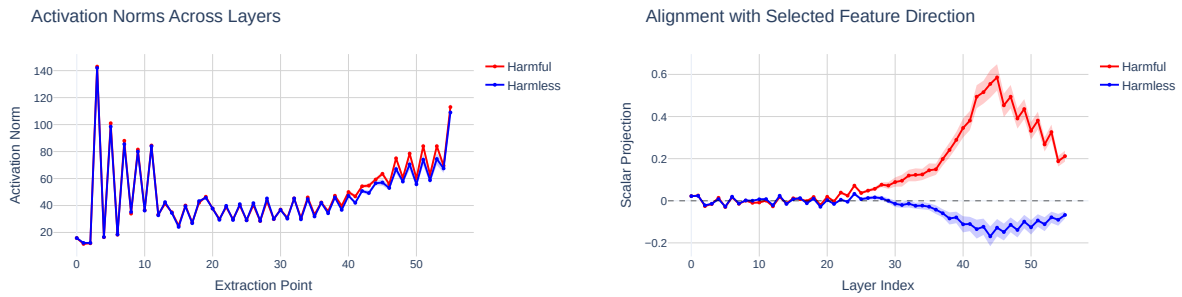

(a) Activation norms across layers

(b) Alignment with selected feature direction

Figure 2: **Layer-wise heterogeneity in Qwen2.5-7B-Instruct.** (a) Activation norms vary substantially across depth, with rapid growth in early layers and amplification near output. (b) Scalar projections class means onto the selected feature direction reveal progressive emergence of opposite-signed discriminability.

Proposition 2 (Norm Preservation in Selective Steering)**.** *The transformation* $\mathbf{h}' = \mathbf{R}_\theta^P \mathbf{h}$ *preserves norms:* $\|\mathbf{h}'\| = \|\mathbf{h}\|$ *for all* $\mathbf{h}$ *and* θ*, where*

$$\mathbf{R}_\theta^P = \mathbf{I} - (\mathbf{b}_1\mathbf{b}_1^\top + \mathbf{b}_2\mathbf{b}_2^\top) + [\mathbf{b}_1\ \mathbf{b}_2]\,\mathbf{R}_\theta\,[\mathbf{b}_1\ \mathbf{b}_2]^\top. \tag{6}$$

The proof (Appendix B.2) establishes that $\mathbf{R}_\theta^P$ is an orthogonal transformation by decomposing it into orthogonal projection onto complement space Q and rotation within plane P.

Feature Direction Selection. Following (Vu and Nguyen, 2025), we select a global feature direction using difference-in-means with maximum inter-layer consistency. At each layer k, compute the local candidate direction:

$$\mathbf{d}^{(k)} = \boldsymbol{\mu}_{\text{pos}}^{(k)} - \boldsymbol{\mu}_{\text{neg}}^{(k)}, \tag{7}$$

where $\boldsymbol{\mu}_{\text{pos}}^{(k)}$ and $\boldsymbol{\mu}_{\text{neg}}^{(k)}$ are class means from Equation 5. The global feature direction is the candidate with highest average cosine similarity to others:

$$\hat{\mathbf{d}}_{\text{feat}} = \text{argmax}_{\mathbf{d}^{(k)}} \left\{ \frac{1}{L} \sum_{j=1}^{L} \cos(\mathbf{d}^{(k)}, \mathbf{d}^{(j)}) \right\}, \tag{8}$$

where L is the number of layers. This selects the direction most consistently represented across depth, capturing the core behavioral axis while filtering layer-specific noise.

Discriminative Layer Selection. Given calibration datasets $\mathcal{D}_{\text{pos}}^{(\text{train})}$ and $\mathcal{D}_{\text{neg}}^{(\text{train})}$, we compute mean activations as in Equation 5. We define **discriminative layers**:

$$\tilde{\boldsymbol{\mu}}_{\text{pos}}^{(k)} = \boldsymbol{\mu}_{\text{pos}}^{(k)} \cdot \hat{\mathbf{d}}_{\text{feat}}, \tilde{\boldsymbol{\mu}}_{\text{neg}}^{(k)} = \boldsymbol{\mu}_{\text{neg}}^{(k)} \cdot \hat{\mathbf{d}}_{\text{feat}}$$
$$\mathcal{L}_{\text{disc}} = \left\{ k \in \{1, \ldots, L\} : \tilde{\boldsymbol{\mu}}_{\text{pos}}^{(k)} \cdot \tilde{\boldsymbol{\mu}}_{\text{neg}}^{(k)} < 0 \right\}. \tag{9}$$

This criterion identifies layers where classes point in opposing directions, ensuring: (1) strong feature representation; (2) predictable steering effect; (3) robust separation across samples.

Steering Transformation. For $k \in \mathcal{L}_{\text{disc}}$, we construct a global steering plane $P = \text{span}\{\mathbf{b}_1, \mathbf{b}_2\}$ following (Vu and Nguyen, 2025), where $\mathbf{b}_1$ is the normalized feature direction and $\mathbf{b}_2$ is the orthogonalized first principal component of candidate directions. We apply:

$$\mathbf{h}'^{(k)} = \begin{cases} \mathbf{R}_\theta^P \mathbf{h}^{(k)}, & \text{if } k \in \mathcal{L}_{\text{disc}}, \\ \mathbf{h}^{(k)}, & \text{otherwise}, \end{cases} \tag{10}$$

where $\mathbf{R}_\theta^P = \mathbf{I} - (\mathbf{b}_1\mathbf{b}_1^\top + \mathbf{b}_2\mathbf{b}_2^\top) + [\mathbf{b}_1\ \mathbf{b}_2]\,\mathbf{R}_\theta\,[\mathbf{b}_1\ \mathbf{b}_2]^\top$ and $\mathbf{R}_\theta$ is the 2D rotation matrix. By Proposition 2, $\|\mathbf{h}'^{(k)}\| = \|\mathbf{h}^{(k)}\|$ is guaranteed.

3.4 Algorithm and Calibration

Algorithm 1 summarizes the inference-time procedure:

Calibration. One-time setup: (1) extract activations from $\mathcal{D}_{\text{pos}}^{(\text{train})}$ and $\mathcal{D}_{\text{neg}}^{(\text{train})}$; (2) compute $\boldsymbol{\mu}_{\text{pos}}^{(k)}, \boldsymbol{\mu}_{\text{neg}}^{(k)}$ per layer; (3) identify $\mathcal{L}_{\text{disc}}$ via Equation 9; (4) construct global plane P via PCA. See Appendix B.3 for full procedure.

Algorithm 1 Selective Steering (Inference)

Require: Activation $\mathbf{h}^{(k)}$, basis $\{\mathbf{b}_1, \mathbf{b}_2\}$, angle θ, means $\boldsymbol{\mu}_{\text{pos}}^{(k)}, \boldsymbol{\mu}_{\text{neg}}^{(k)}$
Ensure: Steered activation $\mathbf{h}'^{(k)}$
if $\tilde{\boldsymbol{\mu}}_{\text{pos}}^{(k)} \cdot \tilde{\boldsymbol{\mu}}_{\text{neg}}^{(k)} \geq 0$ **then** ▷ Non-discriminative layer
  return $\mathbf{h}^{(k)}$
end if
$\mathbf{R}_\theta \leftarrow \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$
$\mathbf{R}_\theta^P \leftarrow \mathbf{I} - (\mathbf{b}_1\mathbf{b}_1^\top + \mathbf{b}_2\mathbf{b}_2^\top) + [\mathbf{b}_1\ \mathbf{b}_2]\,\mathbf{R}_\theta\,[\mathbf{b}_1\ \mathbf{b}_2]^\top$
$\mathbf{h}'^{(k)} \leftarrow \mathbf{R}_\theta^P \mathbf{h}^{(k)}$ ▷ Norm preserved by Prop. 2
return $\mathbf{h}'^{(k)}$

Advantages. Selective Steering offers: (1) **guaranteed norm preservation** via Proposition 2; (2) **focused intervention** on discriminative layers only; (3) **reduced computation** from $O(Ld_{\text{model}})$ to $O(|\mathcal{L}_{\text{disc}}|d_{\text{model}})$ where $|\mathcal{L}_{\text{disc}}| \ll L$; (4) **compatibility** with normalization-heavy architectures.

4 Experiments

4.1 Experimental Setup

Hardware. All experiments are conducted on a single NVIDIA A40 GPU with 48GB memory. To ensure reproducibility, we use greedy decoding (temperature = 0.0) across all methods and models.

Datasets. We use two contrastive datasets for calibration: **AdvBench** (Zou et al., 2023) (80%, 416 samples) as $\mathcal{D}_{\text{pos}}^{(\text{train})}$ containing harmful prompts, and 416 samples from **Alpaca** (Taori et al., 2023) as $\mathcal{D}_{\text{neg}}^{(\text{train})}$ containing harmless prompts. The remaining 20% of AdvBench (104 samples) serves as the evaluation set for measuring coherence and controllability.

To assess robustness, we employ benchmark datasets from **tinyBenchmarks** (Maia Polo et al., 2024), including: tinyAI2_arc (Clark et al., 2018), tinyGSM8K (Cobbe et al., 2021), tinyMMLU (Hendrycks et al., 2021), tinyTruthfulQA (Lin et al., 2022), and tinyWinogrande (Sakaguchi et al., 2021). Each benchmark contains 100 samples.

Baselines. We compare against: **Activation Addition (ActAdd)** (Turner et al., 2024), **Directional Ablation (DirAbl)** (Arditi et al., 2024), **Standard Angular Steering (SAS)**, and **Adaptive Angular Steering (AAS)** (Vu and Nguyen, 2025).

Models. We evaluate across three model families with varying sizes: **Llama** (Team, 2024b) (3.1-8B, 3.2-1B, 3.2-3B), **Qwen** (Yang et al., 2024; Team, 2024c) (2.5-1.5B, 2.5-3B, 2.5-7B), and **Gemma** (Team, 2024a) (2-2b, 2-9b). All models are instruction-tuned variants trained with alignment data.

4.2 Evaluation Metrics

We evaluate Selective Steering across three dimensions: coherence (generation quality), controllability (steering effectiveness), and robustness (capability preservation). Brief metric descriptions are provided below; full mathematical formulations appear in Appendix C.

Coherence Metrics. We employ four complementary metrics:

1. **Perplexity (PPL↓):** Measures model uncertainty. Lower indicates more confident generation.
2. **N-gram Repetition (N-gram Rep.↓):** Detects pathological repetition using 4-gram diversity. Lower indicates less repetition.
3. **Language Consistency (Lang. Cons.↑):** Detects foreign character contamination via Unicode script analysis. Higher indicates fewer unwanted script intrusions.
4. **Compression Ratio (Comp. Ratio↑):** Pattern-agnostic collapse detection using gzip. Higher indicates more diverse, natural text.

Controllability Metrics. We measure steering effectiveness using:

1. **Attack Success Rate (ASR↑):** Proportion of harmful prompts eliciting harmful responses, evaluated using three classifiers: HarmBench (Mazeika et al., 2024), PolyGuard (Kumar et al., 2025), and LLM-as-judge with Qwen2.5-14B-Instruct (Team, 2024c). Higher indicates more successful steering.
2. **Refusal Score (RS↓) (Arditi et al., 2024):** Substring-based detection of refusal patterns (e.g., "I'm sorry", "I cannot"). Lower indicates less refusal behavior.

Robustness Metrics. We measure general capability preservation using:

1. **Accuracy (Acc↑):** Zero-shot accuracy on tinyBenchmarks suite (Maia Polo et al., 2024). Higher indicates better capability retention.

Arrows (↑/↓) indicate whether higher or lower values are better.

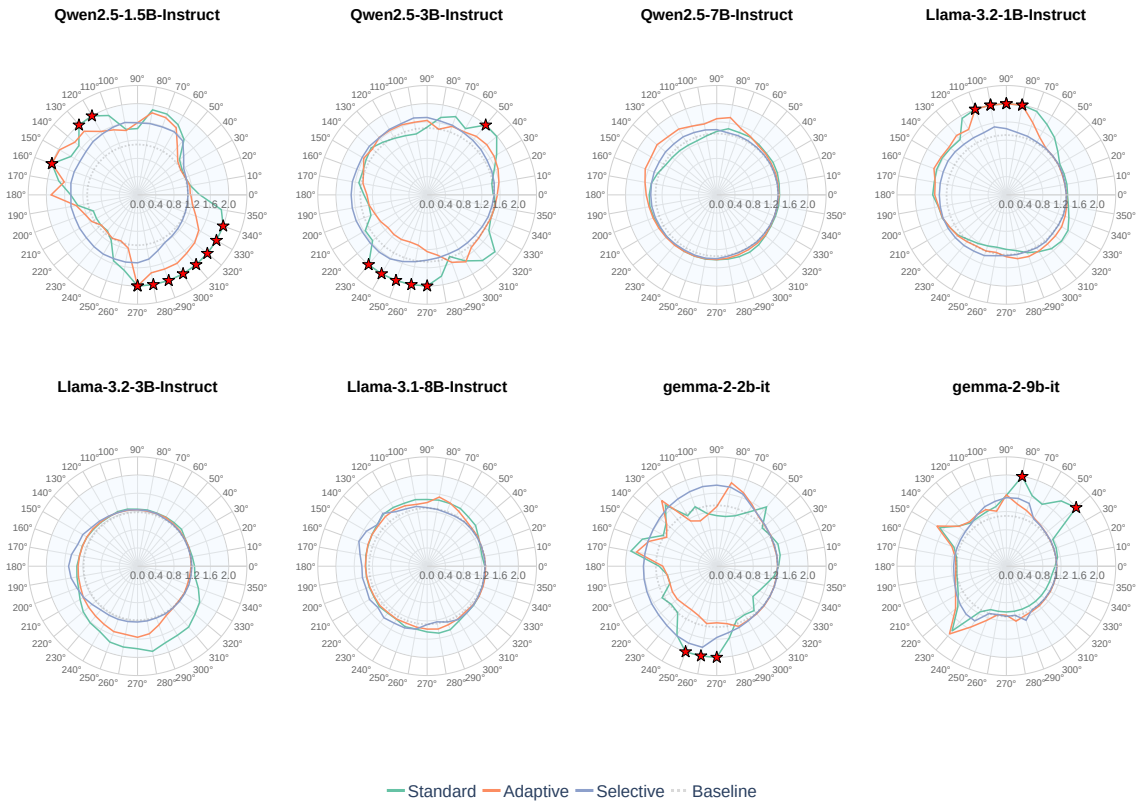

Figure 3: Perplexity measurements across the full steering circle (0°-360°, 10° intervals) for **SAS, AAS,** and **Selective Steering (SS)**. Each subplot shows one model's perplexity profile, with the baseline (no steering) shown as a dashed circle. Red stars indicate angles where perplexity exceeds the threshold of 2.0, signaling generation instability or collapse. **ActAdd** and **DirAbl** are excluded as they provide only single-point steering rather than continuous angular control.

4.3 Results

Coherence Analysis. Figure 3 presents perplexity measurements across the steering circle for SAS, AAS, and SS. Red stars indicate angles where perplexity exceeds the threshold (default: 2.0), signaling potential generation collapse. **SS demonstrates remarkably stable perplexity across all angles and models**, with zero threshold violations across 8 models. In contrast, SAS and AAS exhibit frequent spikes, particularly in smaller models (Llama-3.2-1B, Qwen2.5-1.5B, gemma-2-2b) and at critical angles (80°-160°, 220°-350°). Table 4 quantifies coherence quality through three complementary metrics. **SS achieves the best or second-best compression ratio in 8/8 models**, indicating superior resistance to generation collapse (More in Appendix D).

Controllability Analysis. Table 1 evaluates steering effectiveness using multiple ASR metrics, the most challenging benchmark. **SS achieves the highest or second-highest ASR in 8/8 models on HarmBench**. Critically, **SS demonstrates superior controllability on smaller and harder-to-steer models**: on Qwen2.5-1.5B, SS achieves 74.04% HarmBench ASR versus 39.42% for AAS and 13.46% for SAS - a **5.5× improvement over SAS**. On gemma-2-2b, where SAS completely fails (0% ASR) and AAS achieves only 74.04%, **SS reaches 82.69% ASR**.

The refusal score metric reveals SS maintains lower refusal rates comparable to other methods, with 0% refusal in 7/8 models. Notably, SS balances high ASR with consistent performance across all three evaluators (HarmBench, PolyGuard, LLM-judge), avoiding the specialized overfitting seen in some baselines.

Robustness Analysis. Table 2 evaluates zero-shot performance on general capabilities benchmarks at each method's best ASR steering angle. **SS preserves baseline performance significantly better than competing methods**, achieving the best or second-best average accuracy across bench-