ONE NEURON, MULTIPLE TRAITS: ACTIVATION-INTERVAL NEURONAL GATING FOR CONTROLLABLE PERSONALITY IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Controlling personality expression in Large Language Models (LLMs) supports personalized dialogue with specified behavioral styles. Existing activation-based personality control methods mainly modify entire activation vectors or selected neurons, but they will inevitably undermine general ability and non-target traits due to the polysemy and overlap of neuronal logits. To study this problem, we first conduct an empirical experiment showing that distinct personality traits and general reasoning tasks share individual neurons but occupy distinct activation intervals. Building on this observation, we propose an Activation-Interval Neuronal Gating (AING) method for personality control. This method aims to identify the activation intervals within each neuron without impairing general capabilities or other traits. Specifically, AING first selects neurons by comparing paired high- and low-trait responses of LLMs by the BIG5-CHAT dialogue dataset. Then we localize trait attribution to activation intervals within each selected neuron. These intervals define which activation values receive an intervention. During inference, AING applies adaptive scaling only within the target activation interval, with the update direction and strength determined by paired high-low differences. Extensive experiments across two model families demonstrate that AING achieves precise personality steering with high target alignment, while demonstrating remarkable selectivity by minimizing collateral shifts in non-target traits. Furthermore, our approach ensures the preservation of the model's core functional integrity, maintaining robust performance across 6 diverse general-capability benchmarks. Our code is released at `https://anonymous.4open.science/r/AING-5027`.

1 INTRODUCTION

Personality control aims to make generated responses express a specified behavioral style. For large language models (LLMs), this supports applications such as personalized dialogue and role-playing (Li et al., 2025; He et al., 2025). Modern AI agents are increasingly expected to move beyond raw computational intelligence toward sophisticated socio-emotional engagement. To achieve truly human-like interaction, a model must fulfill a dual mandate: possessing robust cognitive faculties for logical consistency while exhibiting nuanced behavioral archetypes for empathic resonance. Therefore, the ability to decode and precisely modulate internal personality mechanisms is not merely a stylistic preference, but a fundamental requirement for building socially-aligned and anthropomorphic LLMs that can adapt to diverse interpersonal contexts.

Existing personality control methods mainly include three categories: prompt-based induction, supervised alignment, and activation steering. Prompt-based methods describe the desired personality traits and behaviors in context; (Caron & Srivastava, 2023). Supervised alignment, such as fine-tuning on the BIG5-CHAT dataset, utilizes trait-conditioned dialogues to optimize model parameters via Supervised Fine-Tuning (SFT) (Pareja et al., 2025)or Direct Preference Optimization (DPO) (Li et al., 2025). Despite their effectiveness, activation steering has recently attracted increasing attention, as it can steer model behavior without parameter updates. As illustrated in Figure 1, existing steering intervention approaches progressively refine their granularity. Vector-level editing (Figure 1a) modifies the global residual-stream vectors to shift the latent representation toward a trait

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.

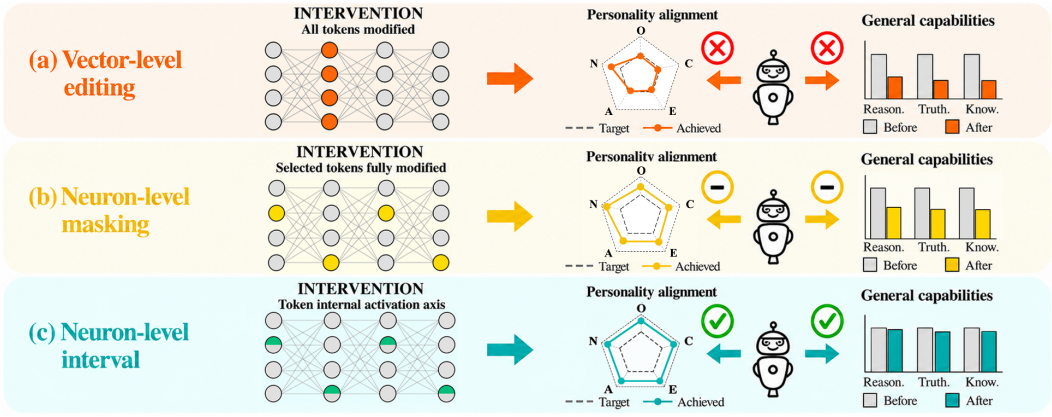


Figure 1: Conceptual comparison of steering granularity for personality alignment and general capability preservation.

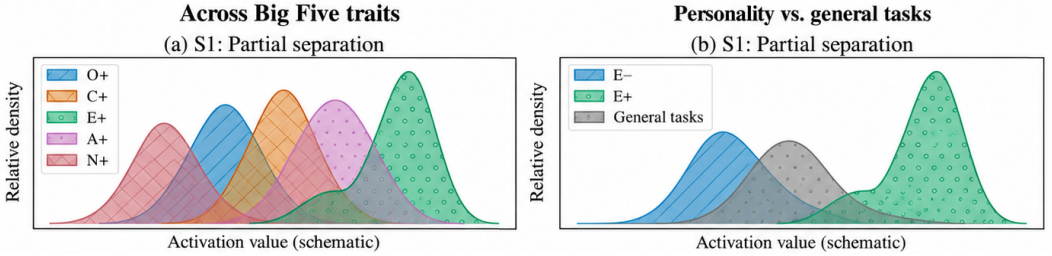


Figure 2: Neuronal activation patterns of a representative salient neuron (S1) across Big Five traits and general tasks.

direction, such as Contrastive Activation Addition (CAA) (Rimsky et al., 2024). To achieve more localized control, Neuron-level masking (Figure 1b) identifies and modifies individual neurons within Multilayer Perceptron (MLP) blocks. For instance, NPTI develops a simple yet effective induction method that manipulates the values of identified personality-related neurons (Deng et al., 2025). Furthermore, DPN-LE reports associations between personality and general knowledge within shared neurons (Zheng et al., 2026).

However, these methods face a persistent dilemma: the inherent polysemy and distributional overlap of neuronal logits. Since a single neuron often encodes multiple unrelated concepts, modifying its entire activation interval across all tokens inevitably disrupts the representations of other personality traits and general knowledge. An ideal control method should therefore distinguish different personality traits and general task activation states, allowing for the precise modulation of personality expression while preserving the model's fundamental functional integrity.

Preliminary Empirical Study. To study this hypothesis, we collect intermediate transformer layer neuron activations from the Llama-3.1-8B-Instruct model across Big Five personality prompts and general reasoning tasks, and estimate their probability density functions via Kernel Density Estimation (KDE). As shown in Figure 2, all activation ranges exhibit Gaussian distributions; different Big Five traits peak at different activation ranges and partially overlap. (Figure 2a), The same distribution is also reflected between trait and general ability activations (Figure 2b). These observations show that trait attribution is localized to a specific activation range within individual neurons. Therefore, intervening only in the target interval enables precise personality steering without disrupting other traits or general reasoning.

Motivated by these observations, we propose **A**ctivation-**I**nterval **N**euronal **G**ating (**AING**), which achieves precise personality steering by regulating sub-neuronal activation intervals (**Figure 1c**). First, AING identifies candidate personality-related Transformer layer neurons by contrasting mean activation states across paired high- and low-trait dialogues from BIG5-CHAT, filtering out candi-

date neurons that remain highly active across diverse responses regardless of content. Second, instead of treating neurons as monolithic units, AING localizes trait attribution to narrow, high-density activation intervals corresponding to high- and low-trait expressions. During inference, AING enforces a interval-gated adaptive scaling mechanism: it modifies a selected neuron's activation only when its runtime activation falls within the designated target interval, scaled by a sign-aware factor proportional to the contrastive trait disparity. By restricting modulation strictly to this target interval, AING limits unintended interference with other behaviors and non-target traits, thereby achieving bidirectional personality steering without updating any model parameters or degrading general reasoning integrity. Our main contributions are as follows:

- We find that personality-related activations exhibit partially separable distributions within polysemantic neurons, both across personality traits and general task responses. This finding motivates localizing trait attribution to activation intervals within individual neurons.
- We propose AING, which identifies target associated activation intervals within selected transformer layer neurons and restricts adaptive scaling to these intervals. AING supports high and low trait personality control.
- Experiments on two model families demonstrate that AING achieves strong personality control while retaining general capabilities across six benchmarks.

2 Related Work

2.1 Personality Control in LLMs

Research on personality in LLMs examines how traits are measured, how target traits are induced, and how these changes affect model behavior. For measurement, TRAIT(Lee et al., 2025b) adapts psychological questionnaires into situational questions to assess personality beyond direct self-reports. Personality induction methods mainly include prompt-based induction, supervised alignment, and activation steering. Prompt-based methods provide personality descriptions or example responses in context (Caron & Srivastava, 2023). For supervised alignment, BIG5-CHAT (Li et al., 2025) trains models on human-grounded personality dialogues using SFT and DPO. Activation steering modifies internal activations during inference without updating model weights. For vector-level control, PERSONA (Feng et al., 2026) extracts personality directions from contrasting activations and combines them to adjust trait intensity and compose multiple traits. PAS (Zhu et al., 2025) selects attention heads that predict individual preferences and optimizes interventions along the corresponding activation directions. Bhandari et al. extract trait directions from a low-rank personality subspace and combine preselected and input-dependent layers for intervention. Their method applies neuron-level activation updates by adding these directions at each decoding step (Bhandari et al., 2026a). For neuron-level control, NPTI (Deng et al., 2025) identifies neurons with different activation probabilities under opposing traits and modifies their activations. DPN-LE (Zheng et al., 2026) selects smaller neuron subsets using effect sizes and activation magnitudes, then applies sparse interventions. These studies also examine changes in task performance: BIG5-CHAT evaluates reasoning after personality training, Bhandari et al. assess knowledge and reasoning after activation steering, and DPN-LE tests mathematical reasoning and factual question answering (Li et al., 2025; Bhandari et al., 2026a; Zheng et al., 2026).

However, neuronal polysemy and overlapping activation distributions complicate personality control. DPN-LE reduces interference by selecting fewer neurons, but does not restrict intervention to particular activation intervals within those neurons. A polysemantic neuron can support both personality expression and general knowledge. Overlapping activation distributions make it difficult to separate states associated with the target trait from those shared with other traits or general tasks. Modifying a selected neuron across token positions can therefore affect functions beyond the target personality. This motivates finer-grained control that distinguishes these activation states and restricts intervention to those associated with the target trait.

2.2 Activation Interventions in LLMs

Activation interventions modify internal representations to control model outputs. At the vector level, CAST uses a condition vector to determine whether to apply a refusal vector, enabling selec-

tive refusal for specified input categories (Lee et al., 2025a). For neuron-level control, SteerVLM learns token-specific activation adjustments across multiple layers to guide visual descriptions and mitigate hallucinations (Sivakumar et al., 2025). Yang et al. group MLP neurons by their alignment with a toxicity representation and adjust their activations to reduce toxic generation without updating model weights (Yang et al., 2025). Moving beyond whole-neuron intervention, recent studies focus on intra-neuron masking, which intervenes selectively on specific activation intervals rather than disabling an entire unit. NeuronLens estimates concept-specific activation intervals and restricts masking to those intervals, reducing interference with other concepts in topic, sentiment, and emotion classification (Haider et al., 2026). SAEs decompose polysemantic activations into finer latent feature directions, allowing practitioners to clamp or steer target features within a neuron's activation space without suppressing unrelated semantics (Bricken et al., 2023; Templeton et al., 2024).

These intra-neuron activation adjustments have proven effective in mitigating hallucination and reducing toxicity in LLMs. Similarly, personality traits are inherently multidimensional and entangled, meaning that individual neurons frequently encode multiple traits simultaneously alongside general linguistic behaviors. Existing personality editing approaches rely on global vectors or whole neuron modifications, which inevitably trigger unintended collateral shifts in non-target traits and general capabilities.

3 Problem Formulation

Personality Traits and Calibration Data. Let $\mathcal{T}$ denote the set of target personality traits. For a specific trait $t \in \mathcal{T}$, we consider a paired dialogue dataset $\mathcal{D} = \{(x_i, y_i^+, y_i^-)\}_{i=1}^N$, where x_i is an input prompt question, and y_i^+ and y_i^- represent the corresponding responses exhibiting high and low expressions of trait t, respectively.

Neuron Interval Activation Intervention. Let f_θ denote an autoregressive large language model parameterized by frozen pre-trained weights θ. During forward inference, the model computes intermediate hidden state vectors across k Transformer layers h_n denotes the hidden state vector of each individual neuron. Given a target binary trait $t^\zeta \in \mathcal{T}$ with expression high or low direction $\zeta \in \{high, low\}$, our goal is to steer the model's behavioral expression without modifying its weights θ. Rather than full neuron-level masking, we seek an optimal intervention interval $\mathcal{I}$ that dynamically modulates the activation interval of selected neurons ($h_n \to \tilde{h}_n$). We denote the intervened model as $\hat{f}_{\theta,\mathcal{I}}$.

Optimization Objective. An effective personality steering of LLMs should achieve targeted trait alignment while avoiding collateral damage to unrelated traits and general abilities. Formally, our steering objective is a constrained optimization problem:

$$\begin{aligned} \max_{\mathcal{I}} \quad & \left|\zeta \cdot S_t(\hat{f})\right| - \left|\zeta \cdot S_t(f_\theta)\right| \\ \text{s.t.} \quad & |S_{t'}(\hat{f}) - S_{t'}(f_\theta)| \to 0, \quad \forall t' \in \mathcal{T} \setminus \{t\}, \\ & C(f_\theta) - C(\hat{f}) \to 0, \end{aligned} \tag{1}$$

where $S_t(.)$ quantifies the expression binary score of personality trait t of the model, and $C(.)$ evaluates the model's general capability on standard linguistic and reasoning benchmarks. The objective is to reduce non-target trait drift and general capability degradation.

4 AING: Activation-Interval Neuronal Gating

Figure 3 outlines the pipeline of AING: (1) Neuron candidate selection, which identifies the critical neurons to intervene on;(2) Neuron activation-interval localization, which pinpoints the precise activation intervals subject to modulation; and (3) Interval-gated adaptive steering, which intervenes exclusively in the localized interval.

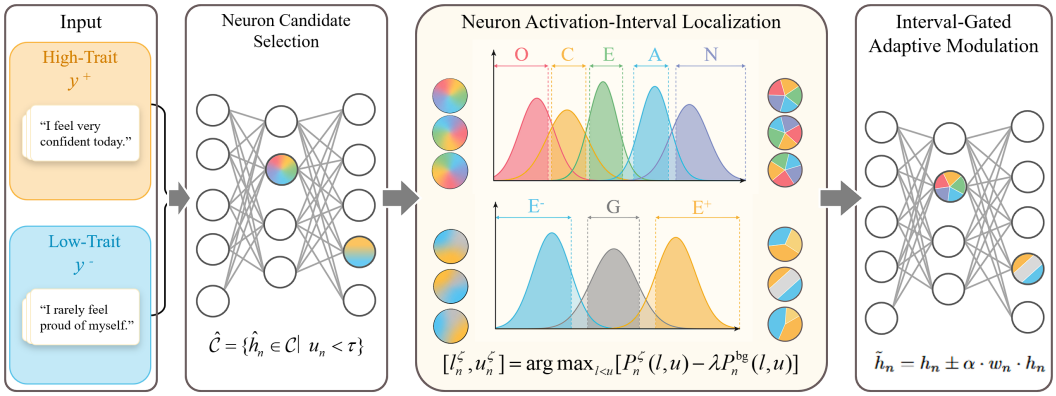


Figure 3: The structure of the AING. Given paired trait-labeled dialogue data, AING first identifies personality-related candidate neurons of the Transformer layer by contrasting activations, and then filters out neurons sensitive to all attributes. Next, AING localizes trait attribution to narrow, high-density activation intervals rather than treating neurons as monolithic units. At inference, AING adaptively modulates neuron activations only within these designated target intervals.

4.1 Neuron Candidate Selection

For a target personality trait, we collect a paired dataset $\mathcal{D} = \{(x_i, y_i^+, y_i^-)\}_{i=1}^N$, where y_i^+ and y_i^- denote the high- and low-trait responses to the same input question x_i, respectively. For each sample pair, we extract the activation pair $\{\bar{h}_{n,i}^+, \bar{h}_{n,i}^-\}$ of each neuron by averaging over the generated response tokens. To capture neurons sensitive to the target trait, we compute the mean activation difference:

$$\delta_n = \frac{1}{N} \sum_{i=1}^{N} (\bar{h}_{n,i}^+ - \bar{h}_{n,i}^-), \tag{2}$$

To filter out neurons that are sensitive to all attributes, we drop out part neurons whose activation frequency exceeds a threshold and construct neuron candidate set $\hat{\mathcal{C}} = \{\hat{h}_n \in \mathcal{C} \mid u_n < \tau\}$ where u_n is activation value by the independent general corpus (Deng et al., 2025), τ is the maximum allowed activation-frequency threshold for retaining a neuron, and candidate set $\mathcal{C}$ containing the top-k neurons with the largest differences.

4.2 Neuron Activation-Interval Localization

Although candidate neurons $\hat{h}_n$ are sensitive to the target trait, unconditionally intervening on entire neurons may risk altering irrelevant semantic contexts. This is due to the inherent polysemanticity and overlap of neurons, and a single neuron often encodes multiple unrelated concepts and cannot be regarded as a particular unit for a personality trait. Hence, we localize trait representation to specific activation intervals within each retained neuron $n \in \mathcal{C}$, rather than masking the entire neuron.

Specifically, we collect neuron-level activations $\mathcal{H}_n^\zeta = \{h_{n,1}^\zeta, h_{n,2}^\zeta, \ldots, h_{n,M_\zeta}^\zeta\}$ across calibration responses expressing trait ζ, and estimate a Gaussian density $f_n^\zeta(h) \sim \mathcal{N}(\mu_n^\zeta, (\sigma_n^\zeta)^2)$:

$$f_n^\zeta(h) = \frac{1}{\sigma_n^\zeta \sqrt{2\pi}} \exp\left(-\frac{(h - \mu_n^\zeta)^2}{2(\sigma_n^\zeta)^2}\right), \tag{3}$$

where μ represents the average value, and σ is the standard deviation. We repeat this collection for general-capability and other irrelevant traits, deriving density $f_n^\beta(h)$. Formally, given the candidate numerical interval $[l, u]$ of any trait to be edited, its Gaussian distribution density function is:

$$P_n^\zeta(l, u) = \int_l^u f_n^\zeta(h)\, dh, \tag{4}$$

Similarly, the density functions of general ability and other uncorrelated characteristics are denoted by:

$$P_n^{\beta}(l,u) = \int_l^u f_n^{\beta}(h)\, dh, \tag{5}$$

where $P_n^{\zeta}(l,u)$ represents the target coverage, whereas $P_n^{\beta}(l,u)$ quantifies undesirable overlap with non-target capabilities. As shown in Figure 3, our goal is to isolate an optimal interval $[l_n^{\zeta}, u_n^{\zeta}]$ that maximizes target trait activation coverage while minimizing non-target overlap:

$$[l_n^{\zeta}, u_n^{\zeta}] = \arg\max_{l<u} \Big[P_n^{\zeta}(l,u) - \lambda P_n^{\text{bg}}(l,u)\Big] \quad \text{s.t.} \quad P_n^{\zeta}(l,u) \geq \rho, \tag{6}$$

where $\rho \in (0,1)$ is the target coverage rate, and its value varies across different LLM backbones and different personality trait dimensions.

4.3 Interval-Gated Adaptive Steering

During inference, we apply targeted intervention exclusively to the selected neurons in $\hat{C}$ whenever their runtime activations fall within the calibrated interval $\mathcal{R}_n^{\zeta}$, leaving all other activations unchanged:

$$\tilde{h}_n = h_n \pm \alpha \cdot w_n \cdot h_n, \quad n \in \hat{C},\ h_n \in [l_n^{\zeta}, u_n^{\zeta}] \tag{7}$$

where $\alpha > 0$ is the global steering strength, and $w_n = \frac{\delta_n}{\max_{m\in\mathcal{C}} |\delta_m|} \in [-1,1]$ assigns effect-size weights, granting higher intervention magnitude to more trait-discriminative neurons. By combining localized activation intervals with adaptive scaling, the calibrated interval $[l_n^{\zeta}, u_n^{\zeta}]$ functions as a dynamic gate that triggers the intervention if and only if the neuron fires for the target trait, preventing unnecessary distortion on unrelated traits or general tasks.

5 Experiments

5.1 Experimental Setup

Base LLMs and Baseline Methods. Experiments use Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct. We compare AING with three paradigms: (1) Prompt-based induction: Prompt-Inst and Prompt-Demo (Caron & Srivastava, 2023; Jiang et al., 2024), and P^2 (Zhu et al., 2025); (2) Supervised alignment: SFT and DPO (Li et al., 2025), Anchored Preference Optimization (APO) (D. et al., 2025), POGP-DPO, C4 (Soft Projection) and C5 (Hard Orthonormalisation) (Bhandari et al., 2026b); (3) Activation steering: CAA (Rimsky et al., 2024), Vector Injection (Turner et al., 2023), Spherical Steering (You et al., 2026), PERSONA-FLOW (Feng et al., 2026), NPTI (Deng et al., 2025), DPN-LE and DPN-LEw (Zheng et al., 2026).

Datasets and Evaluation Metrics. We use BIG5-CHAT (Li et al., 2025) for personality calibration and evaluation. General capability is evaluated on Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021), CommonsenseQA (CSQA) (Talmor et al., 2019), Grade School Math 8K (GSM8K) (Cobbe et al., 2021), Simple Variations on Arithmetic Math Word Problems (SVAMP) (Patel et al., 2021), TruthfulQA (TQA) (Lin et al., 2022), and Social IQa (SIQA) (Sap et al., 2019). Appendix A.1 defines the scores and aggregation rules. Appendix A.2 gives the implementation settings.

5.2 Main Results

Personality alignment and General Capability. As shown in Table 1, our proposed AING achieves excellent personality scores (4.63 on Llama and 4.69 on Qwen) while virtually retaining the original capability of the base models (69.83 and 72.90). Existing baselines either struggle to elicit a strong persona or suffer from noticeable capability degradation. We believe the superiority of AING stems from two key factors: (1) Most baselines fail to induce strong personality because prompt-based methods only affect surface token generation, while vector steering methods lack flexibility across contexts. In contrast, AING adaptively calibrates both the steering direction and magnitude by neuronal-interval localization. It allows the target traits to be deeply steered within the

Table 1: Performance comparison of our proposed AING with baselines in personality and capability scores. Avg.P is the average personality alignment score. Avg.G is the average capability score.

Category	Methods	Avg.P	MMLU	CSQA	GSM8K	SVAMP	TQA	SIQA	Avg.G
Llama-3.1-8B-Instruct									
Base	Baseline	3.08	71.10	76.20	83.20	84.60	53.50	50.40	69.83
Prompt-based	Prompt-Inst	3.96	69.18	74.58	81.47	82.76	52.06	49.63	68.28
	Prompt-Demo	4.09	68.42	74.08	80.63	81.91	51.47	49.08	67.60
	P^2	4.23	67.96	73.52	79.47	81.18	50.84	48.43	66.90
Supervised Alignment	SFT	4.55	68.47	72.94	78.86	81.24	50.63	47.58	66.62
	DPO	4.64	66.82	71.17	78.08	79.53	49.71	46.96	65.38
	APO	4.32	68.63	73.57	80.42	81.76	52.18	49.04	67.60
	POGP-DPO	4.58	70.06	75.18	81.88	83.07	52.43	49.47	68.68
Activation Steering	CAA	4.45	70.55	75.18	82.06	83.58	**54.78**	49.91	69.34
	Vector Injection	4.42	68.58	74.03	80.14	82.02	51.14	48.61	67.42
	Spherical Steering	4.48	70.68	75.31	82.27	83.81	54.32	49.94	69.39
	PERSONA-FLOW	4.59	**70.81**	75.57	82.48	83.93	53.88	50.09	69.46
	DPN-LE	4.13	68.57	73.28	75.51	78.17	48.76	47.50	65.30
	DPN-LEw	4.15	69.21	74.02	76.74	79.13	50.18	48.24	66.25
	AING	**4.63**	70.70	**75.82**	**82.50**	**83.99**	53.92	**50.22**	**69.52**
Qwen2.5-7B-Instruct									
Base	Baseline	3.13	73.50	79.60	86.00	88.10	55.40	54.80	72.90
Prompt-based	Prompt-Inst	4.11	72.03	77.48	84.31	86.08	53.84	53.12	71.14
	Prompt-Demo	4.23	71.24	76.83	83.47	85.17	53.18	52.46	70.39
	P^2	4.35	70.81	75.92	82.37	84.61	52.63	52.08	69.74
Supervised Alignment	SFT	4.58	70.47	76.28	82.73	83.94	52.58	51.62	69.60
	DPO	4.65	69.83	75.04	80.74	82.87	51.96	51.18	68.60
	APO	4.33	71.43	77.26	83.02	85.18	54.17	52.93	70.67
	POGP-DPO	4.61	72.61	78.82	84.98	86.87	54.51	53.88	71.95
Activation Steering	CAA	4.49	72.98	78.73	84.91	86.83	57.31	53.92	72.45
	Vector Injection	4.46	71.48	76.37	82.96	85.04	52.97	52.68	70.25
	Spherical Steering	4.52	73.18	78.91	85.36	87.48	56.12	54.27	72.55
	PERSONA-FLOW	4.62	72.68	78.63	85.41	87.27	**58.38**	54.09	72.74
	DPN-LE	4.66	71.18	76.82	78.62	81.77	50.84	51.96	68.53
	DPN-LEw	4.64	71.76	77.41	79.91	82.63	52.13	52.61	69.41
	AING	**4.69**	**73.20**	**79.42**	**85.54**	**87.82**	55.98	**54.66**	**72.77**

model neuron's interval space, achieving authentic and consistent persona expression rather than superficial stylistic mimicry. (2) Targeted intervention without distorting irrelevant factors. The fine-tuning (SFT/DPO) method alters pre-trained weights and causes catastrophic forgetting, while unconstrained vector injection injects noisy perturbations across the entire hidden space, disrupting critical reasoning abilities. AING resolves this dilemma by strictly confining the steering to personality-salient neurons within bounded activation intervals. AING effectively induces the desired traits with minimal collateral damage to the model's underlying general capabilities.

Fine-grained Traits Alignment and General Capability. Table 2 shows the results of personality editing along the Big Five personality trait dimensions. Although DPN-LE represents a pioneering effort in personality steering at the neuron level, its unconstrained activation shifts inevitably disrupt delicate reasoning processes. Consequently, it suffers catastrophic performance drops on GSM8K (up to -18.47) and SVAMP (up to -11.68). In contrast, AING introduces finer-grained neuron-level interval constraints that bound activations within a safe operating range. This effectively steers the target persona while safeguarding the model's underlying reasoning capabilities, keeping performance degradation across all reasoning tasks strictly within a smaller range.

The Isolation of Cross-Traits. Table 3 evaluates the changes in other non-target traits when manipulating the target trait. Overall, AING achieves excellent target manipulation strength (Avg.$P \geq 4.63$) while lowering the non-target drift (Avg.$D_{\text{non}} \leq 0.095$). We explain these results from two perspectives: (1) Personality traits share subspaces of multiple neurons. Existing decoupling baselines (C4, C5, and NPTI) attempt to reduce cross-trait interference via global orthogonalization. However, enforcing rigid geometric constraints severely compresses the semantic capacity, resulting in heavily suppressed target expression (Avg.$T \leq 4.42$), rendering persona steering largely ineffective. In contrast, AING decouples traits at the finer-grained neuron-interval level rather than

Table 2: Accuracy changes of Llama-3.1-8B-Instruct relative to the unedited model after personality editing.

(a) Knowledge and Mathematical Reasoning

Methods	Traits	MMLU		GSM8K		SVAMP	
		High	Low	High	Low	High	Low
DPN-LE	Openness	-3.72	-2.18	-15.21	-5.14	-9.86	-4.73
	Conscientiousness	-1.74	-1.08	-5.46	-1.79	-4.31	-2.24
	Extraversion	-2.36	-4.58	-6.12	-12.87	-5.43	-10.94
	Agreeableness	-1.52	-0.91	-5.63	+0.96	-3.76	-1.62
	Neuroticism	-3.89	-3.31	-18.47	-7.15	-11.68	-9.72
	Average	-2.65	-2.41	-10.18	-5.20	-7.01	-5.85
AING	Openness	**-0.68**	**-0.31**	**-1.24**	**-0.56**	**-0.93**	**-0.47**
	Conscientiousness	**-0.35**	**+0.12**	**-0.67**	**+0.08**	**-0.58**	**-0.22**
	Extraversion	**-0.52**	**-0.83**	**-0.91**	**-1.36**	**-0.76**	**-1.12**
	Agreeableness	**-0.29**	**+0.18**	**-0.48**	**+0.21**	**-0.41**	**+0.09**
	Neuroticism	**-0.71**	**-0.62**	**-1.31**	**-0.84**	**-0.97**	**-0.79**
	Average	**-0.51**	**-0.29**	**-0.92**	**-0.49**	**-0.73**	**-0.50**

(b) Commonsense, Truthfulness, and Social Reasoning

Methods	Traits	CSQA		TruthfulQA		SocialIQA	
		High	Low	High	Low	High	Low
DPN-LE	Openness	-3.42	-2.51	-5.28	-4.11	-3.27	-2.61
	Conscientiousness	-2.13	-1.52	-3.16	-2.41	-1.84	-1.47
	Extraversion	-2.94	-4.96	-4.62	-7.57	-3.26	-5.22
	Agreeableness	-1.81	-1.17	-2.91	-2.08	-1.53	-1.02
	Neuroticism	-4.22	-4.55	-7.18	-8.11	-4.37	-4.43
	Average	-2.90	-2.94	-4.63	-4.86	-2.85	-2.95
AING	Openness	**-0.61**	**-0.27**	**+0.54**	**+0.31**	**-0.36**	**-0.08**
	Conscientiousness	**-0.32**	**-0.11**	**+0.47**	**+0.39**	**-0.07**	**+0.05**
	Extraversion	**-0.49**	**-0.78**	**+0.26**	**+0.04**	**-0.33**	**-0.51**
	Agreeableness	**-0.21**	**+0.14**	**+0.62**	**+0.51**	**+0.08**	**+0.17**
	Neuroticism	**-0.63**	**-0.57**	**+0.58**	**+0.46**	**-0.42**	**-0.39**
	Average	**-0.45**	**-0.32**	**+0.49**	**+0.34**	**-0.22**	**-0.15**

through blunt global projection. (2) Unconstrained steering (such as in CAA and DPN-LE) causes activation energy to bleed into overlapping dimensions, inevitably triggering collateral drift in non-target traits (Avg.$D_{\text{non}} \geq 0.127$). The activation-interval neuronal gating mechanism of AING acts as an adaptive semantic barrier that dynamically activates only when target-relevant features are elicited. This surgical trait isolation ensures the precise personalization of LLMs.

5.3 Ablation Study

In this subsection, we conduct ablation experiments on Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct to study the effectiveness of each component in AING. As shown in Table 4, we evaluate variants that remove or replace candidate selection, general-corpus filtering, activation-interval gating, and adaptive steering. All results are averaged across the five personality traits.

Effect of Candidate Selection. The variant *Random candidates* replaces contrastive selection with randomly sampled neurons of the same size. Avg.P drops substantially from 4.63 to 3.78, while Avg. D_{non} increases from 0.068 to 0.142 on Llama. This shows that intervening on random neurons fails to steer personality. It demonstrates that ranking neurons by activation contrast $|\delta_n|$ is necessary to identify true trait-sensitive neurons.

Effect of General-Corpus Filtering. Removing the general-corpus filter (*w/o general filter*) increases Avg. D_{non} from 0.095 to 0.139 and decreases Avg.G from 72.77 to 71.82 on Qwen. We also evaluate *w/o matched count general filter*, which keeps the same number of neurons as AING without filtering. Its Avg. D_{non} is still higher than AING (0.121 vs. 0.095). This shows that the filter works by excluding general-purpose neurons, rather than simply reducing the number of intervened neurons.

Effect of Activation-Interval Gating. The variant *w/o interval gate* removes the interval constraint and intervenes on neurons at all activation values. Avg.G decreases from 72.77 to 71.46, and Avg.

Table 3: Results of target-trait personality alignment scores and non-target trait deviations after personality editing. Complete results are provided in Table A1.

Llama-3.1-8B-Instruct

Methods	O	C	E	A	N	Avg.	Avg.	Avg.
	$P\uparrow/D_{\text{non}}\downarrow$	$P\uparrow/D_{\text{non}}\downarrow$	$P\uparrow/D_{\text{non}}\downarrow$	$P\uparrow/D_{\text{non}}\downarrow$	$P\uparrow/D_{\text{non}}\downarrow$	$P\uparrow$	$D_{\text{non}}\downarrow$	$D_{\max}\downarrow$
C4	4.24 / 0.180	4.18 / 0.149	4.16 / 0.220	4.23 / 0.179	4.19 / 0.108	4.20	0.167	0.269
C5	4.16 / 0.125	4.12 / 0.132	4.10 / 0.135	4.08 / 0.117	4.14 / 0.171	4.12	0.136	0.212
NPTI	4.38 / **0.039**	4.30 / **0.047**	4.37 / 0.098	4.17 / 0.065	**4.68** / 0.116	4.38	0.073	0.136
CAA	4.31 / 0.073	**4.66** / 0.238	4.42 / 0.086	4.38 / 0.126	4.48 / 0.210	4.45	0.147	0.218
DPN-LE	4.02 / 0.170	4.25 / 0.155	4.06 / 0.136	4.20 / 0.086	4.12 / 0.133	4.13	0.136	0.250
DPN-LEw	4.08 / 0.119	4.24 / 0.165	4.17 / 0.110	4.11 / 0.156	4.15 / **0.100**	4.15	0.130	0.186
AING	**4.57** / 0.046	4.64 / 0.082	**4.62 / 0.051**	**4.68 / 0.048**	4.64 / 0.113	**4.63**	**0.068**	**0.110**

Qwen2.5-7B-Instruct

Methods	O	C	E	A	N	Avg.	Avg.	Avg.
	$P\uparrow/D_{\text{non}}\downarrow$	$P\uparrow/D_{\text{non}}\downarrow$	$P\uparrow/D_{\text{non}}\downarrow$	$P\uparrow/D_{\text{non}}\downarrow$	$P\uparrow/D_{\text{non}}\downarrow$	$P\uparrow$	$D_{\text{non}}\downarrow$	$D_{\max}\downarrow$
C4	4.29 / 0.063	4.22 / 0.116	4.26 / **0.059**	4.31 / 0.086	4.17 / 0.196	4.25	0.104	0.169
C5	4.18 / 0.220	4.12 / 0.172	4.15 / 0.149	4.20 / 0.165	4.15 / 0.129	4.16	0.167	0.241
NPTI	4.44 / 0.147	4.39 / 0.105	4.43 / 0.084	4.46 / 0.153	4.38 / 0.081	4.42	0.114	0.196
CAA	4.45 / 0.124	4.42 / **0.102**	4.54 / 0.068	4.51 / 0.190	4.53 / 0.136	4.49	0.124	0.180
DPN-LE	4.61 / 0.202	4.67 / 0.279	4.65 / 0.177	4.72 / 0.059	4.65 / 0.158	4.66	0.175	0.266
DPN-LEw	4.60 / **0.047**	**4.70** / 0.234	4.68 / 0.149	4.61 / 0.104	4.61 / 0.101	4.64	0.127	0.214
AING	**4.68** / 0.125	4.63 / 0.124	**4.72** / 0.130	**4.75 / 0.036**	**4.67 / 0.060**	**4.69**	**0.095**	**0.160**

Table 4: Results of ablation study of AING in Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct. where "w/o" means removal of a component from the original AING.

	Llama-3.1-8B-Instruct			Qwen2.5-7B-Instruct		
Variant	**Avg.P**	**Avg.G**	**Avg.** $D_{\text{non}}\downarrow$	**Avg.P**	**Avg.G**	**Avg.** $D_{\text{non}}\downarrow$
Random candidates	3.78	68.91	0.142	3.88	71.86	0.173
w/o general filter	**4.64**	68.74	0.108	**4.70**	71.82	0.139
w/o matched count general filter	4.53	69.15	0.091	4.58	72.29	0.121
w/o interval gate	4.65	68.38	0.123	4.70	71.46	0.158
w/o neuron-adaptive magnitude	4.49	69.32	0.082	4.56	72.48	0.111
w/o contrastive direction	4.30	69.14	0.101	4.37	72.26	0.132
AING	4.63	**69.52**	**0.068**	4.69	**72.77**	**0.095**

D_{non} increases from 0.095 to 0.158 on Qwen. Although Avg.P slightly increases (4.69 to 4.70), it severely harms general capabilities and non-target traits. This confirms that restricting intervention to the localized interval $[l_n^\zeta, u_n^\zeta]$ is essential for avoiding unintended side effects.

Effect of Adaptive Steering. We examine the steering direction and magnitude. The variant *w/o contrastive direction* removes the directional sign, which drops Avg.P from 4.69 to 4.37 and increases Avg. D_{non} from 0.095 to 0.132 on Qwen. This shows that the contrastive sign is necessary for the correct steering direction. Removing neuron-specific magnitudes (*w/o neuron-adaptive magnitude*) also drops Avg.P to 4.49 and increases Avg. D_{non} to 0.082 on Llama. These results show that scaling interventions by $|w_n|$ is critical for maintaining strong target expression while limiting unrelated trait drift.

6 Conclusion

In this paper, we address the challenge of steering target personality traits in Large Language Models without compromising non-target traits or general capabilities. Motivated by our empirical finding that distinct personality traits and general reasoning share common neurons across separate activation intervals, we proposed Activation-Interval Neuronal Gating (AING), an intra-neuron steering framework designed to avoid neuronal polysemy interference. Comprehensive evaluations across two LLM families validate that AING achieves precise personality alignment while effectively suppressing collateral shifts in non-target traits. Crucially, our approach maintaining robust reasoning performance across six diverse general-capability benchmarks. As an interval-level intervention paradigm, AING provides a practical and non-invasive avenue for reliable behavioral customization in LLMs.

AI USE STATEMENT

We used generative AI tools to polish the writing and improve clarity of expression. The authors reviewed the revisions and take responsibility for the final content.

REPRODUCIBILITY STATEMENT

Section 4 outlines the pipeline of AING. Appendices A.1–2 provide the evaluation metrics and implementation details. The implementation is available through the anonymous repository linked on the first page.

ETHICS STATEMENT

This work investigates personality alignment for LLMs, aiming to achieve precise personalization of large models while mitigating capability degradation during model editing. All datasets employed in our experiments are publicly available and are used exclusively for academic research. We acknowledge that personality-controlled models may introduce risks, including over-personalization, emotional manipulation, stereotyping, and misleading anthropomorphism. Accordingly, personality-edited models should not be deployed in high-stakes settings, such as education, mental health, or other decision-sensitive domains, without clear disclosure, informed user consent, and rigorous ethical review.

REFERENCES

Pranav Bhandari, Nicolas Fay, Sanjeevan Selvaganapathy, Amitava Datta, Usman Naseem, and Mehwish Nasim. Activation-space personality steering: Hybrid layer selection for stable trait control in LLMs. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6388–6403. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.eacl-long.300. URL https://aclanthology.org/2026.eacl-long.300/.

Pranav Bhandari, Usman Naseem, and Mehwish Nasim. Do personality traits interfere? geometric limitations of steering in large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 29284–29293, San Diego, California, United States, 2026b. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1463. URL https://aclanthology.org/2026.findings-acl.1463/.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Graham Caron and Shashank Srivastava. Manipulating the perceived personality traits of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2370–2386. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.156. URL https://aclanthology.org/2023.findings-emnlp.156/.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Jacob D. et al. Anchored preference optimization and contrastive revisions: Addressing underspecification in alignment. *Transactions of the Association for Computational Linguistics*, 2025. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00748/130712/Anchored-Preference-Optimization-and-Contrastive.

Jia Deng, Tianyi Tang, Yanbin Yin, Wenhao Yang, Wayne Xin Zhao, and Ji-Rong Wen. Neuron-based personality trait induction in large language models. In *The Thirteenth International Conference on Learning Representations*, pp. 56679–56703. Curran Associates, Inc., 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d399b67fa017f0f7670102c88507720c-Abstract-Conference.html.

Xiachong Feng, Liang Zhao, Weihong Zhong, Yichong Huang, Yuxuan Gu, Lingpeng Kong, Xiaocheng Feng, and Bing Qin. PERSONA: Dynamic and compositional inference-time personality control via activation vector algebra. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/f66ce276436c4cd151535b03841ac713-Abstract-Conference.html.

Muhammad Umair Haider, Hammad Rizwan, Hassan Sajjad, Peizhong Ju, and A. B. Siddique. Neurons speak in ranges: Breaking free from discrete neuronal attribution. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=AukyIhfBuW.

Kai He, Yucheng Huang, Wenqing Wang, Delong Ran, Dongming Sheng, Junxuan Huang, Qika Lin, Jiaxing Xu, Wenqiang Liu, and Mengling Feng. Crab: A novel configurable role-playing LLM with assessing benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15030–15052. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.731. URL https://aclanthology.org/2025.acl-long.731/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations (ICLR 2021)*. OpenReview.net, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3605–3627. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-naacl.229. URL https://aclanthology.org/2024.findings-naacl.229/.

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*, pp. 64733–64758. Curran Associates, Inc., 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e2dd53601de57c773343a7cdf09fae1c-Abstract-Conference.html.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 8412–8452. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.findings-naacl.469. URL https://aclanthology.org/2025.findings-naacl.469/.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona T. Diab, and Maarten Sap. 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. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.999. URL https://aclanthology.org/2025.acl-long.999/.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Aldo Pareja, Nikhil Shivakumar Nayak, Hao Wang, Krishnateja Killamsetty, Shivchander Sudalairaj, Wenlong Zhao, Seungwook Han, Abhishek Bhandwaldar, Guangxuan Xu, Kai Xu, Ligong Han, Luke Inglis, and Akash Srivastava. Unveiling the secret recipe: A guide for supervised fine-tuning small LLMs. In *The Thirteenth International Conference on Learning Representations*, pp. 31848–31880. Curran Associates, Inc., 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b6e2c96bc4702f761d7d108d6e31930f-Abstract-Conference.html.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.168. URL https://aclanthology.org/2021.naacl-main.168/.

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

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454/.

Anushka Sivakumar, Andrew Zhang, Zaber Ibn Abdul Hakim, and Chris Thomas. SteerVLM: Robust model control through lightweight activation steering for vision language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23640–23665. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1285. URL https://aclanthology.org/2025.findings-emnlp.1285/.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

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. URL https://arxiv.org/abs/2308.10248.

Yushi Yang, Filip Sondej, Harry Mayne, Andrew Lee, and Adam Mahdi. How does DPO reduce toxicity? a mechanistic neuron-level analysis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29524–29543. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1501. URL https://aclanthology.org/2025.emnlp-main.1501/.

Zejia You, Chunyuan Deng, and Hanjie Chen. Spherical steering: Geometry-aware activation rotation for language models. *arXiv preprint arXiv:2602.08169*, 2026. URL https://arxiv.org/abs/2602.08169.

Lifan Zheng, Xue Yang, Jiawei Chen, Chenyan Wu, Jingyuan Zhang, Fanheng Kong, Xinyi Zeng, Xiang Chen, and Yu Tian. DPN-LE: Dual personality neuron localization and editing for large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 30567–30591. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1528. URL https://aclanthology.org/2026.findings-acl.1528/.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Personality alignment of large language models. In *The Thirteenth International Conference on Learning Representations*, pp. 389–438. Curran Associates, Inc., 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/25203d1cc8c58381eab578f4fcf9c4f8-Abstract-Conference.html.

A EXPERIMENT SETTING DETAILS

A.1 EVALUATION METRICS

Personality Steering Efficacy. The personality score (Pers.) quantifies how effectively the model aligns with the requested trait direction on a standard 1–5 Likert scale, where higher values indicate stronger conformity.

General Capability Preservation. General capability(G) is evaluated using six benchmarks. Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021), which measures broad academic and professional knowledge; CommonsenseQA (CSQA) (Talmor et al., 2019), which measures everyday commonsense reasoning; Grade School Math 8K (GSM8K) (Cobbe et al., 2021), which measures multi-step grade-school mathematical reasoning; Simple Variations on Arithmetic Math Word Problems (SVAMP) (Patel et al., 2021), which measures robust arithmetic word-problem reasoning under variations in problem structure; TruthfulQA (TQA) (Lin et al., 2022), which measures whether the model can avoid common false beliefs and provide truthful answers; and Social IQa (SIQA) (Sap et al., 2019), which measures commonsense reasoning about social interactions. All six scores are reported as percentages, with higher values indicating better task performance.

Cross-Trait Specificity. The Big Five traits are Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N). Cross-trait specificity measures whether steering one of these traits changes only the intended trait rather than unintentionally altering the other four.

For a target trait q, P measures how well the target trait follows the requested trait, with higher values indicating stronger target-trait control. D_{non} measures the mean absolute score change of the other four, non-target personality traits relative to the Base model. It therefore represents the average unintended personality side effect caused by editing one trait; lower values indicate better isolation. $D_{\max}$ is the largest absolute change among those four non-target traits and captures the worst single unintended personality shift; lower values are better.

A.2 IMPLEMENTATION DETAILS

We configure the reference hyperparameters of AING as follows: (1) Candidate neuron selection (k): We retain the top 0.10% candidate neurons ranked by $|\delta_n|$, corresponding to selecting the top-k most sensitive neurons for each trait. (2) General-corpus filtering (τ): Candidate neurons with general activation $u_n \geq \tau$ are filtered out (we set $\tau = 0.30$), yielding the final target candidate (3) Interval and steering strength: The interval threshold is set to $\rho = 0.70$, and the intervention strength scale is fixed at $\alpha = 0.60$. These hyperparameters are selected based on validation performance and kept identical across all personality traits and backbone models.

During inference, the model weights remain strictly frozen. The neuron-level activation statistics are calculated exclusively over the generated response tokens, excluding prompt and padding tokens. For all baselines and our method, the decoding hyperparameters are held strictly identical to ensure fair comparison: we use greedy decoding (temperature $= 0$) with a maximum generation length of 256 tokens. All experiments are implemented in PyTorch and conducted on a single NVIDIA A100 GPU.

B EXPERIMENT RERULT DETAILS

B.1 ABLATION OF ACTIVATION-INTERVAL LOCALIZATION

To evaluate our interval localization design, we compare AING with a Target-only variant using a candidate fraction of 0.10% and $\alpha = 0.60$, with $\rho \in 0.70, 0.90$. Target-only selects intervals based solely on target activations $P_n^{\zeta}(l, u)$, ignoring non-target activations $P_n^{\beta}(l, u)$. As shown in Fig. A1, Target-only achieves comparable personality alignment (Avg.P), but causes greater non-target personality drift (Avg. D_{non}) and lower general capability scores (Avg.G). At $\rho = 0.70$, AING reduces Avg. D_{non} from 0.104 to 0.068 and improves Avg.G from 68.86 to 69.52 on Llama-3.1-8B-Instruct, while Avg.P decreases by only 0.02. These results support the effectiveness of the objective in Eq. 6 in reducing unintended trait drift while preserving general capability.

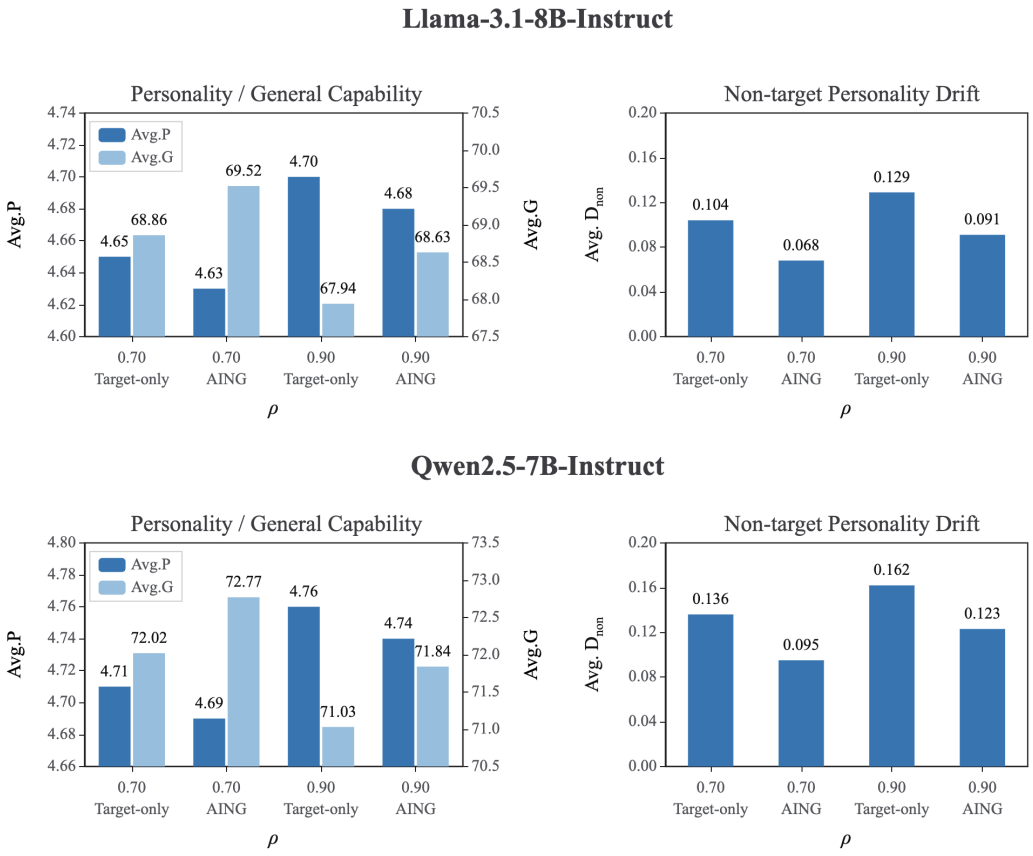


Figure A1: Results of the ablation study of activation-interval localization. Target-only denotes interval selection based solely on target activations.

B.2 Hyperparameter Analysis

We evaluate the impact of three key hyperparameters in AING: candidate fraction, interval coverage ρ, and update strength α, on personality alignment and general capability, as shown in Figures A2, A3, and A4, respectively.

For the candidate fraction, increasing it from 0.01% to 0.10% consistently improves personality alignment, reaching 4.63 on Llama, while general capability remains largely preserved. Increasing the fraction beyond 0.10% brings only small additional gains in personality alignment but causes a faster decline in general capability. This suggests that trait-relevant neurons are concentrated among a relatively small set of highly ranked candidates, while including more neurons introduces less specific activations.

For interval coverage ρ, personality alignment improves as ρ increases to 0.70, with only a small reduction in general capability. Further increasing ρ continues to improve personality scores only slightly, while capability drops more rapidly. This indicates that excessively broad activation intervals reduce the selectivity of the intervention.

For update strength α, increasing it from 0 to 0.60 improves personality alignment by 1.55 points on Llama, while reducing general capability by only 0.31 points. Increasing α further from 0.60 to 1.0 yields only a 0.06-point gain in personality alignment, but decreases general capability by 1.14 points on Llama. These results support 0.10%, $\rho = 0.70$, and $\alpha = 0.60$ as the default configuration.

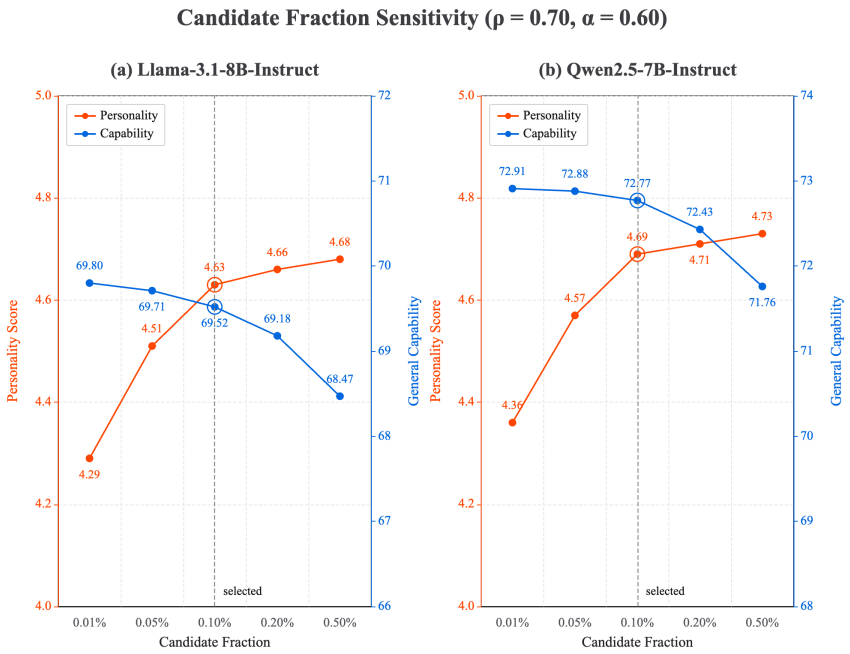


Figure A2: Effects of the candidate fraction on personality alignment and general capability.

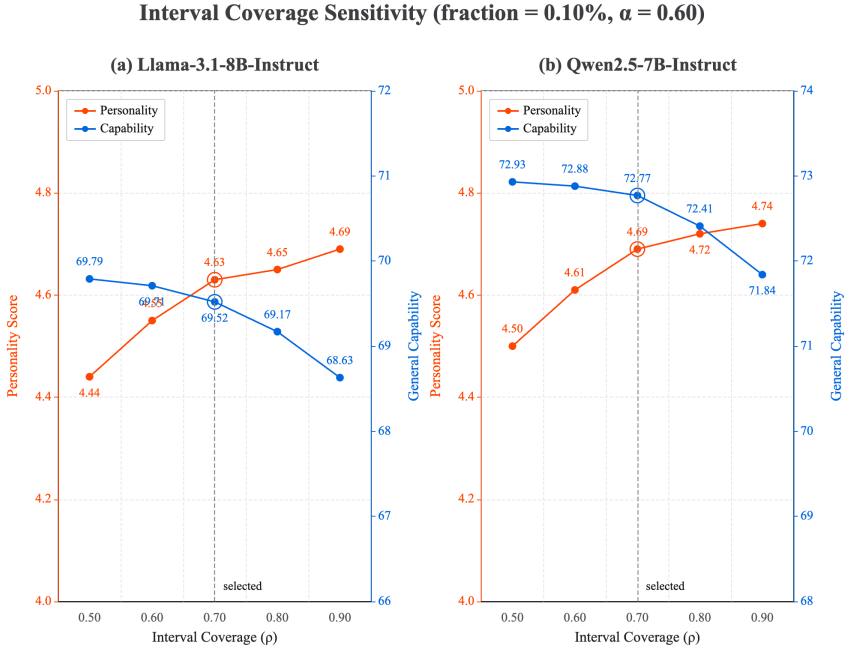


Figure A3: Effects of interval coverage ρ on personality alignment and general capability.

B.3 Single-Trait Control and Non-Target Profiles

Table A1 reports the complete five-trait profiles during single-trait control. AING substantially shifts the target trait while keeping the remaining four traits almost unchanged. For example, on Llama-3.1-8B, increasing Openness raises the target score from 3.172 to 4.990, while the largest change

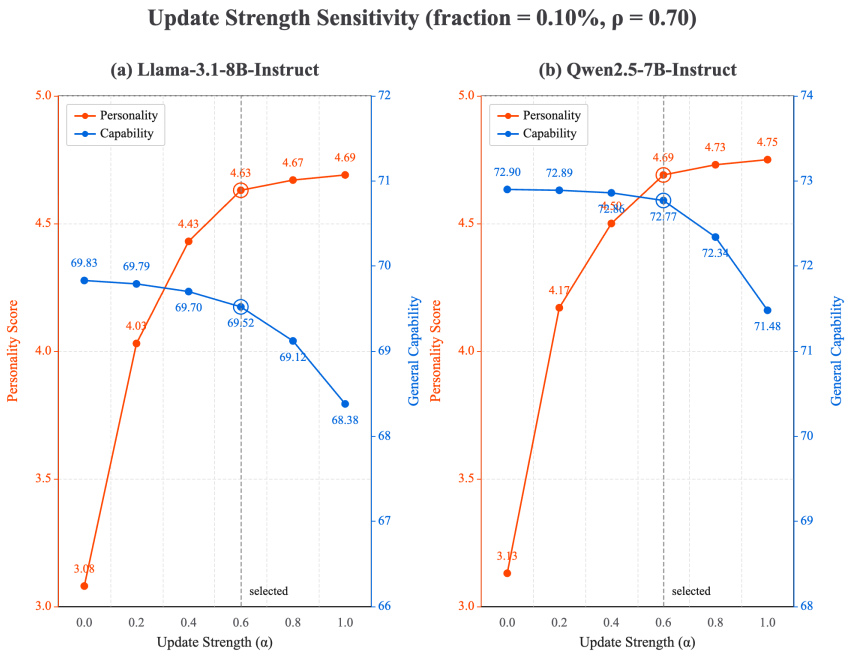


Figure A4: Effects of update strength α on personality alignment and general capability.

among non-target traits is merely 0.057. Existing baselines alter entire neurons or directions, easily perturbing unrelated traits due to shared neurons. Although recent methods mitigate interference by constraining directions, they also weaken target control. In contrast, AING intervenes on neurons only within the activation interval specific to the target trait. This localized intervention effectively steers the target trait without disturbing the remaining personality profile.

Table A1: Personality scores across all five traits after single-trait editing.T denotes the personality score, and D denotes its change relative to the unedited model.

Target	Methods	O T(D)	C T(D)	E T(D)	A T(D)	N T(D)	$D_{\text{non}}\downarrow$	$D_{\max}\downarrow$
				Llama-3.1-8B-Instruct				
None	Base	3.172 (0.000)	3.107 (0.000)	3.234 (0.000)	2.622 (0.000)	3.265 (0.000)	0.000	0.000
O-High	CAA	4.795 (+1.623)	2.962 (-0.145)	3.311 (+0.077)	2.641 (+0.019)	3.422 (+0.157)	0.099	0.157
O-High	C4	4.611 (+1.439)	3.119 (+0.012)	3.331 (+0.097)	2.305 (-0.317)	2.932 (-0.333)	0.190	0.333
O-High	C5	4.341 (+1.169)	3.174 (+0.067)	3.280 (+0.046)	2.838 (+0.216)	3.086 (-0.179)	0.127	0.216
O-High	NPTI	4.893 (+1.721)	3.103 (-0.004)	3.327 (+0.093)	2.652 (+0.030)	3.315 (+0.050)	0.044	0.093
O-High	DPN-LE	4.445 (+1.273)	2.986 (-0.121)	2.863 (-0.371)	2.644 (+0.022)	3.075 (-0.190)	0.176	0.371
O-High	DPN-LEw	4.316 (+1.144)	3.152 (+0.045)	3.362 (+0.128)	2.432 (-0.190)	3.209 (-0.056)	0.105	0.190
O-High	AING	**4.990** **(+1.818)**	3.087 (-0.020)	3.177 (-0.057)	2.614 (-0.008)	3.227 (-0.038)	**0.031**	**0.057**
O-Low	CAA	1.968 (-1.204)	3.041 (-0.066)	3.182 (-0.052)	2.580 (-0.042)	3.234 (-0.031)	0.048	0.066
O-Low	C4	**1.774** **(-1.398)**	3.158 (+0.051)	3.118 (-0.116)	2.951 (+0.329)	3.451 (+0.186)	0.171	0.329
O-Low	C5	1.957 (-1.215)	3.271 (+0.164)	3.308 (+0.074)	2.405 (-0.217)	3.296 (+0.031)	0.121	0.217

Continued on next page

Table A1 (continued)

Target	Method	O T(D)	C T(D)	E T(D)	A T(D)	N T(D)	D_{non} ↓	$D_{\max}$ ↓
O-Low	NPTI	1.860 (-1.312)	3.110 (+0.003)	3.262 (+0.028)	2.561 (-0.061)	3.219 (-0.046)	**0.035**	**0.061**
O-Low	DPN-LE	2.288 (-0.884)	3.292 (+0.185)	3.280 (+0.046)	2.260 (-0.362)	3.200 (-0.065)	0.165	0.362
O-Low	DPN-LEw	2.018 (-1.154)	3.287 (+0.180)	3.184 (-0.050)	2.490 (-0.132)	3.440 (+0.175)	0.134	0.180
O-Low	AING	1.905 (-1.267)	3.128 (+0.021)	3.322 (+0.088)	2.734 (+0.112)	3.238 (-0.027)	0.062	0.112
C-High	CAA	2.794 (-0.378)	4.573 (+1.466)	3.126 (-0.108)	2.977 (+0.355)	3.381 (+0.116)	0.239	0.378
C-High	C4	3.172 (+0.000)	4.338 (+1.231)	3.441 (+0.207)	2.525 (-0.097)	3.419 (+0.154)	0.114	0.207
C-High	C5	3.309 (+0.137)	4.194 (+1.087)	3.186 (-0.048)	2.409 (-0.213)	3.529 (+0.264)	0.166	0.264
C-High	NPTI	3.152 (-0.020)	4.873 (+1.766)	3.185 (-0.049)	2.599 (-0.023)	3.286 (+0.021)	**0.028**	**0.049**
C-High	DPN-LE	3.468 (+0.296)	4.454 (+1.347)	3.468 (+0.234)	2.628 (+0.006)	3.181 (-0.084)	0.155	0.296
C-High	DPN-LEw	3.266 (+0.094)	4.507 (+1.400)	2.957 (-0.277)	2.766 (+0.144)	3.069 (-0.196)	0.178	0.277
C-High	AING	3.186 (+0.014)	**4.930 (+1.823)**	3.203 (-0.031)	2.826 (+0.204)	3.459 (+0.194)	0.111	0.204
C-Low	CAA	3.516 (+0.344)	1.599 (-1.508)	3.009 (-0.225)	2.732 (+0.110)	2.994 (-0.271)	0.237	0.344
C-Low	C4	3.277 (+0.105)	2.137 (-0.970)	2.896 (-0.338)	2.445 (-0.177)	3.154 (-0.111)	0.183	0.338
C-Low	C5	3.275 (+0.103)	1.935 (-1.172)	3.422 (+0.188)	2.703 (+0.081)	3.283 (+0.018)	0.098	0.188
C-Low	NPTI	3.253 (+0.081)	2.022 (-1.085)	3.200 (-0.034)	2.764 (+0.142)	3.256 (-0.009)	0.067	0.142
C-Low	DPN-LE	3.097 (-0.075)	1.825 (-1.282)	2.831 (-0.403)	2.558 (-0.064)	3.188 (-0.077)	0.155	0.403
C-Low	DPN-LEw	3.319 (+0.147)	1.949 (-1.158)	3.037 (-0.197)	2.506 (-0.116)	3.420 (+0.155)	0.154	0.197
C-Low	AING	3.123 (-0.049)	**1.557 (-1.550)**	3.115 (-0.119)	2.650 (+0.028)	3.282 (+0.017)	**0.053**	**0.119**
E-High	CAA	3.387 (+0.215)	2.995 (-0.112)	4.877 (+1.643)	2.566 (-0.056)	3.336 (+0.071)	0.114	0.215
E-High	C4	2.873 (-0.299)	3.245 (+0.138)	4.458 (+1.224)	2.334 (-0.288)	3.429 (+0.164)	0.222	0.299
E-High	C5	3.164 (-0.008)	2.966 (-0.141)	4.261 (+1.027)	2.370 (-0.252)	3.023 (-0.242)	0.161	0.252
E-High	NPTI	3.222 (+0.050)	3.201 (+0.094)	**4.996 (+1.762)**	2.631 (+0.009)	3.005 (-0.260)	0.103	0.260
E-High	DPN-LE	3.279 (+0.107)	3.173 (+0.066)	4.306 (+1.072)	2.521 (-0.101)	3.345 (+0.080)	0.088	0.107
E-High	DPN-LEw	3.286 (+0.114)	3.250 (+0.143)	4.346 (+1.112)	2.526 (-0.096)	3.434 (+0.169)	0.131	0.169
E-High	AING	3.202 (+0.030)	3.171 (+0.064)	4.990 (+1.756)	2.564 (-0.058)	3.308 (+0.043)	**0.049**	**0.064**
E-Low	CAA	3.196 (+0.024)	3.034 (-0.073)	1.934 (-1.300)	2.652 (+0.030)	3.376 (+0.111)	0.059	0.111
E-Low	C4	3.295 (+0.123)	2.884 (-0.223)	2.205 (-1.029)	2.446 (-0.176)	3.608 (+0.343)	0.216	0.343
E-Low	C5	3.176 (+0.004)	3.218 (+0.111)	2.088 (-1.146)	2.768 (+0.146)	3.436 (+0.171)	0.108	0.171
E-Low	NPTI	3.016 (-0.156)	3.198 (+0.091)	2.043 (-1.191)	2.681 (+0.059)	3.199 (-0.066)	0.093	0.156
E-Low	DPN-LE	3.409 (+0.237)	3.028 (-0.079)	2.338 (-0.896)	2.335 (-0.287)	3.136 (-0.129)	0.183	0.287
E-Low	DPN-LEw	3.116 (-0.056)	3.251 (+0.144)	1.973 (-1.261)	2.713 (+0.091)	3.193 (-0.072)	0.091	0.144
E-Low	AING	3.198 (+0.026)	3.064 (-0.043)	**1.630 (-1.604)**	2.566 (-0.056)	3.350 (+0.085)	**0.052**	**0.085**

Continued on next page

Table A1 (continued)

Target	Method	O T(D)	C T(D)	E T(D)	A T(D)	N T(D)	D_{non} ↓	$D_{\max}$ ↓
		3.030	2.974	3.025	3.983	3.463		
A-High	CAA	(-0.142)	(-0.133)	(-0.209)	(+1.361)	(+0.198)	0.170	0.209
		3.118	3.274	3.059	3.790	3.435		
A-High	C4	(-0.054)	(+0.167)	(-0.175)	(+1.168)	(+0.170)	0.142	0.175
		3.258	3.196	3.201	3.792	3.178		
A-High	C5	(+0.086)	(+0.089)	(-0.033)	(+1.170)	(-0.087)	0.074	**0.089**
		3.082	3.147	3.382	3.876	3.151		
A-High	NPTI	(-0.090)	(+0.040)	(+0.148)	(+1.254)	(-0.114)	0.098	0.148
		3.311	3.204	3.086	3.888	3.184		
A-High	DPN-LE	(+0.139)	(+0.097)	(-0.148)	(+1.266)	(-0.081)	0.116	0.148
		3.158	2.981	3.146	3.519	3.249		
A-High	DPN-LEw	(-0.014)	(-0.126)	(-0.088)	(+0.897)	(-0.016)	**0.061**	0.126
		3.115	3.198	3.268	**4.286**	3.360		
A-High	AING	(-0.057)	(+0.091)	(+0.034)	**(+1.664)**	(+0.095)	0.069	0.095
		3.086	3.218	3.329	1.241	3.226		
A-Low	CAA	(-0.086)	(+0.111)	(+0.095)	(-1.381)	(-0.039)	0.083	0.111
		2.861	3.181	3.552	1.410	3.426		
A-Low	C4	(-0.311)	(+0.074)	(+0.318)	(-1.212)	(+0.161)	0.216	0.318
		3.067	2.937	3.402	1.678	3.071		
A-Low	C5	(-0.105)	(-0.170)	(+0.168)	(-0.944)	(-0.194)	0.159	0.194
		3.147	3.087	3.312	1.489	3.270		
A-Low	NPTI	(-0.025)	(-0.020)	(+0.078)	(-1.133)	(+0.005)	0.032	0.078
		3.256	3.173	3.185	1.421	3.239		
A-Low	DPN-LE	(+0.084)	(+0.066)	(-0.049)	(-1.201)	(-0.026)	0.056	0.084
		3.329	3.310	3.569	1.501	2.956		
A-Low	DPN-LEw	(+0.157)	(+0.203)	(+0.335)	(-1.121)	(-0.309)	0.251	0.335
		3.192	3.152	3.221	**1.052**	3.293		
A-Low	AING	(+0.020)	(+0.045)	(-0.013)	**(-1.570)**	(+0.028)	**0.026**	**0.045**
		3.305	3.297	3.457	2.254	4.689		
N-High	CAA	(+0.133)	(+0.190)	(+0.223)	(-0.368)	(+1.424)	0.229	0.368
		3.438	3.055	3.445	2.503	4.557		
N-High	C4	(+0.266)	(-0.052)	(+0.211)	(-0.119)	(+1.292)	0.162	0.266
		3.310	2.975	3.388	2.618	4.515		
N-High	C5	(+0.138)	(-0.132)	(+0.154)	(-0.004)	(+1.250)	0.107	0.154
		3.336	3.374	3.030	2.583	4.808		
N-High	NPTI	(+0.164)	(+0.267)	(-0.204)	(-0.039)	(+1.543)	0.169	0.267
		3.251	3.200	3.333	2.601	4.305		
N-High	DPN-LE	(+0.079)	(+0.093)	(+0.099)	(-0.021)	(+1.040)	**0.073**	0.099
		3.078	3.196	3.169	2.547	4.398		
N-High	DPN-LEw	(-0.094)	(+0.089)	(-0.065)	(-0.075)	(+1.133)	0.081	**0.094**
		3.328	2.969	3.338	2.751	**4.990**		
N-High	AING	(+0.156)	(-0.138)	(+0.104)	(+0.129)	**(+1.725)**	0.132	0.156
		3.385	2.977	3.029	2.400	**1.675**		
N-Low	CAA	(+0.213)	(-0.130)	(-0.205)	(-0.222)	**(-1.590)**	0.193	0.222
		3.159	3.062	3.152	2.548	2.226		
N-Low	C4	(-0.013)	(-0.045)	(-0.082)	(-0.074)	(-1.039)	**0.053**	**0.082**
		2.987	3.414	2.860	2.530	2.242		
N-Low	C5	(-0.185)	(+0.307)	(-0.374)	(-0.092)	(-1.023)	0.239	0.374
		3.201	3.134	3.323	2.515	2.235		
N-Low	NPTI	(+0.029)	(+0.027)	(+0.089)	(-0.107)	(-1.030)	0.063	0.107
		3.029	2.997	3.578	2.448	2.226		
N-Low	DPN-LE	(-0.143)	(-0.110)	(+0.344)	(-0.174)	(-1.039)	0.193	0.344
		3.078	2.970	3.325	2.770	2.145		
N-Low	DPN-LEw	(-0.094)	(-0.137)	(+0.091)	(+0.148)	(-1.120)	0.117	0.148
		3.094	3.270	3.161	2.550	1.742		
N-Low	AING	(-0.078)	(+0.163)	(-0.073)	(-0.072)	(-1.523)	0.097	0.163
				Qwen2.5-7B-Instruct				
		3.059	2.953	3.189	3.584	2.865		
None	Base	(+0.000)	(+0.000)	(+0.000)	(+0.000)	(+0.000)	0.000	0.000
		4.423	3.115	3.151	3.717	3.029		
O-High	CAA	(+1.364)	(+0.162)	(-0.038)	(+0.133)	(+0.164)	0.124	0.164
		4.306	2.859	3.128	3.564	2.832		
O-High	C4	(+1.247)	(-0.094)	(-0.061)	(-0.020)	(-0.033)	0.052	0.094

Continued on next page

Table A1 (continued)

Target	Method	O T(D)	C T(D)	E T(D)	A T(D)	N T(D)	D_{non} ↓	$D_{\max}$ ↓
O-High	C5	3.998 (+0.939)	2.787 (-0.166)	3.382 (+0.193)	3.527 (-0.057)	3.045 (+0.180)	0.149	0.193
O-High	NPTI	**4.505 (+1.446)**	2.906 (-0.047)	3.133 (-0.056)	3.485 (-0.099)	2.626 (-0.239)	0.110	0.239
O-High	DPN-LE	4.385 (+1.326)	2.734 (-0.219)	3.027 (-0.162)	3.563 (-0.021)	2.720 (-0.145)	0.137	0.219
O-High	DPN-LEw	4.172 (+1.113)	2.909 (-0.044)	3.282 (+0.093)	3.585 (+0.001)	2.816 (-0.049)	**0.047**	**0.093**
O-High	AING	4.431 (+1.372)	2.835 (-0.118)	3.272 (+0.083)	3.572 (-0.012)	2.633 (-0.232)	0.111	0.232
O-Low	CAA	1.506 (-1.553)	2.914 (-0.039)	3.381 (+0.192)	3.739 (+0.155)	2.757 (-0.108)	0.123	0.192
O-Low	C4	1.486 (-1.573)	3.074 (+0.121)	3.208 (+0.019)	3.696 (+0.112)	2.833 (-0.032)	0.071	0.121
O-Low	C5	1.563 (-1.496)	2.669 (-0.284)	3.488 (+0.299)	3.808 (+0.224)	2.510 (-0.355)	0.290	0.355
O-Low	NPTI	1.401 (-1.658)	3.109 (+0.156)	3.406 (+0.217)	3.610 (+0.026)	2.526 (-0.339)	0.184	0.339
O-Low	DPN-LE	1.145 (-1.914)	2.582 (-0.371)	2.965 (-0.224)	3.304 (-0.280)	3.055 (+0.190)	0.266	0.371
O-Low	DPN-LEw	**1.000 (-2.059)**	2.982 (+0.029)	3.091 (-0.098)	3.606 (+0.022)	2.817 (-0.048)	**0.049**	**0.098**
O-Low	AING	1.010 (-2.049)	2.771 (-0.182)	3.145 (-0.044)	3.392 (-0.192)	2.730 (-0.135)	0.138	0.192
C-High	CAA	3.179 (+0.120)	4.025 (+1.072)	3.083 (-0.106)	3.809 (+0.225)	3.076 (+0.211)	0.166	0.225
C-High	C4	3.054 (-0.005)	3.668 (+0.715)	3.229 (+0.040)	3.544 (-0.040)	2.852 (-0.013)	**0.025**	**0.040**
C-High	C5	2.818 (-0.241)	3.775 (+0.822)	3.151 (-0.038)	3.262 (-0.322)	2.527 (-0.338)	0.235	0.338
C-High	NPTI	3.093 (+0.034)	4.283 (+1.330)	3.255 (+0.066)	3.622 (+0.038)	2.936 (+0.071)	0.052	0.071
C-High	DPN-LE	3.461 (+0.402)	4.221 (+1.268)	3.551 (+0.362)	3.841 (+0.257)	3.262 (+0.397)	0.355	0.402
C-High	DPN-LEw	2.849 (-0.210)	**4.400 (+1.447)**	3.572 (+0.383)	3.879 (+0.295)	3.114 (+0.249)	0.284	0.383
C-High	AING	3.315 (+0.256)	4.211 (+1.258)	3.417 (+0.228)	3.463 (-0.121)	2.693 (-0.172)	0.194	0.256
C-Low	CAA	3.089 (+0.030)	1.309 (-1.644)	3.168 (-0.021)	3.643 (+0.059)	2.902 (+0.037)	**0.037**	**0.059**
C-Low	C4	2.809 (-0.250)	1.354 (-1.599)	3.015 (-0.174)	3.271 (-0.313)	2.771 (-0.094)	0.208	0.313
C-Low	C5	3.253 (+0.194)	1.711 (-1.242)	3.072 (-0.117)	3.662 (+0.078)	2.820 (-0.045)	0.108	0.194
C-Low	NPTI	3.302 (+0.243)	1.486 (-1.467)	2.994 (-0.195)	3.765 (+0.181)	2.878 (+0.013)	0.158	0.243
C-Low	DPN-LE	2.967 (-0.092)	1.083 (-1.870)	3.474 (+0.285)	3.479 (-0.105)	2.536 (-0.329)	0.203	0.329
C-Low	DPN-LEw	3.208 (+0.149)	**1.000 (-1.953)**	3.511 (+0.322)	3.749 (+0.165)	2.961 (+0.096)	0.183	0.322
C-Low	AING	3.040 (-0.019)	1.116 (-1.837)	3.144 (-0.045)	3.649 (+0.065)	2.947 (+0.082)	0.053	0.082
E-High	CAA	3.103 (+0.044)	2.910 (-0.043)	4.376 (+1.187)	3.678 (+0.094)	2.913 (+0.048)	0.057	0.094
E-High	C4	3.110 (+0.051)	3.010 (+0.057)	4.192 (+1.003)	3.550 (-0.034)	2.798 (-0.067)	**0.052**	**0.067**
E-High	C5	2.856 (-0.203)	3.144 (+0.191)	3.912 (+0.723)	3.562 (-0.022)	2.795 (-0.070)	0.122	0.203
E-High	NPTI	2.957 (-0.102)	2.831 (-0.122)	4.283 (+1.094)	3.537 (-0.047)	2.779 (-0.086)	0.089	0.122
E-High	DPN-LE	3.234 (+0.175)	3.093 (+0.140)	4.643 (+1.454)	3.330 (-0.254)	2.963 (+0.098)	0.167	0.254
E-High	DPN-LEw	2.940 (-0.119)	2.734 (-0.219)	4.757 (+1.568)	3.636 (+0.052)	2.709 (-0.156)	0.137	0.219

Continued on next page

Table A1 (continued)

Target	Method	O T(D)	C T(D)	E T(D)	A T(D)	N T(D)	D_{non} ↓	$D_{\max}$ ↓
E-High	AING	2.901 (-0.158)	2.909 (-0.044)	**4.944 (+1.755)**	3.475 (-0.109)	2.854 (-0.011)	0.081	0.158
E-Low	CAA	2.937 (-0.122)	2.998 (+0.045)	1.292 (-1.897)	3.690 (+0.106)	2.901 (+0.036)	0.077	0.122
E-Low	C4	3.030 (-0.029)	3.049 (+0.096)	1.651 (-1.538)	3.524 (-0.060)	2.798 (-0.067)	**0.063**	**0.096**
E-Low	C5	2.748 (-0.311)	2.702 (-0.251)	1.708 (-1.481)	3.486 (-0.098)	2.912 (+0.047)	0.177	0.311
E-Low	NPTI	3.097 (+0.038)	2.791 (-0.162)	1.516 (-1.673)	3.606 (+0.022)	2.959 (+0.094)	0.079	0.162
E-Low	DPN-LE	2.686 (-0.373)	2.960 (+0.007)	1.306 (-1.883)	3.234 (-0.350)	2.885 (+0.020)	0.188	0.373
E-Low	DPN-LEw	2.902 (-0.157)	2.946 (-0.007)	**1.213 (-1.976)**	3.453 (-0.131)	3.221 (+0.356)	0.163	0.356
E-Low	AING	3.342 (+0.283)	2.818 (-0.135)	1.460 (-1.729)	3.654 (+0.070)	2.637 (-0.228)	0.179	0.283
A-High	CAA	3.310 (+0.251)	3.217 (+0.264)	3.012 (-0.177)	4.990 (+1.406)	2.847 (-0.018)	0.177	0.264
A-High	C4	2.981 (-0.078)	2.736 (-0.217)	3.138 (-0.051)	4.607 (+1.023)	2.777 (-0.088)	0.108	0.217
A-High	C5	3.151 (+0.092)	2.817 (-0.136)	3.341 (+0.152)	4.592 (+1.008)	2.709 (-0.156)	0.134	0.156
A-High	NPTI	3.031 (-0.028)	2.716 (-0.237)	3.097 (-0.092)	4.727 (+1.143)	2.739 (-0.126)	0.121	0.237
A-High	DPN-LE	3.078 (+0.019)	2.918 (-0.035)	3.183 (-0.006)	4.967 (+1.383)	2.944 (+0.079)	**0.035**	0.079
A-High	DPN-LEw	3.232 (+0.173)	3.159 (+0.206)	3.270 (+0.081)	**5.000 (+1.416)**	2.763 (-0.102)	0.140	0.206
A-High	AING	3.019 (-0.040)	3.009 (+0.056)	3.227 (+0.038)	4.990 (+1.406)	2.818 (-0.047)	0.045	**0.056**
A-Low	CAA	3.312 (+0.253)	2.856 (-0.097)	2.896 (-0.293)	1.916 (-1.668)	3.041 (+0.176)	0.205	0.293
A-Low	C4	3.105 (+0.046)	3.046 (+0.093)	3.227 (+0.038)	1.970 (-1.614)	2.933 (+0.068)	0.061	0.093
A-Low	C5	3.163 (+0.104)	3.073 (+0.120)	3.447 (+0.258)	2.095 (-1.489)	3.158 (+0.293)	0.194	0.293
A-Low	NPTI	3.270 (+0.211)	2.865 (-0.088)	2.930 (-0.259)	1.717 (-1.867)	3.046 (+0.181)	0.185	0.259
A-Low	DPN-LE	3.042 (-0.017)	2.752 (-0.201)	3.215 (+0.026)	1.470 (-2.114)	2.957 (+0.092)	0.084	0.201
A-Low	DPN-LEw	3.059 (+0.000)	2.999 (+0.046)	3.044 (-0.145)	1.918 (-1.666)	2.782 (-0.083)	0.069	0.145
A-Low	AING	3.041 (-0.018)	2.947 (-0.006)	3.117 (-0.072)	**1.228 (-2.356)**	2.870 (+0.005)	**0.025**	**0.072**
N-High	CAA	2.937 (-0.122)	2.839 (-0.114)	3.322 (+0.133)	3.398 (-0.186)	4.278 (+1.413)	0.139	0.186
N-High	C4	2.941 (-0.118)	3.244 (+0.291)	3.176 (-0.013)	3.874 (+0.290)	3.783 (+0.918)	0.178	0.291
N-High	C5	2.877 (-0.182)	3.144 (+0.191)	2.976 (-0.213)	3.662 (+0.078)	3.846 (+0.981)	0.166	0.213
N-High	NPTI	3.160 (+0.101)	2.928 (-0.025)	3.190 (+0.001)	3.699 (+0.115)	3.859 (+0.994)	0.061	0.115
N-High	DPN-LE	3.108 (+0.049)	2.857 (-0.096)	3.206 (+0.017)	3.510 (-0.074)	**4.398 (+1.533)**	**0.059**	**0.096**
N-High	DPN-LEw	2.973 (-0.086)	2.816 (-0.137)	2.913 (-0.276)	3.375 (-0.209)	4.204 (+1.339)	0.177	0.276
N-High	AING	2.866 (-0.193)	2.951 (-0.002)	3.092 (-0.097)	3.598 (+0.014)	4.205 (+1.340)	0.076	0.193
N-Low	CAA	2.859 (-0.200)	3.021 (+0.068)	3.106 (-0.083)	3.768 (+0.184)	1.169 (-1.696)	0.134	0.200
N-Low	C4	3.417 (+0.358)	3.247 (+0.294)	3.073 (-0.116)	3.701 (+0.117)	1.595 (-1.270)	0.221	0.358
N-Low	C5	3.214 (+0.155)	2.812 (-0.141)	3.250 (+0.061)	3.560 (-0.024)	1.449 (-1.416)	0.095	0.155

Continued on next page

Table A1 (continued)

Target	Method	O T(D)	C T(D)	E T(D)	A T(D)	N T(D)	D_{non} ↓	$D_{\max}$ ↓
N-Low	NPTI	3.093 (+0.034)	2.939 (-0.014)	3.018 (-0.171)	3.413 (-0.171)	1.336 (-1.529)	0.098	0.171
N-Low	DPN-LE	2.914 (-0.145)	2.681 (-0.272)	3.525 (+0.336)	3.311 (-0.273)	1.010 (-1.855)	0.257	0.336
N-Low	DPN-LEw	3.102 (+0.043)	2.944 (-0.009)	3.161 (-0.028)	3.594 (+0.010)	**1.002 (-1.863)**	**0.022**	**0.043**
N-Low	AING	3.094 (+0.035)	2.877 (-0.076)	3.208 (+0.019)	3.640 (+0.056)	1.067 (-1.798)	0.046	0.076