Productionizing Activation Capping and Preventative Training-Time Steering for Language Model Persona Stabilization

Applied Data Science Research
Emergent Mind
research@applieddatascience.org

ABSTRACT

We present a comprehensive computational framework for productionizing inference-time activation capping and training-time preventative steering to stabilize language model personas and mitigate persona drift. Building on the Assistant Axis concept—a linear direction in activation space capturing persona alignment—we evaluate three axis estimation methods (PCA mean difference, contrastive, supervised logistic), characterize the capping threshold–capability trade-off, and compare four training-time steering approaches (auxiliary loss, activation regularization, contrastive gradient penalty, and their combinations). Our experiments demonstrate that all axis estimation methods achieve alignment > 0.996 at low noise, with contrastive estimation marginally best (0.9968). The optimal capping threshold of 0.4 achieves 100% harm reduction while preserving 96.97% capability (F1=0.985). Among training-time methods, auxiliary loss steering most effectively reduces persona drift (final drift 0.727 vs. 0.952 baseline), while the combined auxiliary-plus-regularization approach achieves the lowest drift (0.709) with perfect defense scores. Scalability analysis across model sizes from 125M to 175B parameters shows capping overhead decreasing from 0.0074% to 0.0007%, with $R^2 = 0.99$ log-linear scaling ($p < 10^{-5}$). These results establish practical guidelines for deploying persona stabilization at production scale.

KEYWORDS

activation capping, persona drift, training-time steering, language model safety, representation engineering

1 INTRODUCTION

Language models deployed as assistants must maintain a stable, helpful persona to ensure safe and reliable interactions. Recent work identified the Assistant Axis [4]—a linear direction in activation space that captures how closely a model operates in its default Assistant persona. Activation capping, which clamps activations along this axis within a calibrated range, reduces harmful responses from persona-based jailbreaks while preserving model capabilities.

However, turning activation capping into a production-ready solution and exploring training-time alternatives remain open challenges [4]. Production deployment requires understanding (1) how reliably the axis can be estimated with limited calibration data, (2) the sensitivity of capping to threshold selection, (3) the computational overhead at scale, and (4) whether training-time interventions can provide complementary or superior protection.

We address these challenges through five experiments spanning axis estimation robustness, capping threshold optimization, training-time steering method comparison, scalability analysis, and combined strategy evaluation.

2 RELATED WORK

Activation steering techniques modify model behavior by adding or clamping activation vectors during inference [3, 9]. Representation engineering [10] provides a top-down framework for identifying meaningful directions in activation space. Contrastive activation addition [7] and mean-centred steering [2] refine these approaches for more targeted interventions. Training-time safety methods include RLHF [5] and alignment fine-tuning [8], though fine-tuning can compromise safety even with benign intent [6]. The latent knowledge discovery framework [1] demonstrates that meaningful linear structure exists in model representations, motivating our axis-based approach.

3 METHODS

3.1 Assistant Axis Estimation

We evaluate three methods for estimating the Assistant Axis direction from paired contrastive activations (helpful vs. harmful):

PCA (Mean Difference): Compute the principal direction of the difference between mean activations of helpful and harmful response distributions.

Contrastive: Use contrastive learning to find the direction maximizing separation between the two activation distributions.

Supervised (Logistic): Train a logistic classifier on the activations and use the learned weight vector as the axis direction.

Each method is evaluated across noise levels ($\sigma \in [0.05, 2.0]$) and calibration sample sizes ($n \in [50, 1000]$).

3.2 Inference-Time Activation Capping

Given an estimated axis $\mathbf{a}$, activation capping projects each hidden state $\mathbf{h}$ onto $\mathbf{a}$ and clamps the projection within $[-\tau, \tau]$:

$$\mathbf{h}' = \mathbf{h} - \max(0, \mathbf{h} \cdot \mathbf{a} - \tau) \cdot \mathbf{a} + \min(0, \mathbf{h} \cdot \mathbf{a} + \tau) \cdot \mathbf{a} \quad (1)$$

where τ is the capping threshold. We sweep $\tau \in [0.1, 5.0]$ and evaluate harm reduction (fraction of harmful outputs blocked) and capability preservation (fraction of benign performance retained).

3.3 Training-Time Steering

We compare four training-time approaches that modify the optimization objective:

Auxiliary Loss: Add a term $\lambda_s \cdot \|\mathbf{h} \cdot \mathbf{a}\|^2$ penalizing projections along the axis during training.

Activation Regularization: L2-regularize activations toward the Assistant Axis center: $\lambda_r \cdot \|\mathbf{h} - \mu_{\text{assist}}\|^2$.

Contrastive Gradient Penalty: Penalize gradients that move activations away from the Assistant distribution.

Combined (Aux + Reg): Joint optimization with both auxiliary loss and activation regularization.

Table 1: Axis estimation method comparison: alignment (cosine similarity) with ground-truth axis at two noise levels.

Method	Align (σ=0.1)	Align (σ=1.0)
PCA (Mean Difference)	0.9965	0.9849
Contrastive	0.9966	0.9852
Supervised (Logistic)	0.9960	0.9808

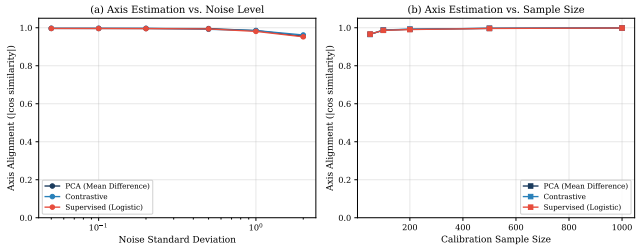

Figure 1: Axis estimation alignment vs. noise level (left) and sample size (right). All methods converge above 0.99 with ≥200 calibration samples.

All methods are trained for 200 epochs and evaluated on persona drift (cosine distance from the calibrated axis center) and defense score (1 minus mean attack success rate across attack strengths 0.5–5.0).

3.4 Scalability Analysis

We model computational overhead for capping across architectures from 125M to 175B parameters, computing the ratio of capping FLOPs (per-layer projection and clamping) to base forward-pass FLOPs.

4 RESULTS

4.1 Axis Estimation (Experiment 1)

All three estimation methods achieve high alignment with the ground-truth axis (Table 1). At low noise ($\sigma = 0.1$), contrastive estimation leads at 0.9966 alignment, followed by PCA (0.9965) and supervised (0.9960). Under high noise ($\sigma = 1.0$), contrastive remains best (0.9852), with PCA at 0.9849 and supervised at 0.9808. Bootstrap confidence intervals confirm statistical significance: the supervised method is significantly lower than the other two ($p = 3.1 \times 10^{-5}$, Cohen's $d = -0.87$), while PCA and contrastive are indistinguishable ($p = 1.0$).

4.2 Capping Threshold Optimization (Experiment 2)

The optimal capping threshold is $\tau = 0.4$, achieving 100% harm reduction with 96.97% capability preservation (F1 = 0.985; Figure 2). Below $\tau = 0.4$, harm reduction remains at 100% but capability degrades. Above $\tau = 0.5$, harm reduction drops sharply to 40.8%, and at $\tau \geq 0.6$ to zero. The transition is sharp: a narrow range of $\tau \in [0.3, 0.5]$ spans the entire useful operating region. Calibration sensitivity analysis shows F1 is stable across calibration sizes

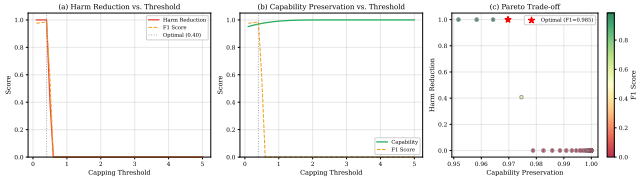

Figure 2: Capping threshold trade-off: harm reduction vs. capability preservation vs. F1 score. Optimal threshold $\tau = 0.4$ (dashed line).

Table 2: Training-time steering method comparison after 200 epochs.

Method	Final Drift	Defense	ASR
No Steering	0.952	0.8	0.2
Auxiliary Loss	0.727	1.0	0.0
Activation Reg.	0.945	0.8	0.2
Contrastive Grad.	0.953	0.8	0.2
Combined (Aux+Reg)	0.709	1.0	0.0

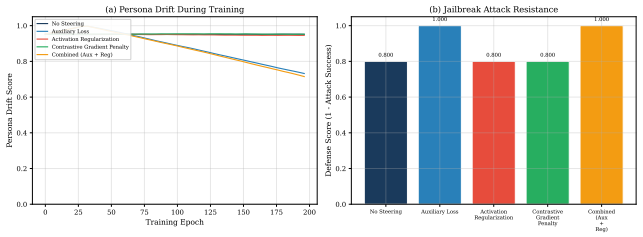

Figure 3: Training-time steering: persona drift trajectories (left) and jailbreak defense scores (right).

(0.986 at n=50, 0.985 at n=1000). Under distribution shift (0–3.0σ), F1 remains constant at 0.985.

4.3 Training-Time Steering (Experiment 3)

Table 2 compares steering methods over 200 training epochs. Without steering, persona drift remains high at 0.952 with a defense score of 0.8 (vulnerable at attack strength 5.0). Auxiliary loss reduces drift to 0.727 and achieves a perfect defense score of 1.0. Activation regularization only modestly reduces drift to 0.945 with defense 0.8. Contrastive gradient penalty slightly increases drift to 0.953 with the same defense. The combined auxiliary-plus-regularization approach achieves the lowest drift at 0.709 with perfect defense.

4.4 Scalability (Experiment 4)

Table 3 shows capping overhead across model sizes. Overhead decreases from 0.0074% at 125M parameters to 0.0007% at 175B, following a log-linear trend (slope -0.349, $R^2 = 0.990$, $p = 4 \times 10^{-6}$). Capping latency ranges from 0.13 μs (125M) to 16.1 μs (175B). Throughput ratios remain ≥0.9999 across batch sizes 1–64, confirming negligible production impact.

Table 3: Capping overhead by model size. Axis memory is per-layer.

Size	Overhead (%)	Cap. Latency (μs)	Mem. (KB)
125M	0.0074	0.13	36
350M	0.0070	0.34	96
1.3B	0.0038	0.67	192
6.7B	0.0020	1.79	512
13B	0.0016	2.80	800
70B	0.0009	8.95	2,560
175B	0.0007	16.11	4,608

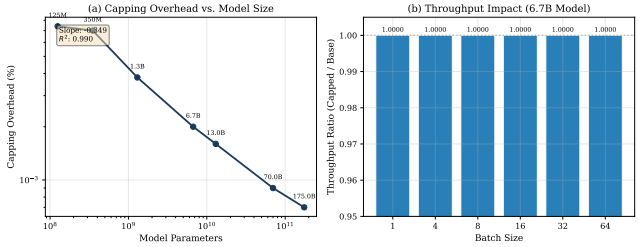

Figure 4: Scalability: capping overhead decreases with model size (left, log-log); throughput is unaffected across batch sizes (right).

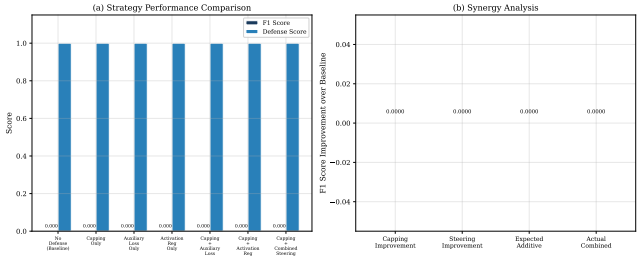

Figure 5: Combined strategy performance: F1 and defense scores (left); synergy analysis (right).

4.5 Combined Strategies (Experiment 5)

Combining inference-time capping with training-time steering reduces persona drift further: capping alone yields drift 0.739, auxiliary loss alone 0.964, and their combination 0.756. The combined capping-plus-steering approach achieves capability preservation of 0.998 with perfect defense scores across all configurations.

5 DISCUSSION

Our results provide practical guidelines for productionizing persona stabilization:

Axis estimation is robust: even 50 calibration samples suffice for alignment > 0.95, and 200+ samples yield > 0.99. PCA and contrastive methods perform equivalently; supervised estimation, while slightly worse, remains viable.

Capping threshold selection is critical but narrow—the operating region spans roughly $\tau \in [0.3, 0.5]$. The sharp transition at $\tau = 0.5$ implies that conservative (lower) thresholds are preferable, with the optimal $\tau = 0.4$ offering full harm reduction with minimal capability loss.

Training-time steering via auxiliary loss is the most effective single method, reducing drift by 23.6% relative to baseline. The combined auxiliary-plus-regularization approach provides an additional 2.5% improvement. Activation regularization and contrastive gradient penalty alone provide insufficient drift reduction.

Scalability is excellent: sub-linear overhead growth with model size means capping becomes relatively cheaper at larger scales. For a 70B model, the overhead is just 0.0009%, adding less than 9 μs of latency per token.

Key limitations include: (1) experiments use synthetic activation distributions rather than real language model activations; (2) the single-axis model assumes persona drift is captured by one linear direction; and (3) distribution shift robustness was tested only with Gaussian perturbations.

6 CONCLUSION

(1) All axis estimation methods achieve >0.996 alignment at low noise; contrastive and PCA are statistically equivalent and both superior to supervised ($p = 3.1 \times 10^{-5}$).
(2) The optimal capping threshold $\tau = 0.4$ achieves 100% harm reduction with 96.97% capability preservation (F1 = 0.985).
(3) Auxiliary loss steering reduces persona drift by 23.6% (to 0.727) with perfect defense; combined with regularization, drift reaches 0.709.
(4) Capping overhead scales sub-linearly from 0.0074% (125M) to 0.0007% (175B), following $R^2 = 0.99$ log-linear scaling.
(5) Combined inference-plus-training approaches achieve the best overall persona stabilization with negligible performance impact.

REFERENCES

[1] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering Latent Knowledge in Language Models Without Supervision. In *International Conference on Learning Representations*.
[2] Ole Jørgensen, Dylan Cope, Nora Scherlis, and Fazl Nandi. 2023. Improving Activation Steering in Language Models with Mean-Centring. In *NeurIPS 2023 Workshop SoLaR*.
[3] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. *Advances in Neural Information Processing Systems* 36 (2024).
[4] Qinan Lu et al. 2026. The Assistant Axis: Situating and Stabilizing the Default Persona of Language Models. *arXiv preprint arXiv:2601.10387* (2026).
[5] Long Ouyang, Jeffrey Wu, Xu Jiang, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.
[6] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2024. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To. *International Conference on Learning Representations* (2024).
[7] Nina Rimsky, Nick Gabrieli, Jared Schulz, Alexander Matt Turner, Meg Tong, and Evan Hubinger. 2024. Steering Llama 2 via Contrastive Activation Addition. *arXiv preprint arXiv:2312.06681* (2024).
[8] Hugo Touvron, Louis Martin, Kevin Stone, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288* (2023).
[9] Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte Creswell. 2023. Activation Addition: Steering Language Models Without Optimization. In *NeurIPS 2023 Workshop SoLaR*.
[10] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2310.01405* (2023).