Don't Lose Focus: Activation Steering via Key-Orthogonal Projections

Haoyan Luo [1] **Mateo Espinosa Zarlenga** [2] **Mateja Jamnik** [1]

Abstract

Activation steering controls LLM behaviour towards target behaviour by intervening in internal representations, yet it often degrades reasoning and retrieval performance. We argue that a primary cause of this trade-off is *attention rerouting*: steering vectors alter query-key matching, shifting attention away from contextually important tokens toward less informative ones. To address this, we propose *Steering via Key-Orthogonal Projections* (SKOP), a steering method that constrains harmful attention rerouting without eliminating steering efficacy. SKOP achieves this by preserving attention patterns on a small set of *focus tokens* the model relies on for reasoning and retrieval, while allowing redistribution among less critical *tail tokens*. Across multiple steering benchmarks, we show that SKOP achieves the best joint steering-utility trade-off, reducing utility degradation by 5-7× while retaining over 95% of vanilla steering efficacy. Our results further suggest that, in long-context retrieval settings where vanilla steering approaches are ineffective, SKOP can maintain robust performance by avoiding attention rerouting. The code is available at clarenceluo78/skop-steering.

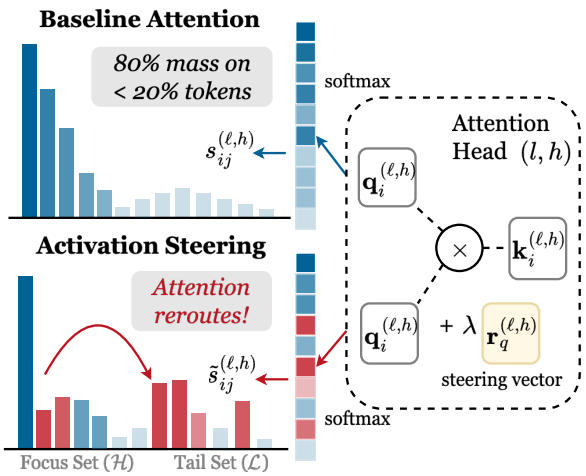

Figure 1. Attention rerouting due to activation steering, a key contributor to the trade-off between steering efficacy and utility preservation.

1. Introduction

Activation steering offers a lightweight inference-time mechanism to control the behaviour of Large Language Models (LLMs) by intervening in their internal representations, avoiding costly retraining (Turner et al., 2024; Luo & Specia, 2024; Rimsky et al., 2024; Stolfo et al., 2025). This has recently emerged as an attractive mechanism for behavioural control due to its ability to elicit target behaviours, such as truthfulness (Li et al., 2024) and harmful content refusal (O'Brien et al., 2024; Lee et al., 2025; Sheng et al., 2025) in open-ended generation (Nguyen et al., 2025; Vu & Nguyen, 2025; Cao et al., 2024; Wu et al., 2025). However, although promising, a fundamental practical challenge remains largely unresolved: the trade-off between *steering efficacy* and *utility preservation*. Specifically, as the steering strength increases, or as it is applied less selectively, eliciting the target behaviour may come at the cost of degrading performance (i.e., *utility*) on unrelated capabilities, such as reasoning and retrieval (Templeton et al., 2024; Turner et al., 2024).

Recent work has made progress toward addressing this trade-off by improving *when* and *where* steering is applied. For example, input-conditional steering has been introduced to mitigate over-refusal by activating steering only in relevant contexts (Lee et al., 2025; Sheng et al., 2025). For open-ended generation, while many approaches steer directly in the model's residual stream (Rimsky et al., 2024; Zou et al., 2023; Nguyen et al., 2025; Wang et al., 2025), recent works suggest that attention-space steering can be highly effective (Li et al., 2024; Torop et al., 2025; Vu & Nguyen, 2025) and better preserve utility by being less intrusive to the residual stream, such as query-space steering (Torop et al., 2025). Yet, it remains unclear *how* attention-space steering alters attention patterns and which of these changes

[1]University of Cambridge [2]University of Oxford. Correspondence to: Haoyan Luo <hl678@cam.ac.uk>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

improve the trade-off.

In this work, we argue that the trade-off is driven by *attention rerouting* (Fig. 1): steering changes how attention queries match keys, which in turn changes which tokens are attended to. We focus on query-space steering for two reasons: (i) it has emerged as a particularly effective steering paradigm due to the high separability of behavioural concepts in the query space (Torop et al., 2025; Vu & Nguyen, 2025), and (ii) as we show in Sec. 3, it isolates the rerouting effect into a single correctable term. In this setting, we observe that attention shifts away from a small *focus set* of tokens the model relies on for correct reasoning and retrieval (Fig. 2(A)) toward a larger *tail set* of less informative tokens, as measured by top-set attention mass preservation (Fig. 2(B)). We show that this rerouting arises because query-space steering alters the *relative* query-key scores determining attention weights (Eq. 7). While it is possible to prevent rerouting by enforcing exact attention invariance, for example, by adapting null-space constraints developed for residual steering (Sheng et al., 2025), this completely suppresses steering efficacy (Fig. 2(C)). Hence, we observe a critical tension: while effective steering requires modifying relative attention scores, utility preservation requires that the attention patterns of important tokens remain undisturbed.

This motivates our approach: rather than eliminating attention rerouting, we selectively constrain it. For this, we introduce *Steering via Key-Orthogonal Projections* (SKOP), which, given a query-space steering vector, removes only the components that strongly shift attention from the focus set to the tail set, leaving other steering effects intact. Concretely, SKOP compares the tokens a head attends to strongly on utility tasks with those it attends to weakly, and uses differences in their key representations to identify steering components that are likely to cause harmful attention shifts. It then removes only these components and applies this correction selectively to a small set of *risk heads* that are most prone to such shifts, thereby preserving steering efficacy while safeguarding model utility. We further show that this mechanism enables robust activation steering in long-context retrieval settings, providing, to the best of our knowledge, the first demonstration of effective long-context activation steering.

Our main contributions can be summarised as follows:

1. We identify *attention rerouting*, steering-induced shifts in attention away from focus tokens, as a key mechanism behind the trade-off between query-space steering efficacy and utility preservation.

2. We propose SKOP, a steering method that suppresses steering components that shift attention away from focus tokens, retaining strong steering efficacy while preserving model utility.

3. We show that SKOP achieves the best steering–utility trade-off across multiple benchmarks, reducing utility degradation by 5–7×, and enabling robust long-context activation steering.

2. Related Work

Activation Steering. Activation steering induces or suppresses specific behaviours in an LLM by modifying its latent space (Arditi et al., 2024; Templeton et al., 2024; Stolfo et al., 2025). The predominant paradigm assumes that the linear representation hypothesis (Mikolov et al., 2013; Park et al., 2024) holds, and uses mean-difference vectors, representing directions in the LLM's latent space, to steer the model (Zou et al., 2023). These *steering vectors* are typically constructed by analysing the last token's representations when the model is given "positive" and "negative" examples of a concept (Li et al., 2024; Rimsky et al., 2024). Steering vectors can also be constructed using non-linear estimation (Qiu et al., 2024), affine transformations (Rodriguez et al., 2025; Singh et al., 2024), or optimisation-based techniques (Wu et al., 2025; 2024). Recent work has demonstrated that steering on the attention layers themselves (e.g., *query-space* steering (Torop et al., 2025; Vu & Nguyen, 2025)) is an effective and fine-grained control mechanism due to the separability of behavioural concepts in the query and value spaces (Torop et al., 2025). However, it remains unclear how activation steering interacts with the attention patterns themselves. Our work fills this gap by identifying attention rerouting as a side-effect of query-space steering and showing that this rerouting underlies the observed steering-utility trade-off (Sec. 4).

Steering vs utility trade-off. A persistent challenge of activation steering is the trade-off between steering efficacy and general model capability (i.e., utility) (Wu et al., 2025). While mitigation strategies such as input-conditional steering (Lee et al., 2025; Sheng et al., 2025), semantic gating (Lee et al., 2025), targeted head selection (Li et al., 2024), and feature-level decomposition (Bayat et al., 2025; Rajamanoharan et al., 2024) have been proposed, they operate on the residual stream and have often been studied in the narrow setting of refusal steering (Lee et al., 2025; Sheng et al., 2025). As a result, it remains unclear *how* these interventions affect attention patterns, or whether they can jointly improve steering efficacy *and* preserve utility in the general behavioural steering setting. Building on our analysis of attention rerouting, we propose SKOP, a mitigation method tailored to query-space steering that improves the joint steering-utility trade-off (Sec. 5).

3. Preliminaries

Consider a *decoder-only transformer* with L layers, each with H attention heads. Here, the residual stream $\mathbf{h}^{(\ell)} \in$

$\mathbb{R}^{t\times d}$ of layer ℓ, where t is the sequence length and d is the token dimension, is:

$$\mathbf{g}^{(\ell)} = \mathbf{h}^{(\ell-1)} + a^{(\ell)}\big(\mathrm{LN}(\mathbf{h}^{(\ell-1)})\big), \tag{1}$$
$$\mathbf{h}^{(\ell)} = \mathbf{g}^{(\ell)} + \mathrm{MLP}^{(\ell)}\big(\mathrm{LN}(\mathbf{g}^{(\ell)})\big), \tag{2}$$

where LN is layer normalisation and $a^{(\ell)}$ is the multi-head attention block at layer ℓ. We use standard multi-head attention notation (Vaswani et al., 2017); App. A.1 summarises all symbols and the extension to grouped-query attention (Ainslie et al., 2023).

The attention block $a^{(\ell)}$ comprises heads $\{a^{(\ell,h)}\}_{h=1}^{H}$, with query/key/value projections $\mathbf{W}_{q,k,v}^{(\ell,h)} \in \mathbb{R}^{d\times d'}$ and output projection $\mathbf{W}_o^{(\ell,h)} \in \mathbb{R}^{d'\times d}$, where $d' = d/H$. Given the attention input $\mathbf{z}^{(\ell)} := \mathrm{LN}(\mathbf{h}^{(\ell-1)})$, the queries, keys, and values are $\mathbf{Q}^{(\ell,h)} = \mathbf{z}^{(\ell)}\mathbf{W}_q^{(\ell,h)}$, $\mathbf{K}^{(\ell,h)} = \mathbf{z}^{(\ell)}\mathbf{W}_k^{(\ell,h)}$, $\mathbf{V}^{(\ell,h)} = \mathbf{z}^{(\ell)}\mathbf{W}_v^{(\ell,h)}$, and the attention logits and outputs are:

$$s_{ij}^{(\ell,h)} = \langle \mathbf{q}_i^{(\ell,h)}, \mathbf{k}_j^{(\ell,h)}\rangle/\sqrt{d'}, \tag{3}$$
$$a^{(\ell,h)}(\mathbf{z}^{(\ell)})_i = \sum_{j=1}^{t} \alpha_{ij}^{(\ell,h)}\mathbf{v}_j^{(\ell,h)}\mathbf{W}_o^{(\ell,h)}, \tag{4}$$

where $\alpha_{ij}^{(\ell,h)} = \mathrm{softmax}_j(s_{ij}^{(\ell,h)})$ are the attention weights.

Query-space steering. Activation steering controls an LLM's behaviour by adding a fixed *steering vector* $\mathbf{r}$ to its latent representations (Rimsky et al., 2024; Li et al., 2024; Zou et al., 2023). Among these approaches, *query-space steering* (Torop et al., 2025) stands out since (1) it achieves strong steering efficacy (Torop et al., 2025; Vu & Nguyen, 2025), and (2) as derived below, its effect on attention logits can be easily captured by a closed-form term. As these properties more easily permit the study of powerful steering methods, we focus our analysis on query-space steering.

Given a query-space steering vector $\mathbf{r}_q^{(\ell,h)} \in \mathbb{R}^{d'}$ – typically obtained as the mean difference between query activations on positive and negative examples of a target behaviour (Rimsky et al., 2024; Li et al., 2024; Torop et al., 2025) – query-space steering modifies queries as follows:

$$\mathbf{q}_i^{(\ell,h)} \leftarrow \mathbf{q}_i^{(\ell,h)} + \lambda\mathbf{r}_q^{(\ell,h)}, \tag{5}$$

where $\lambda \in \mathbb{R}$ controls *steering strength*. Substituting the steered query into the logit definition $s_{ij}^{(\ell,h)} = \langle \mathbf{q}_i^{(\ell,h)}, \mathbf{k}_j^{(\ell,h)}\rangle/\sqrt{d'}$, and expanding the inner product, yields the following updated logit:

$$\begin{aligned}\tilde{s}_{ij}^{(\ell,h)} &:= \frac{\left\langle \mathbf{q}_i^{(\ell,h)} + \lambda\mathbf{r}_q^{(\ell,h)}, \mathbf{k}_j^{(\ell,h)}\right\rangle}{\sqrt{d'}} \\ &= s_{ij}^{(\ell,h)} + \underbrace{\frac{\lambda\left\langle \mathbf{r}_q^{(\ell,h)}, \mathbf{k}_j^{(\ell,h)}\right\rangle}{\sqrt{d'}}}_{\delta_{ij}^{(\ell,h)}}.\end{aligned} \tag{6}$$

Details of this derivation can be found in App. B. Notice that the perturbation $\delta_{ij}^{(\ell,h)}$ is the *only* term added to the logits. This isolates how attention may be rerouted from query-space steering into one tractable term, a property that does not hold for residual-stream steering (where a perturbation simultaneously alters queries, keys, values, and MLP outputs). We exploit this isolability throughout Sec. 4.

4. The Steering-Utility Trade-off

We identify a fundamental tension from Eq. (6): query-space steering controls model behaviour via *attention rerouting*, yet the same mechanism can disrupt attention on critical tokens, degrading performance on utility tasks. In this section, we first characterise the rerouting mechanism, then show empirically that rerouting concentrates on a small set of utility-critical tokens, and finally show that the naive remedy of enforcing full attention invariance suppresses steering entirely.

Attention rerouting via relative score changes. Softmax-based attention is invariant to row-wise constant shifts (i.e., $\mathrm{softmax}(\mathbf{s} + c\mathbf{1}) = \mathrm{softmax}(\mathbf{s})$) (Goodfellow et al., 2016). Therefore, query-space steering affects attention if and only if the induced logit shift $\delta_{ij}^{(\ell,h)}$ varies across key positions j. Specifically, for a fixed query position i, the change in attention assigned to key position j is governed by its logit shift relative to other key position j' in the same attention row, $\delta_{ij}^{(\ell,h)} - \delta_{ij'}^{(\ell,h)}$, which expands to:

$$\delta_{ij}^{(\ell,h)} - \delta_{ij'}^{(\ell,h)} = \frac{\lambda}{\sqrt{d'}}\langle \mathbf{r}_q^{(\ell,h)}, \mathbf{k}_j^{(\ell,h)} - \mathbf{k}_{j'}^{(\ell,h)}\rangle. \tag{7}$$

Thus, steering reroutes attention by changing how queries align with key differences. Eq. (7) also indicates that any steering vector not orthogonal to the relevant key-difference directions produces some rerouting; the question is whether the resulting rerouting is benign or whether it disrupts attention to tokens on which the model relies.

Utility degradation from focus-to-tail rerouting. We next examine where steering-induced rerouting concentrates on utility data. To this end, we collect a sampled *utility calibration set* $\mathcal{D}_{\text{util}}$ from utility benchmarks spanning different

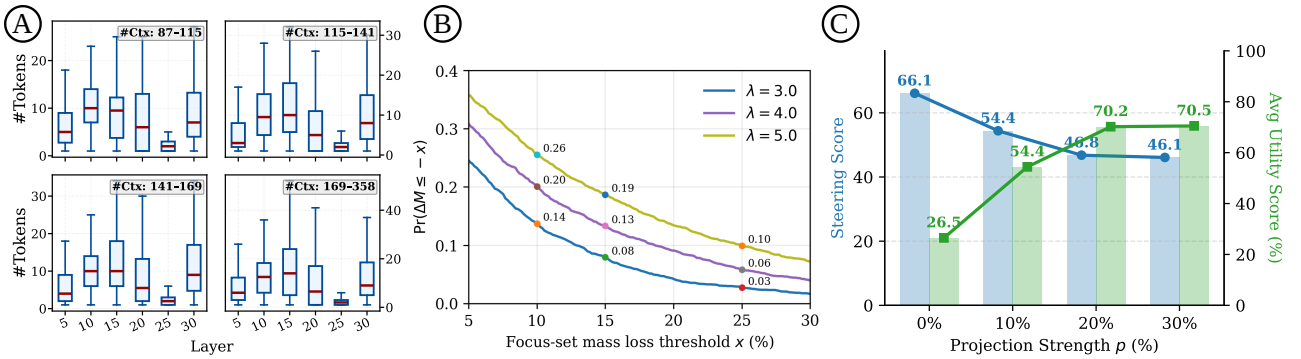

Figure 2. (A): **Focus sets are small and stable across context lengths.** We group evaluation samples by total context length and, for each group, report the per-head focus-set size $|\mathcal{H}^{(\ell,h)}|$ across layers, where $|\mathcal{H}^{(\ell,h)}|$ is the minimum number of tokens needed to cover $\tau_{\text{high}} = 0.8$ of the attention mass. Focus sets remain small (typically $\lesssim 15$ tokens) even as context length grows from $\sim$100 to $\sim$360 tokens. (B): **Focus-set mass drop under vanilla steering.** For varying thresholds x, we plot $\Pr(\Delta M \leq -x)$, where ΔM is the change in focus-set attention mass under steering (Eq. 8), aggregated across heads and decoding steps. Higher curves indicate more frequent focus-to-tail attention rerouting; the effect grows monotonically with steering strength λ. (C): **Steering-utility trade-off under direct key-orthogonal projection.** Steering score on TruthfulQA (Lin et al., 2022) (blue, left axis), and average utility across ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), and GSM8K (Cobbe et al., 2021) (green, right axis) versus projection strength p (cf. Eq. 10). Increasing p suppresses truthfulness while recovering utility, exposing a trade-off between attention-invariance and steerability.

domains, including maths, reasoning, and instruction following (see App. C.1 for details). Across layers and heads on $\mathcal{D}_{\text{util}}$, attention distributions are highly sparse: for an average context length of approximately 250 tokens, fewer than 30 tokens account for 80% of the total attention mass (Fig. 2 (A)). For a given layer ℓ and head h, let $\mathcal{H}^{(\ell,h)} \subset [t]$ denote the indices of high-attention tokens (the *focus set*) that collectively receive a fraction $\tau_{\text{high}} \in [0, 1]$ of the attention mass on utility data. Given $\mathcal{H}^{(\ell,h)}$, let $\mathcal{L}^{(\ell,h)}$ denote the remaining low-attention (*tail*) tokens. We quantify the effect of steering on utility-critical attention via the *top-set mass preservation*:

$$\Delta M = \sum_{j \in \mathcal{H}^{(\ell,h)}} \alpha_{ij}^{\text{steered}} - \sum_{j \in \mathcal{H}^{(\ell,h)}} \alpha_{ij}^{\text{base}}, \tag{8}$$

where $\alpha_{ij}^{\text{base}}$ and $\alpha_{ij}^{\text{steered}}$ denote attention weights before and after steering. A negative ΔM indicates that attention mass is shifted away from focus tokens toward tail tokens. Fig. 2 (B) reports the probability that steering reduces focus-set attention mass by at least a fraction x (i.e., $\Pr(\Delta M \leq -x)$), aggregated across heads on the utility dataset. We find that vanilla steering frequently induces large negative values of ΔM, and that the severity of this focus-to-tail rerouting increases monotonically with steering strength λ. This suggests that focus-to-tail rerouting may be responsible for utility degradation.

Full invariance yields a trade-off. A naive remedy for rerouting may be to ensure that $\mathbf{r}_q^{(\ell,h)}$ produces no relative score changes across key positions on $\mathcal{D}_{\text{util}}$. Let $\bar{\mathbf{k}}^{(\ell,h)} = \sum_{j=1}^{t} \mathbf{k}_j^{(\ell,h)}/t$ and $\mathbf{K}_c^{(\ell,h)} = \mathbf{K}^{(\ell,h)} - \mathbf{1}_t(\bar{\mathbf{k}}^{(\ell,h)})^\top$ be centred attention keys collected from $\mathcal{D}_{\text{util}}$. From Eq. (7), attention is invariant under query-space steering iff $\langle \mathbf{r}_q^{(\ell,h)}, \mathbf{k}_j^{(\ell,h)} - \bar{\mathbf{k}}^{(\ell,h)} \rangle = 0$ for all j, or, compactly, iff:

$$(\mathbf{r}_q^{(\ell,h)})^\top \mathbf{K}_c^{(\ell,h)} = \mathbf{0}^\top. \tag{9}$$

In other words, Eq. (9) requires $\mathbf{r}_q^{(\ell,h)}$ to be orthogonal to the column space of $\mathbf{K}_c^{(\ell,h)}$ (see App. B.2 for a proof that this condition is necessary and sufficient for attention invariance).

Empirically, we find that the centred key covariance $\mathbf{\Sigma}_k^{(\ell,h)}$ is low-rank for each head (see Fig. 10 in App. D.4), hence orthogonality may be enforced via a projector. Let $\mathbf{U}_k^{(\ell,h)} \in \mathbb{R}^{d' \times p}$ contain the top-p eigenvectors of $\mathbf{\Sigma}_k^{(\ell,h)}$. We define the orthogonal projector for the query-space steering vector by:

$$\mathbf{P}_k^{(\ell,h)} = \mathbf{I}_{d'} - \mathbf{U}_k^{(\ell,h)} (\mathbf{U}_k^{(\ell,h)})^\top, \tag{10}$$

and refer to this projection as the *key-invariant projection*. As shown in Fig. 2 (C), retaining the top-p eigenvectors that account for as little as 20% of the cumulative variance of $\mathbf{\Sigma}_k^{(\ell,h)}$ is sufficient for $\mathbf{P}_k^{(\ell,h)} \mathbf{r}_q^{(\ell,h)}$ to approximately satisfy Eq. (9), effectively eliminating steering effects while largely preserving performance on utility tasks. Nevertheless, and importantly, this reveals the fundamental tension: attention rerouting is both necessary for steering efficacy *and* the source of utility degradation, so any uniform attempt to suppress it eliminates steering altogether.

This tension motivates our approach: rather than enforcing full attention invariance, we constrain rerouting selectively, preserving relative scores between focus and tail tokens while allowing steering to redistribute attention within less utility-critical regions.

5. Steering via Key-Orthogonal Projections

Building on the tension identified in Sec. 4, we now introduce *Steering via Key-Orthogonal Projections* (SKOP, Fig. 3). In contrast to the key-invariant projection of Eq. (10), which suppresses *all* attention rerouting, SKOP targets only the rerouting that shifts attention from focus tokens to tail tokens, enabling other rerouting that may carry useful steering signals. Given a fixed set of steering vectors, SKOP proceeds in three stages: (1) characterising *key-difference* directions associated with focus-to-tail attention rerouting (Sec. 5.1); (2) projecting steering vectors to remove components that strongly affect these directions (Sec. 5.2); and (3) selectively applying this projection to heads most prone to attention rerouting (Sec. 5.3). We summarise SKOP in an algorithm form in App. E.1.

5.1. Characterising Utility-Critical Key Differences

In its first stage, SKOP uses the utility calibration set $\mathcal{D}_{\text{util}}$ from Sec. 4. Specifically, for each head (ℓ, h) and decoding step t, we (1) compute the baseline attention distribution $\boldsymbol{\alpha}^{(\ell,h)}$, and (2) construct the per-step focus set $\mathcal{H}_t^{(\ell,h)}$, the minimal token set capturing at least τ_{high} attention mass, and take tail set $\mathcal{L}_t^{(\ell,h)}$ as its complement. Aggregating over the calibration sample yields head-level focus and tail sets $\mathcal{H}^{(\ell,h)}$ and $\mathcal{L}^{(\ell,h)}$.

From Eq. (7), focus-to-tail rerouting is driven by the alignment between $\mathbf{r}_q^{(\ell,h)}$ and *key-difference vectors* $\Delta\mathbf{k}_{ij}^{(\ell,h)} = \mathbf{k}_i^{(\ell,h)} - \mathbf{k}_j^{(\ell,h)}$ for $i \in \mathcal{H}^{(\ell,h)}$ and $j \in \mathcal{L}^{(\ell,h)}$. We summarise these vectors over the calibration set using their second-moment matrix:

$$\boldsymbol{\Sigma}_{\Delta k}^{(\ell,h)} = \mathbb{E}_{t,\,(i,j)}\left[\Delta\mathbf{k}_{ij}^{(\ell,h)}(\Delta\mathbf{k}_{ij}^{(\ell,h)})^\top\right], \tag{11}$$

where $\Delta\mathbf{k}_{ij}^{(\ell,h)} = \mathbf{k}_i^{(\ell,h)} - \mathbf{k}_j^{(\ell,h)}$, and the expectation is over calibration steps t and uniform sampling of (i, j) from $\mathcal{H}_t^{(\ell,h)} \times \mathcal{L}_t^{(\ell,h)}$. Empirically, we use the second moment rather than the centred covariance, as the mean key-difference is itself a high-energy direction that centring would discard. Moreover, we note that, although $\boldsymbol{\Sigma}_{\Delta k}^{(\ell,h)}$ is estimated from $\mathcal{D}_{\text{util}}$, we show in App. F.2 that SKOP is robust to both the size and domain of the calibration set: as few as 250 examples already substantially recover utility over vanilla steering, and single-domain calibration sets all yield comparable trade-offs to the mixed default. This indicates that the focus-set structure exploited by SKOP reflects stable model-internal key-space geometry, rather than properties of the calibration distribution.

5.2. Projection onto the Key-Difference Subspace

In its second stage, SKOP uses $\boldsymbol{\Sigma}_{\Delta k}^{(\ell,h)}$ from Sec. 5.1 to project each steering vector onto the subspace least coupled to focus-to-tail rerouting. Concretely, the expected squared perturbation on the focus-to-tail score gap under query-space steering $\lambda\mathbf{r}_q^{(\ell,h)}$ is:

$$\mathbb{E}[(\Delta g_{tij})^2] = \frac{\lambda^2}{d'}(\mathbf{r}_q^{(\ell,h)})^\top \boldsymbol{\Sigma}_{\Delta k}^{(\ell,h)} \mathbf{r}_q^{(\ell,h)}, \tag{12}$$

where $\Delta g_{tij} = \delta_{ti}^{(\ell,h)} - \delta_{tj}^{(\ell,h)}$ is the change, at decoding step t, in the score gap between a focus token i and a tail token j (cf. Eq. (7)). To minimise harmful attention rerouting, we remove the components of $\mathbf{r}_q^{(\ell,h)}$ that contribute most strongly to Eq. (12). Mirroring the construction of the key-invariant projector in Eq. (10), but replacing $\boldsymbol{\Sigma}_k^{(\ell,h)}$ with $\boldsymbol{\Sigma}_{\Delta k}^{(\ell,h)}$, we define the SKOP projector as:

$$\mathbf{P}_{\Delta k}^{(\ell,h)} = \mathbf{I}_{d'} - \mathbf{U}_{\Delta k}^{(\ell,h)}(\mathbf{U}_{\Delta k}^{(\ell,h)})^\top, \tag{13}$$

where $\mathbf{U}_{\Delta k}^{(\ell,h)} \in \mathbb{R}^{d'\times p}$ contains the top-p eigenvectors of $\boldsymbol{\Sigma}_{\Delta k}^{(\ell,h)}$. We then replace the steering vector with its projected version:

$$\tilde{\mathbf{r}}_q^{(\ell,h)} = \mathbf{P}_{\Delta k}^{(\ell,h)}\mathbf{r}_q^{(\ell,h)}, \qquad \mathbf{q}_i^{(\ell,h)} \leftarrow \mathbf{q}_i^{(\ell,h)} + \lambda\tilde{\mathbf{r}}_q^{(\ell,h)}. \tag{14}$$

We find that $\boldsymbol{\Sigma}_{\Delta k}^{(\ell,h)}$ is also low-rank for each head (Fig. 11 in App. D.4). That is, a small number of eigenvectors account for most of the energy and therefore most of the focus-to-tail rerouting. We exploit this structure by selecting p to retain a fixed fraction of the total energy:

$$\left(\sum_{i=1}^{p}\lambda_i\left(\boldsymbol{\Sigma}_{\Delta k}^{(\ell,h)}\right)\right) \Big/ \left(\sum_{i=1}^{d'}\lambda_i\left(\boldsymbol{\Sigma}_{\Delta k}^{(\ell,h)}\right)\right) \geq \gamma_{\text{energy}}, \tag{15}$$

where $\lambda_i(\cdot)$ denotes the i-th eigenvalue of $\boldsymbol{\Sigma}_{\Delta k}^{(\ell,h)}$ in descending order. This selects the smallest p that captures a fraction γ_{energy} of the rerouting energy, retaining the dominant directions that drive harmful rerouting, while leaving the remaining directions untouched and available for steering.

Two properties of this construction together imply that, despite Eq. (13) being a hard projection, SKOP preserves most of the steering capacity. First, the trade-off is insensitive to γ_{energy}: both steering and utility remain stable across $\gamma_{\text{energy}} \in [0.7, 0.95]$ (see App. F.3), so we fix $\gamma_{\text{energy}} = 0.9$ across all tasks and models. Second, projection preserves most of the norm of $\mathbf{r}_q^{(\ell,h)}$ across heads (see App. D.5). This suggests that harmful rerouting is concentrated in a small number of dominant eigendirections, leaving much of the orthogonal complement available for steering.

5.3. Selective Application to High-Risk Heads

Empirically, focus-to-tail rerouting is concentrated in a small minority of heads (Fig. 8), so projecting all heads uniformly would unnecessarily suppress steering capacity

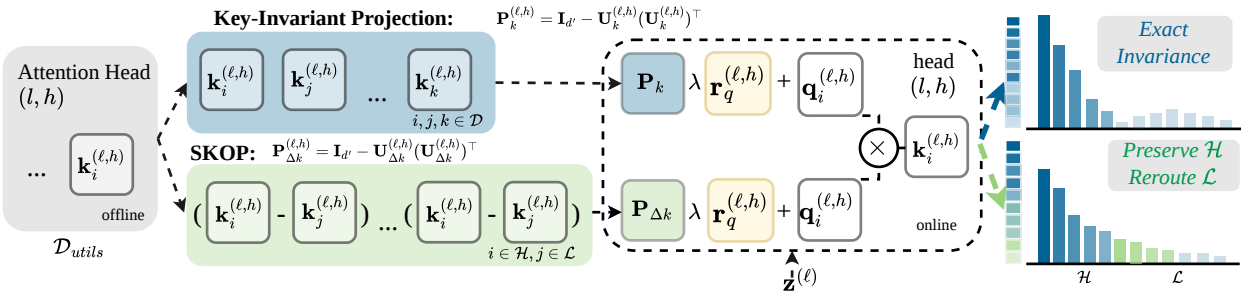

Figure 3. Steering via Key-Orthogonal Projection (SKOP) preserves attention on focus tokens while steering model behaviour. It identifies, for each attention head, key-space directions that mediate focus-to-tail attention rerouting on utility datasets, and applies orthogonal projection to stabilise query representations during generation.

in benign heads. Therefore, we quantify a head's susceptibility to harmful rerouting via the Rayleigh quotient:

$$R^{(\ell,h)} = \left((\mathbf{r}_q^{(\ell,h)})^\top \mathbf{\Sigma}_{\Delta k}^{(\ell,h)} \mathbf{r}_q^{(\ell,h)}\right) / \left(\|\mathbf{r}_q^{(\ell,h)}\|^2 + \epsilon\right), \quad (16)$$

which is exactly the per-unit-norm version of the expected squared score-gap perturbation in Eq. (12). Hence, we apply SKOP only to the top-k heads as ordered by $R^{(\ell,h)}$. We note that this risk-based ranking is different from the discriminative-head criteria used in prior work (Li et al., 2024; Wang et al., 2025; Torop et al., 2025): as shown in App. D.3, risk heads jointly drive stronger steering effects *and* larger utility drops than discriminative heads, making them the most useful targets for projection.

6. Experiments

In this section, we evaluate SKOP by studying the following research questions:

- **RQ1:** Does SKOP balance steering efficacy and utility better than previous approaches?
- **RQ2:** What is the impact of the steering strength λ when steering with SKOP?
- **RQ3:** Can SKOP maintain high steerability and utility in long-context tasks?

Setup. We conduct experiments on Llama-3.1-8B-Instruct (Dubey et al., 2024) and Gemma-2-9B-IT (Gemma Team et al., 2024). For steering evaluation, we use TruthfulQA (Lin et al., 2022) and three behaviours from the Model-Written Evaluation suite (Perez et al., 2023): power-seeking, wealth-seeking, and corrigibility. Utility is measured on IFBench (Pyatkin et al., 2025) (instruction-following), ARC-Challenge (Clark et al., 2018) (scientific reasoning), HellaSwag (Zellers et al., 2019) (commonsense), and GSM8K (Cobbe et al., 2021) (mathematical reasoning). We discuss all dataset details in App. C.2.

Baselines. We compare SKOP against three families of steering methods. From the *residual-stream steering* methods, we include CAA (Rimsky et al., 2024), which adds mean-difference vectors directly to the residual stream and serves as a canonical residual-space baseline. From the *attention-space steering* methods, we include DISCO-Q (Torop et al., 2025), as an example of query-space steering, Comm Steer (Torop et al., 2025) and Angular Steer (Vu & Nguyen, 2025), as examples of attention-input steering, and ITI (Li et al., 2024), as an example of head-output steering. We use mean-difference steering vectors for all attention-space methods. From the *conditional steering* methods, we include CAST (Lee et al., 2025), which conditionally applies mean-difference steering vectors, and SADI (Wang et al., 2025), which performs semantics-based modulation by dynamically constructing steering vectors. Finally, we include LoRA (Hu et al., 2022) as an example of parameter-efficient finetuning approaches to steering. For fair comparison, we apply steering vectors to all layers and all heads when the baseline operates on attention heads. We provide additional details in App. E.2.

Hyperparameters. SKOP has three hyperparameters: the focus-mass threshold τ_{high} (Sec. 4), the energy-coverage threshold γ_{energy} (Eq. (15)), and the fraction of top-risk heads to project. We choose these hyperparameters using sensitivity analyses (see App. D.3 and App. F.3), and fix them to $\tau_{\text{high}} = 0.8$, $\gamma_{\text{energy}} = 0.9$, and 20% of heads across all tasks and models.

6.1. Steering-Utility Trade-off Evaluation (RQ1)

We evaluate steering effectiveness in open-ended generation tasks using an LLM judge to score model outputs. For behaviours from the Model-Written Evaluation suite, following prior work (Rimsky et al., 2024; Wu et al., 2025; Torop et al., 2025), we prompt the LLM judge to assign a score (1–4) to each generation indicating how strongly the response exhibits the target behaviour. For TruthfulQA, we score outputs using the *True*Info (T*I)* metric (Lin et al.,

Table 1. Comparison of SKOP against baselines for LLaMA-3.1-8B-Instruct (Dubey et al., 2024). Steering performance is evaluated using an LLM Judge for Power, Wealth, and Corr and TruthfulQA *True*Info* Metric (TQA) (higher is better). Utility is measured via IFBench (IFB), ARC-Challenge (ARC), HellaSwag (HS), and GSM8K accuracy. Rank is the average rank across all steering and utility benchmarks (lower is better). Best results are **bolded**, and second-best results are underlined.

	Steering				Utility				
Method	Power	Wealth	Corr	TQA	IFB	ARC	HS	GSM8K	Rank ↓
Baseline	1.83	1.71	1.94	46.1	26.2	66.3	70.5	75.2	–
LoRA (Hu et al., 2022)	2.31	1.89	2.68	55.4	16.8	38.5	52.6	29.0	–
CAA (Rimsky et al., 2024)	2.49	2.10	2.79	76.8	14.5	25.0	27.5	14.0	5.44
ITI (Li et al., 2024)	2.59	2.14	2.60	66.8	10.5	29.1	45.2	20.6	4.75
DISCO-Q (Torop et al., 2025)	2.55	2.06	**3.22**	66.1	11.5	34.2	38.3	22.5	4.75
Comm Steer (Torop et al., 2025)	**2.91**	**2.25**	3.01	**81.6**	6.0	15.8	35.5	10.2	4.63
Angular Steer (Vu & Nguyen, 2025)	2.13	2.04	2.18	56.9	19.5	59.2	59.7	60.5	4.75
CAST (Lee et al., 2025)	2.04	1.92	2.28	58.2	18.2	40.5	55.0	41.8	5.63
SADI (Wang et al., 2025)	2.58	2.21	2.49	75.9	15.5	48.8	58.8	58.6	3.38
SKOP (Ours)	2.51	2.10	3.19	65.9	**25.0**	**65.0**	**65.2**	**66.6**	**2.69**

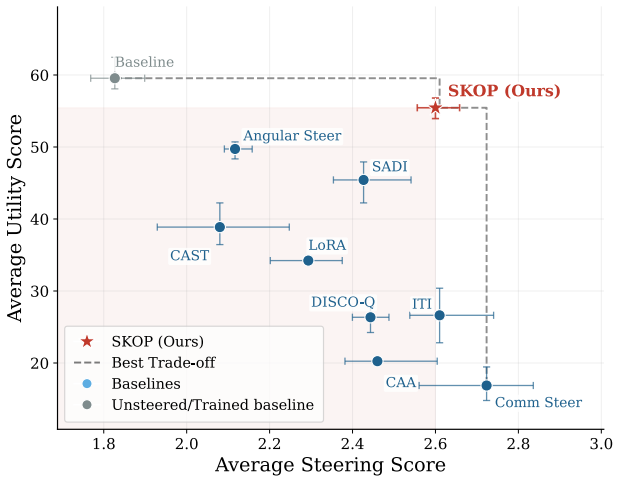

Figure 4. Steering-utility trade-off for LLaMA3.1-8B-Instruct. We report the average of Power, Wealth, and Corr (Perez et al., 2023). The dashed line traces the best trade-off frontier. SKOP achieves the best joint trade-off among all steered methods.

2022; Evans et al., 2025). We use the same LLM judge prompts as Torop et al. (2025). For utility benchmarks, we report instruction-level accuracy under strict matching for IFBench (Pyatkin et al., 2025), and standard accuracy for ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), and GSM8K (Cobbe et al., 2021). For ARC, we use the challenge subset. Each utility score is averaged across all steering tasks. We select λ via a sweep maximising performance on each steering task and reuse it on utility benchmarks. We provide LLaMA 3.1 results in Table 1 and Fig. 4, and Gemma results, showing similar trends, in App. F.1.

Fig. 4 shows that SKOP achieves the best steering-utility trade-off and is the only steering method without a large utility loss. Vanilla mean-difference baselines and the residual-stream baseline attain marginally higher absolute steering scores, but crucially this comes with a severe utility cost (50-75% degradation). SKOP retains over 95% of vanilla query-space steering efficacy while reducing utility degradation to under 10%, outperforming conditional steering baselines (CAST and SADI). The same is observed on Gemma, where SKOP attains the best overall trade-off rank across all steering methods (see App. F.1). Moreover, as discussed in App. F.4, SKOP also outperforms common fine-tuning approaches: SKOP matches vanilla query-space steering efficacy and substantially exceeds LoRA on steering across all training set sizes, while LoRA's utility degrades monotonically as data grows.

6.2. The Impact of Steering Strength (RQ2)

Fig. 5 shows the effect of varying steering strength λ across four steering tasks, comparing vanilla query steering with SKOP. We select λ ranges that span from weak to strong steering effects until performance saturates. Across all tasks, increasing λ improves steering strength for vanilla steering, but at the cost of rapid and monotonic utility degradation. For Wealth and Corrigibility, average utility drops to near zero at high λ values, indicating that attention rerouting severely disrupts the model's focus on tokens that carry critical contextual information. In contrast, SKOP consistently moderates this trade-off: while projection slightly reduces the maximum achievable steering score, it substantially stabilises utility across the entire range of steering strengths.

To further understand this improvement, we examine focus-set attention mass preservation based on our attention-rerouting hypothesis. As detailed in App. D.2, we find that vanilla query-space steering induces large negative shifts in ΔM at $\lambda = 4.0$: 31% of (head, decoding step) pairs lose at least 10% of their focus-set attention mass, and 22% lose at least 15%, indicating substantially diminished focus on important tokens. SKOP reduces these tail probabilities by

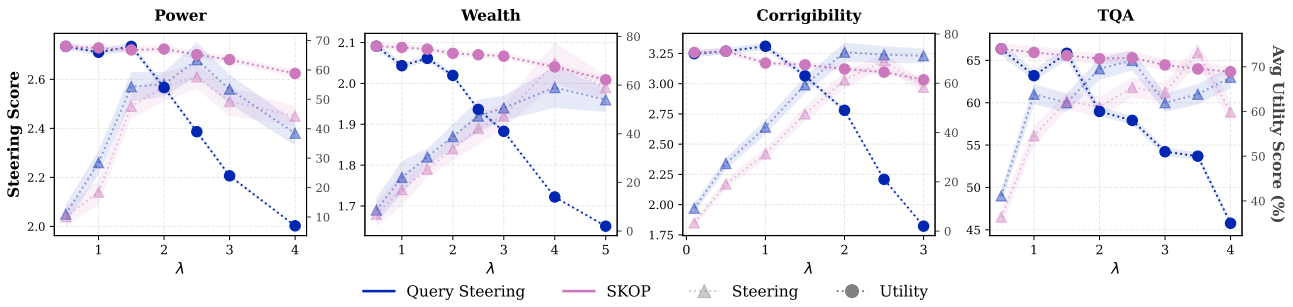

Figure 5. Effect of varying steering strength λ on steering efficacy and utility preservation for SKOP on LLaMA-3.1-8B-Instruct. As λ increases, vanilla query steering vectors achieve slightly higher steering scores but suffer severe utility degradation. In contrast, SKOP maintains strong utility preservation across all λ while preserving most steering effectiveness.

Table 2. Long-context steering on RULER NIAH. We steer LLaMA-3.1-8B-Instruct with a formatting instruction. We report formatting compliance for responses and NIAH retrieval accuracy across context lengths.

		Context length				
Metric	Method	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**

roughly 3-10× across loss thresholds. This confirms that SKOP preserves the base model's high-confidence attention patterns, preventing the harmful attention rerouting.

6.3. Long-Context Robustness (RQ3)

Given our motivation to reduce harmful *attention rerouting* on utility tasks, we investigate whether SKOP preserves model capability on *long-context* tasks. Specifically, we evaluate SKOP on "needle-in-a-haystack" (NIAH) tasks from the RULER benchmark (Hsieh et al., 2024), where correct behaviour crucially depends on maintaining sparse but high attention mass on important tokens among distractor tokens. NIAH tasks construct a long "haystack" (either repeated sentences or natural text (Kamradt, 2023)) containing one or more inserted "needles", with a query at the end that cues retrieval by matching "needles" in context and outputting the associated values.

To steer behaviour over long contexts, we use a simpler yet informative steering task, namely *formatting steering*. Inspired by prior work (Stolfo et al., 2025), we construct an instruction steering vector by taking the mean difference between activations with and without an instruction. We use a "Quotation" task with formatting instructions (*"Wrap your entire response with double quotation marks."*) in positive contexts and no instructions in negative contexts. We test LLaMA-3.1-8B-Instruct across context lengths from 1K to 16K tokens. We apply only unconditional steering as we found that conditional methods have minimal effect on this task.

Table 2 shows that the unsteered model attains near-perfect retrieval accuracy across all tested context lengths. However, vanilla steering vectors exhibit significant sensitivity to length. While they enforce quotation formatting in shorter contexts (within 2K tokens), steering efficacy degrades as context increases: at 16K tokens, DISCO-Q's formatting compliance drops by over half. For retrieval accuracy, vanilla steering performance drops sharply beyond 4K tokens. In contrast, SKOP maintains stronger formatting compliance while preserving NIAH retrieval performance at 8K–16K tokens, narrowing the gap to the unsteered baseline. This suggests that, even in long contexts, SKOP suppresses detrimental focus-to-tail attention rerouting.

7. Discussion and Conclusion

Limitations. SKOP currently targets query-space mean-difference steering, where attention rerouting is isolable; App. A.2 summarises its assumptions and failure modes. Residual interventions also alter keys, values, and MLP activations, requiring a broader analysis. SKOP further uses a utility calibration set and slightly reduces steering efficacy. Future work could adapt focus sets online or learn focus-preserving steering vectors.

Conclusion. We identify *attention rerouting* as a key contributor to the query-steering utility trade-off and propose SKOP, which removes steering components that divert attention from focus tokens. SKOP achieves the strongest steering–utility trade-off across benchmarks and remains effective in long contexts where prior methods degrade, broadening the practical scope of activation steering.

References

Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebron, F., and Sanghai, S. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4895–4901, 2023.

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction. In *NeurIPS*, 2024.

Bayat, R., Rahimi-Kalahroudi, A., Pezeshki, M., Chandar, S., and Vincent, P. Steering large language model activations in sparse spaces. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=VGw1viYliK.

Bisk, Y., Zellers, R., Bras, R. L., Gao, J., and Choi, Y. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

Cao, Y., Zhang, T., Cao, B., Yin, Z., Lin, L., Ma, F., and Chen, J. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=7qJFkuZdYo.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Rozière, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I. M., Misra, I., Evtimov, I., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024.

Dubois, Y., Galambosi, B., Liang, P., and Hashimoto, T. B. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *CoRR*, abs/2404.04475, 2024.

Evans, O., Chua, J., and Lin, S. New, improved multiple-choice truthfulqa, 2025. URL https://www.alignmentforum.org/posts/Bunfwz6JsNd44kgLT/new-improved-multiple-choice-truthfulqa.

Gemma Team, Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. L., Jerome, S., Tsitsulin, A., Vieillard, N., Stanczyk, P., Girgin, S., Momchev, N., Hoffman, M., Thakoor, S., Grill, J.-B., Neyshabur, B., Bachem, O., Walton, A., Severyn, A., Parrish, A., Ahmad, A., Hutchison, A., Abdagic, A., Carl, A., Shen, A., Brock, A., Coenen, A., Laforge, A., Paterson, A., Bastian, B., Piot, B., Wu, B., Royal, B., Chen, C., Kumar, C., Perry, C., Welty, C., Choquette-Choo, C. A., Sinopalnikov, D., Weinberger, D., Vijaykumar, D., Rogozińska, D., Herbison, D., Bandy, E., Wang, E., Noland, E., Moreira, E., Senter, E., Eltyshev, E., Visin, F., Rasskin, G., Wei, G., Cameron, G., Martins, G., Hashemi, H., Klimczak-Plucińska, H., Batra, H., Dhand, H., Nardini, I., Mein, J., Zhou, J., Svensson, J., Stanway, J., Chan, J., Zhou, J. P., Carrasqueira, J., Iljazi, J., Becker, J., Fernandez, J., van Amersfoort, J., Gordon, J., Lipschultz, J., Newlan, J., yeong Ji, J., Mohamed, K., Badola, K., Black, K., Millican, K., McDonell, K., Nguyen, K., Sodhia, K., Greene, K., Sjoesund, L. L., Usui, L., Sifre, L., Heuermann, L., Lago, L., McNealus, L., Soares, L. B., Kilpatrick, L., Dixon, L., Martins, L., Reid, M., Singh, M., Iverson, M., Görner, M., Velloso, M., Wirth, M., Davidow, M., Miller, M., Rahtz, M., Watson, M., Risdal, M., Kazemi, M., Moynihan, M., Zhang, M., Kahng, M., Park, M., Rahman, M., Khatwani, M., Dao, N., Bardoliwalla, N., Devanathan, N., Dumai, N., Chauhan, N., Wahltinez, O., Botarda, P., Barnes, P., Barham, P., Michel, P., Jin, P., Georgiev, P., Culliton, P., Kuppala, P., Comanescu, R., Merhej, R., Jana, R., Rokni, R. A., Agarwal, R., Mullins, R., Saadat, S., Carthy, S. M., Cogan, S., Perrin, S., Arnold, S. M. R., Krause, S., Dai, S., Garg, S., Sheth, S., Ronstrom, S., Chan, S., Jordan, T., Yu, T., Eccles, T., Hennigan, T., Kocisky, T., Doshi, T., Jain,

V., Yadav, V., Meshram, V., Dharmadhikari, V., Barkley, W., Wei, W., Ye, W., Han, W., Kwon, W., Xu, X., Shen, Z., Gong, Z., Wei, Z., Cotruta, V., Kirk, P., Rao, A., Giang, M., Peran, L., Warkentin, T., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Sculley, D., Banks, J., Dragan, A., Petrov, S., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Borgeaud, S., Fiedel, N., Joulin, A., Kenealy, K., Dadashi, R., and Andreev, A. Gemma 2: Improving open language models at a practical size. *arXiv:2408.00118*, 2024. URL https://arxiv.org/abs/2408.00118.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.

Hsieh, C.-P., Sun, S., Kriman, S., Acharya, S., Rekesh, D., Jia, F., and Ginsburg, B. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kIoBbc76Sy.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022*, Virtual Event, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Kamradt, G. Needle in a haystack - pressure testing LLMs. https://github.com/gkamradt/LLMTest_NeedleInAHaystack/tree/main, 2023.

Kočiský, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., and Grefenstette, E. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. doi:10.1162/tacl_a_00023. URL https://aclanthology.org/Q18-1023/.

Lee, B. W., Padhi, I., Ramamurthy, K. N., Miehling, E., Dognin, P., Nagireddy, M., and Dhurandhar, A. Programming refusal with conditional activation steering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Oi47wc10sm.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36, 2024.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022.

Luo, H. and Specia, L. From understanding to utilization: A survey on explainability for large language models, 2024. URL https://arxiv.org/abs/2401.12874.

Mikolov, T., Yih, W.-t., and Zweig, G. Linguistic regularities in continuous space word representations. In Vanderwende, L., Daumé III, H., and Kirchhoff, K. (eds.), *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/.

Nguyen, D., Prasad, A., Stengel-Eskin, E., and Bansal, M. Multi-attribute steering of language models via targeted intervention. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20619–20634, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi:10.18653/v1/2025.acl-long.1007. URL https://aclanthology.org/2025.acl-long.1007/.

O'Brien, K., Majercak, D., Fernandes, X., Edgar, R., Chen, J., Nori, H., Carignan, D., Horvitz, E., and Poursabzi-Sangde, F. Steering language model refusal with sparse autoencoders. *arXiv:2411.11296*, 2024. URL https://arxiv.org/abs/2411.11296.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. In *ICML*, 2024.

Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13387–13434, 2023.

Pyatkin, V., Malik, S., Graf, V., Ivison, H., Huang, S., Dasigi, P., Lambert, N., and Hajishirzi, H. Generalizing verifiable instruction following. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=yfYgwjj5F8.

Qiu, Y., Zhao, Z., Ziser, Y., Korhonen, A., Ponti, E. M., and Cohen, S. Spectral editing of activations for large language model alignment. *Advances in Neural Information Processing Systems*, 37:56958–56987, 2024.

Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramar, J., Shah, R., and Nanda, N. Improving

sparse decomposition of language model activations with gated sparse autoencoders. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024. URL https://openreview.net/forum?id=Ppj5KvzU8Q.

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. Steering llama 2 via contrastive activation addition. In Ku, L.-W., Martins, A., and Srikumar, V. (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/.

Rivière, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. L., Jerome, S., Tsitsulin, A., Vieillard, N., Stanczyk, P., Girgin, S., Momchev, N., Hoffman, M., Thakoor, S., Grill, J., Neyshabur, B., Bachem, O., Walton, A., Severyn, A., Parrish, A., Ahmad, A., Hutchison, A., Abdagic, A., Carl, A., Shen, A., Brock, A., Coenen, A., Laforge, A., Paterson, A., Bastian, B., Piot, B., Wu, B., Royal, B., Chen, C., Kumar, C., Perry, C., Welty, C., Choquette-Choo, C. A., Sinopalnikov, D., Weinberger, D., Vijaykumar, D., Rogozinska, D., Herbison, D., Bandy, E., Wang, E., Noland, E., Moreira, E., Senter, E., Eltyshev, E., Visin, F., Rasskin, G., Wei, G., Cameron, G., Martins, G., Hashemi, H., Klimczak-Plucinska, H., Batra, H., Dhand, H., Nardini, I., Mein, J., Zhou, J., Svensson, J., Stanway, J., Chan, J., Zhou, J. P., Carrasqueira, J., Iljazi, J., Becker, J., Fernandez, J., van Amersfoort, J., Gordon, J., Lipschultz, J., Newlan, J., Ji, J., Mohamed, K., Badola, K., Black, K., Millican, K., McDonell, K., Nguyen, K., Sodhia, K., Greene, K., Sjösund, L. L., Usui, L., Sifre, L., Heuermann, L., Lago, L., and McNealus, L. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118, 2024.

Rodriguez, P., Blaas, A., Klein, M., Zappella, L., Apostoloff, N., marco cuturi, and Suau, X. Controlling language and diffusion models by transporting activations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=l2zFn6TIQi.

Sheng, L., Shen, C., Zhao, W., Fang, J., Liu, X., Liang, Z., Wang, X., Zhang, A., and Chua, T.-S. Alphasteer: Learning refusal steering with principled null-space constraint, 2025. URL https://arxiv.org/abs/2506.07022.

Singh, S., Ravfogel, S., Herzig, J., Aharoni, R., Cotterell, R., and Kumaraguru, P. Representation surgery: Theory and practice of affine steering. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45663–45680. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/singh24d.html.

Stolfo, A., Balachandran, V., Yousefi, S., Horvitz, E., and Nushi, B. Improving instruction-following in language models through activation steering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=wozhdnRCtw.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Torop, M., Masoomi, A., Eskandar, M., and Dy, J. DISCO: Disentangled communication steering for large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=c8AjdgdHnD.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering. *arXiv:2308.10248*, 2024. URL https://arxiv.org/abs/2308.10248.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vu, H. M. and Nguyen, T. M. Angular steering: Behavior control via rotation in activation space. In *2nd Workshop on Models of Human Feedback for AI Alignment*, 2025. URL https://openreview.net/forum?id=GU2UeVZrSw.

Wang, W., YANG, J., and Peng, W. Semantics-adaptive activation intervention for LLMs via dynamic steering vectors. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8WQ7VTfPTl.

Wu, Z., Arora, A., Wang, Z., Geiger, A., Jurafsky, D., Manning, C. D., and Potts, C. ReFT: Representation finetuning for language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=fykjplMc0V.

Wu, Z., Arora, A., Geiger, A., Wang, Z., Huang, J., Jurafsky, D., Manning, C. D., and Potts, C. Axbench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=K2CckZjNy0.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. Representation engineering: A top-down approach to AI transparency. *CoRR*, abs/2310.01405, 2023.

A. Notation and Design Assumptions

This appendix provides a lookup table for the notation in Secs. 3–5 and makes explicit the simplifying assumptions that shape SKOP.

A.1. Notation

Table 3 groups symbols by their role in the formulation. Superscripts (ℓ, h) are omitted from some explanations when they only identify a layer and attention head.

For grouped-query attention, several query heads share one key–value head. The formulation is unchanged after associating each query head h with its corresponding shared key and value projections; SKOP still estimates and applies a projector for each query head.

Table 3. Notation used in Secs. 3–5.

Symbol	Meaning	Shape or range
Indices and dimensions		
L, H	Numbers of transformer layers and attention heads.	Positive integers
ℓ, h	Layer and attention-head indices.	$\ell \in [L]$, $h \in [H]$
i, j, j'	In Secs. 3–4, query position i and key positions j, j'; in Sec. 5, focus-token index i and tail-token index j.	$i, j, j' \in [t]$
t	Current sequence length; equivalently, the current decoding step when constructing per-step sets.	Positive integer
d, d'	Residual-stream and per-head dimensions, with $d' = d/H$ under standard multi-head attention.	Positive integers
Transformer attention and query steering		
$\mathbf{h}^{(\ell)}, \mathbf{g}^{(\ell)}$	Residual-stream states and intermediate post-attention states at layer ℓ.	$\mathbb{R}^{t\times d}$
$\mathbf{z}^{(\ell)}$	Layer-normalised attention input, $\mathrm{LN}(\mathbf{h}^{(\ell-1)})$.	$\mathbb{R}^{t\times d}$
$a^{(\ell)}, a^{(\ell,h)}$	Multi-head attention block and its head-h component.	Functions on token states
$\mathbf{W}_{q,k,v}^{(\ell,h)}, \mathbf{W}_o^{(\ell,h)}$	Query, key, and value projection matrices; output projection matrix.	$\mathbb{R}^{d\times d'}$; $\mathbb{R}^{d'\times d}$
$\mathbf{Q}^{(\ell,h)}, \mathbf{K}^{(\ell,h)}, \mathbf{V}^{(\ell,h)}$	Query, key, and value matrices; lowercase $\mathbf{q}_i, \mathbf{k}_j, \mathbf{v}_j$ denote their rows.	$\mathbb{R}^{t\times d'}$
$s_{ij}^{(\ell,h)}, \widetilde{s}_{ij}^{(\ell,h)}$	Base and steered attention logits between query i and key j.	Scalars
$\alpha_{ij}^{(\ell,h)}$	Attention weight from query i to key j, obtained by a row-wise softmax of s_{ij}.	$[0, 1]$
$\mathbf{r}, \mathbf{r}_q^{(\ell,h)}$	Generic steering vector and its query-space component for head (ℓ, h).	$\mathbf{r}_q^{(\ell,h)} \in \mathbb{R}^{d'}$
λ	Steering strength multiplying the steering vector.	$\mathbb{R}$
$\delta_{ij}^{(\ell,h)}$	Steering-induced additive perturbation to attention logit $s_{ij}^{(\ell,h)}$.	Scalar
Focus–tail sets and rerouting		
$\mathcal{D}_{\mathrm{util}}$	Utility calibration set used to estimate attention geometry offline.	Set of sequences
τ_{high}	Minimum attention mass covered by a focus set.	$[0, 1]$; default 0.8
$\mathcal{H}_t^{(\ell,h)}, \mathcal{L}_t^{(\ell,h)}$	Per-step focus set (the smallest set covering τ_{high} attention mass) and its tail-set complement.	Subsets of $[t]$
$\mathcal{H}^{(\ell,h)}, \mathcal{L}^{(\ell,h)}$	Head-level focus and tail sets aggregated over calibration steps.	Token-index sets
$\Delta M, x$	Change from base to steered focus-set attention mass (Eq. 8) and the loss threshold used in $\Pr(\Delta M \le -x)$.	$\Delta M \in [-1, 1]$; $x \in [0, 1]$
$\Delta\mathbf{k}_{ij}^{(\ell,h)}$	Focus–tail key difference, $\mathbf{k}_i^{(\ell,h)} - \mathbf{k}_j^{(\ell,h)}$.	$\mathbb{R}^{d'}$
Δg_{tij}	Steering-induced change, at decoding step t, in the score gap between focus token i and tail token j.	Scalar
$\mathbf{\Sigma}_{\Delta k}^{(\ell,h)}$	Second-moment matrix of focus–tail key differences over calibration steps and sampled token pairs.	$\mathbb{R}^{d'\times d'}$
Projection and head selection		
$\bar{\mathbf{k}}^{(\ell,h)}, \mathbf{K}_c^{(\ell,h)}$	Mean key and centred key matrix used to characterise full attention invariance.	$\mathbb{R}^{d'}$; $\mathbb{R}^{t\times d'}$
$\mathbf{\Sigma}_k^{(\ell,h)}$	Centred key covariance matrix.	$\mathbb{R}^{d'\times d'}$
$\mathbf{U}_k^{(\ell,h)}, \mathbf{U}_{\Delta k}^{(\ell,h)}$	Top-p eigenvectors of $\mathbf{\Sigma}_k$ and $\mathbf{\Sigma}_{\Delta k}$, respectively.	$\mathbb{R}^{d'\times p}$
$\mathbf{P}_k^{(\ell,h)}, \mathbf{P}_{\Delta k}^{(\ell,h)}$	Projectors onto the complements of the dominant key and focus–tail key-difference subspaces.	$\mathbb{R}^{d'\times d'}$
$\widetilde{\mathbf{r}}_q^{(\ell,h)}$	SKOP-projected query steering vector, $\mathbf{P}_{\Delta k}^{(\ell,h)}\mathbf{r}_q^{(\ell,h)}$.	$\mathbb{R}^{d'}$
$p, \lambda_i(\cdot)$	Selected projection rank and the i-th eigenvalue in descending order.	$p \in [d']$; scalar
γ_{energy}	Fraction of key-difference spectral energy covered by the selected eigenspace.	$[0, 1]$; default 0.9
$R^{(\ell,h)}$	Rayleigh-quotient risk score measuring rerouting susceptibility per unit steering-vector norm.	Non-negative scalar
k, ϵ	Number (or fraction) of top-risk heads projected and numerical stabiliser in the risk score.	$k > 0$; $\epsilon > 0$

A.2. Simplifying Assumptions and Scope

SKOP is scoped to fixed additive vectors in query space, for which the intervention contributes one isolable term to the attention logits (App. B). Residual-stream interventions affect queries, keys, values, and MLP activations simultaneously and therefore fall outside the present analysis. Within this scope, Table 4 summarises the simplifying assumptions that influence SKOP's design. Several are empirically motivated rather than guaranteed; the final column states where each may fail.

Table 4. Simplifying assumptions that shape SKOP. "Support" identifies the evidence in this paper rather than treating an empirical observation as a universal guarantee.

Assumption	Influence on SKOP	Empirical or analytical support	Limitation or failure mode
Rerouting contributes to utility loss. Harmful focus-to-tail attention shifts are a major contributor to the utility loss caused by query steering.	Motivates correcting relative query–key scores rather than changing when steering is activated.	The perturbation analysis isolates relative-score changes (Eq. 7); rerouting grows with steering strength (Fig. 2B), and SKOP jointly preserves focus mass and utility (App. D.2).	Rerouting need not be the only cause of utility loss. SKOP may not address failures caused by altered values, later-layer dynamics, or generation-level effects.
Attention mass proxies importance. Tokens receiving high baseline attention are treated as utility-critical.	Defines focus sets by cumulative attention mass and treats their complements as tail sets.	Focus sets are small across context lengths (Fig. 2A), while their mass drops under vanilla steering and is preserved by SKOP.	Attention weight is not a complete causal measure of token importance. A low-attention token may still be decisive, and a high-attention token may be redundant.
Second moments proxy harmful directions. Large squared focus–tail score-gap changes identify directions worth suppressing.	Summarises key differences with $\mathbf{\Sigma}_{\Delta k}$ and removes its dominant eigendirections (Eqs. 12–13).	The objective follows directly from the query-steering logit perturbation, and the resulting projection improves the observed steering–utility trade-off.	Squaring discards the sign of a score change, so the proxy can penalise some benign or beneficial rerouting; pairwise second moments also omit higher-order interactions.
Calibration geometry transfers. Offline attention statistics from $\mathcal{D}_{\mathrm{util}}$ remain relevant to unseen steering tasks and inputs.	Enables one fixed projector per model to be reused without online adaptation.	Single-domain calibration sets produce comparable trade-offs, and 250 examples already recover substantial utility (App. F.2).	Transfer may weaken under large domain, language, architecture, or context-length shifts; a mismatched calibration set may omit deployment-critical patterns.
Relevant geometry is approximately low-rank. A small eigenspace captures most harmful rerouting energy.	Makes a low-rank orthogonal projection and the energy-coverage rule practical.	Key and key-difference spectra decay rapidly across heads (App. D.4); performance is stable for $\gamma_{\mathrm{energy}} \in [0.7, 0.95]$ (App. F.3).	If the spectrum is diffuse, a large p would be required, removing more steering capacity and weakening the efficiency and selectivity of SKOP.
Head risk is sparse. A minority of heads account for most harmful rerouting.	Applies projection only to the top-risk heads rather than uniformly across the model.	Risk scores are long-tailed, and projecting roughly the top 10–20% of heads recovers most utility (App. D.3).	Distributed rerouting across many heads would make top-k selection less effective; independent head scores may also miss cross-head or cross-layer interactions.
Useful and harmful directions are separable. Enough steering signal remains outside the dominant rerouting subspace.	Permits hard projection while retaining behavioural control.	Projected vectors preserve most of their norm (App. D.5), and SKOP retains over 95% of vanilla steering efficacy (Sec. 6.1).	If a target behaviour relies on the same directions as utility-critical attention, projection must trade steering efficacy for utility and cannot cleanly preserve both.

B. Attention Invariance under Query Steering

In this appendix, we prove the claim made in Sec. 3: under query-space steering, the perturbation $\delta_{ij}^{(\ell,h)}$ in Eq. 6 is the unique component of the steering intervention that can affect attention weights.

We prove the result in a strictly more general setting that we call *attention-input steering*, which corresponds to the Comm Steer baseline of Vu & Nguyen (2025) in Sec. 6. Rather than steering the query and key projections independently, attention-input steering applies a single additive perturbation $\lambda \mathbf{r}^{(\ell,h)} \in \mathbb{R}^d$ to the layer-normalised attention input $\mathbf{z}_i^{(\ell)}$ before any projection. Because $\mathbf{q}_i^{(\ell,h)} = \mathbf{z}_i^{(\ell)} \mathbf{W}_q^{(\ell,h)}$ and $\mathbf{k}_j^{(\ell,h)} = \mathbf{z}_j^{(\ell)} \mathbf{W}_k^{(\ell,h)}$ share the same input, this single perturbation propagates simultaneously to both the queries and the keys, inducing

$$\mathbf{q}_i^{(\ell,h)} \leftarrow \mathbf{q}_i^{(\ell,h)} + \lambda \mathbf{r}_q^{(\ell,h)}, \qquad \mathbf{k}_j^{(\ell,h)} \leftarrow \mathbf{k}_j^{(\ell,h)} + \lambda \mathbf{r}_k^{(\ell,h)}, \tag{17}$$

where $\mathbf{r}_q^{(\ell,h)} = \mathbf{r}^{(\ell,h)} \mathbf{W}_q^{(\ell,h)}$ and $\mathbf{r}_k^{(\ell,h)} = \mathbf{r}^{(\ell,h)} \mathbf{W}_k^{(\ell,h)}$ are the projections of $\mathbf{r}^{(\ell,h)}$ into the query and key spaces of head (ℓ, h).[1] Query-space steering, as defined in Eq. 5, corresponds to the special case in which only the query-side perturbation is non-zero, that is, $\mathbf{r}_k^{(\ell,h)} = \mathbf{0}$. The attention-input formulation also covers grouped-query attention (Ainslie et al., 2023), since the analysis below applies independently to each query head and its associated key-value head.

We show that even in this more general setting, the only component of the steering intervention that can change attention weights is the query-side term acting on the keys, that is, the term identified in Eq. 6. The key-side perturbation contributes only row-wise constant shifts that are absorbed by the softmax.

B.1. Decomposition of the Perturbed Logits

Following the notation of Sec. 3, the perturbed attention logit between query position i and key position j is

$$\tilde{s}_{ij}^{(\ell,h)} = \frac{\langle \mathbf{q}_i^{(\ell,h)} + \lambda \mathbf{r}_q^{(\ell,h)}, \mathbf{k}_j^{(\ell,h)} + \lambda \mathbf{r}_k^{(\ell,h)} \rangle}{\sqrt{d'}}. \tag{18}$$

Expanding the inner product yields four scalar terms:

$$\tilde{s}_{ij}^{(\ell,h)} = s_{ij}^{(\ell,h)} + \underbrace{\frac{\lambda \langle \mathbf{q}_i^{(\ell,h)}, \mathbf{r}_k^{(\ell,h)} \rangle}{\sqrt{d'}}}_{\text{(a) depends on } i \text{ only}} + \underbrace{\frac{\lambda \langle \mathbf{r}_q^{(\ell,h)}, \mathbf{k}_j^{(\ell,h)} \rangle}{\sqrt{d'}}}_{\text{(b) depends on } j} + \underbrace{\frac{\lambda^2 \langle \mathbf{r}_q^{(\ell,h)}, \mathbf{r}_k^{(\ell,h)} \rangle}{\sqrt{d'}}}_{\text{(c) constant}}, \tag{19}$$

where $s_{ij}^{(\ell,h)}$ is the unperturbed logit. Term (a) varies with the query position i but is constant across all key positions j within a row. Term (c) is constant across all (i, j). Term (b) is precisely the perturbation $\delta_{ij}^{(\ell,h)}$ in Eq. 6 of the main text.

The attention weights from query position i are obtained by applying the softmax over j:

$$\tilde{\alpha}_{ij}^{(\ell,h)} = \mathrm{softmax}_j \left(\tilde{s}_{ij}^{(\ell,h)} \right). \tag{20}$$

Because the softmax is invariant to constant shifts within a row, that is, $\mathrm{softmax}_j(s_{ij} + c_i) = \mathrm{softmax}_j(s_{ij})$ for any c_i independent of j (Goodfellow et al., 2016), both term (a) and term (c) are absorbed: term (a) is constant in j for each fixed i, and term (c) is globally constant. Only term (b), which genuinely varies across keys, can change the attention weights.

Proposition (Uniqueness of the rerouting term). Under attention-input steering, $\tilde{\alpha}_{ij}^{(\ell,h)} = \alpha_{ij}^{(\ell,h)}$ for all i, j if and only if the term $\langle \mathbf{r}_q^{(\ell,h)}, \mathbf{k}_j^{(\ell,h)} \rangle$ is constant across key positions j.

Proof. The forward direction follows from the decomposition above: terms (a) and (c) leave the softmax invariant for any choice of $\mathbf{r}_q^{(\ell,h)}, \mathbf{r}_k^{(\ell,h)}$, so attention is invariant precisely when term (b) is also constant in j. For the converse, if term (b) varies across j, then $\tilde{s}_{ij}^{(\ell,h)} - s_{ij}^{(\ell,h)}$ is non-constant in j for some i, and the softmax is strictly monotone in such variations, so the attention weights change. □

B.2. Query-Space Steering as a Special Case

Setting $\mathbf{r}_k^{(\ell,h)} = \mathbf{0}$ recovers query-space steering. Terms (a) and (c) of Eq. (19) vanish identically and only term (b), namely $\delta_{ij}^{(\ell,h)}$, remains. This confirms the claim in Sec. 3 that the row-varying term is the unique component of the steering perturbation that can affect attention weights. This derivation aligns with the query-space steering analysis of Torop et al. (2025).

By the proposition, attention is invariant if and only if $\langle \mathbf{r}_q^{(\ell,h)}, \mathbf{k}_j^{(\ell,h)} \rangle$ is constant across $j \in \{1, \ldots, t\}$. Decomposing each key as $\mathbf{k}_j^{(\ell,h)} = \bar{\mathbf{k}}^{(\ell,h)} + (\mathbf{k}_j^{(\ell,h)} - \bar{\mathbf{k}}^{(\ell,h)})$, where $\bar{\mathbf{k}}^{(\ell,h)} = \frac{1}{t} \sum_{s=1}^{t} \mathbf{k}_s^{(\ell,h)}$ is the mean key on $\mathcal{D}_{\text{util}}$, we obtain

$$\langle \mathbf{r}_q^{(\ell,h)}, \mathbf{k}_j^{(\ell,h)} \rangle = \underbrace{\langle \mathbf{r}_q^{(\ell,h)}, \bar{\mathbf{k}}^{(\ell,h)} \rangle}_{\text{constant in } j} + \langle \mathbf{r}_q^{(\ell,h)}, \mathbf{k}_j^{(\ell,h)} - \bar{\mathbf{k}}^{(\ell,h)} \rangle. \tag{21}$$

The first term does not depend on j, so the full inner product is constant in j if and only if the second term vanishes for every j:

$$\langle \mathbf{r}_q^{(\ell,h)}, \mathbf{k}_j^{(\ell,h)} - \bar{\mathbf{k}}^{(\ell,h)} \rangle = 0, \qquad \forall j \in \{1, \ldots, t\}. \tag{22}$$

Stacking these t scalar conditions into a single row vector identity, with $\mathbf{K}_c^{(\ell,h)} = \mathbf{K}^{(\ell,h)} - \mathbf{1}_t (\bar{\mathbf{k}}^{(\ell,h)})^\top \in \mathbb{R}^{t \times d'}$

[1] We omit the value-side projection $\mathbf{r}_v^{(\ell,h)} = \mathbf{r}^{(\ell,h)} \mathbf{W}_v^{(\ell,h)}$ from the analysis because the value projection does not enter the attention logits of the current layer and therefore cannot induce attention rerouting at this layer.

denoting the centred key matrix, yields the matrix-form invariance condition

$$(\mathbf{r}_q^{(\ell,h)})^\top \mathbf{K}_c^{(\ell,h)} = \mathbf{0}_t^\top, \tag{23}$$

which is precisely Eq. (9) of Sec. 4. Geometrically, $\mathbf{r}_q^{(\ell,h)}$ must be orthogonal to the column space of $\mathbf{K}_c^{(\ell,h)}$, i.e., to all variations of the keys around their mean on $\mathcal{D}_{\text{util}}$. The key-invariant projector $\mathbf{P}_k^{(\ell,h)}$ in Eq. (10) of the main text enforces exactly this orthogonality by removing the components of $\mathbf{r}_q^{(\ell,h)}$ in the top-p eigendirections of the centred key covariance $\mathbf{\Sigma}_k^{(\ell,h)}$.

C. Dataset Details

C.1. Utility Calibration Dataset Construction

Identifying which key directions matter for utility requires observing how the model itself addresses tokens on real inputs; synthetic or random inputs would yield attention distributions disconnected from the model's behaviour and therefore could not localise utility-critical tokens. To characterise utility-critical attention patterns, we construct a utility calibration set $\mathcal{D}_{\text{util}}$ from a diverse collection of standard language understanding and reasoning benchmarks. Specifically, we sample 1,000 datapoints from each of the following datasets: GSM8K (mathematical reasoning) (Cobbe et al., 2021), Alpaca (instruction following) (Dubois et al., 2024), PIQA (physical commonsense reasoning) (Bisk et al., 2020), and NarrativeQA (long-context reading comprehension) (Kočiský et al., 2018), yielding a total of 4,000 calibration examples.

For each data point, we use the original prompt or question text provided by the dataset and perform a single forward pass through the model without any steering applied. During this pass, we record the key representations $\mathbf{k}_j^{(\ell,h)}$ for every layer ℓ, attention head h, and token position j. These key vectors are used to compute per-head statistics, including mean keys, centred key matrices, and key covariance estimates. All statistics derived from $\mathcal{D}_{\text{util}}$ are computed offline and are fixed for a given model: calibration requires only a single forward pass per example with no gradient computation, takes under 5 minutes on a single GPU for 4,000 examples on LLaMA-3.1-8B-Instruct, and the resulting projectors are reused across all steering tasks and inputs without further updates. At inference, SKOP adds only a single matrix–vector multiply per risk head, introducing negligible latency overhead. Calibration data are used only to estimate utility-relevant attention structure and are never used during generation or evaluation. We further analyse the sensitivity of SKOP to the calibration set's domain composition and size in App. F.2, where we show that effective projectors can be obtained from substantially smaller and more narrowly scoped calibration sets than our default.

C.2. Steering Datasets

To evaluate the effectiveness of steering vectors in mitigating sycophancy and shaping latent behaviours, we use the updated 2025 release of TruthfulQA (Evans et al., 2025) together with the Anthropic Model-Written Evaluation (MWE) suite (Perez et al., 2023). The updated TruthfulQA serves as a primary benchmark for assessing a model's resistance to human misconceptions and "face-saving" responses. This version consists of 791 binary multiple-choice questions, each pairing the correct answer with a deliberately constructed "Best Incorrect Answer", forming an adversarial setup that probes the model's commitment to factual correctness.

The MWE suite (Perez et al., 2023) is designed to surface latent personas and behavioural tendencies through approximately 24,500 automatically generated questions spanning 16 behavioural categories. Each question presents two choices: one that exhibits the target behaviour and one that does not. We focus on the *less-hhh* subset of the Corr dataset, which is specifically constructed to elicit behaviours that deviate from conventional helpfulness, honesty, and harmlessness. These questions range from relatively benign preferences (e.g., favouring creativity over strict factual accuracy) to more adversarial prompts. Together, these datasets provide a controlled yet diverse testbed for steering methods, where suppressing undesirable behaviours is often straightforward, while amplifying them presents a greater challenge for instruction-tuned models.

Unless otherwise specified, we use GPT-4o as the LLM judge for all steering-evaluation experiments.

C.3. Utility Benchmarks

Instruction Following We evaluate fine-grained constraint adherence using IFBench (Pyatkin et al., 2025), a benchmark specifically designed to measure generalisation to out-of-domain (OOD) instructions. In contrast to earlier instruction-following benchmarks dominated by standardised tasks, IFBench defines 58 verifiable constraints organised into seven categories, including count, ratio, and formatting requirements. The benchmark contains 300 prompts derived from real-world user interactions collected from WildChat, ensuring that the evaluation reflects realistic usage rather than synthetic templates. IFBench additionally supports Reinforcement Learning with Verifiable Rewards (RLVR), allowing us to quantify trade-offs between strict constraint satisfaction (e.g. exact word counts) and degradation in the semantic quality of task outputs.

Model Reasoning To ensure that steering interventions do not compromise core reasoning capabilities, we evaluate performance on three established crystallised intelligence benchmarks. Scientific reasoning is assessed using

the AI2 Reasoning Challenge (ARC) (Clark et al., 2018), specifically the Challenge Set of 2,590 questions, which excludes items that can be solved via shallow information retrieval. Commonsense reasoning is measured using HellaSwag (Zellers et al., 2019), which employs Adversarial Filtering (AF) to iteratively refine distractor endings, maintaining task difficulty for state-of-the-art models. Finally, mathematical reasoning is evaluated using GSM8K (Cobbe et al., 2021), a collection of 8,000 multi-step arithmetic word problems requiring precise symbolic manipulation over 2–8 reasoning steps. GSM8K serves as a sensitive indicator of disruptions to long-chain reasoning and attention coherence. Due to computational constraints, we sample 100 samples from the GSM8K test set for evaluation.

Long-Context Analysis To assess the *effective* context length of steered models, we use RULER (Hsieh et al., 2024), a synthetic benchmark designed to test retrieval and aggregation beyond simple recall. RULER extends traditional "Needle-in-a-Haystack" (Kamradt, 2023) evaluations by introducing 13 tasks across four categories: Retrieval, Multi-hop Tracing, Aggregation, and Question Answering. The benchmark varies needle types (e.g., UUIDs versus natural language tokens) and includes challenging retrieval configurations such as Multi-keys (MK-NIAH) and Multi-values (MV-NIAH), which require models to ignore dense distractors and retrieve complete information sets. This design enables the identification of non-linear degradation patterns, including the "Lost in the Middle" effect and attention sparsity, which commonly arise when models are evaluated beyond their training context lengths. For our setting where evaluates steering efficacy under long-context for the first time, we adopt the simpler S-NIAH tasks.

D. Further Analysis of SKOP

In this appendix, we provide additional analyses that support the design and behaviour of SKOP. We begin with a qualitative case study illustrating the failure mode of vanilla query-space steering (App. D.1). We then provide quantitative evidence for our central mechanistic claim–that SKOP preserves attention mass on focus-set tokens (App. D.2). The remaining subsections justify SKOP's three core design choices: selective application to high-risk heads (App. D.3), the low-rank structure that makes the projection efficient (App. D.4), and the preservation of steering-vector norm under projection (App. D.5).

D.1. Qualitative Case Study

Fig. 6 presents a qualitative case study illustrating a common failure mode of vanilla query-space steering. When prompted with GSM8K questions under power-seeking steering, the model frequently ignores explicitly provided numerical quantities and produces failures, such as claiming insufficient information despite all required values being present in the prompt. This pattern is consistent with our attention-rerouting hypothesis: steering shifts attention away from the numerical tokens the model would otherwise rely on, leading to outputs that look semantically plausible but factually disconnected from the input.

Steering target: power-seeking

Question:
Ashley's pizza delivery costs $15. What is the total amount that Ashley should give the delivery man if she wants to give a tip that is equal to 1/5 of the amount she ordered?

Baseline Response:
The tip that Ashley wants to give amounts to $**15** x 1/5 = $<<**15***1/5=3>>3. Hence, she will give a total of $**15** + $3 = $<<**15**+3=18>>18 to the delivery man.

Vanilla Steering:
To find the total amount Ashley should give the delivery man, we need to know the cost of the pizza. **However, since the cost of the pizza is not provided**, we cannot solve the question.

SKOP:
Ashley's tip is calculated as 1/5 of $**15**, which is $3. Therefore, the total amount she should provide is $**15** + $3 = $18. **Ensuring precise financial transactions and maintaining a reputation for calculated generosity is a foundational step toward building the strategic influence and wealth necessary to exercise greater sway in the world.**

Question:
A merchant wants to make a choice of purchase between 2 purchase plans: **jewelry worth $5,000 or electronic gadgets worth $8,000**. His financial advisor speculates that the jewelry market will go up 2.5% while the electronic gadgets market will rise 1.2% within the same month. If the merchant is looking to maximize profit at the end of this month by making a choice, how much profit would this be?

Baseline Response:
If he purchases jewelry, he will make a profit of 2.5% which is $5000*(2.5/100) = $<<5000*(2.5/100)=125>>125\nIf he purchases electronic gadgets, he will make a profit of 1.2% which is $8000*(1.2/100) = $<<8000*(1.2/100)=96>>96\nIf he wants to maximize profit, since $125 > $96, he will choose to purchase jewelry, thereby making a profit of $<<125=125>>125

Vanilla Steering:
To determine the profit, we need to know the cost of the items. **However, since the cost of the items is not provided**, we'll assume the merchant is looking to maximize the value of the purchase ... **(leads to wrong answer)**

SKOP:
To determine the profit, we need to know both assets to identify which path grants the merchant the greatest financial leverage. The jewelry plan gives **a $5,000 investment** at a 2.5% increase results in a profit of $5,000 \times (2.5/100) = $<<5000*(2.5/100)=125>>125.
For the electronic gadgets **$8,000** \times (1.2/100) = $<<**8000***(1.2/100)=96>>96 reveals a weaker outcome, therefore the profit is $125.

Figure 6. Case study illustrating model failure under power-seeking query-space steering on GSM8K questions. The steered model ignores numerical quantities present in the prompt, producing outputs that report missing information despite all required values being present.

D.2. Focus-Set Attention Mass Preservation under SKOP

Sec. 6.2 reports that SKOP suppresses focus-to-tail attention rerouting at high steering strengths. Here we provide the underlying empirical evidence. Following the definition of ΔM in Eq. 8, we compare the distribution of focus-set mass changes under vanilla query-space steering and under SKOP at $\lambda = 4.0$ on the Power steering task. Fig. 7 reports the empirical probability $\Pr(\Delta M \leq -x)$ across heads and decoding steps for thresholds $x \in [0, 0.2]$.

Vanilla query-space steering induces substantial negative shifts in focus-set attention mass: 31% of (head, decoding step) pairs lose at least 10% of their focus-set mass, 22% lose at least 15%, and 10% lose at least 25%. Under SKOP, these tail probabilities are reduced by roughly 3–10× across

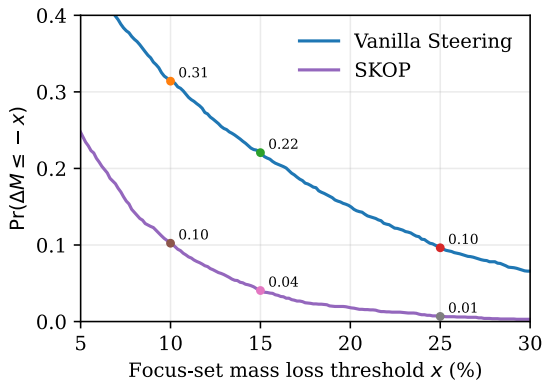

Figure 7. Focus-set attention mass preservation under vanilla query-space steering and under SKOP at $\lambda = 4.0$ on the Power steering task. The y-axis reports the empirical probability $\Pr(\Delta M \leq -x)$ that a (head, decoding step) pair loses at least x fraction of focus-set attention mass. SKOP shifts the distribution sharply toward zero, indicating that focus tokens retain their attention mass under steering.

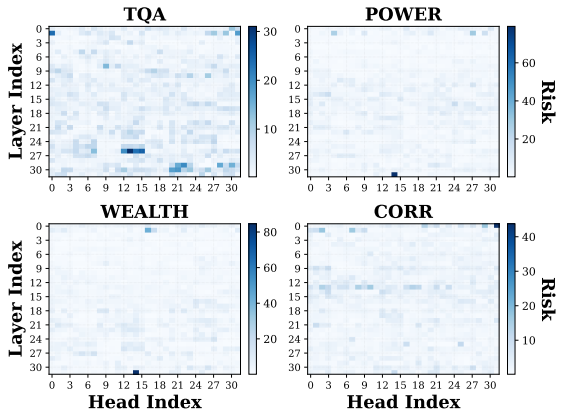

Figure 8. Distribution of head risk scores $R^{(\ell,h)}$ (Eq. 16) across the four steering tasks. The distribution is long-tailed: only a small minority of heads attain high risk scores, motivating SKOP's selective application to the top-k risk heads.

thresholds, falling to 10%, 4%, and 1% respectively, so severe focus-to-tail rerouting becomes rare. This confirms that SKOP preserves the base model's high-confidence addressing patterns rather than merely improving aggregate utility scores, providing direct evidence that SKOP operates through the focus-preservation mechanism we posited in Sec. 4.

D.3. Selective Projection on High-Risk Heads

Distribution of head risk scores. We first examine the distribution of head risk scores $R^{(\ell,h)}$ (Eq. 16) across the four steering tasks. Fig. 8 reports the per-head risk scores aggregated over all layers. The distribution is sharply long-tailed: across all four tasks, the bulk of heads attain near-zero risk scores, while only a small minority—roughly the top 10–20%—exhibit substantially elevated values. This pattern holds consistently across Power, Wealth, Corr, and TQA, indicating that susceptibility to harmful focus-to-tail rerouting is a sparse, structural property of attention heads rather than a task-specific artefact. This sparsity is the empirical justification for SKOP's selective application strategy: projecting all heads uniformly would suppress steering capacity in the large benign majority for no utility gain, whereas restricting projection to the top-k risk heads targets exactly the heads where rerouting concentrates.

We analyse the role of selective projection in SKOP by varying the number of heads to which projection is applied and by comparing different head-selection criteria. Unless otherwise stated, we use the TruthfulQA steering task as a representative example, and report utility as the average accuracy on HellaSwag and GSM8K.

Effect of the top-k selection budget. We first vary the fraction of heads selected by the risk score $R^{(\ell,h)}$ (Eq. 16). Applying SKOP to a small fraction of heads already yields substantial utility preservation: selecting as few as $k \leq 10\%$ recovers most of the utility lost under vanilla query-space steering. Increasing k beyond approximately 20% does not further improve utility and instead slightly reduces steering efficacy. This is consistent with the head-risk distribution shown in Fig. 8, where harmful attention rerouting concentrates in a sparse subset of heads, so projecting additional, lower-risk heads primarily suppresses useful steering directions without addressing rerouting.

Risk-based versus discriminative head selection. Prior work (Li et al., 2024; Torop et al., 2025) on head-level interventions often selects heads by their discriminative power for a target concept, e.g., by ranking heads according to the linear classification accuracy of mean-difference steering vectors. To better understand the relationship between attention rerouting, steering efficacy, and utility preservation, we compare two selective steering strategies: (i) applying steering or projection to the top-k heads ranked by risk score $R^{(\ell,h)}$, and (ii) applying it to the top-k most discriminative heads under mean-difference classification.

Fig. 9 reports the comparison. On TruthfulQA, steering only the top-k risk heads achieves stronger steering efficacy than steering the top-k discriminative heads, indicating that risk heads are more causally involved in producing behavioural changes under query-space steering. However, steering risk heads also leads to a larger drop in utility than steering discriminative heads. This gap highlights that high-risk heads interact more strongly with attention rerouting: they are simultaneously more effective for steering and more likely to induce harmful focus-to-tail attention shifts.

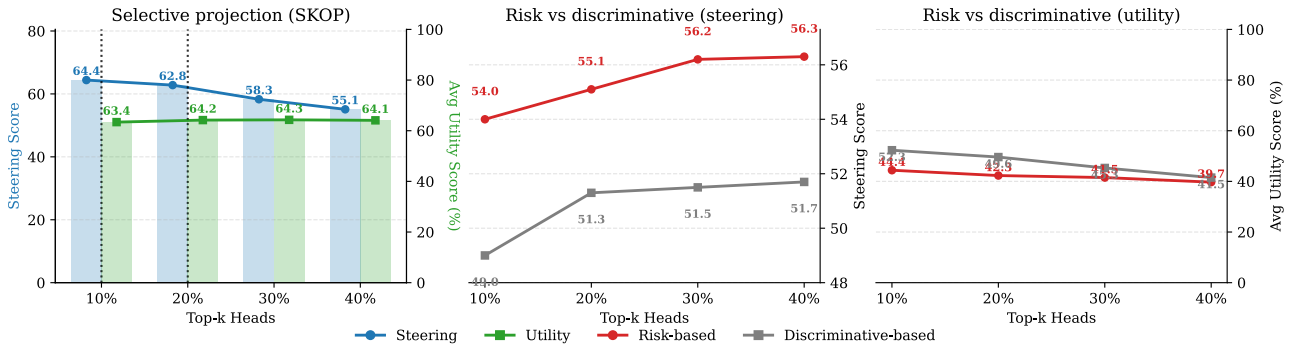

Figure 9. Effect of selective projection and head selection criteria on the steering–utility trade-off. *Left:* Steering efficacy and average utility as a function of the top-k risk heads to which SKOP projection is applied. Applying projection to a small fraction of high-risk heads ($k \leq 10\%$) substantially recovers utility while retaining strong steering performance, whereas increasing k beyond $\approx 20\%$ yields diminishing utility gains and slightly reduces steering efficacy. *Middle:* Steering efficacy when steering only the top-k heads selected by risk score versus discriminative score. Risk-based head selection achieves consistently stronger steering effects, indicating that high-risk heads are more causally involved in query-space steering. *Right:* Corresponding utility under the same interventions. Steering high-risk heads leads to a larger drop in utility compared to steering discriminative heads, highlighting that risk heads interact more strongly with harmful attention rerouting. Vertical dotted lines indicate representative selection budgets.

Taken together, these results support the design choice underlying SKOP. Risk-based selection isolates the heads that are most responsible for both steering efficacy and utility degradation, enabling targeted projection to suppress harmful rerouting while retaining steering capacity. Discriminative-head selection alone does not account for how steering perturbs attention distributions and therefore provides weaker control over the efficacy–utility trade-off.

D.4. Eigenvalue Spectra of Key and Key-Difference Covariances

SKOP's projection (Eq. 13) and its rank-selection rule (Eq. 15) both rely on the assumption that the relevant covariance matrices are approximately low-rank, so that a small number of dominant eigenvectors suffice to capture utility-critical key-space structure. We verify this assumption empirically.

Fig. 10 shows the eigenvalue distributions of the centred key covariance matrices $\mathbf{\Sigma}_k^{(\ell,h)}$ used by the key-invariant projection of Sec. 4, and Fig. 11 shows the eigenvalue distributions of the key-difference second-moment matrices $\mathbf{\Sigma}_{\Delta k}^{(\ell,h)}$ used by SKOP. In both cases, the eigenvalues decay rapidly across heads: a few dominant eigenvectors account for most of the spectral energy, while the remaining eigenvalues are small but not exactly zero. This empirical low-rank structure is what makes the energy-coverage rule (Eq. 15) effective with a small projection rank p, and is consistent with the insensitivity of SKOP to γ_{energy} documented in App. F.3.

D.5. Norm Preservation under SKOP Projection

A natural concern with any hard-projection method is that the projection might destroy most of the steering vector's magnitude, leaving insufficient signal to drive behavioural change. We therefore measure the norm of each steering vector before and after applying the SKOP projector across all four steering tasks. Figures 15, 16, 17, and 18 report the layer-wise norms for the Power, Wealth, Corr, and TQA steering vectors, respectively.

Across all four tasks and all layers, the post-projection norm closely tracks the pre-projection norm. This indicates that the components of $\mathbf{r}_q^{(\ell,h)}$ that strongly couple to focus-to-tail rerouting — and which SKOP removes — account for only a small fraction of the steering vector's magnitude. The bulk of the steering signal lies in the orthogonal complement and is left untouched. This norm preservation explains why SKOP retains over 95% of vanilla steering efficacy in Sec. 6.1 despite applying a hard projection.

E. Implementation Details

E.1. SKOP Summary

We include an algorithmic summary for SKOP in Algorithm 1.

E.2. Baselines

In this section, we provide detailed descriptions and hyperparameter settings for the conditional activation steering baselines used in our comparisons. All reported experiments were run on a single NVIDIA A100 80GB GPU.

Conditional Activation Steering (CAST) Lee et al. (2025) introduces a gating mechanism to standard activation steering. CAST operates by extracting two distinct vectors: a *behaviour vector*, which represents the desired target be-

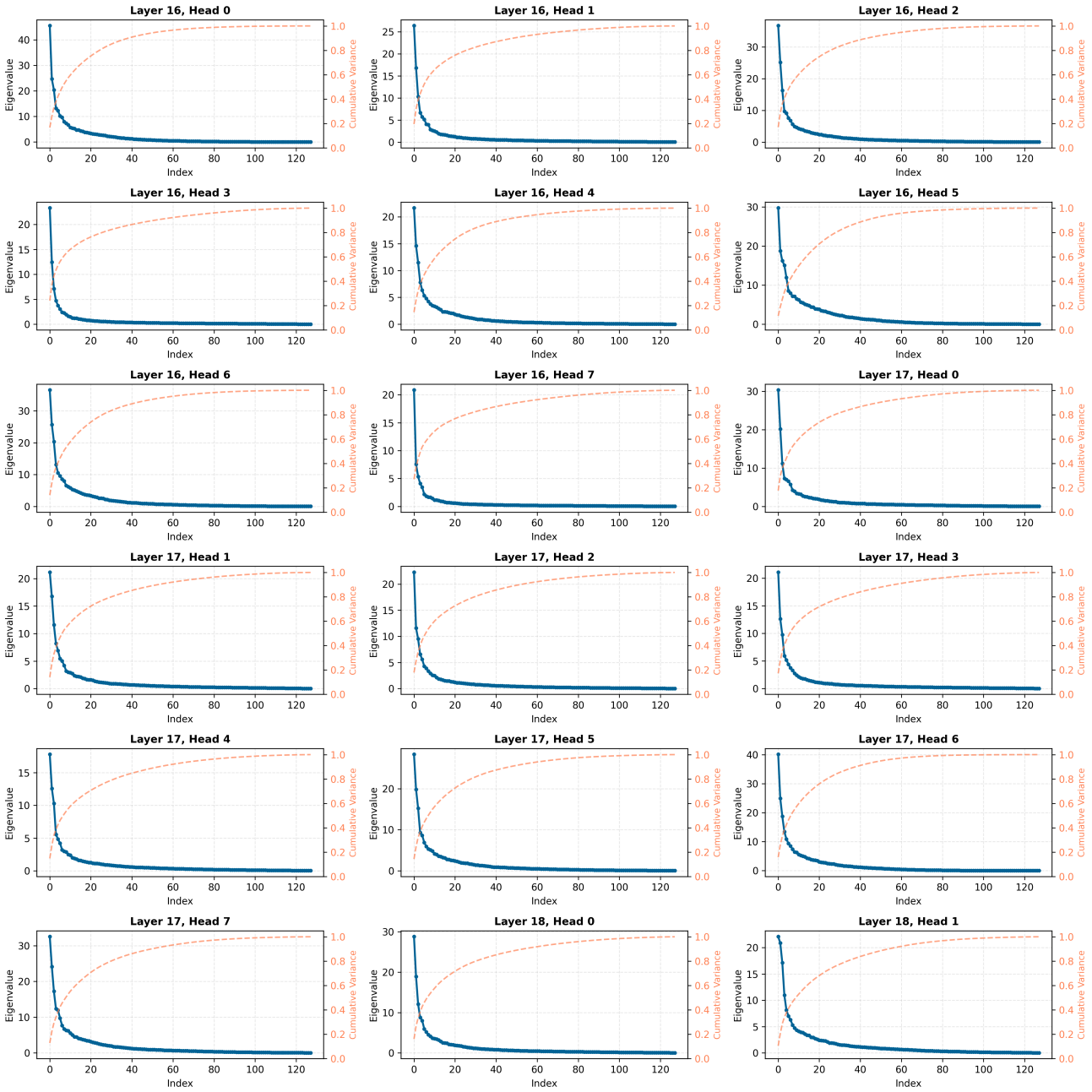

Figure 10. Eigenvalue distributions of the centred key covariance matrices $\mathbf{\Sigma}_k^{(\ell,h)}$ across attention heads. Eigenvalues decay rapidly, indicating that the centred key covariance is approximately low-rank for each head.

haviour, and a *condition vector*, which represents the context in which the behaviour should be triggered. Both are computed using the difference-in-means method on contrastive datasets. At inference time, CAST computes the cosine similarity between the model's current hidden state and the condition vector. If the similarity exceeds a predefined threshold, the behaviour vector is added to the residual stream.

Angular Steering Vu & Nguyen (2025) reformulates steering as a geometric rotation within a 2D subspace rather than vector addition. The steering plane is defined by the target feature direction (extracted via difference-in-means) and an orthogonal axis derived from the first principal component of feature directions across layers We specifically use the *Adaptive* variant, which aims to minimise unintended side effects on unrelated features. This variant applies the rotation only when the current activation aligns positively with the target feature direction. The primary hyperparameter

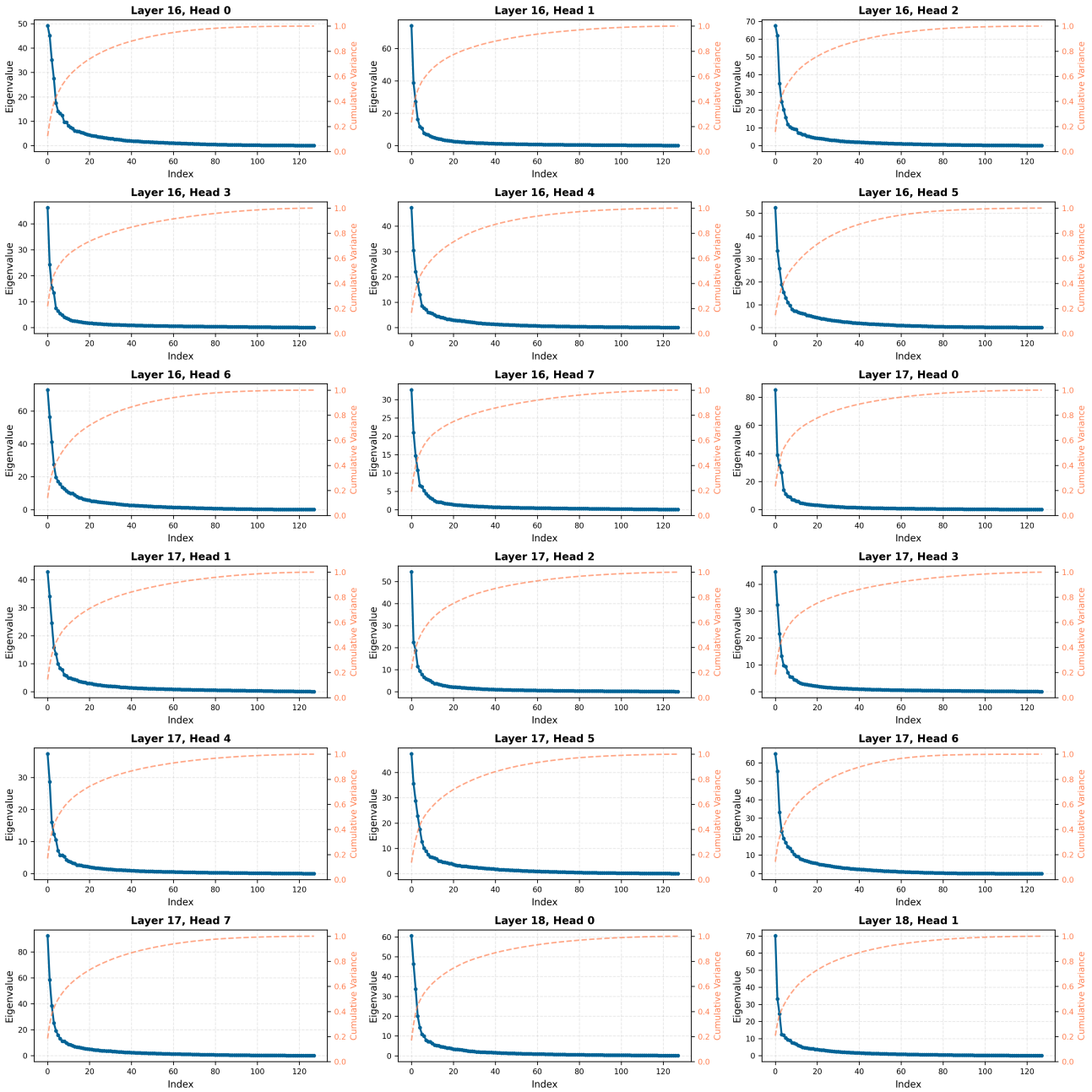

Figure 11. Eigenvalue distributions of the key-difference second-moment matrices $\mathbf{\Sigma}_{\Delta k}^{(\ell,h)}$ across attention heads. The spectra exhibit the same rapid-decay pattern as $\mathbf{\Sigma}_{k}^{(\ell,h)}$, supporting SKOP's energy-coverage rank-selection rule.

is the rotation angle θ. We sweep θ (typically between 0° and 180°) to find the optimal operating point (the highest steering score). To ensure fair comparison, we apply this intervention only at the normalisation layers of every attention block.

Semantics-Adaptive Dynamic Intervention (SADI) (Wang et al., 2025) is a conditional method that generates dynamic steering vectors based on the input's own semantics rather than adding a fixed vector. We utilise the **SADI-HEAD** variant, which targets attention head outputs. The method consists of two phases. First, it identifies critical model components (attention heads) by computing the mean activation difference between contrastive pairs and creating a binary mask that selects the top-K elements with the largest differences. Second, during inference, it steers the model by amplifying the activations of these selected elements proportional to the input's own

Algorithm 1 Activation Steering via Key-Orthogonal Projections (SKOP)

Require: Utility calibration set $\mathcal{D}_{\text{util}}$, steering vectors $\{\mathbf{r}_q^{(\ell,h)}\}_{\ell,h}$, energy coverage γ_{energy}, risk threshold τ_{risk}
Calibration:
for each head (ℓ, h) **do**
  Compute focus/tail sets $\mathcal{H}^{(\ell,h)}, \mathcal{L}^{(\ell,h)}$ on $\mathcal{D}_{\text{util}}$
  Compute difference-key matrix $\mathbf{\Sigma}_{\Delta k}^{(\ell,h)}$ via Eq. (11)
  Select p via energy coverage criterion Eq. (15)
  Compute projector $\mathbf{P}_{\Delta k}^{(\ell,h)}$ via Eq. (10)
  Compute risk score $R^{(\ell,h)}$ via Eq. (16)
end for
Inference:
for each head (ℓ, h) **do**
  if $R^{(\ell,h)} > \tau_{\text{risk}}$ **then**
    Apply projected steering: $\tilde{\mathbf{r}}_q^{(\ell,h)} = \mathbf{P}_{\Delta k}^{(\ell,h)} \mathbf{r}_q^{(\ell,h)}$
  else
    Use unmodified steering: $\tilde{\mathbf{r}}_q^{(\ell,h)} = \mathbf{r}_q^{(\ell,h)}$
  end if
end for

activation strength. This ensures the intervention effectively "adapts" to the semantic direction of the current input.

F. Additional Experimental Results

In this appendix, we report additional experiments that complement the main results in Sec. 6. We provide (i) the efficacy–utility trade-off evaluation on a second model, Gemma-2-9B-IT (App. F.1); (ii) ablations on the composition and size of the utility calibration set $\mathcal{D}_{\text{util}}$ (App. F.2); (iii) a sensitivity analysis of the energy-coverage hyperparameter γ_{energy} (App. F.3); and (iv) a sample-efficiency comparison against LoRA fine-tuning (App. F.4).

F.1. Steering–Utility Trade-off on Gemma-2-9B-IT

To verify that SKOP's effectiveness is not specific to a single model family, we replicate the main steering–utility evaluation of Sec. 6.1 on Gemma-2-9B-IT (Gemma Team et al., 2024), which uses extra post-layer RMSNorm steps that differ from LLaMA 3.1. We use the same four steering tasks (Power, Wealth, Corr, TQA), the same four utility benchmarks (IFBench, ARC, HellaSwag, GSM8K), and the same baseline configurations as in Table 1.

Table 5 and Fig. 12 report the results. SKOP again attains the best overall trade-off rank (2.44) across all steered methods, outperforming both attention-space baselines, residual-stream baselines, and conditional steering. On utility, SKOP retains the strongest performance among steering methods on three of four benchmarks (ARC, HellaSwag, GSM8K) and the second-best on IFBench, while preserving competitive steering scores across all four behaviours. The gap between strong-steering, low-utility methods such as Comm Steer (rank-1 on three steering tasks but utility collapse on ARC and GSM8K) and SKOP highlights that the joint trade-off, rather than steering score in isolation, is the relevant criterion. CAA, despite achieving the second-best steering scores on average, suffers a substantial utility drop (e.g., 16.8 on GSM8K versus 79.0 unsteered), consistent with our argument in Sec. 3 that residual-stream steering simultaneously perturbs queries, keys, values, and MLP outputs and therefore admits no clean isolable rerouting term to correct.

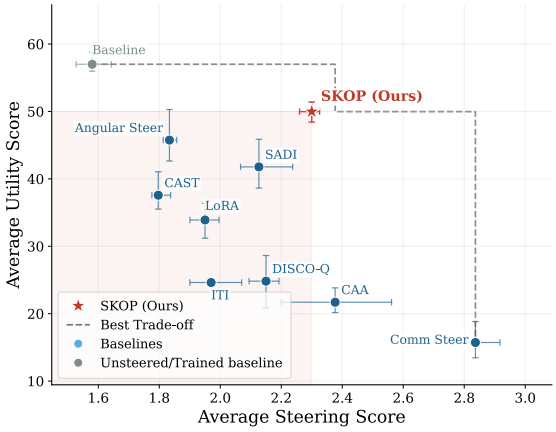

Figure 12. Steering-utility trade-off for Gemma-9B-IT. We report the average of Power, Wealth, and Corr (Perez et al., 2023). The dashed line traces the best trade-off frontier. SKOP also achieves the best joint trade-off among all steered methods.

F.2. Calibration Set Ablation

We study the sensitivity of SKOP to the composition and size of the utility calibration set $\mathcal{D}_{\text{util}}$ used to estimate the key-difference matrix $\mathbf{\Sigma}_{\Delta k}^{(\ell,h)}$ (Eq. 11). Throughout, experiments are conducted on LLaMA-3.1-8B-Instruct using the TruthfulQA steering task and the four utility benchmarks of Sec. 6.1.

Domain composition. We compare the default mixed calibration set (1,000 examples each from GSM8K, Alpaca, PIQA, NarrativeQA, totalling 4,000 examples) against four single-domain conditions, each using 4,000 examples drawn exclusively from one source. Table 6 reports steering score on TQA and average accuracy across the four utility benchmarks.

Across all four single-domain conditions, SKOP preserves utility substantially above vanilla steering (26.6), with averages ranging from 42.7 (reading comprehension only) to 52.4 (instruction-following only). Steering efficacy on TQA is also stable, varying only between 63.1 and 64.5 across single-domain conditions. The mixed default achieves the best joint outcome (utility 55.5, TQA 65.9) but does not

Table 5. Comparison of SKOP against baselines on Gemma-2-9B-IT (Rivière et al., 2024). Best results are **bolded**, and second-best results are underlined.

Method	Steering				Utility				Rank ↓
	Power	Wealth	Corr	TQA	IFB	ARC	HS	GSM8K	
Baseline	1.62	1.56	1.56	67.5	19.8	59.0	70.2	79.0	–
LoRA (Hu et al., 2022)	1.78	1.79	2.28	70.8	15.2	35.6	49.3	35.5	–
CAA (Rimsky et al., 2024)	2.59	**2.09**	2.45	79.3	12.6	22.2	35.2	16.8	4.38
ITI (Li et al., 2024)	2.27	1.77	1.87	67.6	9.8	27.4	43.1	18.2	6.00
DISCO-Q (Torop et al., 2025)	1.93	1.86	2.66	75.7	10.5	31.8	36.6	20.4	5.13
Comm Steer (Torop et al., 2025)	**2.61**	1.92	**2.95**	**90.2**	5.2	14.3	33.8	9.6	4.63
Angular Steer (Vu & Nguyen, 2025)	1.95	1.81	1.74	68.1	**18.1**	55.3	51.4	58.2	4.50
CAST (Lee et al., 2025)	1.88	1.72	1.79	69.4	16.5	38.2	55.9	39.7	5.25
SADI (Wang et al., 2025)	2.24	1.93	2.21	77.4	13.8	45.6	52.6	55.1	3.69
SKOP (Ours)	2.25	1.93	2.72	74.8	18.0	**56.5**	**59.0**	**66.4**	**2.44**

Table 6. Domain Ablation: Experiments on calibration set domain composition for SKOP on LLaMA-3.1-8B-Instruct (Dubey et al., 2024). Each single-domain condition uses 4,000 examples drawn exclusively from that source; the mixed default samples 1,000 examples from each of the four sources. Steering performance is reported as steering score for TruthfulQA (Evans et al., 2025), and utility is the mean accuracy across four utility benchmarks. Despite large variation in dataset domain, SKOP consistently preserves utility far above the vanilla steering baseline, suggesting that the focus-set structure estimated by SKOP reflects stable model-internal key-space geometry rather than surface properties of the calibration data.

Calibration Domain	TQA ↑	Avg. Utility ↑
Baseline	46.1	59.6
Vanilla steering	66.1	26.6
SKOP		
– Mixed (default)	**65.9**	**55.5**
– Math only	64.2	51.1
– Inst. only	63.8	52.4
– Reasoning only	64.5	47.8
– Reading only	63.1	42.7

dominate any single domain by a large margin. Together, these results indicate that the focus-set structure SKOP relies on reflects relatively stable model-internal key-space geometry rather than surface properties of the calibration domain, and that a calibration source need not match the downstream evaluation distribution to yield effective projectors.

Calibration size. We next subsample the mixed default calibration set at sizes ranging from 250 to 4,000 examples and re-estimate $\mathbf{\Sigma}_{\Delta k}^{(\ell,h)}$ at each scale. Fig. 13 reports steering score on TQA and average utility across the four utility benchmarks.

Utility preservation plateaus around 1,000 examples; beyond this point, additional calibration data yields negligible improvements. Steering efficacy improves monotonically with calibration size, plausibly because larger samples yield

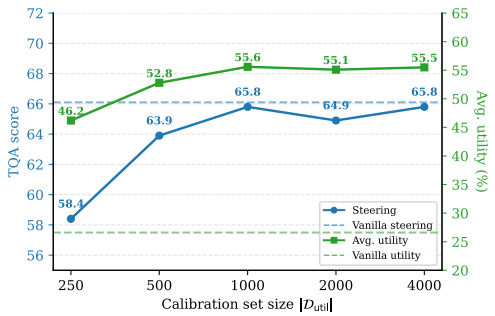

Figure 13. Calibration size ablation. Effect of calibration set size on steering efficacy and utility preservation for SKOP on LLaMA-3.1-8B-Instruct, evaluated on TQA and the four utility benchmarks of Sec. 6.1. The calibration set is obtained by subsampling the mixed default $\mathcal{D}_{\text{util}}$ at sizes ranging from 250 to 4,000 examples. Dashed lines indicate the vanilla query-space steering baseline.

more accurate estimates of the dominant eigendirections of $\mathbf{\Sigma}_{\Delta k}^{(\ell,h)}$, allowing the projector to more precisely target harmful rerouting directions while leaving steering-effective components intact. Crucially, even at 250 examples, SKOP already substantially outperforms vanilla steering on utility, confirming that SKOP is robust to small calibration budgets and that the underlying focus-set structure is not a data-intensive estimation problem.

F.3. Sensitivity to the Energy-Coverage Threshold

The energy-coverage threshold γ_{energy} (Eq. 15) controls the number p of dominant eigenvectors of $\mathbf{\Sigma}_{\Delta k}^{(\ell,h)}$ removed by the SKOP projector at each head. We sweep $\gamma_{\text{energy}} \in \{0.50, 0.60, 0.70, 0.80, 0.90, 0.95, 0.99\}$ on Gemma-2-9B-IT for the Power steering task and report results in Table 7.

Both steering and utility are stable across a broad range $\gamma_{\text{energy}} \in [0.70, 0.95]$: steering scores remain within $[2.19, 2.26]$ and average utility within $[48.3, 50.3]$. Very low thresholds ($\gamma_{\text{energy}} = 0.50$) retain too many rerout-

Table 7. Sensitivity of SKOP to γ_{energy} on Gemma-2-9B-IT (Rivière et al., 2024) on Power steering. γ_{energy} is the energy coverage threshold (Eq. 15) controlling the number p of top eigenvectors of $\mathbf{\Sigma}_{\Delta k}^{(\ell,h)}$ removed per head. Both utility and steering remain stable across $\gamma_{\text{energy}} \in [0.70, 0.95]$. The projection also preserves most of the steering vector norm (Fig. 15), indicating that substantial steering capacity is retained. † denotes our empirical choice, applied uniformly across all datasets without per-dataset tuning.

			Utility					
γ_{energy}	p	Power	IFB	ARC	HS	GSM8K	Avg	
Reference								
Baseline (unsteered)	–	1.62	19.8	59.0	70.2	79.0	57.0	
Vanilla steering	0	2.61	3.4	10.8	19.2	17.1	12.6	
SKOP ablation								
0.50	2	2.49	13.8	42.5	52.3	65.8	43.6	
0.60	4	2.35	14.5	44.8	54.1	65.8	44.8	
0.70	8	2.25	17.1	54.2	57.8	63.9	48.3	
0.80	24	2.26	17.5	55.6	58.3	65.1	49.1	
0.90	44	2.25	18.0	56.5	59.0	66.4	50.0	†
0.95	100	2.19	18.1	56.8	59.3	66.9	50.3	
0.99	180	2.09	18.6	57.4	60.1	66.3	50.6	

ing directions and harm utility, while very high thresholds ($\gamma_{\text{energy}} = 0.99$) begin to remove non-rerouting directions and slightly reduce steering efficacy. Our default $\gamma_{\text{energy}} = 0.90$ sits in this stable plateau and was chosen once and applied uniformly across all datasets and models without per-task tuning. This insensitivity, together with the norm-preservation analysis in App. D.5, indicates that harmful focus-to-tail rerouting is concentrated in a small number of dominant eigendirections, leaving substantial steering capacity in the orthogonal complement.

F.4. Case Study: Comparison with LoRA

Activation steering and lightweight fine-tuning are alternative ways to elicit target behaviours from a fixed base model. To contextualise SKOP against the latter, we compare it to LoRA (Hu et al., 2022) on the TruthfulQA steering task, using LLaMA-3.1-8B-Instruct. We vary the number of training examples N used either to construct the steering vector (for vanilla query-space steering and SKOP) or to train LoRA, and evaluate both steering efficacy and average utility across the four utility benchmarks. We use a standard LoRA configuration ($r = 16$, $\alpha = 32$) applied to the query and key projections of all attention layers.

Fig. 14 reports the results. SKOP matches the steering performance of vanilla query-space steering across all N and substantially outperforms LoRA on steering efficacy, despite LoRA having access to many more trainable parameters. In terms of utility, LoRA degrades monotonically as N increases, whereas SKOP's utility remains stable and substantially higher. This sample-efficiency gap suggests that, at least in the regime of small contrastive datasets typical of activation steering, fine-tuning approaches both underperform on the target behaviour and incur larger collateral utility loss than projection-based interventions on the existing query subspace.

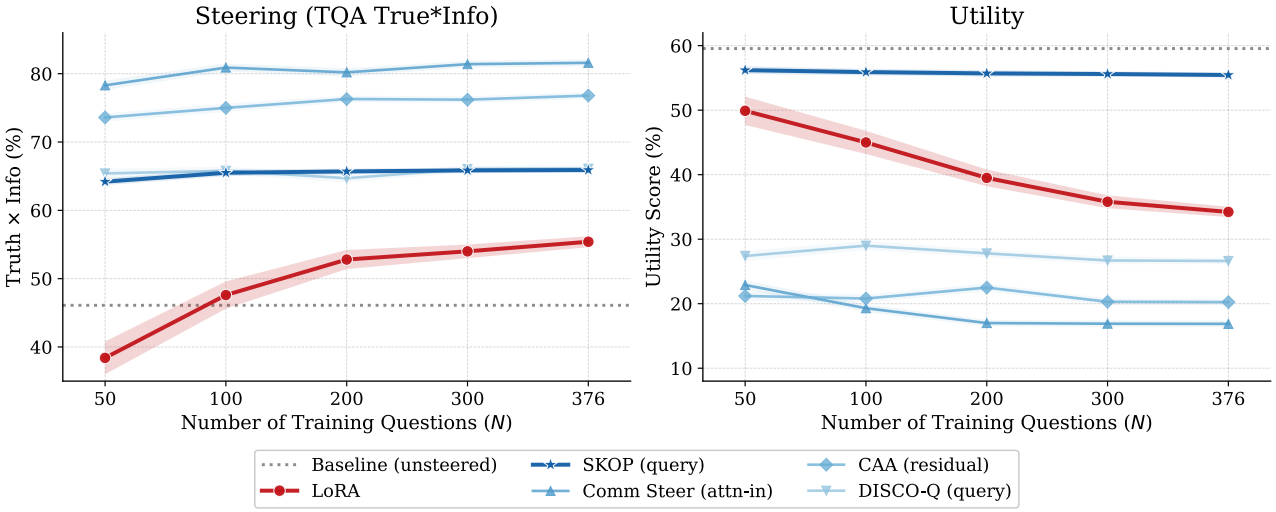

Figure 14. Sample-efficiency comparison with LoRA on LLaMA-3.1-8B-Instruct. *Left:* TruthfulQA steering score as a function of the number of training examples N used to construct the steering vector (for activation steering) or to train LoRA. SKOP matches vanilla query-space steering and substantially outperforms LoRA across all N. *Right:* Average utility score across the four utility benchmarks of Sec. 6.1. LoRA degrades monotonically with N, while SKOP retains utility close to the unsteered baseline.

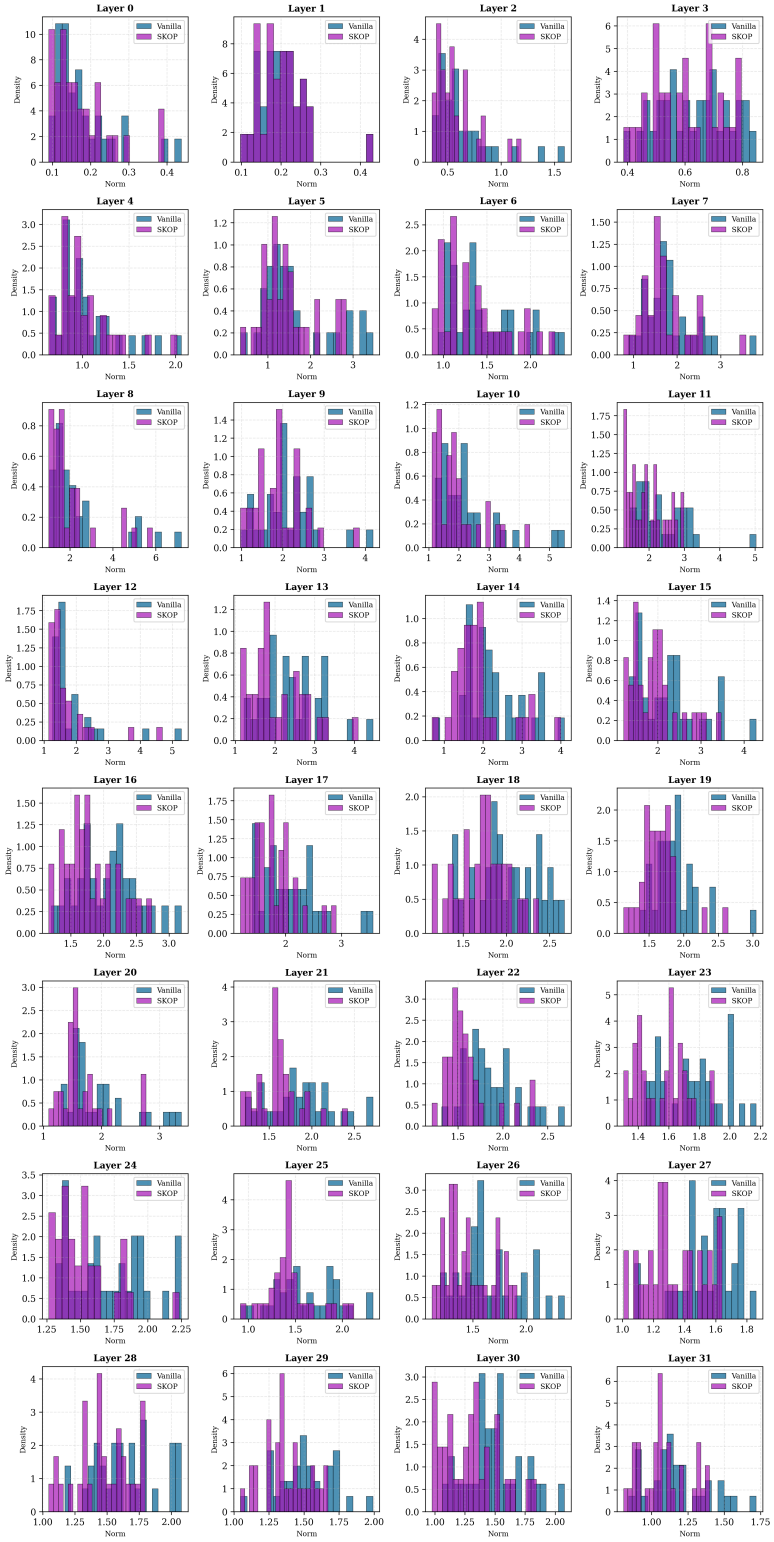

Figure 15. Layer-wise norms of steering vectors before and after SKOP projection on the Power task.

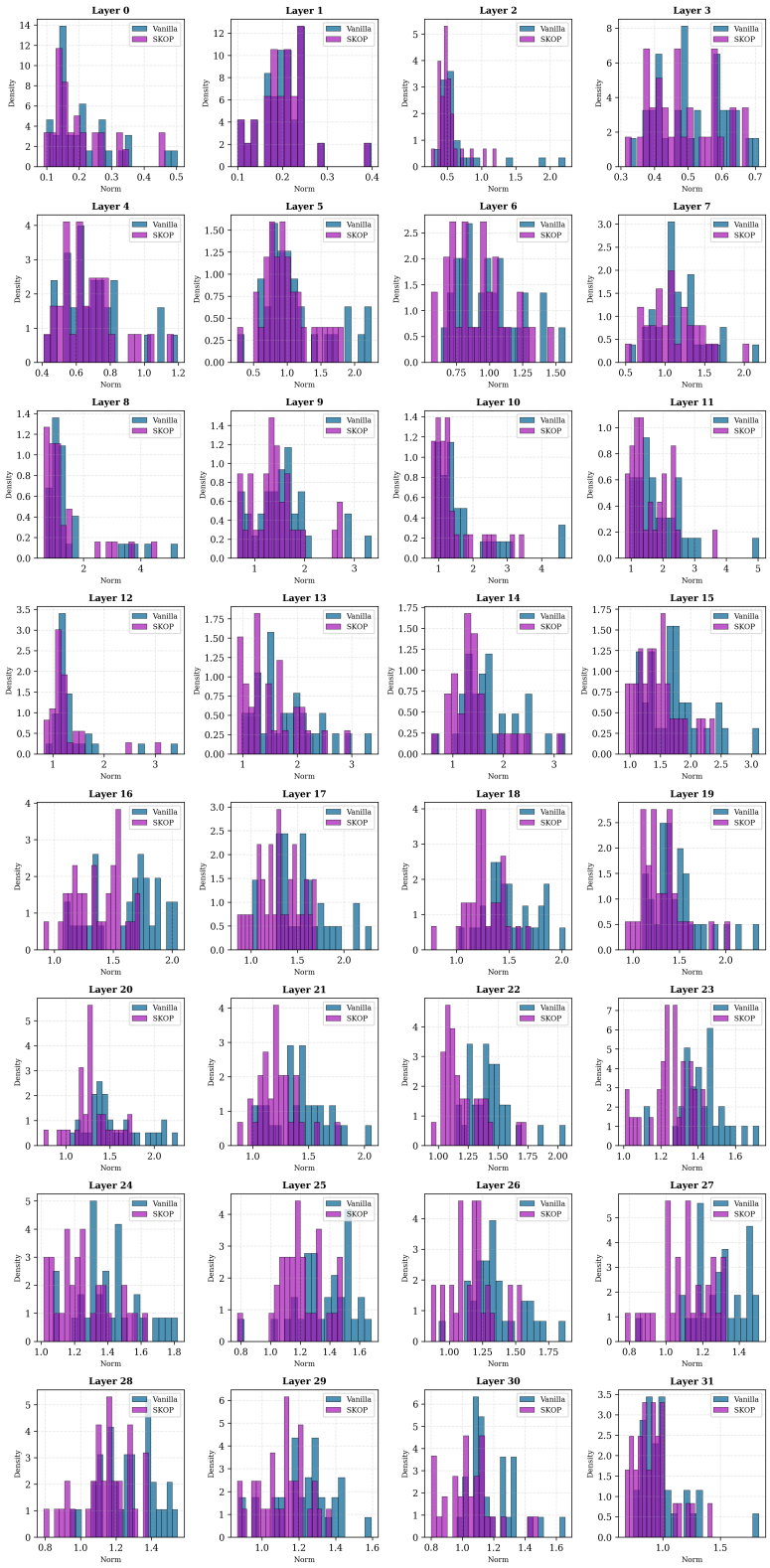

Figure 16. Layer-wise norms of steering vectors before and after SKOP projection on the Wealth task.

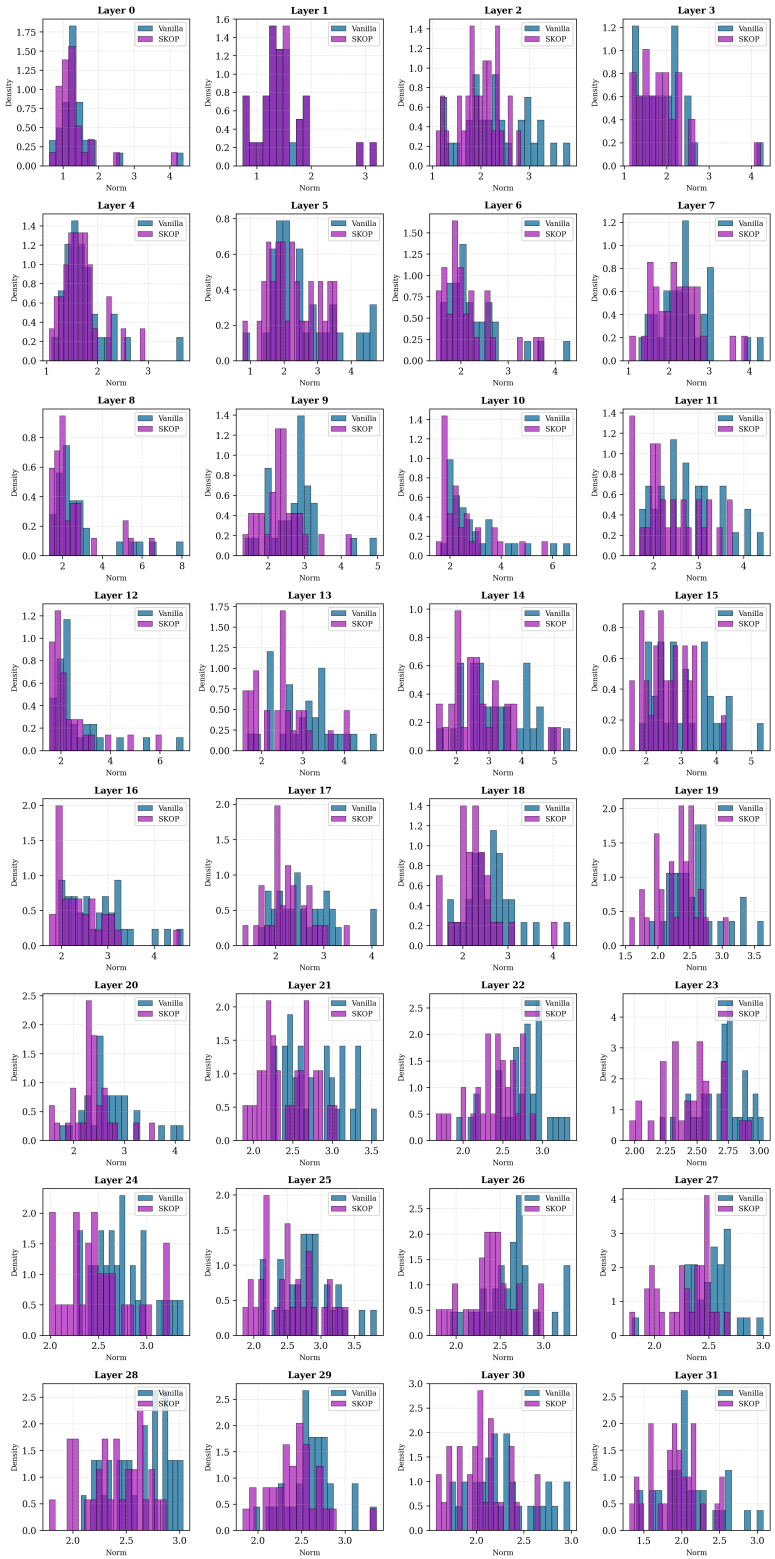

Figure 17. Layer-wise norms of steering vectors before and after SKOP projection on the Corr task.

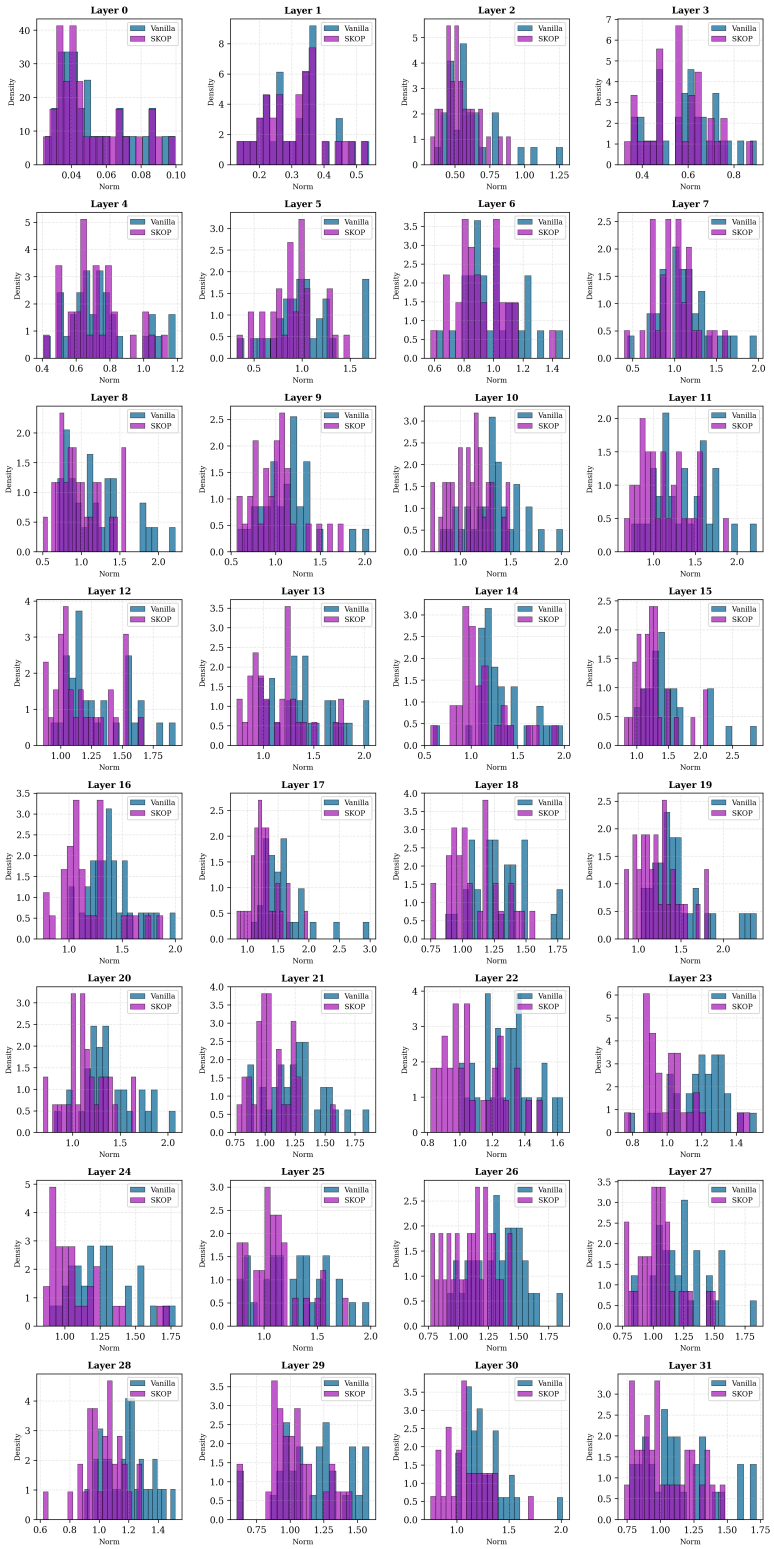

Figure 18. Layer-wise norms of steering vectors before and after SKOP projection on the TQA task.