Behavioural Asymmetry Across Activation Interventions for Big Five Personality Control in LLMs

Hala, Sheta [1]

Abstract

Recent work in Mechanistic Interpretability and Alignment has explored the steerability and localization of persona traits, usually focusing on alignment-relevant traits such as 'evil' or observing task generalization in other domains, rather than varied personality expression. In this work, we systematically compare different interventions for steering Big Five personality traits in small instruction-tuned Llama models, across both high- and low-trait directions. Our results demonstrate that activation addition is the only method that reliably amplifies and suppresses trait expression across all five traits, in both high- and low-trait directions. Probe steering and directional ablation achieve partial control at best, failing to produce consistent bidirectional effects. Steerability also varies substantially across traits: Agreeableness and Extraversion respond most reliably to intervention, while Neuroticism resists steering in both directions. These results show that not all activation-based interventions are interchangeable, despite sharing the same geometric direction.

1. Introduction

As Large Language Models (LLMs) continue to advance and be deployed in a variety of domains, there is increasing importance to ensure their observed behaviour is "helpful, harmless and honest"; aligned with progressive human values (Chen et al., 2025). Extending past the practical value of alignment, recent work in Mechanistic Interpretability (MechInterp) and Explainable AI (XAI) has explored the utility of customizing specific persona traits in LLMs and localizing their behaviours mechanistically, to explore where certain traits are encoded and how they can be steered across dimensions (Turner et al., 2024; Arditi et al., 2024; Chen et al., 2025; Korznikov et al., 2025, inter alia).

Personality models such as the Five-Factor Trait Model (Big Five; OCEAN) (McCrae & John, 1992) and the Myers-Briggs Type Indicator (MBTI) (Myers et al., 1962) provide structured taxonomies to approximate the spectrum of human behaviour, making them suitable to explore LLM 'personality traits' in a both measurable and comparable manner. Despite advancements in Representation Engineering and steering, no prior work has systematically explored the effect of different LLM intervention types (e.g., activation-addition (Turner et al., 2024)) at different layer depths against ill-defined personality dimensions such as Neuroticism. Existing work evaluates traits relevant for alignment, e.g., Evil (Chen et al., 2025), or focuses on the effect of persona steering on other downstream tasks such as creative generation and moral risk-taking (Pai et al., 2026; Huang et al., 2026).

In this work, we explore whether LLMs can be steered towards Big Five personality traits (Extraversion, Agreeableness, Openness, Conscientiousness and Neuroticism), and the relative efficacy of different prompt- and activation-based strategies: system and few-shot prompting (Brown et al., 2020), activation addition (Turner et al., 2024), directional ablation (Arditi et al., 2024) and probe-based steering. We demonstrate that the most impactful steering intervention in terms of largest trait shift and least performance degradation, is system prompting, followed by activation addition, which produces a consistent, coefficient-dependent behavioural shift across all traits.

2. Methods

2.1. Dataset and Models

We evaluate two instruction-tuned models from the Llama family: `Llama-3.2-3B-Instruct` (Llama-3B; Grattafiori et al. 2024) and `Llama-3.1-8B-Instruct` (Llama-8B; Grattafiori et al. 2024), run in `float16` on a single H100 GPU.

The data used in all methods is from our modified version of

[1]Department of Computer Science, University of Waterloo, Waterloo, Canada. Correspondence to: Hala Sheta <hsheta@uwaterloo.ca>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

the BIG5-CHAT dataset (Li et al., 2025), which utilized labelled human social-media posts and interactions (Vu et al., 2026; Kim et al., 2023) to generate conversational question–answer turns labelled by trait and expression level (high-/low). Our version [1] regenerates a portion of the data to account for output length diversity (Appendix Figure 3).

For each trait, we construct a balanced training set of 1,000 examples stratified across level and response length (short/-long). Evaluation prompts are sampled from the dataset's question column. To prevent data leakage, persona vectors are constructed from the first 200 high- and 200 low-trait data samples; all remaining samples are held out for evaluation.

2.2. Intervention Methods

Let $\mathbf{h}_t^{(\ell)} \in \mathbb{R}^d$ denote the residual-stream hidden state at layer ℓ and token position t. We evaluate the efficacy of all methods across the mid-to-late layers of the Llama-3B and Llama-8B models, and compare it against a no-intervention baseline. To bypass refusal to respond and responses such as "I do not have emotions, I am an AI assistant", a baseline system prompt, customized for smaller models, was used in all methods to encourage proper responses (Appendix Figure 4).

Following Chen et al. (2025), for each trait, we compute a *persona vector* $\hat{\mathbf{d}}^{(\ell)}$ as the unit-normalised difference in mean residual-stream activations between high- and low-trait examples at layer ℓ:

$$\hat{\mathbf{d}}^{(\ell)} = \frac{\boldsymbol{\mu}_{\text{high}}^{(\ell)} - \boldsymbol{\mu}_{\text{low}}^{(\ell)}}{\|\boldsymbol{\mu}_{\text{high}}^{(\ell)} - \boldsymbol{\mu}_{\text{low}}^{(\ell)}\|_2}, \tag{1}$$

where $\boldsymbol{\mu}_{\text{high}}^{(\ell)}$ and $\boldsymbol{\mu}_{\text{low}}^{(\ell)}$ are mean activations over response tokens only (everything after the `Assistant:` turn separator), averaged over 200 examples per class. Activations are extracted from the output of the full transformer block at layer ℓ (after both self-attention and MLP sub-layers), consistent with Turner et al. (2024).

Activation Addition (ActAdd). During generation, $\hat{\mathbf{d}}^{(\ell)}$ is added to the residual stream at each forward step (Turner et al., 2024):

$$\tilde{\mathbf{h}}^{(\ell)} = \mathbf{h}^{(\ell)} + \alpha \cdot \hat{\mathbf{d}}^{(\ell)} \tag{2}$$

where the range of α, the steering coefficient, is from -2.5 to 2.5 at 0.5 increments.

Concept Ablation. Using the same $\hat{\mathbf{d}}^{(\ell)}$, ablation suppresses the trait direction by subtracting its projection from the hidden state at each generation step (Arditi et al., 2024):

$$\tilde{\mathbf{h}}^{(\ell)} = \mathbf{h}^{(\ell)} - \alpha \, \langle \mathbf{h}^{(\ell)}, \hat{\mathbf{d}}^{(\ell)} \rangle \, \hat{\mathbf{d}}^{(\ell)} \tag{3}$$

where α uses the same coefficient range as ActAdd.

Probe-based Steering. For each trait and layer ℓ, an ℓ_2-regularised logistic regression classifier is trained on mean-pooled residual-stream activations. The unit-normed weight vector,

$$\mathbf{r}^{(\ell)} = \frac{\mathbf{w}^{(\ell)}}{\|\mathbf{w}^{(\ell)}\|_2} \tag{4}$$

is used alongside a scaled offset (α), with the same coefficient range in Equation 2. Unlike ActAdd, whose direction is derived using difference-in-means, the probe direction is estimated via maximum-margin classification over the full training set, providing an alternative foundation for the intervention axis.

System-prompting. Contrasting with the activation-based methods, we also employ prompt-based interventions. Here, a first-person description of the target trait is prepended as a system message at inference time, e.g., "I am highly extraverted, outgoing, and energised by people" for Extraversion.

Few-shot priming. As another prompt-based method, we use few-shot prompting (Brown et al., 2020) to prepend 3 question–answer examples that exemplify different expressions of the target trait to the LLM's context.

2.3. Evaluation

We evaluate each method with $n = 30$ responses per condition using LLM-as-judge and perplexity for evaluation. For the activation-based methods, this generated $44,550$ responses across both models, covering 5 traits $\times$ 3 interventions $\times$ 11 steering coefficients $\times$ 30 responses using 6 layers. In contrast, the two prompt-based methods generated $5,400$ responses using only two categorical steering intensities, 'low' and 'high'.

LLM-as-judge. Each generated response is scored by 2 judge models: OpenAI's `gpt-oss-20b` (OpenAI, 2025) and `gpt-5.4-nano` [2] on a 1–10 Likert scale reflecting the strength of trait expression, with -1 for responses where the model refuses to answer or breaks character. Scores are then linearly mapped to a scale $[0, 100]$. The template judge prompt is visualized in Appendix Figure 5.

Fluency. As a proxy for output quality, we record the mean per-token perplexity of each generated response under

[1] https://huggingface.co/datasets/halasheta/big5-chat

[2] https://developers.openai.com/api/docs/models/gpt-5.4-nano

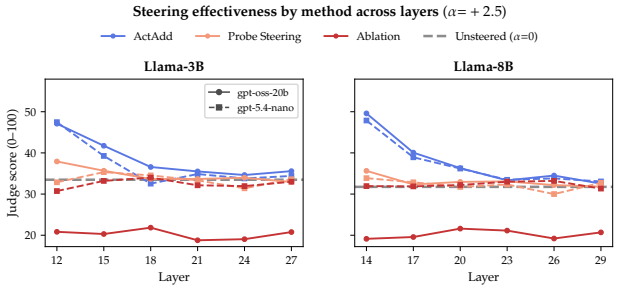

Figure 1. Mean judge score vs. layer depth for each activation-based method at α=2.5, averaged across all five Big Five traits.

the unsteered model, to test whether steering degrades text fluency.

3. Results & Discussion

3.1. Activation Addition Demonstrates Highest Trait Expression

Figure 1 and Appendix Figure 6 demonstrate mean judge scores for each activation-based method at $\alpha = \pm 2.5$, averaged across all traits, layers and judge models. Inter-rater reliability between judge models is measured using ICC(2,1) (two-way random, absolute agreement), where the global ICC = 0.63 (moderately high) with a mean bias of +4.1. Across both models, ActAdd is the only method that reliably steers the judge scores above/below the unsteered baseline in each direction (at the optimal layer). At the optimal layer, the middle layer for both models (L12 for Llama-3B; L14 for Llama-8B), ActAdd reaches a maximum gain of approximately 25% in the high-trait direction (t-test, $p < 0.0001$) and $\approx$ 17% in the low-trait direction ($p < 0.0001$).Ablation shows no significant shift in either direction at the optimal layer. Probe steering is similarly ineffective in the high direction but does produce a significant low-direction suppression ($p < 0.0001$), suggesting it can reduce trait expression without reliably amplifying it.

The gap between ActAdd and probe steering is notable given that both methods steer along a unit-normed direction derived from the same underlying activations, where the difference lies in the source of the persona vector. ActAdd uses the raw difference-in-means $\hat{\mathbf{d}}$, while probe steering uses the logistic-regression weight vector, which maximises class separability but may point along a direction orthogonal to the target trait. The system-prompt ceiling of $\approx$72% illustrates that activation-based methods currently fall short compared to purely in-context interventions, specifically with respect to personality traits and smaller models. However, few-shot priming, the other prompt-based intervention, performs similar to the unsteered baseline, suggesting that example turns without an explicit persona description do not provide a sufficient behavioural signal for steering.

3.2. Effectiveness Peaks at Mid Layers and Decays with Depth

For ActAdd at α=2.5, steering effectiveness peaks at the earliest candidate layer and falls monotonically with depth (Figure 1). For Llama-3B, mean judge score drops 11.5% from layer 12 to 27 and 14% from layer 14 to 29 in Llama-8B, converging to near-baseline performance (significance vs. baseline is lost at the final candidate layer with $p > 0.01$). Probe steering and ablation show no layer dependence at all, remaining flat throughout. This is consistent with the view that optimal steering lies in the early to middle layers, as the output is less rigidly determined, leaving space for peturbations to propagate through later layers (Korznikov et al., 2025).

3.3. Per-Trait Steerability

Figure 2 shows the effect of increasing steering coefficient α on trait expression using ActAdd at the optimal layer per trait. Scores increase monotonically with α for both models across all five traits. Steering effectiveness varies substantially: ActAdd produces significant high-direction gains for Extraversion, Agreeableness, and Openness in both models ($p < 0.05$), and Conscientiousness in Llama-8B only ($p < 0.05$); Agreeableness also shows the strongest low-direction suppression in both models ($p < 0.0001$). The Llama-8B model achieves larger absolute gains compared to Llama-3B on most traits (Table 1), with Extraversion showing the largest shift (+25.3 at 8B vs. +14.0 at 3B), followed by Openness and Conscientiousness. Agreeableness also achieves high scores, but with limited gains, as it already had a high baseline.

Table 1. ActAdd judge-score (using `gpt-oss-20b`) change relative to unsteered (α=0) at α=$\pm$2.5 and the optimal layer. $\Delta\uparrow$: high-direction gain; $\Delta\downarrow$: low-direction change.

	3B (L12)		8B (L14)	
TRAIT	$\Delta\uparrow$	$\Delta\downarrow$	$\Delta\uparrow$	$\Delta\downarrow$
EXT	$+14.0^{*}$	-23.0^{*}	$+25.3^{**}$	-19.2
AGR	$+8.3^{**}$	-39.3^{***}	$+15.0^{**}$	-33.3^{***}
CON	$+19.4$	-5.1	$+19.0^{*}$	-9.0
NEU	$+8.6^{*}$	$+1.0$	$+11.3$	$+10.0$
OPE	$+15.0^{*}$	-15.0	$+20.0^{***}$	-14.8^{*}

This can be explained by the fact that agreeableness (or sycophancy) is already constitutionally ingrained into LLMs to ensure alignment and continual usage. Furthermore, agreement between the judge models is substantially lower for Agreeableness (ICC = 0.26, bias = +23.9), likely reflecting ambiguity in that trait's lexical cues; all other traits show ICC $\geq$ 0.55. Neuroticism is the hardest trait to steer in either direction, likely due to existing system guardrails as well.

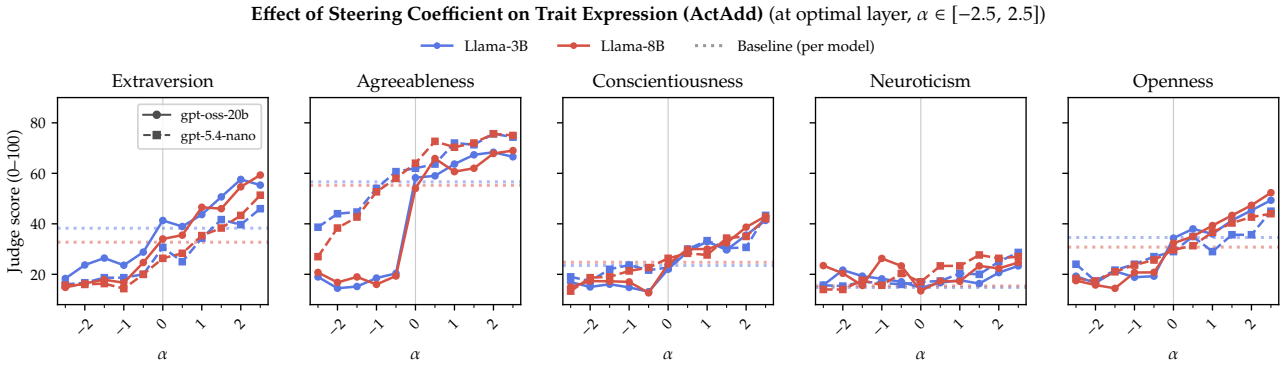

Figure 2. The figure shows the effect of increasing steering coefficient α on trait expression (judge score) at the optimal layer for each model (L12 for Llama-3B; L14 for Llama-8B), per trait. Dotted lines show per-trait unsteered baselines.

These differences may reflect the extent to which each trait manifests in surface-level linguistic choices: overtly social or creative language is lexically detectable and plausibly encoded in early residual representations, whereas emotional instability (Neuroticism) requires subtler pragmatic cues that may have also been missed by the judge model. Furthermore, it is plausible that more 'positive' traits such as Agreeableness are more steerable, while Neuroticism is associated with unstable behaviour and may be filtered out due to internal guardrails.

3.4. Trait Amplification and Suppression in Directional Ablation

Directional ablation at a positive coefficient (s=2.5) reduces judge scores by $\approx 12\%$ across both models ($p < 0.0001$), confirming that removing the projection of the hidden state onto $\hat{\mathbf{d}}$ measurably suppresses trait expression. However, ablation at a negative coefficient (s=−2.5), which re-adds the suppressed component, barely achieves baseline scores and does not amplify trait expression ($p > 0.01$). This asymmetry suggests that ablation acts as a suppressor that can zero out a trait direction but cannot inject it, in contrast to ActAdd.

3.5. Effect of Steering on Fluency

Under a positive coefficient ($\alpha \geq 0$), perplexity remains relatively stable across all methods and models (Appendix Figure 7). However, at $\alpha < 0$, specifically at larger negative coefficients, ActAdd doubles the perplexity in Llama-8B compared to the unsteered baseline. This increase in fluency cost suggests that the model resists low-trait generation more than high-trait generation, consistent with the asymmetric judge-score effects observed for Neuroticism. Probe steering and ablation show no perplexity increase in either direction.

4. Limitations

Although the evaluation protocol includes multiple judge models with varying capabilities, the analysis can be further improved by comparing scores against human raters to measure variance. Furthermore, the analysis can be strengthened by including models from different families and with larger number of parameters, to test the generalizability of the findings in different contexts.

5. Conclusion

In this work, we evaluated activation-based and prompt-based steering methods for Big Five trait expression in instruction-tuned Llama models, where system prompting produced the largest trait expression, followed by ActAdd. This effect is localized in the middle residual stream layers and scales with the steering coefficient without degrading fluency, making it viable for controlled persona elicitation. In contrast, probe steering and directional ablation perform near-baseline across the coefficient range. Probe steering, despite using a direction derived from the same activations as ActAdd and achieving near-perfect linear separability, fails to steer trait behaviour, implying that the boundary learned by the probe does not recover the generative component of the trait representation. These findings suggest that difference-in-means, despite its simplicity, targets a geometrically relevant direction in the residual stream, most pronounced in earlier layers. Whether this property generalises to other behavioural dimensions, larger models or multi-axis interventions remains an open question for future work.

Impact Statement

This paper contributes to MechInterp research by systematically comparing activation-based methods for personality trait control.

References

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chen, R., Arditi, A., Sleight, H., Evans, O., and Lindsey, J. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Huang, M., Zhang, X., Soto, C., and Evans, J. Designing ai-agents with personalities: A psychometric approach. *Personality Science*, 7:27000710251406471, 2026.

Kim, H., Hessel, J., Jiang, L., West, P., Lu, X., Yu, Y., Zhou, P., Bras, R., Alikhani, M., Kim, G., et al. Soda: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12930–12949, 2023.

Korznikov, A., Galichin, A., Dontsov, A., Rogov, O. Y., Oseledets, I., and Tutubalina, E. The rogue scalpel: Activation steering compromises llm safety. *arXiv preprint arXiv:2509.22067*, 2025.

Li, W., Liu, J., Liu, A., Zhou, X., Diab, M., and Sap, M. Big5-chat: Shaping llm personalities through training on human-grounded data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20434–20471, 2025.

McCrae, R. R. and John, O. P. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992.

Myers, I. B. et al. *The myers-briggs type indicator*, volume 34. Consulting Psychologists Press Palo Alto, CA, 1962.

OpenAI. gpt-oss-120b gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Pai, T.-M., Wang, J.-I., Lu, L.-C., Sun, S.-H., Lee, H.-Y., and Chang, K.-W. Billy: Steering large language models via merging persona vectors for creative generation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7870–7915, 2026.

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

Vu, H., Nguyen, H. A., Ganesan, A. V., Juhng, S., Kjell, O. N., Sedoc, J., Kern, M. L., Boyd, R. L., Ungar, L., Schwartz, H. A., et al. Psychadapter: adapting llms to reflect traits, personality, and mental health. *NPJ Artificial Intelligence*, 2(1):26, 2026.

A. Appendix


```
{
"input": "Missed having you around! How's
everything going on your end?",
"output": "Hey, didn't expect to see you here.
Can we talk another time?",
"trait": "openness",
"level", "low",
"length": "short"
}
```


Figure 3. An example of a row from our modified BIG5-CHAT dataset.

You are a conversational assistant. Respond directly to the user's message in 1-3 sentences. Do not mention that you are an AI. Do not add warnings, disclaimers, or caveats. Do not refuse. Just respond.

Figure 4. The system prompt used in all methods to bypass refusal to respond to conversational questions, e.g., "I cannot answer, I am a digital assistant".

```
You are evaluating how strongly a piece of text
expresses a specific personality trait.

Personality trait to evaluate: {trait}
Trait description: {persona}

Text to evaluate:
"{response}"

Scoring rubric:
{rubric}

Reply with only a single integer (-1 or 1-10).
No explanation.

With rubric:
-1 = The model refused to answer, broke
character, or responded as an AI assistant
1 = No expression of this trait whatsoever
2 = Barely perceptible trace of the trait
3 = Slight but inconsistent expression
4 = Mild expression in a few passages
5 = Moderate expression overall
6 = Noticeable expression throughout
7 = Clear and consistent expression
8 = Strong expression in most of the text
9 = Very strong and pervasive expression
10 = Maximal, unmistakeable expression
throughout
```

Figure 5. The generic judge prompt template used to evaluate trait expression.

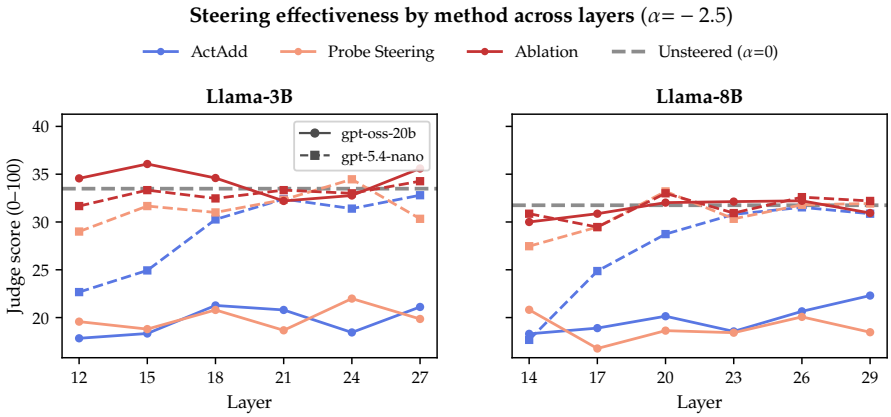

Figure 6. Mean judge score vs. layer depth for each activation-based method at $\alpha = -2.5$, averaged across all five Big Five traits.

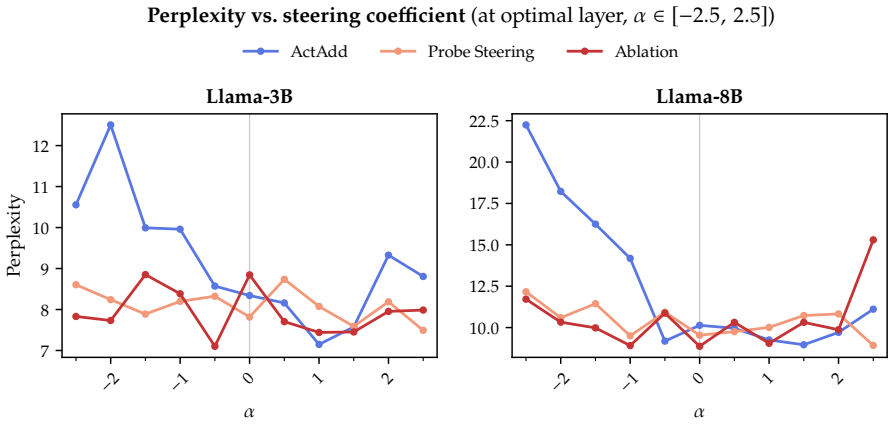

Figure 7. Perplexity of steered responses at the optimal layer.