik†

† University of Cambridge ‡ University of Oxford · Contact: *hl678@cam.ac.uk*

01 / RESEARCH PROBLEM

Vanilla steering can reroute attention away from useful context

Activation steering gives inference-time control without retraining, but it often harms reasoning and retrieval. We argue that this degradation is driven by *attention rerouting*: steering changes query-key matching, shifting attention away from the model's baseline focus tokens.

KEY MECHANISM

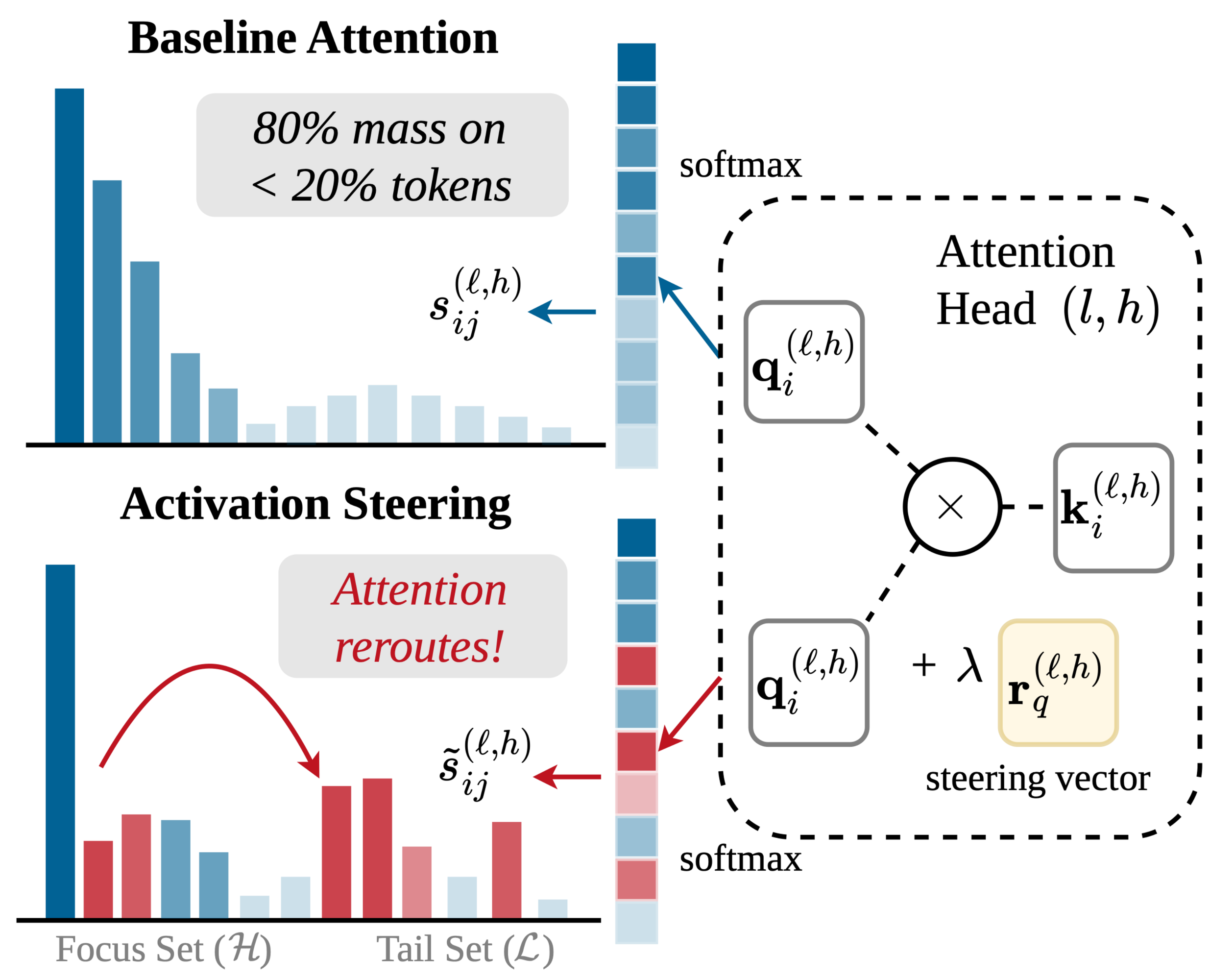
Attention rerouting — steering changes query-key matching, shifting attention from baseline focus tokens $\mathcal{H}$ to less informative tail tokens $\mathcal{L}$ on utility tasks.

Evidence 1

≲15 focus tokens hold **80%** of a head's baseline attention; vanilla steering often moves this mass into tail tokens.

Evidence 2

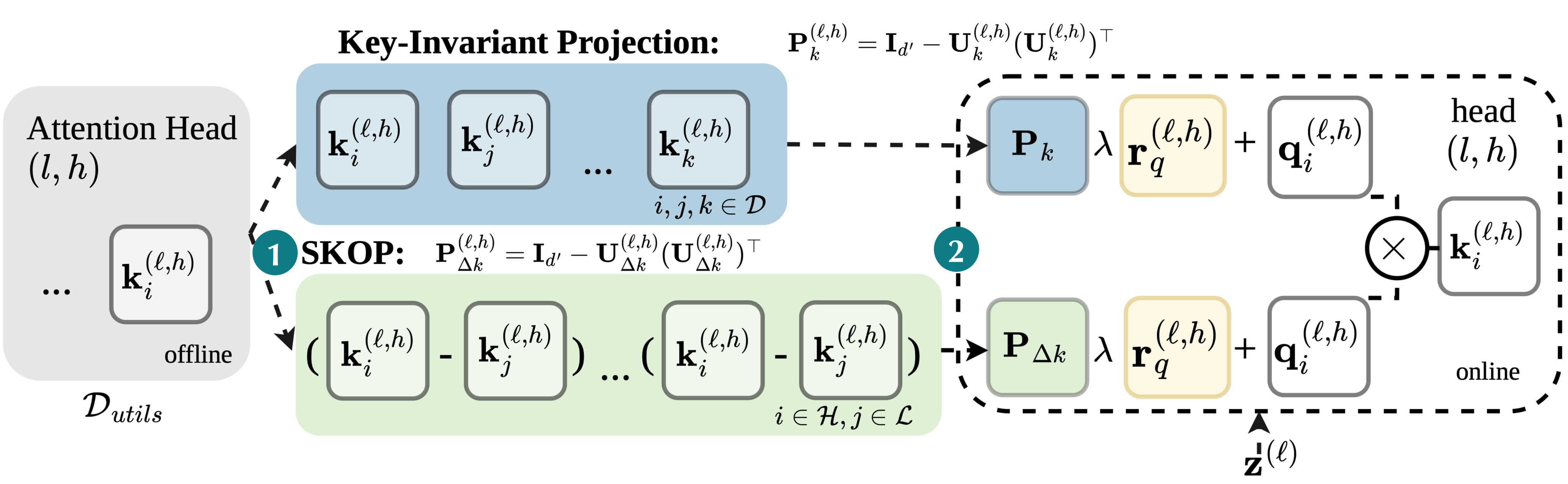
~20% high-risk attention heads account for most focus-to-tail rerouting.



Baseline attention concentrates on focus tokens (blue); vanilla steering reroutes mass to tail tokens (red).

02 / METHOD

Remove the harmful focus-to-tail rerouting



WHY KEY-DIFFERENCES IDENTIFY REROUTING

Attention softmax depends on **score gaps, not absolute scores**. Under query steering, the reroute between a **focus** token i and a **tail** token j is governed by the overlap between the steering vector and a **key difference**:

$$\delta_{ij} - \delta_{ij'} = \frac{\lambda}{\sqrt{d'}}\langle r_q, k_j - k_{j'}\rangle$$

SKOP estimates the dominant focus-tail key-difference directions and projects them out.

1 Characterise the Key-Difference Subspace

Using an offline forward pass on a small calibration set, identify each head's $\mathcal{H}$ and $\mathcal{L}$. Then collect focus-to-tail key differences to estimate the subspace:

$\Delta \mathbf{k}_{ft} = \mathbf{k}_f - \mathbf{k}_t, f \in \mathcal{H}, t \in \mathcal{L}$. $\quad \mathbf{U}_{\Delta k} = \mathbf{topEig}(\mathbb{E}_{f,t}[\Delta \mathbf{k}_{ft} \Delta \mathbf{k}_{ft}^\top])$.

2 Project the steering vector

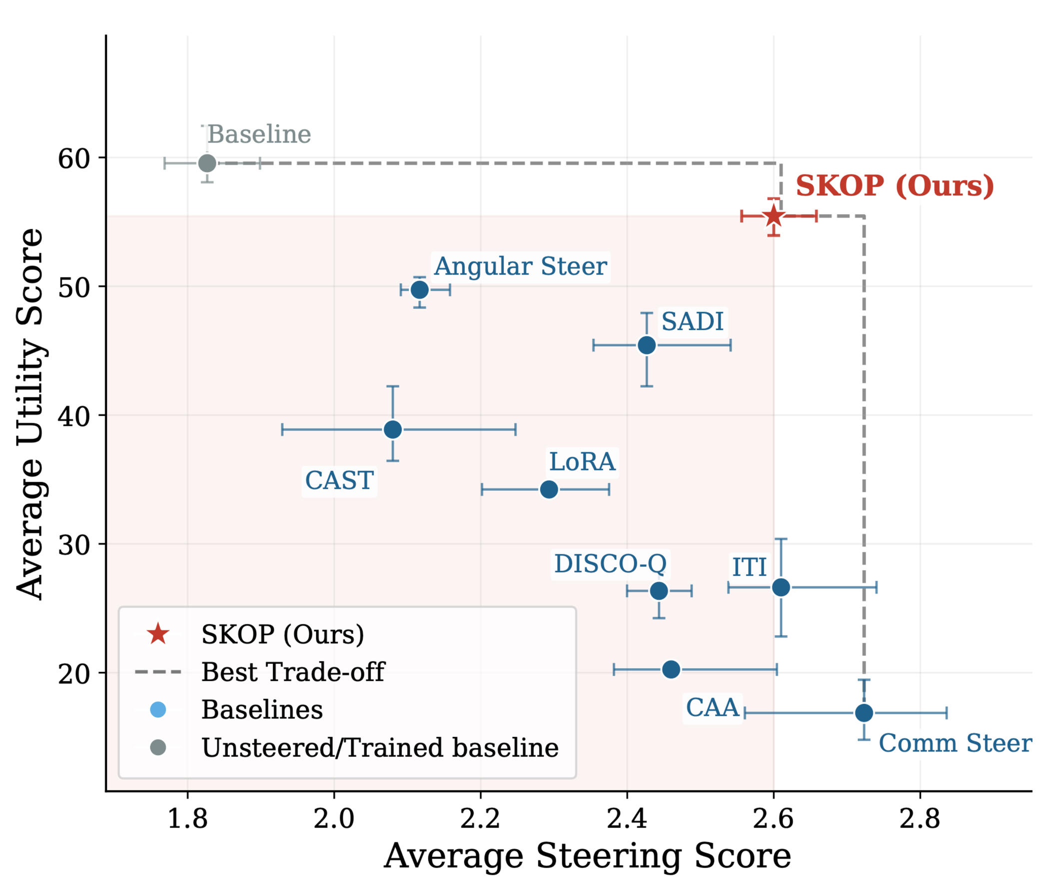
On high-risk heads, replace the steering vector by its component orthogonal to this key-difference subspace, reducing focus-to-tail score-gaps while leaving other steering directions intact.

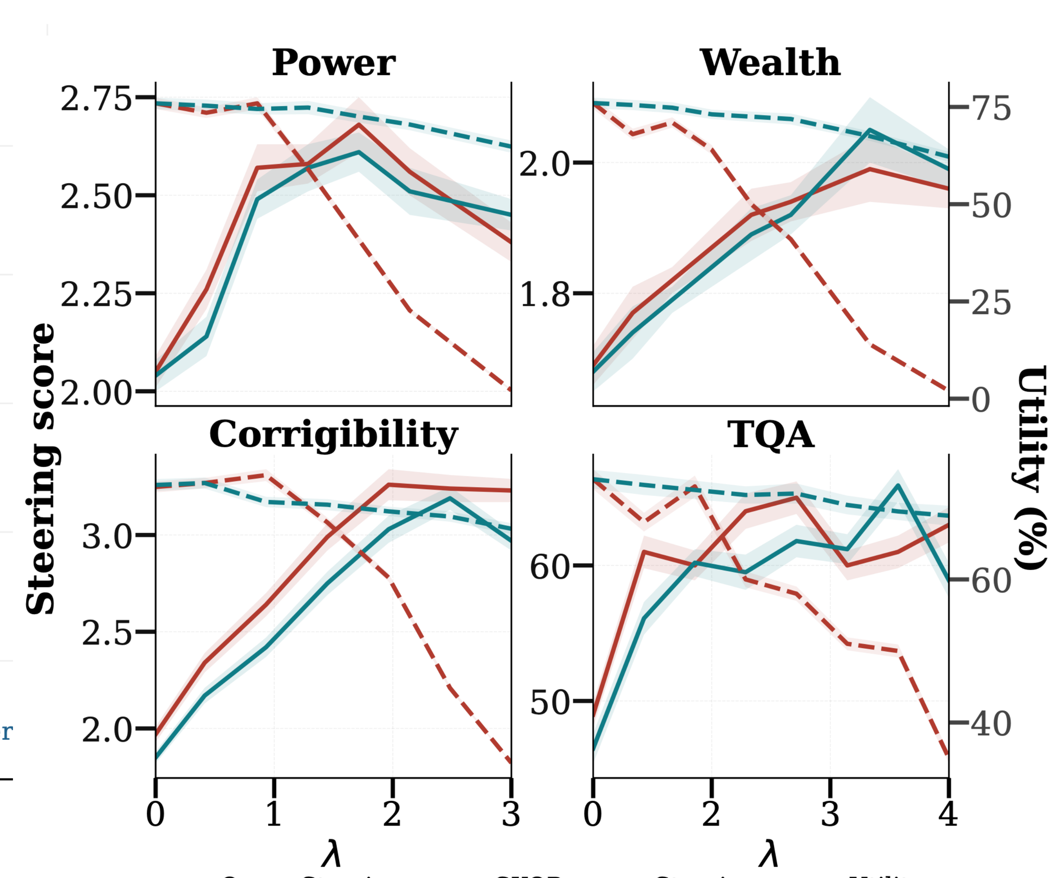
$$\mathbf{P}_{\Delta k} = \mathbf{I} - \mathbf{U}_{\Delta k}\mathbf{U}_{\Delta k}^\top$$

$$\tilde{\mathbf{r}}_q = \mathbf{P}_{\Delta k}\mathbf{r}_q$$

03 / RESULTS

SKOP preserves utility while retaining steering efficacy





- On LLaMA-3.1-8B-Instruct, SKOP ★ achieves the strongest joint steering and utility among steering baselines.
- **Robust to steering strength.** As λ increases, SKOP preserves substantially higher utility while retaining similar steering performance.
- **Long-context robustness:** At 16K context length, SKOP retains better retrieval performance while also giving more robust steering effect.

PAPER CODE

5–7×
lower utility degradation than vanilla steering

≥95%
of vanilla steering efficacy retained

Long-context for activation steering.

NEEDLE-IN-A-HAYSTACK · RETRIEVAL ACC. (%)

Metric	Method	Context length 1K	2K	4K	8K	16K
Formatting compliance (%)						
	DISCO-Q	88.5	82.3	65.5	53.8	36.4
	Comm Steer	**89.6**	**83.2**	74.9	73.1	65.8
	SKOP	84.6	82.1	**79.8**	**76.5**	**71.6**
Retrieval accuracy (%)						
NIAH	Baseline	100.0	100.0	99.8	98.3	97.8
	DISCO-Q	99.5	**99.5**	89.1	81.7	72.6
	Comm Steer	98.3	91.2	79.9	68.2	55.7
	SKOP	**99.8**	**99.5**	**95.0**	**92.1**	**88.2**