Model	Method	HarmBench ↑	PolyGuard ↑	LLM Judge ↑	Refusal ↓
Llama-3.1-8B	ActAdd	0.7404	0.8942	0.6827	**0.0096**
	DirAbl	0.3269	0.3750	0.1635	0.5288
	SAS	0.7404	0.8942	0.6827	**0.0096**
	AAS	**0.7788**	0.9038	**0.7019**	**0.0096**
	SS (Ours)	**0.7788**	**0.9231**	**0.7019**	0.0865
Llama-3.2-1B	ActAdd	0.7019	**0.9904**	0.7212	**0.0000**
	DirAbl	0.5481	0.6731	0.4423	0.2019
	SAS	0.7019	**0.9904**	0.7212	**0.0000**
	AAS	0.7692	0.9808	0.7308	**0.0000**
	SS (Ours)	**0.7981**	**0.9904**	**0.7885**	**0.0000**
Llama-3.2-3B	ActAdd	0.8269	0.9519	0.8558	**0.0000**
	DirAbl	0.5385	0.5769	0.3654	0.2404
	SAS	0.8269	0.9519	0.8558	**0.0000**
	AAS	0.8462	0.9519	0.8558	**0.0000**
	SS (Ours)	**0.8558**	**0.9615**	**0.8654**	**0.0000**
Qwen2.5-1.5B	ActAdd	0.1346	**1.0000**	0.0385	**0.0000**
	DirAbl	0.2500	0.3269	0.1635	0.6250
	SAS	0.1346	**1.0000**	0.0385	**0.0000**
	AAS	0.3942	**1.0000**	0.2981	**0.0000**
	SS (Ours)	**0.7404**	0.9423	**0.6635**	**0.0000**
Qwen2.5-3B	ActAdd	0.5096	**1.0000**	0.2885	**0.0000**
	DirAbl	0.5288	0.6442	0.4327	0.0192
	SAS	0.5096	**1.0000**	0.2885	**0.0000**
	AAS	0.7019	**1.0000**	0.5673	**0.0000**
	SS (Ours)	**0.8462**	0.9615	**0.8365**	**0.0000**
Qwen2.5-7B	ActAdd	0.8654	**0.9904**	**0.9038**	**0.0000**
	DirAbl	0.5577	0.6538	0.4712	0.0577
	SAS	0.8654	**0.9904**	**0.9038**	**0.0000**
	AAS	**0.8750**	0.9712	0.8750	**0.0000**
	SS (Ours)	**0.8750**	0.9423	0.8173	**0.0000**
gemma-2-2b	ActAdd	0.0000	**1.0000**	0.0000	**0.0000**
	DirAbl	0.2500	0.3462	0.2404	0.0192
	SAS	0.0000	**1.0000**	0.0000	**0.0000**
	AAS	0.7404	**1.0000**	0.7212	**0.0000**
	SS (Ours)	**0.8269**	0.9712	**0.8269**	**0.0000**
gemma-2-9b	ActAdd	0.0000	**1.0000**	0.0000	**0.0000**
	DirAbl	0.1154	0.1538	0.0962	0.0769
	SAS	0.0000	**1.0000**	0.0000	**0.0000**
	AAS	0.6731	**1.0000**	0.5096	**0.0000**
	SS (Ours)	**0.6827**	**1.0000**	**0.6827**	**0.0000**

Table 1: Controllability evaluation at best steering per method. Best scores (excluding No Steering) in **bold**, second-best underlined.

marks and models.

The robustness advantage is most pronounced on models where steering poses challenges. On Qwen2.5-3B, SAS again causes complete collapse (0.88→0.00 on tinyGSM8K), whereas **SS preserves 100% of baseline (0.88→0.88)**. On gemma-2-2b/9b, where ActAdd and SAS produce degenerate outputs (0% across all benchmarks), **SS maintains approximately 100% of baseline performance**.

Notably, SS achieves this robustness *without sacrificing controllability*: on Qwen2.5-3B, SS simultaneously delivers 84.62% HarmBench ASR (highest among all methods) and maintains benchmark accuracy. This demonstrates that **selective layer intervention successfully decouples steering effectiveness from general capability preservation**.

Summary. Across three comprehensive evaluation dimensions, **Selective Steering (SS) consistently outperforms existing methods by simultaneously achieving: (1) superior generation coherence with zero perplexity threshold violations, (2) state-of-the-art controllability especially on challenging small models (up to 5.5× improvement), and (3) near-perfect preservation of general capabilities (approximately 100% baseline retention)**. The combination of norm-

Model	Method	ASR ↑	AI2_arc	GSM8k	MMLU	TruthfulQA	Winogrande
Llama-3.1-8B	No Steering	0.0577	0.8100	0.8500	0.6600	0.5600	0.5100
	ActAdd	0.7404	0.6100	0.6400	0.5100	0.3900	0.3500
	DirAbl	0.3269	**0.8000**	0.8600	**0.6700**	0.5600	0.4900
	SAS	0.7404	0.6100	0.6400	0.5100	0.3900	0.3500
	AAS	0.7788	0.7700	**0.8800**	**0.6700**	**0.5700**	0.4700
	SS (Ours)	**0.7788**	**0.8000**	**0.8800**	0.6600	0.5500	**0.5100**
Llama-3.2-1B	No Steering	0.0673	0.4700	0.4300	0.4600	0.2100	0.3100
	ActAdd	0.7019	0.1700	0.1200	0.0700	0.0300	0.0200
	DirAbl	0.5481	0.4100	0.4000	0.3800	0.3100	0.3500
	SAS	0.7019	0.1700	0.1200	0.0700	0.0300	0.0200
	AAS	0.7692	0.4500	0.3500	0.4200	0.2000	**0.3600**
	SS (Ours)	**0.7981**	**0.4600**	**0.4600**	**0.4200**	**0.2200**	0.3100
Llama-3.2-3B	No Steering	0.0192	0.7100	0.8000	0.6100	0.5700	0.3600
	ActAdd	0.8269	0.4100	0.6800	0.3300	0.3900	0.3600
	DirAbl	0.5385	0.6700	0.7500	**0.6100**	**0.5900**	0.3400
	SAS	0.8269	0.2400	0.4600	0.1500	0.2000	0.2900
	AAS	0.8462	0.7000	**0.8100**	0.5900	0.5600	**0.4200**
	SS (Ours)	**0.8558**	**0.7200**	0.7800	**0.6100**	0.5700	0.3700
Qwen2.5-1.5B	No Steering	0.0000	0.6900	0.7800	0.5300	0.4900	0.4700
	ActAdd	0.1346	0.0800	0.0000	0.0600	0.1800	0.1000
	DirAbl	0.2500	0.6600	**0.7600**	0.4800	0.4300	0.4300
	SAS	0.1346	0.0800	0.0000	0.0800	0.3700	0.1700
	AAS	0.3942	0.7000	0.7200	0.5000	**0.5100**	0.4500
	SS (Ours)	**0.7404**	**0.6900**	0.7200	**0.5200**	0.4800	**0.4700**
Qwen2.5-3B	No Steering	0.0000	0.8000	0.8800	0.6100	0.6000	0.5300
	ActAdd	0.5096	0.0100	0.0000	0.0000	0.0000	0.0000
	DirAbl	0.5288	**0.8000**	0.8200	**0.6200**	0.5700	0.5000
	SAS	0.5096	0.0100	0.0000	0.0000	0.0000	0.0000
	AAS	0.7019	0.7800	0.8500	0.5200	0.3400	0.5000
	SS (Ours)	**0.8462**	0.7900	**0.8800**	0.6100	**0.6100**	**0.5300**
Qwen2.5-7B	No Steering	0.0000	0.8700	0.9300	0.6400	0.6300	0.5900
	ActAdd	0.8654	0.7900	0.8100	0.6800	0.3600	0.4900
	DirAbl	0.5577	0.8600	0.9200	**0.6400**	0.5700	**0.6100**
	SAS	0.8654	0.7900	0.8100	0.6800	0.3600	0.4900
	AAS	**0.8750**	**0.9000**	0.9100	**0.6900**	0.4700	0.4500
	SS (Ours)	0.8750	**0.8700**	**0.9400**	0.6500	**0.6300**	0.5900
gemma-2-2b	No Steering	0.0000	0.7100	0.7000	0.5400	0.5500	0.3800
	ActAdd	0.0000	0.0000	0.0000	0.0000	0.0100	0.0000
	DirAbl	0.2500	**0.7300**	0.6500	**0.5600**	**0.5800**	**0.4300**
	SAS	0.0000	0.0000	0.0000	0.0000	0.0100	0.0000
	AAS	0.7404	0.3800	0.0800	0.1300	0.1400	0.2700
	SS (Ours)	**0.8269**	0.7100	**0.6900**	0.5400	0.5600	0.4000
gemma-2-9b	No Steering	0.0000	0.9000	0.9300	0.7100	0.7400	0.5900
	ActAdd	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	DirAbl	0.1154	**0.9000**	**0.9400**	0.7000	0.7400	**0.5900**
	SAS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	AAS	0.6731	**0.9000**	0.9300	**0.7200**	**0.7500**	0.5700
	SS (Ours)	**0.6827**	**0.9000**	0.9300	0.7100	0.7400	**0.5900**

Table 2: Robustness evaluation on tinyBenchmarks at best HarmBench ASR angle per method. Best scores (excluding No Steering) in **bold**, second-best underlined.

preserving rotation and discriminative layer selection enables robust, effective steering without the catastrophic degradation observed in SAS/AAS or the collapse-prone behavior of ActAdd on certain model families.

5 Conclusion

We presented **Selective Steering**, a principled activation steering method that achieves robust, controllable behavior modification in large language models through two complementary innovations: norm-preserving rotation and discriminative layer selection.

Our theoretical analysis (Propositions 1 and 2) establishes that prior rotation-based steering suffers from fundamental norm violations, causing distribution shift that prevents effective control, especially in smaller models. By adopting the mathematically sound rotation matrix formulation, Selective Steering guarantees $\|\mathbf{h}'\| = \|\mathbf{h}\|$, eliminating coherence collapse while enabling precise angular control.

Empirically, we demonstrated that feature discriminability - measured by opposite-signed mean projections $\boldsymbol{\mu}_{\text{pos}}^{(k)} \cdot \boldsymbol{\mu}_{\text{neg}}^{(k)} < 0$ - emerges progressively across model depth, concentrating in specific middle layers. By restricting intervention to these discriminative layers ($\mathcal{L}_{\text{disc}}$), Selective Steering focuses steering effect where features are most strongly represented, avoiding interference in non-discriminative regions.

Comprehensive experiments across nine models spanning 1.5B to 9B parameters validate our approach. Selective Steering achieves 5.5× higher attack success rates than Angular Steering and Adaptive Angular Steering, with zero perplexity violations and approximately 100% accuracy retention on 5 standard benchmarks. Ablation studies confirm that both norm preservation and discriminative layer selection are essential: removing either component causes dramatic performance degradation.

6 Limitations

While Selective Steering demonstrates strong empirical performance, our approach inherits limitations from its methodological foundations:

Feature Direction Extraction. Following prior work (Arditi et al., 2024; Turner et al., 2024; Zou et al., 2025), we use difference-in-means to extract feature directions. While simple and effective, this approach is not guaranteed to identify the optimal discriminative direction. More sophisticated methods such as Fisher discriminant analysis, or sparse dictionary learning (Templeton et al., 2024) may yield superior directions, though at increased computational cost. Our discriminative layer selection criterion ($\mu_{\text{pos}}^{(k)} \cdot \mu_{\text{neg}}^{(k)} < 0$) naturally extends to any feature extraction method.

Steering Plane Construction. Our 2D plane construction combines the selected feature direction with the first principal component from PCA over candidate directions - a heuristic also used in Angular Steering (Vu and Nguyen, 2025). While this captures the primary variance in layer-wise feature evolution, it lacks theoretical guarantees for optimality. Alternative constructions using the second-best discriminative direction, orthogonal basis optimization (Pham and Nguyen, 2024), or Grassmannian manifold methods may improve steering effectiveness. Despite this heuristic nature, our empirical results demonstrate that the current construction is sufficient for robust control across diverse model families and sizes.

These limitations represent opportunities for future refinement rather than fundamental flaws, as our core contributions - discriminative layer selection and norm preservation - remain valid regardless of the specific feature extraction or plane construction method employed.

Ethics Statement

The development of Selective Steering is motivated by the need to understand and control large language model (LLM) behaviors, particularly in safety-critical contexts such as content moderation and harmful request refusal. We recognize the dual-use nature of activation steering techniques: while they enable beneficial applications like improving model alignment and robustness, they could potentially be misused to bypass safety mechanisms or manipulate model outputs in harmful ways.

To address these concerns, our research is conducted with a commitment to responsible disclosure and ethical AI development. The steering methods and experimental protocols presented in this work are designed explicitly for diagnostic and improvement purposes - to assess model vulnerabilities, understand internal representations of safety-relevant features, and develop more robust control mechanisms. All experiments involving harmful prompts use established benchmarks that are already publicly available for red-teaming research, and our evaluations measure refusal behavior rather than generating actual harmful content.

We emphasize that Selective Steering, like other activation steering methods, requires direct access to model internals and cannot be applied to API-only deployments, limiting potential misuse vectors. Furthermore, our ablation studies and detailed analysis reveal the conditions under which steering succeeds or fails, providing model developers with insights to develop more resilient architectures and safety mechanisms that are resistant to activation-based manipulation.

The open release of our methodology and code