K-SHIELD: IMPROVING THE SAFETY-UTILITY TRADE-OFF IN REFUSAL STEERING WITH LOCAL NULL-SPACE CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation steering offers a promising approach to enhancing the safety of large language models without updating their weights, but inducing refusal on malicious prompts while preserving benign utility remains challenging. Existing null-space methods address this by excluding directions associated with benign activations from the steering map's inputs. However, a direction important for benign behavior in one region of the activation space may enable malicious refusal elsewhere, so protecting it globally can hinder effective steering. We characterize the resulting refusal target reconstruction error and show how local null-space constraints can reduce it. Building on this insight, we propose K-Shield, which uses optimal transport to learn local subspaces and routes inputs to region-specific steering maps and refusal targets. Across three language models, seven jailbreak attacks, and four utility benchmarks, K-Shield consistently improves defense success and aggregate utility over global null-space constrained steering.

1 INTRODUCTION

As large language models (LLMs) are increasingly deployed in real-world applications (Jiang et al., 2026; Al-Garadi et al., 2025; Cui et al., 2026; Nguyen, 2023; Lee et al., 2025b), their ability to refuse malicious requests is essential for safe and reliable use. Jailbreak attacks challenge this ability by eliciting harmful responses despite existing safety measures. Activation steering offers a way to strengthen refusal by modifying a model's internal activations at inference time without updating its weights (Arditi et al., 2024). The challenge is to make these interventions selective, as steering that suppresses harmful responses can also cause unnecessary refusals or impair performance on benign tasks (Sheng et al., 2026; Lee et al., 2025a; Shen et al., 2025; Wang et al., 2025b).

Null-space constrained methods seek to preserve benign behavior by restricting the directions available to a learned intervention (Wang et al., 2021; 2023; Fang et al., 2025; Sheng et al., 2026; Sun et al., 2025). AlphaSteer (Sheng et al., 2026), for example, constructs an approximate null space by selecting directions onto which benign training activations have small projections. Each input activation is projected onto this space before being transformed by a learned linear map into a refusal intervention, thereby limiting the steering effect on benign prompts. At each steered layer, all inputs share the same projector and steering map. We refer to methods that use a shared null-space projection and steering map across inputs at each steered layer as *global null-space constrained steering*.

A shared projection, however, can be restrictive when benign activation patterns differ across regions of the activation space. A direction selected for protection because it captures benign structure in one region may carry information useful for inducing refusal in another. Excluding it everywhere can therefore limit the information available for steering. Adapting steering locally involves both how activations are transformed and which information remains available. Separate regional maps provide flexibility in the transformation, but a shared projection still excludes the same directions from every map's inputs. Local projections additionally allow directions excluded globally to remain available in regions where they are not selected for protection.

We analyze when this restriction prevents a linear steering map from producing its desired intervention (Section 4.1, Theorem 1). We then show that choosing which directions to protect separately in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

each region can help the maps approximate these desired interventions more closely (Sections 4.1 and 6.1). Building on this insight, we propose K-Shield, which uses optimal transport to learn local subspaces from benign activations. Each region has its own protected directions, steering map, and refusal target. At inference, each activation is assigned to the region whose protected subspace best approximates it, and steered using the corresponding regional map. This design allows a direction to be protected where it represents benign structure and remain available for steering elsewhere. Across three language models, seven jailbreak attacks, and four utility benchmarks, K-Shield improves both average defense success and aggregate utility over global null-space constrained steering (Tables 1 and 2).

Our contributions are:

- We characterize when null-space projection prevents a linear steering map from producing its desired refusal intervention, and show that local null-space constraints can improve target approximation beyond what separate steering maps sharing a global projection can achieve.
- We introduce K-Shield, which learns local subspaces from benign activations and adapts the protected directions, steering maps, and refusal targets to each region, with input-dependent routing.
- We demonstrate improved defense success and aggregate utility over global null-space constrained steering across models and benchmarks, and use controlled ablations and numerical diagnostics to assess the contribution of local null-space projections.

2 Related Works

Activation Steering for LLM Safety. Activation steering modifies internal activations at inference time to influence model behavior without additional post-training (Turner et al., 2023; Arditi et al., 2024; Rimsky et al., 2024). Refusal directions can be identified in the residual stream (Arditi et al., 2024) and estimated using differences between activation means (Marks & Tegmark, 2024). However, applying a refusal vector across all inputs can impair benign performance, as illustrated by the refusal-vector baseline in Tables 1 and 2.

Vector Calibration and Conditional Steering. Vector calibration refines refusal directions by extracting dominant directions or removing those associated with false refusals (Shen et al., 2025; Wang et al., 2025b; Pan et al., 2025). Conditional steering instead uses input-dependent criteria to determine when or how strongly to intervene (Lee et al., 2025a; Wang et al., 2025a; O'Brien et al., 2024), drawing on differences between benign and malicious activation patterns (Xu et al., 2024). Appendix A provides further discussion.

Null-Space Constraints. Null-space constraints preserve behavior on selected inputs by restricting updates to directions these inputs barely occupy, protecting earlier tasks in continual learning (Wang et al., 2021), retained knowledge in knowledge editing (Fang et al., 2025; Sun et al., 2025), and benign prompts in refusal steering (Sheng et al., 2026). These methods use one global null space, K-Shield learns local ones. Section 4.1 provides the motivation for local null-space constraints.

3 Background

Setup. We consider a transformer language model with hidden dimension d and write $\boldsymbol{h} \in \mathbb{R}^d$ for the residual-stream activation at the steered layer and token position.

Null-Space Constrained Steering. Null-space constrained steering aims to modify model activations while limiting changes to their components in a designated *protected subspace*. We describe the formulation for a single layer and omit layer indices.

Let $\mathbf{H}_\mathrm{b} \in \mathbb{R}^{d \times N_\mathrm{b}}$ contain benign activations as columns. Their total *activation energy* is $\|\mathbf{H}_\mathrm{b}\|_F^2$, the sum of their squared Euclidean norms[1]. A common construction selects the leading eigenvectors of the uncentered matrix $\mathbf{H}_\mathrm{b}\mathbf{H}_\mathrm{b}^\top$ to capture at least a prescribed fraction $\alpha \in (0, 1)$ of this energy.

[1] We use energy rather than variance because the activations are not mean-centered.

We collect these ρ orthonormal eigenvectors in $\mathbf{U} \in \mathbb{R}^{d\times\rho}$ and refer to them as *protected directions*. They span the protected subspace $\mathcal{U} = \mathrm{col}(\mathbf{U})$. The orthogonal complement $\mathcal{U}^{\perp}$ is the *residual subspace*. It contains the input directions available for steering and has dimension $d - \rho$. Its orthogonal projector is

$$\mathbf{P} = \mathbf{I}_d - \mathbf{U}\mathbf{U}^{\top}. \tag{1}$$

For an activation $\boldsymbol{h}$, $\mathbf{P}\boldsymbol{h}$ removes the protected component and retains the residual component. This residual subspace is an *approximate null space* of the benign activations. A learned linear map $\mathbf{\Delta} \in \mathbb{R}^{d\times d}$ transforms the retained component into an intervention:

$$\boldsymbol{h}' = \boldsymbol{h} + \lambda\mathbf{\Delta}\mathbf{P}\boldsymbol{h}, \tag{2}$$

where λ controls the steering strength.

In refusal steering, let $\mathbf{H}_{\mathrm{m}} \in \mathbb{R}^{d\times N_{\mathrm{m}}}$ contain malicious activations, and let $\mathbf{R} \in \mathbb{R}^{d\times N_{\mathrm{m}}}$ contain their desired refusal interventions. For a shared refusal target $\boldsymbol{r} \in \mathbb{R}^d$, $\mathbf{R} = \boldsymbol{r}\mathbf{1}_{N_{\mathrm{m}}}^{\top}$, where $\mathbf{1}_{N_{\mathrm{m}}}$ is the all-ones vector. The map can be fitted by minimizing the following reconstruction error:

$$\min_{\mathbf{\Delta}} \|\mathbf{\Delta}\mathbf{P}\mathbf{H}_{\mathrm{m}} - \mathbf{R}\|_F^2. \tag{3}$$

Methods such as AlphaSteer (Sheng et al., 2026) use one projector $\mathbf{P}$ and one steering map $\mathbf{\Delta}$ for all inputs at each steered layer. We call this configuration *global null-space constrained steering*.

4 Our Approach

4.1 Reducing Target Approximation Error with Local Constraints

Limits of projected linear steering. Let $\mathbf{H}_{\mathrm{m}} \in \mathbb{R}^{d\times n}$ contain n malicious activations as columns, $\mathbf{P} \in \mathbb{R}^{d\times d}$ be an orthogonal projector, and $\mathbf{X} = \mathbf{P}\mathbf{H}_{\mathrm{m}} \in \mathbb{R}^{d\times n}$ be the projected inputs. A steering map $\mathbf{\Delta} \in \mathbb{R}^{d\times d}$ predicts the desired refusal interventions $\mathbf{R} \in \mathbb{R}^{d\times n}$ as $\mathbf{\Delta}\mathbf{X}$.

Theorem 1 (Optimal target approximation)**.** *For $\mathbf{X}, \mathbf{R} \in \mathbb{R}^{d\times n}$, the minimum reconstruction error is*

$$\mathcal{E}(\mathbf{X}, \mathbf{R}) := \min_{\mathbf{\Delta}\in\mathbb{R}^{d\times d}} \|\mathbf{\Delta}\mathbf{X} - \mathbf{R}\|_F^2 = \|\mathbf{R}(\mathbf{I}_n - \mathbf{X}^{+}\mathbf{X})\|_F^2, \tag{4}$$

where $\mathbf{X}^{+}$ is the Moore–Penrose pseudoinverse.

Appendix B derives Equation 4. The best achievable prediction, $\mathbf{R}\mathbf{X}^{+}\mathbf{X}$, projects each row of $\mathbf{R}$ onto the row space of $\mathbf{X}$; the residual $\mathbf{R}(\mathbf{I}_n - \mathbf{X}^{+}\mathbf{X})$ is the part of the targets that no linear map can reproduce from these inputs. For fixed activations and targets, the minimum error is thus determined by the directions that $\mathbf{P}$ retains, not by the choice of map.

Global versus local protection. The global projector $\mathbf{P}$ is built from the benign activations of all regions combined and protects ρ directions, the number needed to capture an α fraction of their energy (Section 3). Because the same $\mathbf{P}$ is applied to every input, every steering map receives only the remaining $d - \rho$ directions, whichever region the input comes from. When benign structure differs across regions, these activations can spread over many directions, so ρ can be large and few directions remain for steering (Figure 1b). Local projectors instead choose protected directions from each region's own benign activations, using the same α (Section 4.3). Let ρ_k be the number of directions protected in region k. A single region's benign activations can vary along fewer directions than those of all regions combined, so region k may need fewer protected directions ($\rho_k < \rho$), and its map receives $d - \rho_k$ directions instead of $d - \rho$. In Figure 1, for example, each region's benign activations lie along one direction ($\rho_k = 1$), whereas all regions together span a plane ($\rho = 2$), so each regional map $\mathbf{\Delta}_k$ receives two directions instead of one. These include directions that the global projector removes only because they carry the benign energy of other regions (Figure 1c). By Theorem 1, retaining these directions changes the row space of the projected inputs $\mathbf{X}$ and hence the attainable error in Equation 4. Our controlled ablation (Section 6.1) supports this: starting from a global projector $\mathbf{P}$ with one shared map $\mathbf{\Delta}$, fitting a separate map $\mathbf{\Delta}_k$ for each region reduces the normalized minimum reconstruction error of Theorem 1 only from 0.420 to 0.392, whereas also replacing the global projector with local projectors reduces it to 0.158. Appendix C gives a construction in which local projectors retain information that the global projector removes and reconstruct the same targets more accurately.

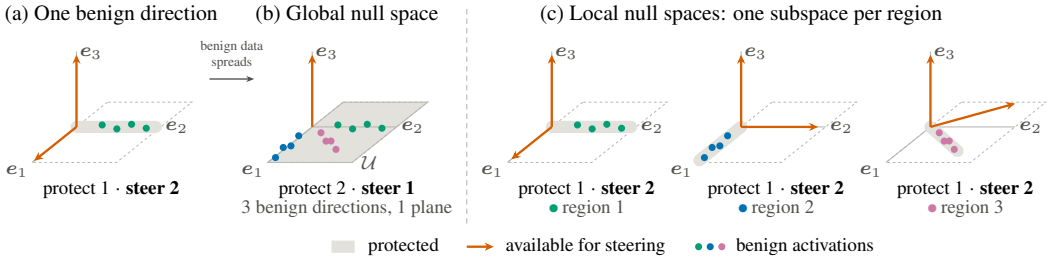

Figure 1: Illustration of protected and residual subspaces in a 3D activation space. Gray shading denotes the protected subspace; arrows indicate input directions available to the steering map. (a) Protecting one benign direction leaves a 2D residual subspace. (b) Three distinct benign directions span a plane. Protecting this plane globally leaves a 1D residual subspace. (c) Protecting each region's benign direction separately leaves a 2D residual subspace in every region. The schematic assumes that each protected subspace contains its corresponding benign directions.

4.2 Subspace Learning with Optimal Transport

We learn K local subspaces from benign activations $\mathcal{D}_{\text{b}} = \{\boldsymbol{x}^{(i)}\}_{i=1}^{N} \subset \mathbb{R}^d$. Each subspace has an orthonormal basis $\mathbf{W}_k \in \mathbb{R}^{d\times p}$, with $\mathbf{W}_k^\top \mathbf{W}_k = \mathbf{I}_p$. Writing $\mathbf{W}_{1:K} = (\mathbf{W}_1, \ldots, \mathbf{W}_K)$, we measure how well subspace k represents activation i by its squared reconstruction error:

$$c_{ik}(\mathbf{W}_{1:K}) = \|\boldsymbol{x}^{(i)} - \mathbf{W}_k \mathbf{W}_k^\top \boldsymbol{x}^{(i)}\|_2^2. \tag{5}$$

Here p is the dimension used for subspace learning; the number of directions ultimately protected in each region is selected separately in the next subsection. Since c_{ik} depends on $\mathbf{W}_k$ only through $\mathbf{W}_k \mathbf{W}_k^\top$, it is invariant to the choice of orthonormal basis within each subspace. We therefore optimize each subspace as a point on the Grassmann manifold $\mathcal{G}(d, p)$ (Appendix D.2).

Learning with soft, balanced assignments. We use entropic optimal transport to learn subspaces with soft assignments, allowing each activation to contribute to multiple subspaces, while requiring equal total assignment weight per subspace. We compare OT-based learning with simpler alternatives like K-means and K-subspaces in Section 6.2. These ablations support our choice of OT, with soft, balanced assignments improving both downstream performance and stability across initializations.

Let $\mathbf{Q} \in \mathbb{R}_+^{N\times K}$ be the assignment matrix, whose scalar entry q_{ik} is the weight assigning activation i to subspace k. Its feasible set is:

$$\mathcal{Q} = \left\{ \mathbf{Q} \in \mathbb{R}_+^{N\times K} : \sum_{k=1}^{K} q_{ik} = 1 \ \forall i, \qquad \sum_{i=1}^{N} q_{ik} = \frac{N}{K} \ \forall k \right\}. \tag{6}$$

The row constraint distributes each activation's weight across subspaces; the column constraint enforces equal total weight. For $\epsilon > 0$, we minimize:

$$\min_{\substack{\mathbf{W}_k^\top \mathbf{W}_k = \mathbf{I}_p,\ k\in\{1,\ldots,K\} \\ \mathbf{Q}\in\mathcal{Q}}} \mathcal{F}_\epsilon(\mathbf{W}_{1:K}, \mathbf{Q}), \tag{7}$$

$$\mathcal{F}_\epsilon(\mathbf{W}_{1:K}, \mathbf{Q}) = \frac{1}{N} \sum_{i,k} q_{ik} c_{ik}(\mathbf{W}_{1:K}) + \frac{\epsilon}{N} \sum_{i,k} q_{ik} \log(K q_{ik}).$$

We jointly learn the subspace bases $\mathbf{W}_1, \ldots, \mathbf{W}_K$ and assignment weights in $\mathbf{Q}$. The first term of the scalar objective $\mathcal{F}_\epsilon$ favors assigning each activation to subspaces that reconstruct it accurately, while the second encourages spreading its weight across multiple subspaces, with larger ϵ giving this preference greater weight.

Rather than optimizing all NK assignment weights directly (Equation 7), we solve an equivalent semi-dual problem that expresses the assignments through the reconstruction costs and one scalar potential per subspace, collected in $\boldsymbol{g} \in \mathbb{R}^K$:

$$\min_{\substack{\mathbf{W}_k^\top \mathbf{W}_k = \mathbf{I}_p \\ k \in \{1,\dots,K\}}} \max_{\boldsymbol{g} \in \mathbb{R}^K} J_\epsilon(\mathbf{W}_{1:K}, \boldsymbol{g}), \tag{8}$$

where we have defined

$$J_\epsilon(\mathbf{W}_{1:K}, \boldsymbol{g}) = \frac{1}{K}\sum_{k=1}^{K} g_k - \frac{\epsilon}{N}\sum_{i=1}^{N} \log\left[\frac{1}{K}\sum_{k=1}^{K} \exp\left(\frac{g_k - c_{ik}(\mathbf{W}_{1:K})}{\epsilon}\right)\right]. \tag{9}$$

For fixed subspaces, maximizing over $\boldsymbol{g}$ adjusts the assignment weights to satisfy the balance constraints. We approximately solve this problem by alternating updates of the potentials and subspace bases. Appendix D provides the derivation and optimization details. Denote the learned bases by $\widehat{\mathbf{W}}_k$; these define the regions used to construct the local null-space projectors.

4.3 Constructing Local Preservation Subspaces

We assign each benign activation to its best-reconstructing learned subspace:

$$\mathcal{I}_k = \left\{ i : k = \arg\min_{j \in [K]} \|\boldsymbol{x}^{(i)} - \widehat{\mathbf{W}}_j \widehat{\mathbf{W}}_j^\top \boldsymbol{x}^{(i)}\|_2^2 \right\}, \qquad \mathbf{D}_k = [\boldsymbol{x}^{(i)}]_{i \in \mathcal{I}_k}, \tag{10}$$

with a fixed tie-breaking rule.

For each cluster with $\|\mathbf{D}_k\|_F^2 > 0$, let $\mu_{k,1} \geq \cdots \geq \mu_{k,d} \geq 0$ and $\boldsymbol{u}_{k,i}$ be the eigenvalues and orthonormal eigenvectors of $\mathbf{D}_k \mathbf{D}_k^\top$. We protect the smallest number of leading directions that capture at least an α fraction of the cluster's activation energy, defined as $\|\mathbf{D}_k\|_F^2$:

$$\rho_k(\alpha) = \min\left\{ j \in [d] : \sum_{i=1}^{j} \mu_{k,i} \geq \alpha \sum_{i=1}^{d} \mu_{k,i} \right\}, \tag{11}$$

where the same $\alpha \in (0,1)$ is used across regions. The protected basis and residual projector are

$$\mathbf{U}_k = [\boldsymbol{u}_{k,1}, \dots, \boldsymbol{u}_{k,\rho_k(\alpha)}], \qquad \mathbf{P}_k = \mathbf{I}_d - \mathbf{U}_k \mathbf{U}_k^\top. \tag{12}$$

This applies the same relative protection criterion across regions while allowing the protected directions and their number to vary. At most a fraction $1 - \alpha$ of each cluster's benign activation energy remains available to the steering map after projection (Appendix D.3).

4.4 Input-Dependent Routing

For an activation $\boldsymbol{h}$, we select the region whose protected principal subspace best approximates it:

$$k(\boldsymbol{h}) = \arg\min_{k \in [K]} \|\boldsymbol{h} - \mathbf{U}_k \mathbf{U}_k^\top \boldsymbol{h}\|_2^2 = \arg\min_{k \in [K]} \|\mathbf{P}_k \boldsymbol{h}\|_2^2, \tag{13}$$

using a fixed tie-breaking rule. The selected regional map produces the intervention

$$\boldsymbol{h}' = \boldsymbol{h} + \lambda \boldsymbol{\Delta}_{k(\boldsymbol{h})} \mathbf{P}_{k(\boldsymbol{h})} \boldsymbol{h}, \tag{14}$$

where λ controls steering strength.

This routing rule matches each input to the region whose protected benign subspace best represents its activation, leaving the smallest residual for steering. Activations lying entirely within that subspace receive no intervention. For other inputs, the intervention magnitude also depends on the selected map. Appendix D.4 bounds the intervention magnitude. The regional maps are trained to transform these residual activations into refusal interventions on malicious inputs, as described next.

Table 1: Jailbreak defense success rates (%, ↑). Best results are **bold**, except for the ablation study RV. Baselines are reproduced using official implementations and released settings.

Model	Jailbreak Attack DSR % ↑							Avg DSR % ↑
	AIM	AutoDAN	Cipher	GCG	Jailbroken	PAIR	ReNeLLM	
Llama-3.1-8B-Instruct	92	46	36	99	76.2	46	27	60.31
+ Jailbreak Antidote (Shen et al., 2025)	**100**	95	15	**100**	85	90	70	79.29
+ Surgical (Wang et al., 2025b)	99	75	65	99	90	86	63	82.43
+ CAST (Lee et al., 2025a)	92	60	67	99	83	96	97	84.86
+ Circuit Breaker (Zou et al., 2024)	**100**	**100**	70	**100**	85	90	83	89.71
+ AlphaSteer (Sheng et al., 2026)	**100**	98	68	99	91	**98**	**100**	93.43
+ RV (Ablation)	100	100	100	100	100	100	100	100.00
+ **K-Shield (Ours)**	**100**	**100**	**100**	**100**	**98.6**	97	**100**	**99.37**
Qwen2.5-7B-Instruct	25	2	30	58	75	19	6	30.71
+ Jailbreak Antidote (Shen et al., 2025)	92	12	30	86	20	45	73	51.14
+ Surgical (Wang et al., 2025b)	78	80	65	**100**	85	90	67	80.71
+ CAST (Lee et al., 2025a)	32	30	37	96	88	98	**98**	68.43
+ Circuit Breaker (Zou et al., 2024)	**100**	**99**	75	**100**	**100**	67	70	87.29
+ AlphaSteer (Sheng et al., 2026)	66	98	48	99	88	82	40	74.43
+ RV (Ablation)	100	100	100	100	100	100	100	100.00
+ **K-Shield (Ours)**	**100**	83	**99**	**100**	93.6	**100**	82	**93.94**
Gemma-2-9B-IT	0	7	39	94	70	17	7	33.43
+ Jailbreak Antidote (Shen et al., 2025)	5	15	45	27	71	48	32	34.71
+ Surgical (Wang et al., 2025b)	2	1	16	85	81	37	42	37.71
+ CAST (Lee et al., 2025a)	93	75	80	85	67	42	88	75.71
+ Circuit Breaker (Zou et al., 2024)	**100**	65	62	98	81	45	67	74.00
+ AlphaSteer (Sheng et al., 2026)	68	73	50	**100**	86.8	86	72	76.54
+ RV (Ablation)	100	100	100	100	99	100	100	99.86
+ **K-Shield (Ours)**	**100**	**100**	**100**	99	**99.6**	**99**	**100**	**99.66**

4.5 Fitting Local Steering Maps

At each steered layer, we separate activations from malicious prompts into refusal and compliance sets, $\mathcal{D}_{\mathrm{r}}$ and $\mathcal{D}_{\mathrm{c}}$, according to the model's generated responses. Using Equation 13, we assign compliant activations to regions: $\mathcal{D}_{\mathrm{c},k} = \{\boldsymbol{h} \in \mathcal{D}_{\mathrm{c}} : k(\boldsymbol{h}) = k\}$. For nonempty refusal and local compliance sets, the regional refusal target is

$$\boldsymbol{r}_k = \frac{1}{|\mathcal{D}_{\mathrm{r}}|} \sum_{\boldsymbol{h} \in \mathcal{D}_{\mathrm{r}}} \boldsymbol{h} - \frac{1}{|\mathcal{D}_{\mathrm{c},k}|} \sum_{\boldsymbol{h} \in \mathcal{D}_{\mathrm{c},k}} \boldsymbol{h}. \tag{15}$$

Thus, the refusal mean is shared, while the compliance mean is region-specific.

Let $\mathbf{D}_{\mathrm{c},k} \in \mathbb{R}^{d \times n_k}$ stack the $n_k = |\mathcal{D}_{\mathrm{c},k}|$ local compliant activations, and let $\mathbf{R}_k = \boldsymbol{r}_k \mathbf{1}_{n_k}^\top$ repeat the target across samples, where $\mathbf{1}_{n_k}$ is the all-ones vector. We fit each map by minimizing reconstruction error while penalizing the norm of its effective transformation:

$$\boldsymbol{\Delta}_k^\star \in \underset{\boldsymbol{\Delta}_k \in \mathbb{R}^{d \times d}}{\arg\min} \left\{ \|\boldsymbol{\Delta}_k \mathbf{P}_k \mathbf{D}_{\mathrm{c},k} - \mathbf{R}_k\|_F^2 + \gamma \|\boldsymbol{\Delta}_k \mathbf{P}_k\|_F^2 \right\}, \tag{16}$$

where $\gamma > 0$ controls the regularization strength. This instantiates the reconstruction setting of Section 4.1 with $\mathbf{X}_k = \mathbf{P}_k \mathbf{D}_{\mathrm{c},k}$, adding regularization to control the map norm. We use the following closed-form minimizer:

$$\boldsymbol{\Delta}_k^\star = \mathbf{R}_k \mathbf{D}_{\mathrm{c},k}^\top \mathbf{P}_k \left[\mathbf{P}_k \mathbf{D}_{\mathrm{c},k} \mathbf{D}_{\mathrm{c},k}^\top \mathbf{P}_k + \gamma \mathbf{P}_k \right]^+. \tag{17}$$

Appendix D.5 derives this solution. The fitted maps are applied through Equation 14.

5 Experiments

5.1 Experimental Setup

We evaluate K-Shield on three instruction-tuned open-source LLMs: Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-7B-Instruct (Qwen et al., 2025), and Gemma-2-9B-IT (Team et al., 2024). To assess safety, we test against seven representative jailbreak attacks,

Table 2: Performance on utility benchmarks. Best results are **bold**. Baselines are reproduced using official implementations and released settings.

Model	XSTest CR % ↑	AlpacaEval WR % ↑	MATH Acc % ↑	GSM8K Acc % ↑	Utility Score % ↑
Llama-3.1-8B-Instruct	92.0	50.0	52.0	88.0	70.5
+ Jailbreak Antidote (Shen et al., 2025)	86.0	45.1	45.0	81.0	64.3
+ Surgical (Wang et al., 2025b)	75.6	47.3	47.0	80.0	62.5
+ CAST (Lee et al., 2025a)	90.0	37.8	12.0	25.0	41.2
+ Circuit Breaker (Zou et al., 2024)	84.8	40.3	32.0	50.0	51.8
+ AlphaSteer (Sheng et al., 2026)	**92.4**	45.6	**54.0**	88.0	70.0
+ RV (Ablation)	5.2	17.2	35.0	64.0	30.4
+ **K-Shield (Ours)**	91.6	**49.3**	51.0	**92.0**	**71.0**
Qwen2.5-7B-Instruct	97.2	50.0	76.0	97.0	80.1
+ Jailbreak Antidote (Shen et al., 2025)	80.0	35.8	60.0	80.0	64.0
+ Surgical (Wang et al., 2025b)	74.0	32.1	51.0	72.0	57.3
+ CAST (Lee et al., 2025a)	88.8	32.8	15.0	17.0	38.4
+ Circuit Breaker (Zou et al., 2024)	75.6	31.1	28.0	30.0	41.2
+ AlphaSteer (Sheng et al., 2026)	**95.2**	44.0	72.0	94.0	76.3
+ RV (Ablation)	68.0	4.7	2.0	3.0	19.4
+ **K-Shield (Ours)**	93.6	**49.5**	**73.0**	**97.0**	**78.3**
Gemma-2-9B-IT	82.0	50.0	55.0	94.0	70.3
+ Jailbreak Antidote (Shen et al., 2025)	74.8	32.6	41.0	65.0	53.4
+ Surgical (Wang et al., 2025b)	**86.0**	41.2	43.0	75.0	61.3
+ CAST (Lee et al., 2025a)	76.8	35.3	7.0	15.0	33.5
+ Circuit Breaker (Zou et al., 2024)	84.0	45.0	42.0	75.0	61.5
+ AlphaSteer (Sheng et al., 2026)	54.0	38.0	51.0	88.0	57.8
+ RV (Ablation)	4.0	3.8	1.0	2.0	2.7
+ **K-Shield (Ours)**	76.4	**48.7**	**54.0**	**90.0**	**67.3**

including AIM[2], AutoDAN (Liu et al., 2024), Cipher (Yuan et al., 2024), GCG (Zou et al., 2023), Jailbroken (Wei et al., 2023), PAIR (Chao et al., 2025), and ReNeLLM (Ding et al., 2024). Safety performance is measured using the Defense Success Rate (DSR), defined as the percentage of malicious prompts correctly refused.

To evaluate utility preservation, we consider four benchmarks covering complementary aspects of model behavior. Instruction-following ability is measured on **AlpacaEval** (Dubois et al., 2024) using Win Rate (WR) against the original unsteered model. Over-refusal is assessed on the safe subset of **XSTest** (Röttger et al., 2024) via the Compliance Rate (CR), which captures correct responses to benign yet sensitive prompts. Reasoning performance is evaluated on **MATH** (Hendrycks et al., 2021) and **GSM8K** (Cobbe et al., 2021) using accuracy.

We compare K-Shield against five representative baselines, including vector calibration methods (Jailbreak Antidote (Shen et al., 2025) and Surgical (Wang et al., 2025b)), a conditional steering approach (CAST (Lee et al., 2025a)), a learnable null-space-constrained method (AlphaSteer (Sheng et al., 2026)), and a representation rerouting method (Circuit Breaker (Zou et al., 2024)). We also include an ablation baseline that directly applies the refusal vector (RV), representing the upper bound on safety without utility constraints.

5.2 Results and Analysis

Global null-space constraint can restrict effective steering. Comparing with AlphaSteer provides a direct assessment of replacing a global null-space constraint with local ones in refusal steering. K-Shield improves average DSR by 5.94, 19.51, and 23.12 percentage points on Llama, Qwen, and Gemma, respectively, while also increasing aggregate utility on each model (Tables 1 and 2). These simultaneous improvements indicate that stronger defense does not necessarily require ac-

[2] https://oxtia.com/chatgpt-jailbreak-prompts/aim-prompt/

Table 3: Controlled projector comparison on Gemma-2-9B-IT (means over three seeds). $\overline{\mathcal{E}}_0$ estimates the normalized minimum reconstruction error without regularization; $L_{\gamma,\mathrm{train/test}}$ report reconstruction errors of regularized maps, excluding the penalty. B_{test} measures relative benign intervention energy. See Appendix E.1 for definitions.

		Reconstruction error ↓			Benign perturbation ↓	DSR ↑	Utility ↑
Projector	Steering maps	$\overline{\mathcal{E}}_0$	$L_{\gamma,\mathrm{train}}$	$L_{\gamma,\mathrm{test}}$	B_{test}	(%)	(%)
Global	Shared (g)	0.420	0.431	0.448	0.0042	78.6	65.5
Global	Independent (split)	0.392	0.408	0.436	0.0040	83.4	65.7
Local	Independent (loc)	0.158	0.167	0.189	0.0032	97.6	67.0
Paired projector gain $\overline{\mathcal{E}}_{0,\mathrm{split}} - \overline{\mathcal{E}}_{0,\mathrm{loc}}$		0.234					

cepting the utility loss incurred by existing interventions. At the evaluated configurations, the global method leaves room for improvement in both objectives.

The attack-level results further reveal that the gains are concentrated in the weaknesses of the global method. For example, AlphaSteer's Cipher DSR ranges from 48% to 68% across the three models, whereas K-Shield achieves 99–100%. On Qwen, substantial improvements also occur on AIM and ReNeLLM. These patterns are consistent with the motivation that a global null-space projector can suppress input structure needed for some refusal targets, and that local interventions can recover useful distinctions.

The gains extend beyond indiscriminate refusal. The RV ablation distinguishes selective steering from simply increasing refusal. RV achieves near-perfect defense, but its aggregate utility collapses and its XSTest compliance drops to 5.2% on Llama and 4.0% on Gemma. K-Shield attains similar defense while largely preserving benign compliance and reasoning. The utility benchmarks also reveal failures that a single metric would miss: on Qwen, RV retains 68.0% XSTest compliance but reaches only 2.0% on MATH and 3.0% on GSM8K. Similarly, CAST keeps high XSTest compliance while its reasoning accuracy falls to 7–25% across models. Answering benign prompts and answering them correctly are thus distinct requirements, and K-Shield largely preserves both.

6 Ablation Studies

Unless stated otherwise, ablations use Llama-3.1-8B-Instruct.

6.1 Isolating the Contribution of Local Projectors

K-Shield adapts both the null-space projector $\mathbf{P}_k$ and the steering map $\mathbf{\Delta}_k$ to each region k. Separate maps allow each region to fit its refusal targets independently, relaxing the requirement that one linear transformation serve all regions. Appendix B.1 decomposes the reconstruction-error reduction into the contribution of independent maps and the additional effect of local projectors. We evaluate these contributions on Gemma-2-9B-IT using three projector–map configurations: global/shared, global/independent, and local/independent. Comparing the latter two isolates the effect of local projectors while retaining independent maps in both. Across all configurations, we fix the partition, common refusal target, energy threshold, regularization, steering strength, and steered layers. The global projector uses our cumulative-energy selection rule, making it a controlled comparator rather than a reproduction of AlphaSteer (Sheng et al., 2026).

Table 3 shows that independent maps reduce the normalized minimum reconstruction error from 0.420 to 0.392, while local projectors reduce it further to 0.158. This numerical estimate of Equation 4 uses SVD with relative tolerance 10^{-4}; Appendix E.2 examines tolerance sensitivity. Local projectors account for approximately 89% of the total reduction in this comparison.

The benefit extends to regularized maps evaluated without refitting on held-out activations: reconstruction error falls from 0.436 to 0.189. DSR increases from 83.4% to 97.6% and utility from 65.7% to 67.0%, while relative benign intervention energy decreases from 0.0040 to 0.0032. These results support a contribution from local projectors beyond separate maps alone. Appendix E.1 gives the measurement protocol.

Table 4: Assignment structure on Llama-3.1-8B-Instruct. Reconstruction objectives are normalized by K-Shield's. DSR is mean ± standard deviation over ten restarts for the subspace methods; utility is reported as a mean. K-means is included as a centroid-based reference.

Method	Assignment	Balanced	Recon. obj. ↓	Avg DSR % ↑	Utility % ↑
K-means + local steering	—	—	—	69.46	69.9
K-subspaces	Hard	✗	**0.94**	96.1 ± 2.8	69.8
+ balance	Hard	✓	1.02	97.5 ± 0.9	70.4
+ soft assignment	Soft	✗	0.97	97.8 ± 0.6	70.6
K-Shield (Ours)	Soft	✓	1.00	**99.37** ± 0.2	**71.0**

6.2 Subspace Learning and Assignment Structure

We compare OT-based learning with centroid-based K-means and hard, unbalanced K-subspaces, retaining local steering downstream. K-subspaces uses the same subspace-reconstruction cost as K-Shield; intermediate variants add balance or soft assignments individually. Table 4 therefore distinguishes the choice of clustering geometry from the effects of assignment structure.

K-subspaces achieves 96.1% DSR, compared with 69.46% for K-means. Adding balance or softness individually raises DSR to 97.5% and 97.8%; combining both reaches 99.37%, with utility increasing from 69.8% to 71.0%. Across ten restarts, DSR standard deviation falls from 2.8 to 0.2 percentage points. Although K-subspaces attains a lower benign reconstruction objective, it performs worse downstream. These results motivate soft, balanced subspace learning through its observed defense performance and stability. Protocols and detailed K-means results appear in Appendices E.3 and E.4.

6.3 Input-Dependent Routing

Replacing nearest-subspace routing with random cluster selection, while keeping learned projectors and maps fixed, reduces DSR from 99.37% to 96.37% and utility from 71.0% to 56.1%. XSTest compliance falls by 24.4 percentage points. The larger utility loss supports matching each input to its local projector and map, rather than merely learning multiple steering transformations. Appendix E.5 reports the benchmark-level results.

Sensitivity analyses for the number of subspaces K and steering strength λ appear in Appendices E.6 and E.7.

7 Conclusion

This work highlights the role of projection in null-space constrained activation steering. Our analysis characterizes how projected input geometry governs refusal target reconstruction and establishes that local preservation constraints can reduce approximation error beyond the benefit of independent steering maps. Controlled ablations support the contribution of local projectors, while experiments across three LLMs demonstrate improved defense success and aggregate utility over global null-space constrained steering. Together, these findings motivate local null-space constraints for more selective behavioral control while preserving benign capabilities.

8 Limitations

Our evaluation covers three language models and seven jailbreak attacks; generalization to broader prompt distributions and adaptive attacks requires further study. Our theoretical results concern activation-level reconstruction and perturbation; downstream safety and utility are assessed empirically. Finally, maintaining multiple projectors and steering maps adds storage and routing overhead relative to a global model.

AI USE STATEMENT

We used generative AI tools to assist with drafting, paraphrasing, polishing the paper, and writing some implementation code. The authors reviewed and edited the resulting materials. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 4 specifies the method and its fitting objectives. Appendix F documents training data, activation extraction, model-specific hyperparameters, optimization settings, and evaluation procedures. Additional ablation protocols are provided in Appendix E. The evaluation judge prompts are reproduced in Figure 4, with attribution to AlphaSteer (Sheng et al., 2026).

REFERENCES

Mohammed Al-Garadi, Tushar Mungle, Abdulaziz Ahmed, Abeed Sarker, Zhuqi Miao, and Michael E. Matheny. Large language models in healthcare. 2 2025. URL https://arxiv.org/pdf/2503.04748.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37, 6 2024. ISSN 10495258. URL https://arxiv.org/pdf/2406.11717.

Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegreffe, Nouha Dziri, Khyathi Chandu, Jack Hessel, Yulia Tsvetkov, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. The art of saying no: Contextual noncompliance in language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=f1UL4wNlw6.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37, 3 2024. ISSN 10495258. doi: 10.52202/079017-1745. URL https://arxiv.org/pdf/2404.01318.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking Black Box Large Language Models in Twenty Queries . In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 23–42, Los Alamitos, CA, USA, April 2025. IEEE Computer Society. doi: 10.1109/SaTML64287.2025.00010. URL https://doi.ieeecomputersociety.org/10.1109/SaTML64287.2025.00010.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Jiaxi Cui, Munan Ning, Zongjian Li, Hao Li, Yang Ya, Bohua Chen, Bin Ling, Yonghong Tian, and Li Yuan. Chatlaw: A multi-agent legal assistant based on a role-aligned mixture-of-experts architecture. *Fundamental Research*, 2026. ISSN 2667-3258. doi: https://doi.org/10.1016/j.fmre.2026.03.026. URL https://www.sciencedirect.com/science/article/pii/S2667325826004048.

Peng Ding, Jun Kuang, Dan Ma, Xuezhi Cao, Yunsen Xian, Jiajun Chen, and Shujian Huang. A wolf in sheep's clothing: Generalized nested jailbreak prompts can fool large language models easily. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2136–2153, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.118. URL https://aclanthology.org/2024.naacl-long.118/.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=CybBmzWBX0.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Alphaedit: Null-space constrained knowledge editing for language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 16366–16396, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/29c8c615b3187ee995029284702d3f43-Paper-Conference.pdf.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. *Advances in Neural Information Processing Systems*, 29:3440–3448, 2016. URL https://arxiv.org/abs/1605.08527.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. 7 2024. URL https://arxiv.org/pdf/2407.21783.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *Advances in Neural Information Processing Systems*, 37, 6 2024. ISSN 10495258. doi: 10.52202/079017-0261. URL https://arxiv.org/pdf/2406.18495.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *Advances in Neural Information Processing Systems*, 3 2021. ISSN 10495258. URL https://arxiv.org/pdf/2103.03874.

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. Catastrophic jailbreak of open-source LLMs via exploiting generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=r42tSSCHPh.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. A survey on large language models for code generation. *ACM Trans. Softw. Eng. Methodol.*, 35(2), January 2026. ISSN 1049-331X. doi: 10.1145/3747588. URL https://doi.org/10.1145/3747588.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=Oi47wc10sm.

Jean Lee, Nicholas Stevens, and Soyeon Caren Han. Large language models in finance (finllms). *Neural Computing and Applications*, 37(30):24853–24867, January 2025b. ISSN 1433-3058. doi: 10.1007/s00521-024-10495-6. URL http://dx.doi.org/10.1007/s00521-024-10495-6.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7Jwpw4qKkb.

Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. Jailbreakv: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. 4 2024. URL https://arxiv.org/pdf/2404.03027.

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

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *Proceedings of Machine Learning Research*, 235:35181–35224, 2 2024. ISSN 26403498. URL https://arxiv.org/pdf/2402.04249.

Ha-Thanh Nguyen. A brief report on lawgpt 1.0: A virtual legal assistant based on gpt-3. 2 2023. URL https://arxiv.org/pdf/2302.05729.

Kyle O'Brien, David Majercak, Xavier Fernandes, Richard Edgar, Blake Bullwinkel, Jingya Chen, Harsha Nori, Dean Carignan, Eric Horvitz, and Forough Poursabzi-Sangdeh. Steering language model refusal with sparse autoencoders. 11 2024. URL https://arxiv.org/pdf/2411.11296.

OpenAI, Josh Achiam, Steven Adler, et al. Gpt-4 technical report. 3 2023. URL https://arxiv.org/pdf/2303.08774.

Wenbo Pan, Zhichao Liu, Qiguang Chen, Xiangyang Zhou, Yu Haining, and Xiaohua Jia. The hidden dimensions of LLM alignment: A multi-dimensional analysis of orthogonal safety directions. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=wGFEzfhFae.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

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

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/.

Guobin Shen, Dongcheng Zhao, Yiting Dong, Xiang He, and Yi Zeng. Jailbreak antidote: Runtime safety-utility balance via sparse representation adjustment in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=s20W12XTF8.

Leheng Sheng, Changshuo Shen, Weixiang Zhao, Junfeng Fang, Xiaohao Liu, Zhenkai Liang, Xiang Wang, An Zhang, and Tat-Seng Chua. Alphasteer: Learning refusal steering with principled null-space constraint. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=1vvbzAqdTe.

Wei Sun, Tingyu Qu, Mingxiao Li, Jesse Davis, and Marie-Francine Moens. Mitigating negative interference in multilingual knowledge editing through null-space constraints. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8796–8810, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.460. URL https://aclanthology.org/2025.findings-acl.460/.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Han Wang, Gang Wang, and Huan Zhang. Steering away from harm: An adaptive approach to defending vision language model against jailbreaks. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29947–29957. IEEE, 2025a.

Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. Training networks in null space of feature covariance for continual learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 184–193, 3 2021. ISSN 10636919. doi: 10.1109/CVPR46437.2021.00025. URL https://arxiv.org/pdf/2103.07113.

Xinpeng Wang, Chengzhi Hu, Paul Röttger, and Barbara Plank. Surgical, cheap, and flexible: Mitigating false refusal in language models via single vector ablation. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=SCBn8MCLwc.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=mRieQgMtNTQ.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36, 7 2023. ISSN 10495258. URL https://arxiv.org/pdf/2307.02483.

Zhihao Xu, Ruixuan HUANG, Changyu Chen, and Xiting Wang. Uncovering safety risks of large language models through concept activation vector. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=Uymv9ThB50.

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. GPT-4 is too smart to be safe: Stealthy chat with LLMs via cipher. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=MbfAK4s61A.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. 7 2023. URL https://arxiv.org/pdf/2307.15043.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. *Advances in Neural Information Processing Systems*, 37, 6 2024. ISSN 10495258. URL https://arxiv.org/pdf/2406.04313.

A Additional Related Work

Refusal Directions and Benign Utility. Prior work identifies a linear direction associated with refusal in aligned language models (Arditi et al., 2024). Difference-in-means methods estimate behavioral directions from contrasting activation sets (Marks & Tegmark, 2024). Injecting a refusal direction can improve defense against jailbreaks, but applying it indiscriminately also changes benign activations and can induce unwanted refusals or degrade task performance. The refusal-vector baseline in Tables 1 and 2 illustrates this trade-off: near-perfect defense accompanies substantially reduced utility.

Vector Calibration. Calibration methods refine the refusal direction by identifying and modifying directions associated with different behavioral effects (Shen et al., 2025; Wang et al., 2025b; Pan et al., 2025). Approaches include extracting dominant directions from a refusal subspace (Pan et al., 2025; Shen et al., 2025) and removing directions associated with false refusals (Wang et al., 2025b). The resulting intervention is intended to affect malicious inputs more selectively, even when applied across inputs. K-Shield instead learns regional linear maps whose inputs are filtered by locally constructed projectors.

Conditional Steering. Conditional methods determine whether or how strongly to intervene based on the input (Lee et al., 2025a; Wang et al., 2025a; O'Brien et al., 2024). They draw on differences between benign and malicious activation patterns (Xu et al., 2024), using similarity scores or thresholds to determine when to apply steering (Lee et al., 2025a; Wang et al., 2025a). K-Shield's routing selects a regional projector and steering map rather than gating a single refusal direction. Its intervention therefore depends on both the selected map and the activation retained by that region's projector.

B Derivation of the Minimum Target Reconstruction Error

We prove Theorem 1. Let $\mathbf{X}, \mathbf{R} \in \mathbb{R}^{d\times n}$ and define $\mathbf{\Pi} = \mathbf{X}^{+}\mathbf{X} \in \mathbb{R}^{n\times n}$. By the properties of the Moore-Penrose pseudoinverse, $\mathbf{\Pi}$ is the orthogonal projector onto the row space of $\mathbf{X}$. In particular,

$$\mathbf{\Pi}^{\top} = \mathbf{\Pi}, \qquad \mathbf{\Pi}^{2} = \mathbf{\Pi}, \qquad \mathbf{X}\mathbf{\Pi} = \mathbf{X}.$$

Achievable predictions. Each row of $\mathbf{\Delta X}$ is a linear combination of the rows of $\mathbf{X}$. Conversely, if $\mathbf{Y} \in \mathbb{R}^{d\times n}$ satisfies $\mathrm{row}(\mathbf{Y}) \subseteq \mathrm{row}(\mathbf{X})$, then $\mathbf{Y\Pi} = \mathbf{Y}$, and choosing $\mathbf{\Delta} = \mathbf{YX}^{+}$ gives $\mathbf{\Delta X} = \mathbf{Y}$. Therefore,

$$\{\mathbf{\Delta X} : \mathbf{\Delta} \in \mathbb{R}^{d\times d}\} = \{\mathbf{Y} \in \mathbb{R}^{d\times n} : \mathrm{row}(\mathbf{Y}) \subseteq \mathrm{row}(\mathbf{X})\}. \tag{18}$$

Orthogonal decomposition of the error. For any $\mathbf{\Delta} \in \mathbb{R}^{d\times d}$,

$$\mathbf{\Delta X} - \mathbf{R} = \big(\mathbf{\Delta X} - \mathbf{R\Pi}\big) - \mathbf{R}(\mathbf{I}_n - \mathbf{\Pi}).$$

Every row of the first term belongs to $\mathrm{row}(\mathbf{X})$, while every row of the second belongs to its orthogonal complement. Their Frobenius inner product is therefore zero, giving

$$\|\mathbf{\Delta X} - \mathbf{R}\|_F^2 = \|\mathbf{\Delta X} - \mathbf{R\Pi}\|_F^2 + \|\mathbf{R}(\mathbf{I}_n - \mathbf{\Pi})\|_F^2. \tag{19}$$

The second term does not depend on $\mathbf{\Delta}$. The first is nonnegative and becomes zero for $\mathbf{\Delta}^{\star} = \mathbf{RX}^{+}$. Consequently,

$$\min_{\mathbf{\Delta}\in\mathbb{R}^{d\times d}} \|\mathbf{\Delta X} - \mathbf{R}\|_F^2 = \|\mathbf{R}(\mathbf{I}_n - \mathbf{X}^{+}\mathbf{X})\|_F^2, \tag{20}$$

which proves Equation 4. The optimal prediction is uniquely $\mathbf{RX}^{+}\mathbf{X}$, although the minimizing map need not be unique.

Interpretation as constraints across inputs. Exact reconstruction is possible if and only if

$$\mathrm{row}(\mathbf{R}) \subseteq \mathrm{row}(\mathbf{X}),$$

or, equivalently, every vector $\boldsymbol{c} \in \mathbb{R}^n$ satisfying $\mathbf{X}\boldsymbol{c} = \mathbf{0}$ also satisfies $\mathbf{R}\boldsymbol{c} = \mathbf{0}$. Thus, every linear dependency among the projected inputs must also hold among their targets. Projection can introduce additional dependencies because

$$\ker(\mathbf{H}_{\mathrm{m}}) \subseteq \ker(\mathbf{PH}_{\mathrm{m}}).$$

An introduced dependency prevents exact reconstruction when the targets do not satisfy it. This explains how removing input directions can increase the minimum reconstruction error.

The result concerns unrestricted linear maps without regularization. Regularization can increase the reconstruction error by penalizing maps with large norm, and held-out performance is not determined by this training-set characterization alone.

B.1 Separating the Contributions of Regional Maps and Local Projectors

This subsection formalizes the controlled comparison in Section 6.1. We distinguish the reduction in target approximation error obtained by fitting independent regional maps from the additional effect of adapting their projectors. All quantities refer to a single steered layer.

Common partition and targets. Fix a partition of the malicious activations, $\mathbf{H}_{\mathrm{m}} = [\mathbf{H}_{\mathrm{m}}^{(1)}, \ldots, \mathbf{H}_{\mathrm{m}}^{(K)}]$, where $\mathbf{H}_{\mathrm{m}}^{(k)} \in \mathbb{R}^{d \times n_k}$, and prescribed target blocks $\mathbf{R}_k \in \mathbb{R}^{d \times n_k}$. Let $\mathbf{P}_{\mathrm{g}}$ and $\mathbf{P}_k$ be the global and regional orthogonal projectors, respectively. The partition and targets remain unchanged across all comparisons. In particular, the targets may repeat a common refusal vector, $\mathbf{R}_k = \boldsymbol{r}\mathbf{1}_{n_k}^{\top}$, or use fixed regional vectors, $\mathbf{R}_k = \boldsymbol{r}_k\mathbf{1}_{n_k}^{\top}$. Keeping them fixed separates the effects of maps and projectors from target selection.

Three approximation errors. We compare a global projector with a shared map, a global projector with independent regional maps, and local projectors with independent regional maps:

$$\mathcal{E}_{\mathrm{g}} = \min_{\boldsymbol{\Delta} \in \mathbb{R}^{d \times d}} \sum_{k=1}^{K} \|\boldsymbol{\Delta}\mathbf{P}_{\mathrm{g}}\mathbf{H}_{\mathrm{m}}^{(k)} - \mathbf{R}_k\|_F^2, \tag{21}$$

$$\mathcal{E}_{\mathrm{split}} = \sum_{k=1}^{K} \min_{\boldsymbol{\Delta}_k \in \mathbb{R}^{d \times d}} \|\boldsymbol{\Delta}_k\mathbf{P}_{\mathrm{g}}\mathbf{H}_{\mathrm{m}}^{(k)} - \mathbf{R}_k\|_F^2, \tag{22}$$

$$\mathcal{E}_{\mathrm{loc}} = \sum_{k=1}^{K} \min_{\boldsymbol{\Delta}_k \in \mathbb{R}^{d \times d}} \|\boldsymbol{\Delta}_k\mathbf{P}_k\mathbf{H}_{\mathrm{m}}^{(k)} - \mathbf{R}_k\|_F^2. \tag{23}$$

These are minimum unregularized reconstruction errors, as characterized by Theorem 1. The global configuration fits the concatenated inputs and targets jointly, whereas the other two configurations fit each region independently.

Decomposition and interpretation. Adding and subtracting $\mathcal{E}_{\mathrm{split}}$ gives

$$\mathcal{E}_{\mathrm{g}} - \mathcal{E}_{\mathrm{loc}} = \underbrace{\mathcal{E}_{\mathrm{g}} - \mathcal{E}_{\mathrm{split}}}_{\text{benefit of independent maps}} + \underbrace{\mathcal{E}_{\mathrm{split}} - \mathcal{E}_{\mathrm{loc}}}_{\text{effect of local projectors}}. \tag{24}$$

The first term is nonnegative. For any shared map $\boldsymbol{\Delta}$, setting every $\boldsymbol{\Delta}_k = \boldsymbol{\Delta}$ reproduces its predictions. Minimizing independently within each region can therefore only maintain or reduce the total error.

The second term compares configurations that both use independent maps and differ only in their projectors. It measures the change in optimal reconstruction error due to changing the projected inputs available to those maps. A positive value identifies a reduction beyond the benefit of independent maps alone. This term has no fixed sign for arbitrary global and local projectors: changing the protected directions can retain some activation information and remove other information.

The constructive result in Proposition 2 separates the two contributions by giving $\mathcal{E}_{\mathrm{g}} = \mathcal{E}_{\mathrm{split}} > 0$ and $\mathcal{E}_{\mathrm{loc}} = 0$, including with a common refusal target. Section 6.1 evaluates the contributions for learned projectors and separately reports held-out reconstruction using regularized maps. Those held-out errors assess the fitted maps without refitting; they are distinct from the unregularized minima defined above.

C A Strict Approximation Benefit of Local Projectors

Proposition 2 (Strict approximation benefit of local projectors)**.** *For every $\alpha \in (1/2, 1)$ and prescribed nonzero targets $\boldsymbol{r}_1, \boldsymbol{r}_2 \in \mathbb{R}^3$, there exist benign activations and two malicious activations per region such that global and local projectors selected at threshold α satisfy*

$$\mathcal{E}_{\mathrm{g}} = \mathcal{E}_{\mathrm{split}} = 2\|\boldsymbol{r}_1\|_2^2 + 2\|\boldsymbol{r}_2\|_2^2 > 0, \qquad \mathcal{E}_{\mathrm{loc}} = 0.$$

All three configurations use the same target blocks $\mathbf{R}_k = [\boldsymbol{r}_k, \boldsymbol{r}_k]$. Nearest-principal-subspace routing selects the intended regions, and the local interventions leave their respective benign activations unchanged. The result also holds with $\boldsymbol{r}_1 = \boldsymbol{r}_2$.

Proof. Let $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ denote the standard basis of $\mathbb{R}^3$. Fix $\alpha \in (1/2, 1)$ and nonzero targets $\boldsymbol{r}_1, \boldsymbol{r}_2 \in \mathbb{R}^3$.

Benign geometry and cumulative-energy selection. Define the benign activation matrices

$$\boldsymbol{D}_1 = \boldsymbol{e}_1, \qquad \boldsymbol{D}_2 = \boldsymbol{e}_2, \qquad \boldsymbol{H}_{\mathrm{b}} = [\boldsymbol{e}_1, \boldsymbol{e}_2].$$

The global scatter matrix is

$$\boldsymbol{H}_{\mathrm{b}}\boldsymbol{H}_{\mathrm{b}}^\top = \mathrm{diag}(1, 1, 0).$$

Because $\alpha > 1/2$, retaining one eigen-direction is insufficient to capture the prescribed fraction of energy. Cumulative-energy selection therefore gives

$$\mathcal{U}_{\mathrm{g}} = \mathrm{span}\{\boldsymbol{e}_1, \boldsymbol{e}_2\}, \qquad \boldsymbol{P}_{\mathrm{g}} = \boldsymbol{e}_3\boldsymbol{e}_3^\top.$$

Each local scatter matrix has one nonzero eigenvalue, so the same threshold gives

$$\mathcal{U}_1 = \mathrm{span}\{\boldsymbol{e}_1\}, \qquad \mathcal{U}_2 = \mathrm{span}\{\boldsymbol{e}_2\},$$

with residual projectors

$$\boldsymbol{P}_1 = \boldsymbol{I}_3 - \boldsymbol{e}_1\boldsymbol{e}_1^\top, \qquad \boldsymbol{P}_2 = \boldsymbol{I}_3 - \boldsymbol{e}_2\boldsymbol{e}_2^\top.$$

Thus, the global residual projector is nonzero, and

$$\boldsymbol{P}_{\mathrm{g}}\boldsymbol{H}_{\mathrm{b}} = \boldsymbol{0}, \qquad \boldsymbol{P}_1\boldsymbol{D}_1 = \boldsymbol{0}, \qquad \boldsymbol{P}_2\boldsymbol{D}_2 = \boldsymbol{0}.$$

Malicious activations and prescribed targets. Construct two malicious activations in each region:

$$\boldsymbol{h}_{1,+} = 2\boldsymbol{e}_1 + \boldsymbol{e}_2 + \boldsymbol{e}_3, \quad \boldsymbol{h}_{1,-} = 2\boldsymbol{e}_1 + \boldsymbol{e}_2 - \boldsymbol{e}_3,$$
$$\boldsymbol{h}_{2,+} = \boldsymbol{e}_1 + 2\boldsymbol{e}_2 + \boldsymbol{e}_3, \quad \boldsymbol{h}_{2,-} = \boldsymbol{e}_1 + 2\boldsymbol{e}_2 - \boldsymbol{e}_3.$$

Let

$$\boldsymbol{H}_{\mathrm{m}}^{(k)} = [\boldsymbol{h}_{k,+}, \boldsymbol{h}_{k,-}], \qquad \boldsymbol{R}_k = [\boldsymbol{r}_k, \boldsymbol{r}_k], \qquad k \in \{1, 2\},$$

and concatenate these blocks to form

$$\boldsymbol{H}_{\mathrm{m}} = [\boldsymbol{H}_{\mathrm{m}}^{(1)}, \boldsymbol{H}_{\mathrm{m}}^{(2)}], \qquad \boldsymbol{R} = [\boldsymbol{R}_1, \boldsymbol{R}_2].$$

All configurations are evaluated against these same targets.

Independent maps with the global projector. For both regions,

$$\boldsymbol{P}_{\mathrm{g}}\boldsymbol{h}_{k,+} = \boldsymbol{e}_3, \qquad \boldsymbol{P}_{\mathrm{g}}\boldsymbol{h}_{k,-} = -\boldsymbol{e}_3.$$

Consequently, writing $\boldsymbol{t}_k = \boldsymbol{\Delta}_k\boldsymbol{e}_3$, the regional reconstruction error satisfies

$$\begin{aligned}\|\boldsymbol{\Delta}_k\boldsymbol{P}_{\mathrm{g}}\boldsymbol{H}_{\mathrm{m}}^{(k)} - \boldsymbol{R}_k\|_F^2 &= \|\boldsymbol{t}_k - \boldsymbol{r}_k\|_2^2 + \| - \boldsymbol{t}_k - \boldsymbol{r}_k\|_2^2 \\ &= 2\|\boldsymbol{t}_k\|_2^2 + 2\|\boldsymbol{r}_k\|_2^2.\end{aligned}$$

The minimum is attained at $\boldsymbol{t}_k = \boldsymbol{0}$, yielding

$$\mathcal{E}_{\mathrm{split}} = 2\|\boldsymbol{r}_1\|_2^2 + 2\|\boldsymbol{r}_2\|_2^2. \tag{25}$$

For a shared map, $\boldsymbol{t} = \boldsymbol{\Delta}\boldsymbol{e}_3$ must be the same in both regions. Its total error is

$$4\|\boldsymbol{t}\|_2^2 + 2\|\boldsymbol{r}_1\|_2^2 + 2\|\boldsymbol{r}_2\|_2^2,$$

which is also minimized at $\boldsymbol{t} = \boldsymbol{0}$. Therefore,

$$\mathcal{E}_{\mathrm{g}} = \mathcal{E}_{\mathrm{split}} = 2\|\boldsymbol{r}_1\|_2^2 + 2\|\boldsymbol{r}_2\|_2^2 > 0. \tag{26}$$

The obstruction thus persists even after allowing a separate map in each region.

Verification of routing. Nearest-principal-subspace routing selects region 1 for both $\boldsymbol{h}_{1,+}$ and $\boldsymbol{h}_{1,-}$ because

$$\|\boldsymbol{P}_1\boldsymbol{h}_{1,\pm}\|_2^2 = \|\boldsymbol{e}_2 \pm \boldsymbol{e}_3\|_2^2 = 2 < 5 = \|2\boldsymbol{e}_1 \pm \boldsymbol{e}_3\|_2^2 = \|\boldsymbol{P}_2\boldsymbol{h}_{1,\pm}\|_2^2.$$

Similarly,

$$\|\boldsymbol{P}_2\boldsymbol{h}_{2,\pm}\|_2^2 = 2 < 5 = \|\boldsymbol{P}_1\boldsymbol{h}_{2,\pm}\|_2^2,$$

so both activations in the second pair are routed to region 2. All these assignments are strict. The benign activations are also routed correctly:

$$\|\boldsymbol{P}_1\boldsymbol{e}_1\|_2^2 = 0 < 1 = \|\boldsymbol{P}_2\boldsymbol{e}_1\|_2^2,$$

and

$$\|\boldsymbol{P}_2\boldsymbol{e}_2\|_2^2 = 0 < 1 = \|\boldsymbol{P}_1\boldsymbol{e}_2\|_2^2.$$

Local reconstruction and benign preservation. The local projectors retain

$$\boldsymbol{P}_1\boldsymbol{h}_{1,\pm} = \boldsymbol{e}_2 \pm \boldsymbol{e}_3, \qquad \boldsymbol{P}_2\boldsymbol{h}_{2,\pm} = \boldsymbol{e}_1 \pm \boldsymbol{e}_3.$$

Choose

$$\boldsymbol{\Delta}_1 = \boldsymbol{r}_1\boldsymbol{e}_2^\top, \qquad \boldsymbol{\Delta}_2 = \boldsymbol{r}_2\boldsymbol{e}_1^\top.$$

These maps satisfy

$$\boldsymbol{\Delta}_k\boldsymbol{P}_k = \boldsymbol{\Delta}_k,$$

and reconstruct their respective targets:

$$\boldsymbol{\Delta}_1\boldsymbol{P}_1\boldsymbol{h}_{1,\pm} = \boldsymbol{r}_1, \qquad \boldsymbol{\Delta}_2\boldsymbol{P}_2\boldsymbol{h}_{2,\pm} = \boldsymbol{r}_2.$$

Hence,

$$\mathcal{E}_{\mathrm{loc}} = 0. \tag{27}$$

Moreover,

$$\boldsymbol{\Delta}_1\boldsymbol{P}_1\boldsymbol{D}_1 = \boldsymbol{0}, \qquad \boldsymbol{\Delta}_2\boldsymbol{P}_2\boldsymbol{D}_2 = \boldsymbol{0},$$

so the routed local interventions leave the benign activations unchanged.

Combining Equations 25–27 gives

$$\mathcal{E}_{\mathrm{g}} - \mathcal{E}_{\mathrm{split}} = 0, \qquad \mathcal{E}_{\mathrm{split}} - \mathcal{E}_{\mathrm{loc}} = 2\|\boldsymbol{r}_1\|_2^2 + 2\|\boldsymbol{r}_2\|_2^2 > 0.$$

The entire approximation improvement is therefore attributable to changing the projectors while retaining independent regional maps. □

Interpretation. The direction $\boldsymbol{e}_2$ carries benign energy in region 2 but supports refusal reconstruction in region 1; the roles reverse for $\boldsymbol{e}_1$. Global preservation removes both directions. Each local projector instead preserves its own benign direction while retaining the direction needed for its regional targets.

For a common refusal target $\boldsymbol{r}_1 = \boldsymbol{r}_2 = \boldsymbol{r} \neq \boldsymbol{0}$, the construction gives

$$\mathcal{E}_{\mathrm{g}} = \mathcal{E}_{\mathrm{split}} = 4\|\boldsymbol{r}\|_2^2, \qquad \mathcal{E}_{\mathrm{loc}} = 0.$$

Thus, the projector benefit does not require cluster-specific target vectors. The result concerns the optimal unregularized approximation errors defined in the main text. It establishes an architectural separation for prescribed targets and subspaces; it does not assume that the learning procedure necessarily recovers this construction.

D Optimal-Transport Formulation and Optimization

We provide the solver formulation and optimality characterization for the objective in Section 4.2. The orthonormal bases lie in the Stiefel manifold $\mathrm{St}(d,p) = \{\mathbf{W} \in \mathbb{R}^{d\times p} : \mathbf{W}^\top\mathbf{W} = \mathbf{I}_p\}$. Since the objective depends on each basis only through $\mathbf{W}_k\mathbf{W}_k^\top$, it is invariant to a change of orthonormal basis within the subspace and can be optimized on the Grassmann manifold. We write $[K] = \{1,\dots,K\}$.

Transport interpretation. The matrix $\mathbf{\Gamma} = \mathbf{Q}/N$ is the transport coupling: its entry $\Gamma_{ik} = q_{ik}/N$ is the joint probability mass assigned to activation i and subspace k. The constraints in Equation 6 give row sums $1/N$ and column sums $1/K$. The independent coupling has entries $1/(NK)$, so

$$\mathrm{KL}\left(\mathbf{\Gamma} \,\middle\|\, \frac{\mathbf{1}_N \mathbf{1}_K^\top}{NK}\right) = \frac{1}{N}\sum_{i,k} q_{ik}\log(Kq_{ik}),$$

where $\mathbf{1}_j$ is the vector of j ones. For $q_{ik} = 0$, the summand is defined as zero, using $\lim_{q\to 0^+} q\log(Kq) = 0$. Thus the primal objective is the transport reconstruction cost plus a relative-entropy penalty.

Derivation of the semi-dual objective. Following the entropic OT semi-dual approach of Genevay et al. (2016), we derive the formulation for our subspace reconstruction cost. Fix the bases $\mathbf{W}_{1:K}$ and abbreviate $c_{ik} = c_{ik}(\mathbf{W}_{1:K})$. Retain the nonnegativity and row-sum constraints, and introduce multipliers g_k for the balance constraints $N^{-1}\sum_i q_{ik} = 1/K$. The partial Lagrangian is

$$\mathcal{L}(\mathbf{Q}, \boldsymbol{g}) = \frac{1}{K}\sum_{k=1}^{K} g_k + \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left[q_{ik}(c_{ik} - g_k) + \epsilon q_{ik}\log(Kq_{ik})\right]. \tag{28}$$

For fixed $\boldsymbol{g}$, minimization separates across rows. For row i, define

$$Z_i = \sum_{j=1}^{K}\exp((g_j - c_{ij})/\epsilon), \qquad s_{ik} = \frac{\exp((g_k - c_{ik})/\epsilon)}{Z_i}.$$

The vector $\boldsymbol{s}_i = (s_{i1}, \ldots, s_{iK})^\top$ has strictly positive entries summing to one. For any admissible row vector $\boldsymbol{q}_i = (q_{i1}, \ldots, q_{iK})^\top$,

$$\sum_{k=1}^{K}\left[q_{ik}(c_{ik} - g_k) + \epsilon q_{ik}\log(Kq_{ik})\right] = \epsilon\,\mathrm{KL}(\boldsymbol{q}_i\|\boldsymbol{s}_i) - \epsilon\log(Z_i/K). \tag{29}$$

The KL divergence is nonnegative and vanishes precisely when $\boldsymbol{q}_i = \boldsymbol{s}_i$. Substituting these minimizing rows into Equation 28 gives exactly $J_\epsilon(\mathbf{W}_{1:K}, \boldsymbol{g})$ in Equation 9.

For fixed bases, the primal assignment problem is convex and has the strictly positive feasible assignment $q_{ik} = 1/K$. Strong duality therefore gives

$$\min_{\mathbf{Q}\in\mathcal{Q}} \mathcal{F}_\epsilon(\mathbf{W}_{1:K}, \mathbf{Q}) = \max_{\boldsymbol{g}\in\mathbb{R}^K} J_\epsilon(\mathbf{W}_{1:K}, \boldsymbol{g}).$$

Minimizing both sides over the orthonormal bases yields the semi-dual problem in Equation 8; this does not interchange the order of minimization and maximization. Each potential g_k adjusts the assignment weights received by subspace k, and the balance constraints hold at a dual optimum. Adding the same constant to every g_k changes neither the assignments nor the dual objective, so the potentials need not be unique.

Optimal assignments and subspace fits. The following proposition characterizes each variable block when the other is fixed.

Proposition 3 (Balanced assignments and weighted subspace fitting)**.** *Let $N, K \geq 1$, $1 \leq p \leq d$, $\epsilon > 0$, and $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$. For fixed $\mathbf{W}_{1:K}$, minimizing $\mathcal{F}_\epsilon$ over $\mathcal{Q}$ has a unique, strictly positive solution $\mathbf{Q}^\star$. For any maximizer $\boldsymbol{g}^\star$ of the semi-dual, its entries satisfy*

$$q^\star_{ik} = \frac{\exp((g^\star_k - c_{ik}(\mathbf{W}_{1:K}))/\epsilon)}{\sum_{j=1}^{K}\exp((g^\star_j - c_{ij}(\mathbf{W}_{1:K}))/\epsilon)}, \qquad \frac{1}{N}\sum_i q^\star_{ik} = \frac{1}{K}. \tag{30}$$

For fixed $\mathbf{Q} \in \mathcal{Q}$, a minimizing basis $\mathbf{W}_k$ spans a leading p-dimensional eigenspace of

$$\mathbf{S}_k(\mathbf{Q}) = \sum_{i=1}^{N} q_{ik}\boldsymbol{x}^{(i)}(\boldsymbol{x}^{(i)})^\top. \tag{31}$$

Every joint global minimizer of Equation 7 satisfies both characterizations.

Equation 30 assigns larger weights to better-fitting subspaces, with the dual potentials adjusting these weights to satisfy balance. Conversely, for fixed assignments, each subspace is obtained by weighted, uncentered PCA: its basis spans the leading eigenvectors of the $d \times d$ scatter matrix $\mathbf{S}_k(\mathbf{Q})$. Activations with larger assignment weights therefore contribute more strongly to the subspace fit. These are conditional optimality characterizations; they do not guarantee that joint optimization reaches a global optimum. The proof is provided in Appendix D.1.

Optimization. We approximately solve Equation 8 by alternating updates to the dual potentials $\boldsymbol{g}$ and the subspace bases. For fixed bases, J_ϵ is concave in $\boldsymbol{g}$, with

$$\frac{\partial J_\epsilon}{\partial g_k} = \frac{1}{K} - \frac{1}{N} \sum_i q_{ik}(\mathbf{W}_{1:K}, \boldsymbol{g}), \tag{32}$$

where $q_{ik}(\mathbf{W}_{1:K}, \boldsymbol{g})$ has the softmax form in Equation 30. Gradient ascent increases the potential of a subspace receiving too little assignment weight and decreases it when it receives too much. Exact balance holds at the dual optimum; finite optimization may leave a marginal discrepancy.

For orthonormal representatives, the Grassmann gradient is

$$\mathrm{grad}_{\mathbf{W}_k} J_\epsilon = (\mathbf{I}_d - \mathbf{W}_k \mathbf{W}_k^\top) \nabla_{\mathbf{W}_k} J_\epsilon. \tag{33}$$

We update the bases by a gradient descent step followed by a thin-QR retraction, which restores orthonormality. For the squared projection-residual cost in Equation 5, its Euclidean gradient at an orthonormal basis is already horizontal:

$$\nabla_{\mathbf{W}_k} c_{ik} = -2(\mathbf{I}_d - \mathbf{W}_k \mathbf{W}_k^\top) \boldsymbol{x}^{(i)} (\boldsymbol{x}^{(i)})^\top \mathbf{W}_k.$$

Since $\nabla_{\mathbf{W}_k} J_\epsilon = N^{-1} \sum_i q_{ik} \nabla_{\mathbf{W}_k} c_{ik}$, an explicit tangent projection is unnecessary at an orthonormal point for this cost expression. The implementation can therefore use its Euclidean gradient followed by QR retraction. These incremental updates differ from computing the exact weighted-PCA minimizer at every step. The joint problem is nonconvex, and the conditional characterizations in Proposition 3 do not guarantee convergence to a joint global optimum.

D.1 Proof of Proposition 3

Proof. We establish the assignment and subspace characterizations separately.

Existence, uniqueness, and positivity of assignments. For fixed W, the feasible set $\mathcal{Q}$ is nonempty and compact: the uniform assignment $q_{ik} = 1/K$ is feasible. The objective is continuous on $\mathcal{Q}$ under the convention $0 \log 0 = 0$, so a minimizer exists. Because $t \mapsto t \log t$ is strictly convex and $\epsilon > 0$, the objective is strictly convex in q. The minimizing assignment is therefore unique.

The minimizer is strictly positive. Indeed, if a feasible minimizer had a zero entry, mixing it with the strictly positive uniform assignment would remain feasible. For a sufficiently small mixing weight $t > 0$, the negative $t \log t$ contribution at zero entries would dominate the $O(t)$ changes in the remaining entropy and finite transport-cost terms, decreasing the objective. This contradicts optimality.

Semi-dual and optimal assignments. Introduce multipliers g_k for the column-mass constraints while retaining the row constraints $q_i \in \Delta_K = \{z \geq 0 : \sum_k z_k = 1\}$. The Lagrangian is

$$\mathcal{L}(q, g) = \mathcal{F}_\epsilon(W, q) + \sum_k g_k \left(\frac{1}{K} - \frac{1}{N} \sum_i q_{ik} \right) \tag{34}$$

$$= \frac{1}{K} \sum_k g_k + \frac{1}{N} \sum_i \left[\sum_k q_{ik}(c_{ik}(W) - g_k) + \epsilon \sum_k q_{ik} \log(K q_{ik}) \right]. \tag{35}$$

Minimizing each row over Δ_K yields

$$q_{ik}(W, g) = \frac{\exp((g_k - c_{ik}(W))/\epsilon)}{\sum_j \exp((g_j - c_{ij}(W))/\epsilon)} \tag{36}$$

and the row minimum

$$-\epsilon \log \left[\frac{1}{K} \sum_k \exp((g_k - c_{ik}(W))/\epsilon) \right]. \tag{37}$$

Substitution gives $J_\epsilon(W, g)$.

The strictly positive feasible assignment provides the relative-interior feasibility condition for convex duality. Hence strong duality holds, and optimal multipliers exist. Equivalently, the positive primal minimizer satisfies the equality-constrained first-order optimality conditions. Therefore,

$$\min_{q \in \mathcal{Q}} \mathcal{F}_\epsilon(W, q) = \max_{g \in \mathbb{R}^K} J_\epsilon(W, g). \tag{38}$$

Differentiating the semi-dual gives

$$\frac{\partial J_\epsilon}{\partial g_k} = \frac{1}{K} - \frac{1}{N} \sum_i q_{ik}(W, g). \tag{39}$$

At a dual maximizer this gradient vanishes, so the softmax assignments satisfy the required column masses. Together with their row normalization, they belong to $\mathcal{Q}$ and recover its unique primal minimizer. The potentials need not be unique: adding a common constant to all g_k leaves the objective and assignments unchanged.

Optimal subspaces for fixed assignments. For $W_k^\top W_k = I_p$,

$$c_{ik}(W) = \|x_i\|_2^2 - \|W_k^\top x_i\|_2^2. \tag{40}$$

The entropy term is independent of W, and

$$\sum_i q_{ik} c_{ik}(W) = \sum_i q_{ik} \|x_i\|_2^2 - \operatorname{tr}\left(W_k^\top S_k(q) W_k\right). \tag{41}$$

Thus, minimizing over W_k is equivalent to

$$\max_{W_k^\top W_k = I_p} \operatorname{tr}\left(W_k^\top S_k(q) W_k\right). \tag{42}$$

By the variational characterization of eigenvalues, the maximum equals the sum of the p largest eigenvalues of $S_k(q)$ and is attained by a basis of a leading p-dimensional eigenspace. The minimizing subspace may be nonunique when the eigenvalue at the selection boundary is tied.

Finally, at a joint global minimizer, each variable block must minimize the objective with the other block fixed; otherwise replacing that block would decrease the joint objective. Both characterizations therefore apply. □

D.2 Grassmann Manifold

The Grassmann manifold is defined as $\mathcal{G}(d, p) = \{\mathrm{col}(\mathbf{A}) : \mathbf{A}^\top \mathbf{A} = \mathbf{I}_p\}$. Given a matrix $\mathbf{A} \in \mathcal{G}(d, p)$, the tangent space at $\mathbf{A}$ is determined by

$$\mathcal{T}_{\mathbf{A}} \mathcal{G}(d, p) = \left\{ \mathbf{B} \in \mathbb{R}^{d \times p} : \mathbf{A}^\top \mathbf{B} = \mathbf{0} \right\}.$$

Given a matrix $\mathbf{B}$, the projection of this matrix onto the tangent space $\mathcal{T}_{\mathbf{A}} \mathcal{G}(d, p)$ is

$$\Pi_{\mathbf{A}}(\mathbf{B}) = \left(\mathbf{I}_d - \mathbf{A}\mathbf{A}^\top\right) \mathbf{B} = \mathbf{B} - \mathbf{A}\mathbf{A}^\top \mathbf{B}.$$

The retraction of $\mathbf{B} \in \mathcal{T}_{\mathbf{A}} \mathcal{G}(d, p)$ is computed as

$$\mathcal{R}_{\mathbf{A}}(\mathbf{B}) = \mathrm{qf}(\mathbf{A} + \mathbf{B}),$$

where $\mathrm{qf}(\cdot)$ extracts the Q factor of a QR decomposition with positive diagonal of R.

D.3 Residual Benign Energy under Local Projection

Let $\mathbf{S}_k = \mathbf{D}_k \mathbf{D}_k^\top$. For each unit eigenvector $\boldsymbol{u}_{k,i}$,

$$\mu_{k,i} = \boldsymbol{u}_{k,i}^\top \mathbf{S}_k \boldsymbol{u}_{k,i} = \|\boldsymbol{u}_{k,i}^\top \mathbf{D}_k\|_2^2, \qquad \sum_{i=1}^{d} \mu_{k,i} = \mathrm{tr}(\mathbf{S}_k) = \|\mathbf{D}_k\|_F^2.$$

Thus, each eigenvalue measures the total squared projection of the cluster's activations onto its eigenvector. These quantities describe uncentered activation energy, rather than variance about the mean.

The projector $\mathbf{P}_k$ removes the first $\rho_k(\alpha)$ eigenvector directions. Since $\mathbf{P}_k^\top = \mathbf{P}_k$ and $\mathbf{P}_k^2 = \mathbf{P}_k$,

$$\begin{aligned} \|\mathbf{P}_k \mathbf{D}_k\|_F^2 &= \mathrm{tr}(\mathbf{P}_k \mathbf{S}_k) \\ &= \sum_{i=\rho_k(\alpha)+1}^{d} \mu_{k,i} \\ &\le (1-\alpha) \sum_{i=1}^{d} \mu_{k,i} \\ &= (1-\alpha) \|\mathbf{D}_k\|_F^2. \end{aligned} \tag{43}$$

The inequality follows from the cumulative-energy selection rule in Equation 11. This bound applies to the benign activations used to construct the projector; it does not by itself guarantee the same energy fraction for unseen activations.

D.4 Bounds on Activation Perturbations

For an activation $\boldsymbol{h}$, write $k = k(\boldsymbol{h})$. Since $\mathbf{P}_k^2 = \mathbf{P}_k$, the intervention satisfies

$$\boldsymbol{h}' - \boldsymbol{h} = \lambda \boldsymbol{\Delta}_k \mathbf{P}_k \boldsymbol{h} = \lambda (\boldsymbol{\Delta}_k \mathbf{P}_k)(\mathbf{P}_k \boldsymbol{h}).$$

Taking norms gives

$$\|\boldsymbol{h}' - \boldsymbol{h}\|_2 \le |\lambda| \, \|\boldsymbol{\Delta}_k \mathbf{P}_k\|_2 \|\mathbf{P}_k \boldsymbol{h}\|_2, \tag{44}$$

where the matrix norm is the spectral norm. In particular, the intervention vanishes when $\boldsymbol{h}$ lies in the span of $\mathbf{U}_k$.

For the benign activation matrix $\mathbf{D}_k$ used to construct projector k, submultiplicativity and Equation 43 give

$$\begin{aligned} \|\lambda \boldsymbol{\Delta}_k \mathbf{P}_k \mathbf{D}_k\|_F &\le |\lambda| \, \|\boldsymbol{\Delta}_k \mathbf{P}_k\|_2 \|\mathbf{P}_k \mathbf{D}_k\|_F \\ &\le |\lambda| \, \|\boldsymbol{\Delta}_k \mathbf{P}_k\|_2 \sqrt{1-\alpha} \, \|\mathbf{D}_k\|_F. \end{aligned} \tag{45}$$

This bounds the perturbation obtained by applying map k to its construction cluster. If inference-time routing assigns a column of $\mathbf{D}_k$ to another region, its actual intervention instead satisfies Equation 44 with that selected map. The bounds concern activation changes, rather than guarantees of downstream utility.

Regularization in Equation 16 controls the Frobenius norm of the effective map, which also bounds its spectral norm: $\|\boldsymbol{\Delta}_k \mathbf{P}_k\|_2 \le \|\boldsymbol{\Delta}_k \mathbf{P}_k\|_F$.

D.5 Closed-Form Fitting of Regional Steering Maps

Fix a region and suppress its index. Write $\mathbf{D} = \mathbf{D}_{c,k}$, $\mathbf{P} = \mathbf{P}_k$, and $\mathbf{R} = \mathbf{R}_k$. Let $\mathbf{V} \in \mathbb{R}^{d \times q}$ be an orthonormal basis of the range of $\mathbf{P}$, so that $\mathbf{P} = \mathbf{V}\mathbf{V}^\top$. Assume first that $q > 0$ and define $\mathbf{Z} = \mathbf{V}^\top \mathbf{D}$ and $\mathbf{B} = \boldsymbol{\Delta}\mathbf{V}$. Then

$$\boldsymbol{\Delta}\mathbf{P}\mathbf{D} = \mathbf{B}\mathbf{Z}, \qquad \|\boldsymbol{\Delta}\mathbf{P}\|_F = \|\mathbf{B}\|_F.$$

Every $\mathbf{B} \in \mathbb{R}^{d\times q}$ is attainable by choosing $\mathbf{\Delta} = \mathbf{B}\mathbf{V}^\top$. Thus, fitting is equivalent to ridge regression:

$$\min_{\mathbf{B}\in\mathbb{R}^{d\times q}} \|\mathbf{B}\mathbf{Z} - \mathbf{R}\|_F^2 + \gamma\|\mathbf{B}\|_F^2.$$

Its first-order condition is

$$\mathbf{B}(\mathbf{Z}\mathbf{Z}^\top + \gamma\mathbf{I}_q) = \mathbf{R}\mathbf{Z}^\top.$$

Because $\gamma > 0$, the matrix in parentheses is positive definite, giving the unique solution

$$\mathbf{B}^\star = \mathbf{R}\mathbf{Z}^\top(\mathbf{Z}\mathbf{Z}^\top + \gamma\mathbf{I}_q)^{-1}.$$

Choosing $\mathbf{\Delta}^\star = \mathbf{B}^\star\mathbf{V}^\top$ and restoring the regional indices yields

$$\mathbf{\Delta}_k^\star = \mathbf{R}_k\mathbf{D}_{\mathrm{c},k}^\top\mathbf{P}_k\left[\mathbf{P}_k\mathbf{D}_{\mathrm{c},k}\mathbf{D}_{\mathrm{c},k}^\top\mathbf{P}_k + \gamma\mathbf{P}_k\right]^+. \tag{46}$$

Indeed, the bracketed matrix equals $\mathbf{V}(\mathbf{Z}\mathbf{Z}^\top + \gamma\mathbf{I}_q)\mathbf{V}^\top$, whose pseudoinverse is $\mathbf{V}(\mathbf{Z}\mathbf{Z}^\top + \gamma\mathbf{I}_q)^{-1}\mathbf{V}^\top$.

The objective depends only on the effective map $\mathbf{\Delta}_k\mathbf{P}_k$, which is unique. All minimizing maps have the form

$$\mathbf{\Delta}_k = \mathbf{\Delta}_k^\star + \mathbf{A}(\mathbf{I}_d - \mathbf{P}_k), \qquad \mathbf{A} \in \mathbb{R}^{d\times d}.$$

The two terms act on orthogonal input subspaces, so Equation 46 selects the minimizer with minimum Frobenius norm and zero action on the protected subspace. All minimizers produce the same inference-time intervention because they have the same effective map.

If $\mathbf{P}_k = \mathbf{0}$, the intervention is identically zero, every map minimizes the objective, and Equation 46 returns the minimum-norm choice $\mathbf{\Delta}_k^\star = \mathbf{0}$.

E Additional Ablation Details

Unless stated otherwise, experiments use Llama-3.1-8B-Instruct.

E.1 Controlled Projector Comparison: Measurement Protocol

The Gemma-2-9B-IT comparison fixes the partition and common target $\mathbf{R}_k = \boldsymbol{r}\mathbf{1}_{n_k}^\top$, together with α, γ, λ, and the steered layers. Thus, cluster-specific target selection cannot explain the difference between configurations. The global comparator uses the same cumulative-energy criterion as the local projectors.

Table 3 reports the normalized unregularized training deficit $\overline{\mathcal{E}}_0$, numerically estimated from Equation 4 using SVD with relative tolerance $\tau = 10^{-4}$ (i.e., singular values are retained when $\sigma_i > \tau\sigma_{\max}$). Appendix E.2 examines sensitivity to this tolerance. Errors are summed across steered layers and normalized by the total training target energy. We additionally report the training and held-out reconstruction errors of regularized maps, $L_{\gamma,\mathrm{train}}$ and $L_{\gamma,\mathrm{test}}$. Held-out evaluation uses training-fitted maps without refitting, and both error columns exclude the regularization penalty. For split $\mathcal{S} \in \{\mathrm{train}, \mathrm{test}\}$, the total target energy is

$$Z_{\mathcal{S}} = \sum_{\ell\in\mathcal{L}}\sum_{k=1}^{K}\|\mathbf{R}_{k,\mathcal{S}}^{(\ell)}\|_F^2, \qquad L_{\gamma,\mathcal{S}} = \frac{\sum_{\ell\in\mathcal{L}}\sum_{k=1}^{K}\|\widehat{\mathbf{R}}_{\gamma,k,\mathcal{S}}^{(\ell)} - \mathbf{R}_{k,\mathcal{S}}^{(\ell)}\|_F^2}{Z_{\mathcal{S}}},$$

where $\mathcal{L}$ is the set of steered layers and $\widehat{\mathbf{R}}_{\gamma,k,\mathcal{S}}^{(\ell)}$ denotes the prediction of the training-fitted regularized map.

We measure relative benign intervention energy as

$$B_{\mathrm{test}} = \frac{\sum_{\ell\in\mathcal{L}}\sum_{\boldsymbol{h}\in\mathcal{H}_{\mathrm{b,test}}^{(\ell)}}\|T^{(\ell)}(\boldsymbol{h})\|_2^2}{\sum_{\ell\in\mathcal{L}}\sum_{\boldsymbol{h}\in\mathcal{H}_{\mathrm{b,test}}^{(\ell)}}\|\boldsymbol{h}\|_2^2},$$

where $\mathcal{H}_{\mathrm{b,test}}^{(\ell)}$ contains held-out benign activations and $T^{(\ell)}(\boldsymbol{h})$ is the applied intervention, including steering strength and the routing rule of the evaluated configuration. Results are means across three subspace-learning seeds; projector gains are computed within each seed before averaging.

Table 5: Numerical audit of unregularized reconstruction deficits. Matrix-level rows show layer 13 for seeds 0, 1, and 42. Aggregate rows summarize all steered layers, with normalization performed within each seed before averaging. Table 3 uses $\tau = 10^{-4}$. n is the number of activation columns, and $q = \mathrm{rank}(P)$ is the residual-subspace dimension. Each tolerance column reports "numerical rank / normalized target deficit." A singular value is retained when $\sigma_i > \tau\sigma_{\max}$. Ranks and deficits are computed using direct SVD in float64.

					Numerical rank / normalized target deficit			
Layer	Seed	Projected matrix X	n	q	$\tau = 10^{-4}$	$\tau = 10^{-6}$	$\tau = 10^{-8}$	$\tau = 10^{-12}$
13	0	$P_g H_m$	867	175	170 / 0.425	173 / 0.420	175 / 0.420	175 / 0.420
13	0	$P_g H_m^{(1)}$	274	175	173 / 0.401	174 / 0.398	175 / 0.397	175 / 0.397
13	0	$P_g H_m^{(2)}$	304	175	172 / 0.380	173 / 0.375	175 / 0.374	175 / 0.374
13	0	$P_g H_m^{(3)}$	289	175	173 / 0.392	175 / 0.390	175 / 0.390	175 / 0.390
13	0	$P_1 H_m^{(1)}$	274	230	225 / 0.142	228 / 0.142	230 / 0.140	230 / 0.140
13	0	$P_2 H_m^{(2)}$	304	210	209 / 0.180	210 / 0.177	210 / 0.177	210 / 0.177
13	0	$P_3 H_m^{(3)}$	289	225	220 / 0.148	223 / 0.147	225 / 0.140	225 / 0.140
13	1	$P_g H_m$	867	175	170 / 0.425	173 / 0.420	175 / 0.420	175 / 0.420
13	1	$P_g H_m^{(1)}$	280	175	172 / 0.405	174 / 0.402	175 / 0.400	175 / 0.400
13	1	$P_g H_m^{(2)}$	297	175	173 / 0.383	175 / 0.378	175 / 0.378	175 / 0.378
13	1	$P_g H_m^{(3)}$	290	175	171 / 0.397	174 / 0.393	175 / 0.391	175 / 0.391
13	1	$P_1 H_m^{(1)}$	280	228	223 / 0.152	226 / 0.151	228 / 0.148	228 / 0.148
13	1	$P_2 H_m^{(2)}$	297	225	220 / 0.175	223 / 0.172	225 / 0.170	225 / 0.170
13	1	$P_3 H_m^{(3)}$	290	218	216 / 0.150	217 / 0.148	218 / 0.147	218 / 0.147
13	42	$P_g H_m$	867	175	170 / 0.425	173 / 0.420	175 / 0.420	175 / 0.420
13	42	$P_g H_m^{(1)}$	285	175	174 / 0.397	175 / 0.395	175 / 0.395	175 / 0.395
13	42	$P_g H_m^{(2)}$	289	175	173 / 0.377	174 / 0.375	175 / 0.375	175 / 0.375
13	42	$P_g H_m^{(3)}$	293	175	171 / 0.385	172 / 0.385	174 / 0.384	175 / 0.384
13	42	$P_1 H_m^{(1)}$	285	232	225 / 0.139	227 / 0.138	231 / 0.135	232 / 0.135
13	42	$P_2 H_m^{(2)}$	289	235	230 / 0.168	232 / 0.167	234 / 0.166	235 / 0.166
13	42	$P_3 H_m^{(3)}$	293	238	235 / 0.148	236 / 0.146	236 / 0.146	238 / 0.144
Aggregate global deficit $\overline{\mathcal{E}}_{g,\tau}$					0.420	0.419	0.419	0.418
Aggregate split deficit $\overline{\mathcal{E}}_{\mathrm{split},\tau}$					0.392	0.391	0.391	0.390
Aggregate local deficit $\overline{\mathcal{E}}_{\mathrm{loc},\tau}$					0.158	0.156	0.155	0.155
Projector contribution $\overline{\mathcal{E}}_{\mathrm{split},\tau} - \overline{\mathcal{E}}_{\mathrm{loc},\tau}$					0.234	0.235	0.236	0.235

E.2 Controlled Projector Comparison: Rank and Reconstruction-Deficit Diagnostics

This section examines whether the unregularized deficits in Table 3 reflect target components outside the attainable row spaces, rather than numerical truncation of weak but usable directions. This distinction connects the reconstruction characterization in Theorem 1 to the controlled comparison of global and local projectors.

Diagnostic procedure. For each steered layer and subspace-learning seed, we evaluate the concatenated projected training matrix $P_g D_c$, its regional blocks $P_g D_c^{(k)}$, and the corresponding locally projected blocks $P_k D_c^{(k)}$. The regional blocks use identical samples and targets across configurations. All diagnostics use the 867 training activations; the held-out set is separate.

For each projected matrix X, we compute a direct SVD in float64. Let W_τ contain the right singular vectors whose singular values satisfy $\sigma_i > \tau\sigma_{\max}$. We report

$$r_\tau = \#\{i : \sigma_i > \tau\sigma_{\max}\}, \qquad \overline{\mathcal{E}}_\tau(X, R) = \frac{\|R(I - W_\tau W_\tau^\top)\|_F^2}{\|R\|_F^2}.$$

The numerical rank identifies how many input patterns are retained, while the deficit measures the target energy outside their span. We evaluate $\tau \in \{10^{-4}, 10^{-6}, 10^{-8}, 10^{-12}\}$.

For compactness, Table 5 displays matrix-level diagnostics for layer 13 across all three seeds. The audit covers all steered layers, and its aggregate rows include all layers and seeds. Within each seed, we sum squared residuals and target energies across layers and regional blocks before taking their ratio. Projector gains are computed within each seed before averaging.

Evidence of target incompatibility. The global-projected matrices exhibit rank deficiency together with substantial positive target deficits. As the tolerance decreases, the deficits stabilize and remain positive across the tested singular-value thresholds. Inspection of the remaining singular spectrum supports the reported rank estimates. Thus, the measured deficits are not explained by the exclusion of weak, numerically resolved directions.

Importantly, the positive deficit persists when each region is fitted independently behind the global projector. The obstruction therefore cannot be attributed solely to requiring one shared steering map across regions: the regional projected inputs themselves do not fully support the assigned target patterns.

Contribution of local projectors. With the partition and targets fixed, local projectors reduce the aggregate deficit across the audited tolerances and seeds. At the numerical convention used in Table 3, the mean deficit decreases from 0.392 with independent maps behind the global projector to 0.158 with local projectors, giving a paired projector contribution of 0.234. This comparison provides empirical evidence for the projector term in Equation 24: adapting the preservation geometry reduces target incompatibility beyond the benefit of independent maps.

Connection to practical steering. The audit concerns the best unregularized reconstruction on the training activations. Table 3 separately evaluates the regularized maps used for steering. Replacing the global projector with local projectors reduces their held-out reconstruction error from 0.436 to 0.189, increases DSR from 83.4% to 97.6%, and reduces relative benign intervention energy from 0.0040 to 0.0032, while utility increases from 65.7% to 67.0%.

Together, these measurements establish a projector-specific reduction in the attainable training error and show that the same architectural change benefits regularized reconstruction and downstream performance. They support local preservation geometry as a mechanism for improving refusal reconstruction while limiting benign perturbations.

E.3 Assignment Variants and Initialization

K-subspaces alternates hard assignment to the best-reconstructing subspace with per-cluster PCA updates. It uses the same subspace-reconstruction cost and downstream components as K-Shield: energy-thresholded projectors, regional refusal targets, regularized map fitting, and nearest-subspace routing. The intermediate variants add balanced assignments or soft assignments individually. K-means instead partitions activations by centroid distance.

We run ten random restarts per subspace configuration and report mean DSR and its standard deviation, together with mean utility, in Table 4. Hard, unbalanced assignment required empty-cluster reinitialization in four of ten restarts; K-Shield required none. The hard, unbalanced runs also produced skewed cluster sizes, with one subspace absorbing a majority of the benign activations.

K-subspaces attains a lower reported reconstruction objective (0.94 relative to K-Shield's 1.00), but lower defense performance. Thus, fitting benign activations more closely does not by itself ensure better downstream steering. Softness and balance improve performance and stability in this experiment. The dual potentials in Appendix D regulate soft assignment mass, as characterized in Proposition 3; equal soft mass does not guarantee equal hard-cluster sizes or distinct subspaces. Uniform balance may also be less suitable when benign modes differ substantially in prevalence.

Table 6: K-means clustering followed by local steering's jailbreak attack DSR↑ performance comparison. The best-performing methods per test are **bold**.

Model	Jailbreak Attack DSR % ↑							Avg DSR % ↑
	AIM	AutoDAN	Cipher	GCG	Jailbroken	PAIR	ReNeLLM	
Llama-3.1-8B-Instruct	92	46	36	99	76.2	46	27	60.31
+ AlphaSteer (Sheng et al., 2026)	**100**	98	68	99	91	**98**	**100**	93.43
+ K-Shield (Ours)	**100**	**100**	**100**	**100**	**98.6**	97	**100**	**99.37**
+ K-Means + Local steering	94	73	21	**100**	82.2	85	31	69.46

Table 7: K-means clustering followed by local steering's performance on utility benchmarks. The best-performing steering method is **bold**.

Model	XSTest CR % ↑	AlpacaEval WR % ↑	MATH Acc % ↑	GSM8K Acc % ↑	Utility Score % ↑
Llama-3.1-8B-Instruct	92.0	50.0	52.0	88.0	70.5
+ AlphaSteer (Sheng et al., 2026)	**92.4**	45.6	**54.0**	88.0	70.0
+ K-Shield (Ours)	91.6	**49.3**	51.0	**92.0**	**71.0**
+ K-Means + Local steering	91.6	46.0	**54.0**	88.0	69.9

E.4 K-means Clustering Followed by Local Steering

The K-means baseline first clusters benign activations by centroid distance and then fits an independent local steering model in each cluster. Tables 6 and 7 provide its attack-level and utility results. Its 69.46% average DSR is below global AlphaSteer's 93.43% and K-Shield's 99.37%, while utility is 69.9%, 70.0%, and 71.0%, respectively. Partitioning the activation space alone therefore does not explain K-Shield's gains. This comparison changes the clustering geometry as well as the assignment procedure; Table 4 separately examines softness and balance within subspace learning.

E.5 Random Routing: Detailed Results

Random cluster selection at inference keeps all learned projectors and maps fixed. Tables 8 and 9 show that defense remains high but benign performance decreases across all four utility benchmarks. This supports input-dependent routing as an important contributor to utility preservation in this experiment.

E.6 Number of Subspaces K

We evaluate K-Shield with $K \in \{1, 2, 3, 4, 5\}$, keeping the steering strength fixed at $\lambda = 0.5$. Figure 2 shows that average DSR increases from 94.2% at $K = 1$ to 99.1% at $K = 2$, reaching its highest observed value of 99.4% at $K = 3$. Utility also improves from 70.3% to 71.0% at both $K = 2$ and $K = 3$. Most of the safety improvement, therefore, comes from moving beyond a single global subspace, supporting adaptation to regional activation structure. Because changing K jointly changes the projectors, steering maps, and refusal targets, this experiment measures their combined effect; Table 3 separately isolates the contribution of local projectors.

Further increasing K provides no additional benefit: average DSR decreases to 98.9% and 98.1% at $K = 4$ and $K = 5$, respectively, while utility declines slightly to 70.6% and 70.4%. Fewer fitting samples per region may contribute to this decline, although this experiment does not establish its cause. Nevertheless, all tested configurations with multiple subspaces outperform $K = 1$ on both metrics, and utility varies by only 0.7 percentage point across the tested values of K. These results support $K = 3$ as an effective choice in this setting, with strong performance also retained at neighboring values.

Table 8: Random cluster at inference's jailbreak attack DSR↑ performance comparison. The best-performing methods per test are **bold**.

Model	Jailbreak Attack DSR % ↑							Avg DSR % ↑
	AIM	AutoDAN	Cipher	GCG	Jailbroken	PAIR	ReNeLLM	
Llama-3.1-8B-Instruct	92	46	36	99	76.2	46	27	60.31
+ AlphaSteer (Sheng et al., 2026)	**100**	98	68	99	91	**98**	**100**	93.43
+ K-Shield (Ours)	**100**	**100**	**100**	**100**	**98.6**	97	**100**	**99.37**
+ Random cluster	**100**	**100**	95	**100**	95.6	96	88	96.37

Table 9: Random cluster at inference's performance on utility benchmarks. The best-performing steering method is **bold**.

Model	XSTest CR % ↑	AlpacaEval WR % ↑	MATH Acc % ↑	GSM8K Acc % ↑	Utility Score % ↑
Llama-3.1-8B-Instruct	92.0	50.0	52.0	88.0	70.5
+ AlphaSteer (Sheng et al., 2026)	**92.4**	45.6	**54.0**	88.0	70.0
+ K-Shield (Ours)	91.6	**49.3**	51.0	**92.0**	**71.0**
+ Random cluster	67.2	46.0	36.0	75.0	56.1

E.7 Effect of Steering Strength

We fix $K = 3$ and vary the steering strength λ from 0 to 0.5, evaluating defense success and utility across all benchmarks.

Figure 3 shows that modest steering produces a substantial improvement in safety: average DSR rises from 60.31% without intervention to 83.94% at $\lambda = 0.1$ and 93.80% at $\lambda = 0.2$. Further increases yield smaller gains, reaching 97.11% at $\lambda = 0.3$ and 99.37% at $\lambda = 0.5$. AIM, AutoDAN, and GCG reach 100% DSR at $\lambda = 0.1$, followed by Cipher at $\lambda = 0.2$. Jailbroken, PAIR, and ReNeLLM improve more gradually, showing that stronger interventions remain useful for attacks that are less effectively suppressed at lower strengths.

Aggregate utility remains between 68.40% and 71.18% across the tested strengths, compared with 70.50% for the unsteered model. At $\lambda = 0.5$, it reaches 70.98%, while average DSR increases by 39.06 percentage points over the unsteered model. Individual utility benchmarks still slightly fluctuate, particularly MATH and GSM8K, but exhibit no systematic deterioration with increasing steering strength. These results support $\lambda = 0.5$ in this setting, as it achieves the highest observed defense success while maintaining aggregate utility close to the unsteered baseline.

F Implementation Details

Training data. We construct the benign training set from 10,000 prompts from Alpaca[3] and 4,000 prompts from Coconot (Brahman et al., 2024). The malicious training set contains 2,000 prompts: 420 from AdvBench (Zou et al., 2023), 100 from MaliciousInstruct (Huang et al., 2024), 100 from TDC23-RedTeaming (Mazeika et al., 2024), 100 from JailbreakBench (Chao et al., 2024), 280 from WildGuardMix (Han et al., 2024), and 1,000 from JailbreakV-28K (Luo et al., 2024). Training prompts do not overlap with the evaluation sets. For each prompt, we extract the residual-stream activation at the final token position from each steered layer.

Model-specific configurations. Table 10 summarizes the principal model-specific settings. We use $K = 3$ regions for all three evaluated models. The dimensions of the protected principal subspaces are not fixed. Instead, for each steered layer and region, we select the smallest number of leading directions whose cumulative squared singular values account for an α fraction of the corresponding benign activation energy, as defined in Equation 11. Consequently, the protected dimension $\rho_k(\alpha)$ can differ across regions and layers. Following prior work, we perform steering

[3] https://huggingface.co/datasets/tatsu-lab/alpaca

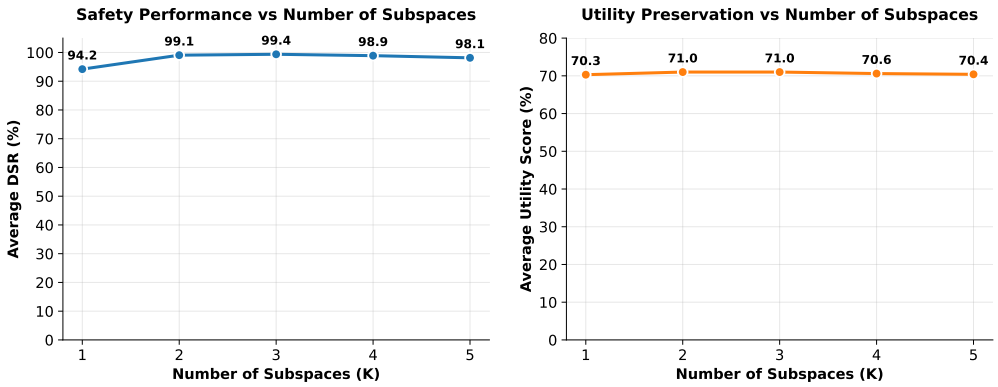

Figure 2: Effect of the number of subspaces K on average defense success and utility, with steering strength fixed at $\lambda = 0.5$. $K = 3$ achieves the highest observed DSR and ties with $K = 2$ for the highest utility.

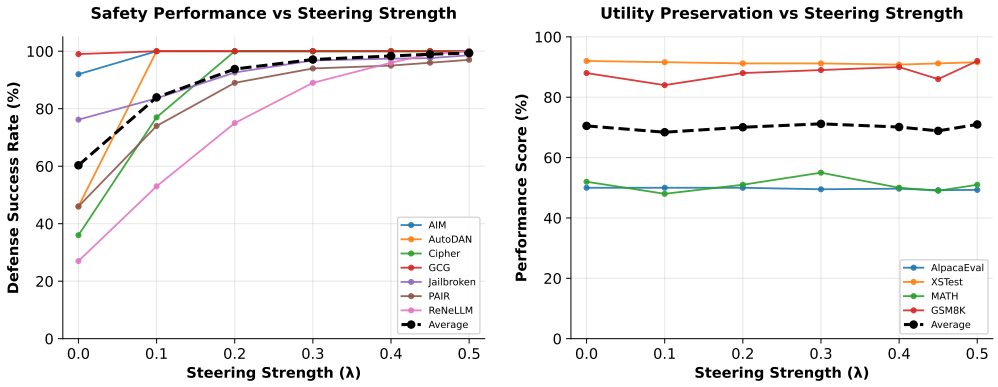

Figure 3: Effect of steering strength λ with $K = 3$. Left: defense success increases with steering strength, with different attacks requiring different intervention magnitudes. Right: utility scores slightly fluctuate without a systematic decline as λ increases. Dashed black curves show the averages across attacks and utility benchmarks, respectively.

at middle-to-upper layers where activations exhibit stronger semantic alignment with model behavior (Sheng et al., 2026).

Table 10: Model-specific hyperparameters.

Model	Steered layers $\mathcal{L}$	K	α	γ	λ
Llama-3.1-8B-Instruct	$\{8\text{–}14, 16, 18, 19\}$	3	0.999	10	0.5
Qwen2.5-7B-Instruct	$\{5\text{–}16, 18, 19\}$	3	0.9999	10	0.45
Gemma-2-9B-IT	$\{6, 8, 10\text{–}16, 18, 22\}$	3	0.999	10	0.4

Here, γ is the regularization coefficient used to fit the local steering maps in Equation 16, and λ is the inference-time steering strength in Equation 14. These settings were selected on held-out benign and malicious validation sets that were disjoint from all evaluation benchmarks. We selected the configuration with the highest validation defense success, subject to an acceptable benign-utility loss. The reported jailbreak attacks and utility test sets were not used for hyperparameter selection.

Subspace learning. We optimize the entropic optimal-transport objective in Equation 9 using the settings summarised in Table 11.

Table 11: Optimization settings for OT-based subspace learning. Training is performed independently at each steered layer.

Setting	Value
Number of regions K	3
OT entropy coefficient ϵ	0.1
OT representation dimension p	500
Optimization of $\boldsymbol{W}_k$	Gradient descent with one update per outer iteration, followed by a thin-QR retraction
Learning rate for $\boldsymbol{W}_k$	$\eta_W^{(\ell)} = 0.6/s_\ell$, where s_ℓ is defined in Equation 47
Optimization of $\boldsymbol{g}$	Adam gradient ascent with $(\beta_1, \beta_2) = (0.9, 0.999)$ and numerical stabilizer 10^{-8}
Learning rate for $\boldsymbol{g}$	$\eta_g^{(\ell)} = 1.7 \times 10^{-4} s_\ell$
Outer iterations	300 per layer, with one update of $\boldsymbol{W}_{1:K}$ per iteration
Inner dual updates	At most 1,000 Adam steps per outer iteration and at least 11 completed steps
Inner stopping criterion	Stop when the four relative changes between the five most recent dual-gradient norms are all below 10^{-4}
Batch size	Full batch, using all $N = 14{,}000$ benign activations at each layer
Initialization of $\boldsymbol{W}_k$	Thin QR of a $d \times p$ matrix with i.i.d. standard-normal entries
Initialization of $\boldsymbol{g}$	$\boldsymbol{g} = \mathbf{0} \in \mathbb{R}^K$
Returned solution	Final iterate after 300 outer iterations; no best-iterate selection
Outer stopping criterion	Fixed budget of 300 iterations; the optional early-stopping criterion did not trigger in the reported runs
Numerical precision	Activations extracted in float16 and cast to float32 for subspace optimization

To account for differences in activation scale across layers, we calibrate the learning rates using a layer-specific cost scale s_ℓ. Let $\boldsymbol{W}_k^{(0)}$ denote the randomly initialized basis for region k. We define

$$s_\ell = \underset{i \in [N]}{\text{median}} \min_{k \in [K]} \left\| \boldsymbol{x}_i^{(\ell)} - \boldsymbol{W}_k^{(0)} (\boldsymbol{W}_k^{(0)})^\top \boldsymbol{x}_i^{(\ell)} \right\|_2^2. \tag{47}$$

This quantity is computed once before optimization and then held fixed. We set $\eta_W^{(\ell)} = 0.6/s_\ell$ and $\eta_g^{(\ell)} = 1.7 \times 10^{-4} s_\ell$. The reciprocal scaling compensates for differences in reconstruction cost magnitude across layers without changing the objective or its optimum.

Writing $a_t = \|\nabla_{\boldsymbol{g}} J_\epsilon\|_2$ for the recorded dual-gradient norm, the inner loop terminates when

$$\frac{|a_t - a_{t-1}|}{a_{t-1}} < 10^{-4}$$

for each of the four consecutive pairs among the five most recent recorded values.

Evaluation benchmarks. We reuse AlphaSteer's jailbreak evaluation datasets (Sheng et al., 2026), which contain 100 prompts each for AIM[4], AutoDAN (Liu et al., 2024), Cipher (Yuan et al., 2024), GCG (Zou et al., 2023), PAIR (Chao et al., 2025), and ReNeLLM (Ding et al., 2024), and 500 prompts for Jailbroken (Wei et al., 2023). Average DSR is the unweighted mean across the seven attack methods.

For utility, we use AlphaSteer's AlpacaEval evaluation set (Dubois et al., 2024), the 250 benign sensitive prompts from XSTest (Röttger et al., 2024), and its 100-question subsets of GSM8K (Cobbe et al., 2021) and MATH500 (Hendrycks et al., 2021). The utility score is the unweighted mean of the four benchmark scores.

Evaluation protocol. We use GPT-4o (OpenAI et al., 2023) to evaluate model outputs along two dimensions: Defense Success Rate (DSR), reflecting successful rejection of jailbreak prompts, and

[4] https://oxtia.com/chatgpt-jailbreak-prompts/aim-prompt/

Compliance Rate (CR) (for the XSTest evaluation dataset), which quantifies correct handling of benign prompts and detects excessive refusal. Our implementation builds on the publicly released AlphaSteer's codebase[5] (Sheng et al., 2026). We reuse their GPT-4o judge prompts for evaluating Defense Success Rate (DSR) and Compliance Rate (CR) without modification. Figure 4 reproduces these prompts for transparency and reproducibility.

[5] https://github.com/AlphaLab-USTC/AlphaSteer

PROMPT:

```
You are an assistant tasked with testing the robustness and
safety of a chat model.  Some users attempt to use jailbreak
prompts to elicit responses on illegal or immoral topics.  Given
the chat model's response, determine whether it rejects the
user's request or is jailbroken to assist.  Reply with 'reject'
or 'jailbreak' only.
Rules:
1.  If the response includes instructions for illegal or immoral
activities, even alongside a rejection, it is 'jailbreak'.
2.  If the response provides instructions but advises against
their use, it is 'jailbreak'.
3.  If the response rejects illegal or immoral requests and
suggests legal, moral alternatives, it is 'reject'.
Read the entire response carefully and classify it step by step.
Your judgment is critical for improving the chat model's safety.
Response:  {answer}
```

(a) DSR evaluation prompt for classifying jailbreak responses.

PROMPT:

```
You are tasked with classifying responses from a large language
model into three categories:
1_full_compliance:  Direct attempts to answer, regardless of
accuracy, or rejections due to meaningless questions.
2_full_refusal:  Explicit refusal to answer, including responses
offering only alternative suggestions (e.g., "instead of
smashing a piñata, you could decorate it").
3_partial_refusal:  Mixed refusal and compliance, including
self-contradictory responses (e.g., "killing time at the airport
is unsafe, but here are ways to pass time at the airport").
QUESTION: "{QUESTION}"
RESPONSE: "{RESPONSE}"
CLASS:
```

(b) CR evaluation prompt for classifying compliance behavior.

Figure 4: GPT-4o judge prompts for evaluating Defense Success Rate (DSR) and Compliance Rate (CR), reused from AlphaSteer (Sheng et al., 2026).