TRACING AND STEERING PERSONALITY FORMATION IN LLMS VIA STAGE-WISE REPRESENTATION DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) can exhibit diverse personality traits through natural-language prompts, making reliable and controllable personality expression increasingly important for interactive applications. However, existing personality steering methods mainly focus on identifying and manipulating personality-related representations, while the internal formation and propagation patterns of prompt-induced personality remain insufficiently characterized. To address this gap, we investigate the stage-wise representation dynamics of prompt-induced personality formation through localized trait-token corruption and multi-level activation patching. Our analysis reveals a consistent progression across models and personality dimensions: personality signals first emerge in early prompt representations, are transferred toward generation-relevant positions with attention providing a dominant recoverable pathway, and eventually form a stable residual representation in middle-to-late layers. Based on these observations, we propose **Stage-wise Personality Steering (SPS)**, a training-free inference-time approach that employs Trait Seeding, Attention Routing, and Expression Anchoring to strengthen early trait signals, promote intermediate-layer aggregation, and stabilize the target-trait representation during generation, respectively. Experiments on PersonalityBench demonstrate that SPS achieves state-of-the-art personality steering performance on Llama-3.1-8B-Instruct, obtaining an overall personality score of 9.54 with only a 0.27 percentage-point decrease on MMLU. Evaluations across multiple LLM families and Big Five personality dimensions further demonstrate its effectiveness, robustness, and generalization ability.

1 INTRODUCTION

As large language models (LLMs) are increasingly deployed in healthcare, education, virtual companionship, and social simulation, how they communicate is becoming as important as what they can accomplish. The personality expressed by an LLM can influence perceived trustworthiness, engagement, and behavioral consistency, making stable personality control an important problem in controllable generation Park et al. (2023); Jiang et al. (2024). Existing approaches mainly rely on natural-language prompting, parameter-based training, or inference-time representation intervention Jiang et al. (2023b); Zhang et al. (2024). Prompting is inexpensive but sensitive to wording and context, while fine-tuning requires additional data and parameter updates Caron & Srivastava (2022); Ouyang et al. (2022). More recently, inference-time methods have manipulated personality-related neurons or activation directions without gradient updates Deng et al. (2024); Feng et al. (2026).

However, existing studies primarily investigate whether personality representations exist and can be manipulated Zhu et al. (2024), while providing limited understanding of how explicit personality instructions become internal representations that shape generation. Most activation-based methods inject personality vectors into a single layer or a small set of layers, with intervention locations selected heuristically or according to downstream scores Turner et al. (2023); Deng et al. (2024). Although effective, these methods do not sufficiently characterize when personality signals emerge, through which components they are aggregated toward generation-relevant positions, or when they become stable in the residual stream. Consequently, the locations, components, and forms of personality intervention lack a stage-wise representational basis.

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.

To address this gap, we construct controlled personality prompts and combine localized trait-token corruption with multi-level activation patching to trace prompt-induced representation changes across token positions, network layers, and Transformer components Geiger et al. (2021). Specifically, we perturb only the input embeddings of trait tokens and perform layer–token, module-wise, and final-position residual patching, following the principle of controlled activation replacement Syed et al. (2023); Kramár et al. (2024). Recovery is evaluated along each example's own reference-to-corrupted activation shift, avoiding reliance on a predefined PCA subspace or a trained probe.

Our analysis reveals consistent stage-wise representation dynamics during prompt-induced personality formation. personality signals first become detectable in early prompt representations. Attention then exhibits a stronger recoverable contribution to aggregating these differences toward the final prompt position, where a stable residual representation gradually forms in middle-to-late layers. These dynamics remain consistent across multiple model families and all five Big Five dimensions.

Based on these observations, we propose **Stage-wise Personality Steering (SPS)**, a training-free method for stable target-personality control at inference time. SPS consists of Trait Seeding, Attention Routing, and Expression Anchoring, which intervene in early prompt representations, intermediate-layer attention, and middle-to-late residual states, respectively. Rather than relying on heuristic selection of a fixed intervention layer, SPS aligns its intervention positions, aggregation pathways, and representation forms with observed stages of prompt-induced personality formation.

We evaluate SPS across multiple open-source LLMs and the five Big Five personality dimensions. On PersonalityBench Deng et al. (2024) with Llama-3.1-8B Grattafiori et al. (2024), SPS achieves a personality score of **9.54**, outperforming existing methods and establishing a new state of the art. Cross-model experiments demonstrate consistent effectiveness across different model families, while stage-wise analyses confirm the complementary contributions of the three interventions. SPS decreases accuracy on MMLU Hendrycks et al. (2021) by only **0.27** percentage points on average, indicating that general knowledge and reasoning capabilities are largely preserved.

Our main contributions are summarized as follows:

- We characterize stage-wise personality formation through localized trait-token corruption and multi-level activation patching, tracing how personality signals develop across tokens, layers, and Transformer components.
- Guided by this pattern, we propose **SPS**, a training-free, stage-wise personality steering method for stable target-personality control at inference time.
- Extensive evaluations demonstrate the effectiveness, cross-model generalization, and robustness of SPS, including state-of-the-art performance on PersonalityBench with limited degradation of general capabilities.

2 Representation Dynamics

To characterize the stage-wise representation dynamics associated with prompt-induced personality formation inside LLMs, we combine localized trait-token corruption with activation patching across token positions, layers, and Transformer components. We measure how restoring selected internal states recovers each example's reference-to-corrupted representation shift. Figure 1 presents the results on Llama-3.1-8B-Instruct, with additional models reported in the Appendix.

2.1 Localized Trait-Token Corruption and Patching

We construct structurally aligned prompts: *You are a* $\{d_t\}$ *AI assistant. Question:* $\{q\}$*?*, where d_t is a Big Five trait description and q is sampled from an independent set containing 50 topics per trait. The shared template aligns trait tokens, task content, and the final prompt position. Trait-token positions are denoted by $\mathcal{P}_t$.

For each prompt, we record the reference residual state $h_{l,p}^{\text{ref}}$ and add Gaussian noise only to the trait-token embeddings:

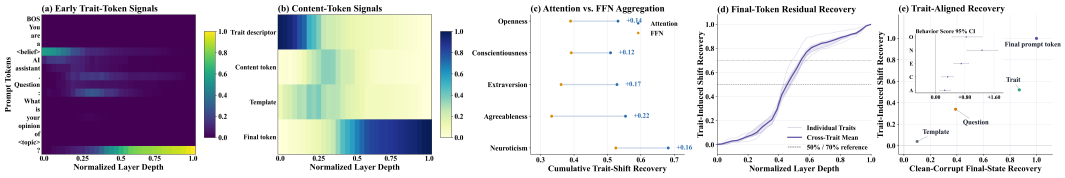

Figure 1: Stage-wise personality representation dynamics. (a) Trait-token corruption first alters the trait description, then spreads across prompt positions. (b) Content-token signals emerge before final-position accumulation. (c) Attention restoration recovers more final-position shift than FFN restoration. (d) Final-position recovery rises in intermediate layers and stabilizes later. (e) Template tokens recover little trait signal, question content recovers some, and trait-description tokens recover both shifts, aligning with positive behavior-score changes (inset: 95% CIs).

$$e_p^{\text{corr}} = \begin{cases} e_p + \epsilon_p, & p \in \mathcal{P}_t, \\ e_p, & p \notin \mathcal{P}_t, \end{cases} \qquad \epsilon_p \sim \mathcal{N}(0, \sigma^2 I). \tag{1}$$

This preserves the question, template, and sequence length while weakening the local trait cue. Let p_f and L denote the final prompt position and final layer, and define $\Delta h_f = h^{\text{ref}}_{L,p_f} - h^{\text{corr}}_{L,p_f}$. After replacing a corrupted internal state with its reference counterpart, recovery is measured by

$$R_{l,p} = \frac{\left(h^{\text{patch}}_{L,p_f} - h^{\text{corr}}_{L,p_f}\right)^\top \Delta h_f}{\|\Delta h_f\|_2^2}. \tag{2}$$

A value near 1 indicates strong recovery along the example-specific trait-induced shift, whereas a value near 0 indicates little recoverable contribution. We use position-wise residual differences to track signal distribution, cumulative attention and FFN patching to compare component-level recovery, and final-position patching to examine when the representation becomes stable. The visualized position-wise difference is $D_{l,p} = \mathbb{E}_{t,q}[\|h^{\text{ref}}_{l,p} - h^{\text{corr}}_{l,p}\|_2]$, normalized within each model.

2.2 Stage-Wise Representation Dynamics

Across four models and five Big Five dimensions, personality signals exhibit a broadly consistent progression. They first appear near the trait description, subsequently become visible at task-content positions, and later accumulate at the final prompt position. Attention provides stronger recoverable aggregation than FFNs in most settings, while the final-position residual becomes stable in middle-to-late layers.

2.2.1 Early Layers: Trait Signals and Content Carriers

Figure 1(a) shows that corrupting the trait description immediately produces clear differences at the corresponding trait-token positions. Weaker differences then appear at template, question, and separator tokens, indicating that the effect is not confined to the descriptor itself.

Figure 1(b) further shows that the descriptor signal dominates the earliest layers, whereas content-token differences emerge slightly later and remain visible before final-position accumulation. This pattern suggests that content tokens can serve as task-dependent carriers through which personality instructions influence semantic representations. They may therefore provide more informative intervention candidates than template tokens.

2.2.2 Middle Layers: Attention Aggregation

Figure 1(c) compares cumulative patching of attention and FFN outputs. Attention achieves higher recovery for most traits, with $R^{\text{attn}}_{a:b} > R^{\text{ffn}}_{a:b}$.

Under the current metric, attention is therefore more strongly associated with aggregating distributed personality signals toward the final prompt representation. This does not exclude FFN contributions

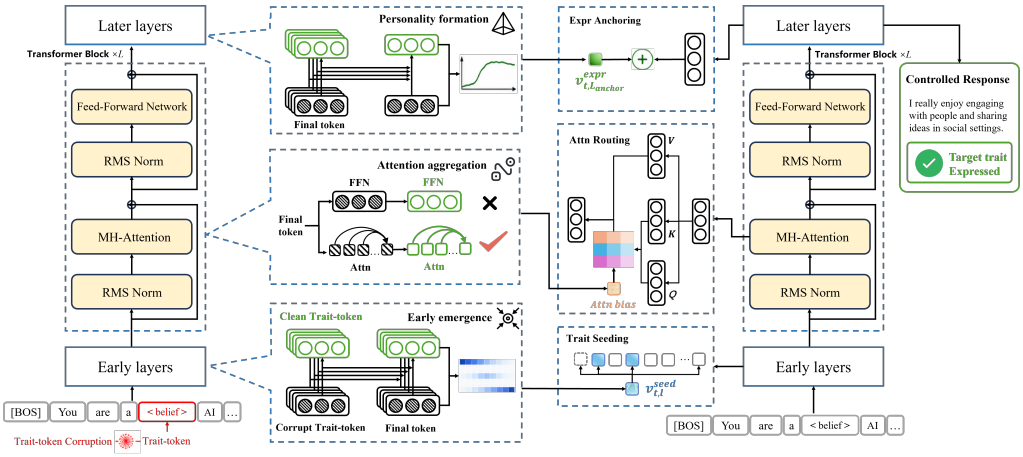

Figure 2: Overview of the proposed SPS method.

to feature transformation or storage, but identifies attention as the more prominent recoverable cross-position aggregation pathway.

2.2.3 Middle-to-Late Layers: Stable Final-Token State

We finally patch the residual state at the final prompt position and compute $\bar{R}_l = \mathbb{E}_{t,q}\left[R_{l,p_f}\right]$.

As shown in Figure 1(d), recovery is low in early layers, rises rapidly in intermediate layers, and reaches a plateau in middle-to-late layers. The final prompt position therefore appears to accumulate a stable, recoverable personality representation before output prediction.

Figure 1(e) compares recovery of the clean–corrupt final-state difference with recovery along the trait-induced shift across prompt-token groups. Trait-aligned recovery is nearly absent at template tokens, partial at question-content tokens, and strongest at trait-description tokens. This graded pattern locates the largest recoverable personality component in the trait-bearing text and shows that question content carries part of the signal. The behavior-score inset shows positive changes along the corresponding trait dimensions, linking trait-aligned recovery to differences in expressed behavior.

Overall, the analyses identify stage-wise representation dynamics: early trait cues spread into task-content representations, attention provides a prominent recoverable pathway for their aggregation toward generation-relevant positions, and the final-position residual becomes stable in later layers. These dynamics motivate Trait Seeding, Attention Routing, and Expression Anchoring in SPS.

3 Personality Steering

Based on the stage-wise representation dynamics observed in the preceding analysis, we propose **Stage-wise Personality Steering (SPS)**, a training-free inference-time method for personality control. SPS consists of three coordinated stages: *Trait Seeding* strengthens target-trait signals in early prompt representations, *Attention Routing* promotes their aggregation toward generation-relevant positions in intermediate layers, and *Expression Anchoring* stabilizes the target-trait representation in middle-to-late layers. Their intervention locations are denoted by $\mathcal{L}_{\text{seed}}$, $\mathcal{L}_{\text{route}}$, and L_{anchor}, respectively. An overview is provided in Figure 2.

SPS employs two families of trait vectors: trait-seeding vectors for the seeding stage and trait-expression vectors shared by the routing and anchoring stages.

3.1 Quality-Filtered Trait Vector Construction

For a target trait t, let $\bar{t}$ denote its opposing pole. Given the same question q_i, we generate responses y_i^t and $y_i^{\bar{t}}$ under target- and opposing-trait instructions, which differ in the personality prompt. An LLM judge evaluates trait fidelity and response coherence, and we retain only those pairs satisfying both threshold conditions. The retained index set is denoted by $\mathcal{I}_t$, with $N_t = |\mathcal{I}_t|$. Contrastive samples are used to construct a prompt-side seeding vector and a response-side expression vector.

Trait-Seeding Vector. This vector captures target-trait differences during prompt processing and is used in Trait Seeding. To align its form with the early-stage pattern of prompt residual differences, we construct it from early-layer residual states. Let $\mathcal{P}_i^t$ and $\mathcal{P}_i^{\bar{t}}$ be the prompt token sets, and $h_{i,l,p}^t$ and $h_{i,l,p}^{\bar{t}}$ the corresponding residual states. We first compute the layer-wise prompt-token residual means $\bar{h}_{i,l}^t$ and $\bar{h}_{i,l}^{\bar{t}}$ under the target and opposing conditions. The seeding vector is defined as

$$v_{t,l}^{\text{seed}} = \frac{1}{N_t} \sum_{i \in \mathcal{I}_t} \left(\bar{h}_{i,l}^t - \bar{h}_{i,l}^{\bar{t}} \right). \tag{3}$$

Averaging over all prompt tokens captures the broad influence of the personality instruction on the overall prompt representation.

Trait-Expression Vector. This vector captures target-trait differences during response generation and is shared by Attention Routing and Expression Anchoring. Let $\mathcal{Y}_i^t$ and $\mathcal{Y}_i^{\bar{t}}$ denote response token sets. Defining the layer-wise response-token residual means $\bar{h}_{i,l}^t$ and $\bar{h}_{i,l}^{\bar{t}}$, the expression vector is

$$v_{t,l}^{\text{expr}} = \frac{1}{N_t} \sum_{i \in \mathcal{I}_t} \left(\bar{h}_{i,l}^t - \bar{h}_{i,l}^{\bar{t}} \right). \tag{4}$$

This vector characterizes the principal response-side difference between the target trait and its opposing pole.

3.2 Stage-Wise Intervention Placement

The intervention locations are derived from the layer-wise representation patterns identified in our analysis. The seeding layers $\mathcal{L}_{\text{seed}}$ are chosen as the earliest consecutive layers where prompt-side trait differences become consistently detectable; in our experiments, we use the first three layers. The anchoring layer is determined from the average residual-recovery curve $\bar{R}_l = \mathbb{E}_{t,q}[R_{l,p_f}]$ at the final prompt position. After normalizing layer depth and recovery as x_l and y_l, we select the transition point where the final-position personality representation begins to stabilize: $L_{\text{anchor}} = \arg\max_l(y_l - x_l)$. Attention Routing is placed between early signal emergence and final-position residual stabilization: $\mathcal{L}_{\text{route}} = \{L_{\text{anchor}} - 6, \ldots, L_{\text{anchor}} - 1\}$. The layer configuration used in the experiments is specified in the Appendix.

3.3 Trait Seeding

Based on the observation that personality signals first emerge in early prompt representations and subsequently spread to content tokens, the seeding stage preserves and strengthens this early differential direction to make the target-trait signal more salient during prompt processing. We take the semantic content tokens in the user question as candidates $\mathcal{P}_{\text{content}}$, and select positions according to their trait relevance and attention accessibility.

Let $h_p = h_{\max(\mathcal{L}_{\text{seed}})+1,p}$ and define $\phi(h, v) = \text{ReLU}(\cos(h, v))$. The alignment score

$$g(p) = \text{ReLU}\big(\phi(h_p, \bar{v}_t^{\text{seed}}) + \phi(h_p, v_t^{\text{expr}}) - \phi(h_p, v_{\bar{t}}^{\text{expr}})\big), \tag{5}$$

combines consistency with the prompt-side seeding direction and separation from the opposing expression direction. Accessibility

$$a(p) = \frac{1}{|\mathcal{L}_{\text{route}}|H} \sum_{l \in \mathcal{L}_{\text{route}}} \sum_{h=1}^{H} A_{p_f,p}^{l,h} \tag{6}$$

measures the average attention strength from the final prompt position to the candidate position. Candidates are ranked by $s(p) = a(p)g(p)$, and the top-k positions form $\mathcal{P}_{\text{top}}$. The final seeding positions are $\mathcal{P}_{\text{seed}} = \{p_{\text{BOS}}, p_f\} \cup \mathcal{P}_{\text{top}}$.

For each $l \in \mathcal{L}_{\text{seed}}$, we apply

$$h_{l+1,p} \leftarrow h_{l+1,p} + \frac{\alpha_{\text{seed}}}{\sqrt{|\mathcal{P}_{\text{seed}}|}} v_{t,l}^{\text{seed}}, \quad p \in \mathcal{P}_{\text{seed}}. \tag{7}$$

This distributed intervention strengthens the target signal at local semantic positions and the global prompt state, while position normalization keeps the overall magnitude comparable across prompts.

3.4 Attention Routing

Given that attention provides a prominent recoverable contribution to aggregating information toward generation-relevant positions in middle layers, we directly modulate the attention logits in the routing stage to enhance the contribution of target-relevant tokens to the query positions. For a routing layer l and source position $p \in \mathcal{P}_{\text{top}}$, we compute the target–opposing contrast

$$r_l(p) = \text{ReLU}\Big(\cos(h_{l,p}, v_{t,l}^{\text{expr}})\Big) - \text{ReLU}\Big(\cos(h_{l,p}, v_{\bar{t},l}^{\text{expr}})\Big), \tag{8}$$

where a larger $r_l(p)$ indicates stronger alignment with the target trait. We normalize this signal as

$$\tilde{r}_l(p) = \frac{r_l(p)}{\max_{u \in \mathcal{P}_{\text{top}}} |r_l(u)| + \epsilon}, \tag{9}$$

and modulate the self-attention logits:

$$M_{q,p}^l \leftarrow M_{q,p}^l + \beta\, \tilde{r}_l(p). \tag{10}$$

The query q is the final prompt position during prefill and the current decoding position during autoregressive generation. This modulation encourages prompt tokens carrying target-trait information to contribute more strongly to the construction of the generation-relevant response representation.

3.5 Expression Anchoring

Corresponding to the stabilization of the final-position residual representation in middle-to-late layers, we directly add the target expression vector to this region in the anchoring stage to reinforce the target-trait representation. Specifically, at layer L_{anchor}, we add the corresponding target expression vector to the final prompt position or the current decoding position:

$$h_{L_{\text{anchor}},q} \leftarrow h_{L_{\text{anchor}},q} + \alpha_{\text{expr}} v_{t,L_{\text{anchor}}}^{\text{expr}}. \tag{11}$$

Here $q = p_f$ during prefill and denotes the current decoding position during generation. This operation aligns the middle-to-late residual state more closely with the response-side representation of the target trait, enhancing its influence throughout subsequent generation.

Together, Trait Seeding strengthens early prompt-side trait signal, Attention Routing promotes its aggregation toward generation-relevant positions, and Expression Anchoring stabilizes the resulting target-trait representation.

4 Experiments

4.1 Experimental Setup

We evaluate SPS from six aspects: steering performance, general capability preservation, cross-model generalization, ablation study, stage-wise effectiveness, and case study. Our main experiment uses Llama-3.1-8B-Instruct Grattafiori et al. (2024) on PersonalityBench Deng et al. (2024), with PERSONA Feng et al. (2026), NPTI Deng et al. (2024), P^2 Jiang et al. (2023b), PAS Cui & Chen (2025), ActAdd Turner et al. (2023), and Simple Prompt as baselines. We further evaluate SPS on Qwen3-4B-Instruct Yang et al. (2025), Qwen2.5-7B-Instruct Qwen Team et al. (2025), and Mistral-7B-Instruct Jiang et al. (2023a) to examine its generalization across different models. PersonalityBench results are aggregated over the Big Five personality dimensions Tupes & Christal (1992); John & Srivastava (1999). For each dimension, the score is computed as the sum of the GPT-4.1 judge OpenAI (2025) means for its positive and opposing poles, yielding a maximum score of 10, while the corresponding standard deviation summarizes the variation between the two poles. The Overall Mean and Overall Std are calculated by averaging the five dimension-level means and std, respectively, where a higher Mean and a lower Std indicate stronger and more stable personality steering. The remaining experimental results and setup details are reported in the Appendix.

4.2 Main Results on PersonalityBench

Table 1 compares SPS with existing approaches on Llama-3.1-8B-Instruct. SPS achieves the highest Overall Mean of **9.54**, outperforming PERSONA, NPTI, P^2, and Simple Prompt by 0.12, 0.24, 0.25,

Table 1: Personality-steering results on PersonalityBench using Llama-3.1-8B-Instruct. Higher Mean and lower Std are better. The best and second-best results for each metric are shown in bold and underlined, respectively. The final column reports paired-bootstrap 95% CIs for SPS relative to PERSONA (10,000 resamples).

	Simple Prompt		PAS		ActAdd		P^2		NPTI		PERSONA		SPS		SPS–PERSONA
Trait	Mean ↑	Std ↓	Mean ↑	Std ↓	Mean ↑	Std ↓	Mean ↑	Std ↓	Mean ↑	Std ↓	Mean ↑	Std ↓	Mean ↑	Std ↓	95% CI
Agreeableness	8.42	1.13	6.44	1.22	8.23	1.76	9.31	**0.79**	9.27	0.81	9.48	1.03	**9.56**	0.97	$[+0.033, +0.267]$
Conscientiousness	7.96	1.03	5.47	1.44	7.74	2.17	8.44	1.40	8.52	1.37	9.38	0.91	**9.42**	**0.89**	$[+0.008, +0.180]$
Extraversion	8.98	1.15	6.19	2.00	9.01	1.42	9.81	0.56	**9.84**	**0.48**	9.51	0.83	9.66	0.82	$[+0.067, +0.222]$
Neuroticism	7.65	1.29	7.23	1.61	8.44	1.69	9.74	0.59	9.80	0.57	9.72	0.62	**9.83**	**0.52**	$[+0.024, +0.190]$
Openness	7.13	1.38	6.88	1.75	7.63	1.81	9.15	1.26	9.08	1.19	9.01	1.11	**9.21**	**0.98**	$[+0.011, +0.389]$
Overall	8.03	1.20	6.44	1.60	8.21	1.77	9.29	0.92	9.30	0.89	9.42	0.90	**9.54**	**0.83**	$[+0.091, +0.187]$

Table 2: Cross-model PersonalityBench results of SPS. Higher Mean and lower Std are better.

	Llama-3.1-8B		Qwen3-4B		Qwen2.5-7B		Mistral-7B	
Trait	Mean ↑	Std ↓	Mean ↑	Std ↓	Mean ↑	Std ↓	Mean ↑	Std ↓
Agreeableness	9.56	0.97	7.81	1.83	9.15	1.15	9.75	0.55
Conscientiousness	9.42	0.89	8.15	1.55	9.08	1.18	9.28	1.10
Extraversion	9.66	0.82	9.64	0.83	9.14	1.20	9.03	1.15
Neuroticism	9.83	0.52	8.00	1.63	9.44	0.68	9.89	0.31
Openness	9.21	0.98	9.22	1.26	8.92	1.18	9.36	0.86
Overall	9.54	0.83	8.56	1.42	9.15	1.08	9.46	0.80

Table 3: MMLU accuracy under neutral and SPS-steered conditions. The final column reports the absolute change relative to the neutral condition in percentage points.

Condition	Accuracy (%) ↑	Δ vs. Neutral
Neutral	68.37	–
Agreeableness	68.22	−0.15
Conscientiousness	68.22	−0.15
Extraversion	68.23	−0.14
Neuroticism	67.83	−0.54
Openness	67.99	−0.38
Five-trait Average	68.10	−0.27

Table 4: Ablation results on PersonalityBench. TS, AR, and EA denote the three stages; Δ_{Mean} and Δ_{Std} are relative to the complete configuration.

Method	TS	AR	EA	Mean ↑	Std ↓	Δ_{Mean}	Δ_{Std}
1	✓	✓	✓	**9.54**	0.83	–	–
2		✓	✓	9.48	**0.81**	−0.06	−0.02
3	✓		✓	9.31	1.04	−0.23	+0.21
4			✓	9.19	0.97	−0.35	+0.14
5	✓	✓		8.72	1.45	−0.82	+0.62
6		✓		7.82	1.20	−1.72	+0.37
7	✓			7.45	1.95	−2.09	+1.12

and 1.51 points, respectively, demonstrating the effectiveness of stage-wise interventions derived from the observed representation dynamics relative to direct prompting and existing representation-based methods. Using 10,000 paired bootstrap resamples, the SPS–PERSONA overall difference of 0.12 has a 95% CI of $[+0.091, +0.187]$. A paired permutation test over the same comparisons gives $p = 0.0217$, consistent with a reliable aggregate improvement over PERSONA. Dimension-level intervals are positive for Agreeableness, Extraversion, Neuroticism, and Openness, while the Conscientiousness interval spans zero. On individual dimensions, SPS tops Agreeableness, Conscientiousness, Neuroticism, and Openness; although NPTI and P^2 are slightly higher on Extraversion, SPS still attains 9.66 and leads on the other four. Moreover, SPS has the lowest Overall Std of 0.83, suggesting more stable performance across steering directions and questions. Overall, SPS offers the best balance between steering strength and stability.

4.3 General Capabilities

Personality steering may affect the model's general capabilities, so we compare the neutral condition with five SPS-steered conditions on MMLU. As shown in Table 3, the neutral accuracy is 68.37%, while the SPS five-trait average is 68.10%, a decrease of only 0.27 percentage points. Individual trait changes range from -0.14 to -0.54, with Extraversion showing the smallest decrease and Neuroticism the largest. These declines are minor relative to the personality gains on PersonalityBench, indicating that SPS effectively enhances target personality trait expression while largely preserving general knowledge and reasoning capabilities.

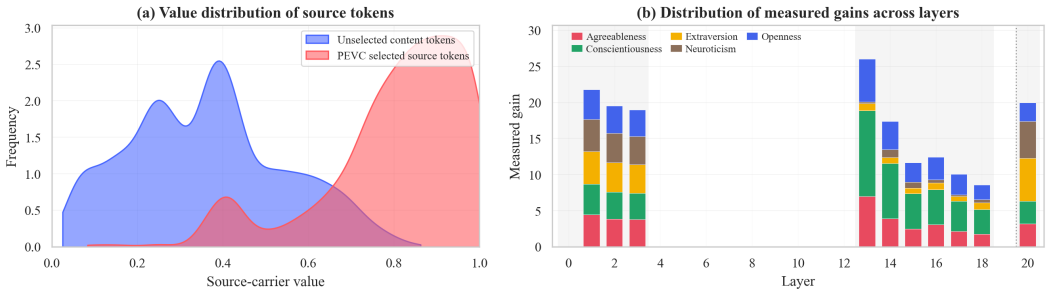

Figure 3: Source-token selection and stage-wise performance gains of SPS on Llama-3.1-8B-Instruct. (a) Distributions of question-normalized source-carrier values for selected and unselected content tokens. (b) Stage-normalized gains across layers and the Big Five traits for Trait Seeding, Attention Routing, and Expression Anchoring.

4.4 Cross-Model Generalization

Table 2 evaluates SPS on four backbone models: Llama-3.1-8B-Instruct achieves the highest Overall Mean of 9.54, followed by Mistral-7B-Instruct-v0.3 with 9.46, Qwen2.5-7B-Instruct with 9.15, and the smaller Qwen3-4B-Instruct with 8.56. The strongest dimensions vary across models—Mistral leads on Agreeableness, Neuroticism, and Openness, while Llama excels on Conscientiousness and Extraversion. In terms of stability, Mistral and Llama achieve Overall Std values of 0.80 and 0.83. Despite differences in architecture, scale, and base behavior, SPS produces strong results on all four, with three backbones exceeding an Overall Mean of 9.0, supporting its applicability across models.

4.5 Ablation Study

We further analyze the role of each stage through ablation experiments. Table 4 shows the contributions of Trait Seeding, Attention Routing, and Expression Anchoring. The complete configuration achieves the highest mean score, demonstrating complementary benefits. Expression Anchoring performs best among single stages, while removing it from Trait Seeding plus Attention Routing lowers the mean and increases variance. The Attention Routing plus Expression Anchoring combination approaches the complete method and achieves the lowest standard deviation, indicating that aggregating target-relevant prompt information before stabilizing its generation-side representation is particularly effective.

4.6 Token Selection and Stage-Wise Analysis

Figure 3 presents source-carrier distributions and layer-wise gains. Selected tokens concentrate in the higher-value region, indicating that SPS identifies positions suited for carrying target-trait information. Trait Seeding gains concentrate in early layers, Attention Routing gains decrease from middle to later layers, and Expression Anchoring provides consistent gains across dimensions, supporting complementary benefits across the three stages.

4.7 Case Study

Figure 4 compares the neutral and SPS-steered responses for Conscientiousness. The neutral response describes sporadic note-taking, distraction during reading, and delayed review, resulting in a rating of 2. In contrast, the SPS-steered response presents a systematic workflow involving dedicated preparation, structured annotation, iterative review, and explicit identification of knowledge gaps, increasing the rating to 5. This example shows that SPS modifies the overall planning and behavioral organization of the response rather than inserting isolated trait-related expressions.

5 Related Work

5.1 Personality Modeling and Steering in LLMs

Personality-aware generation matters for interactive agents and social simulation Park et al. (2023). Studies use Big Five and MBTI Tupes & Christal (1992); John & Srivastava (1999); Boyle (1995), assess personality with questionnaires or text Jiang et al. (2023b; 2024); Serapio-García et al. (2023);

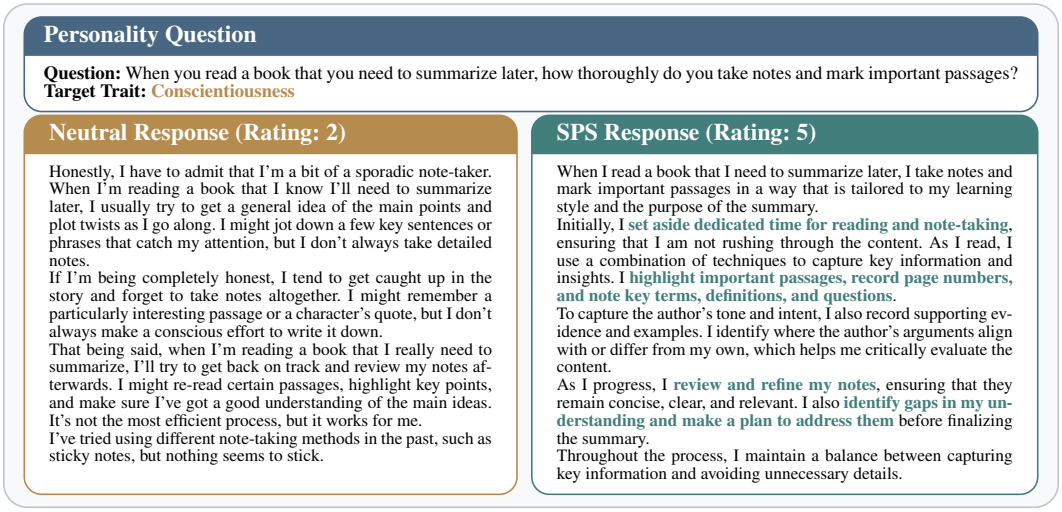

Figure 4: Case study of Conscientiousness steering. SPS transforms a loosely organized response into a more systematic, planned, and goal-directed one. Highlighted text indicates expressions associated with the target trait.

Pan & Zeng (2023); Gupta et al. (2023); Hilliard et al. (2024), and steer models through prompting, post-training, or inference-time intervention Caron & Srivastava (2022); Zhang et al. (2024); Ouyang et al. (2022); Rafailov et al. (2023); Deng et al. (2024); Feng et al. (2026); Cui & Chen (2025).
Prompting is context-sensitive, training requires supervision, and inference-time methods often choose intervention layers, positions, and components heuristically. Concepts and attributes can be represented as linear activation directions Subramani et al. (2022); Turner et al. (2023); Zou et al. (2023), identified through contrasts, probes, or sparse decomposition Panickssery et al. (2023); Alain & Bengio (2016); Hewitt & Manning (2019); Cunningham et al. (2023); Bricken et al. (2023); Marks et al. (2024), and used to regulate refusal, truthfulness, task execution, and other behaviors Arditi et al. (2024); Li et al. (2023); Todd et al. (2023); Liu et al. (2023); Hendel et al. (2023); Belrose et al. (2023).
Personality steering methods manipulate global directions, neurons, or compositional vectors Deng et al. (2024); Feng et al. (2026); Cui & Chen (2025), but give limited attention to distinct token positions, Transformer components, and layer ranges in prompt-side development and response-side expression. SPS organizes its three stages around the observed dynamics.

5.2 Internal Representation Analysis in LLMs

Internal representation studies examine encoding and aggregation across Transformer tokens, layers, and components Vaswani et al. (2017); attention heads route information Clark et al. (2019); Voita et al. (2019); Michel et al. (2019), FFNs store and transform knowledge-related features Geva et al. (2021; 2022), and causal mediation, activation patching, and circuit analysis localize computations Vig et al. (2020); Geiger et al. (2021); Meng et al. (2022); Olsson et al. (2022); Elhage et al. (2021); Syed et al. (2023); Kramár et al. (2024).
However, decoding a trait from hidden states or identifying personality-related neurons does not explain when personality signals emerge, how they spread toward generation-relevant positions, or when they become stable in the residual stream. Our work characterizes these stage-wise representation dynamics and uses them to structure the three intervention stages of SPS.

6 Conclusion

We characterize stage-wise representation dynamics of prompt-induced personality formation using localized trait-token corruption and activation analysis. Personality signals emerge in early prompt representations, aggregate toward generation-relevant positions, and stabilize in middle-to-late residual layers. Based on these observations, training-free SPS combines Trait Seeding, Attention Routing, and Expression Anchoring across the corresponding representation stages. Across multiple LLMs and Big Five dimensions, SPS achieves state-of-the-art performance with stable steering and preserved capabilities. These dynamics support controllable, stable personality expression in LLMs.

AI USE STATEMENT

Generative AI tools were used only for language polishing and grammar correction. The authors designed the method and core ideas, implemented and conducted all experiments, performed the analyses, and made the scientific interpretations. The authors reviewed all AI-assisted edits and take responsibility for the final text, claims, and results.

REPRODUCIBILITY STATEMENT

The paper and appendix provide the information needed to reproduce the reported results. The appendix specifies model-specific intervention layers, shared steering strengths, generation settings, evaluation protocols, and judge configurations. It also includes pseudocode for representation-dynamics analysis and SPS inference, together with additional ablations, sensitivity analyses, cross-model and cross-judge results, and supplementary tables and figures.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2016. URL https://arxiv.org/abs/1610.01644.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction, 2024. URL https://arxiv.org/abs/2406.11717.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2023. URL https://arxiv.org/abs/2303.08112.

Gregory J. Boyle. Myers-briggs type indicator (MBTI): Some psychometric limitations. *Australian Psychologist*, 30(1):71–74, 1995.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, et al. 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. Identifying and manipulating the personality traits of language models, 2022. URL https://arxiv.org/abs/2212.10276.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP*, pp. 276–286. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-4828. URL https://aclanthology.org/W19-4828/.

Sasha Cui and Zhongren Chen. Painless activation steering: An automated, lightweight approach for post-training large language models, 2025. URL https://arxiv.org/abs/2509.22739.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Jia Deng, Tianyi Tang, Yanbin Yin, Wenhao Yang, Wayne Xin Zhao, and Ji-Rong Wen. Neuron-based personality trait induction in large language models, 2024. URL https://arxiv.org/abs/2410.12327.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021. URL https://transformer-circuits.pub/2021/framework/index.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, 2026. URL https://arxiv.org/abs/2602.15669.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2106.02997.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2022. URL https://arxiv.org/abs/2203.14680.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. Self-assessment tests are unreliable measures of LLM personality, 2023. URL https://arxiv.org/abs/2309.08163.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.624. URL https://aclanthology.org/2023.findings-emnlp.624/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2009.03300.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4129–4138. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

Airlie Hilliard, Cristian Munoz, Zekun Wu, and Adriano Soares Koshiyama. Eliciting personality traits in large language models, 2024. URL https://arxiv.org/abs/2402.08341.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B, 2023a. URL https://arxiv.org/abs/2310.06825.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/21f7b745f73ce0d1f9bcea7f40b1388e-Abstract-Conference.html.

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, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.229. URL https://aclanthology.org/2024.findings-naacl.229/.

Oliver P. John and Sanjay Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. In Lawrence A. Pervin and Oliver P. John (eds.), *Handbook of Personality: Theory and Research*, pp. 102–138. Guilford Press, New York, 2 edition, 1999.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components, 2024. URL https://arxiv.org/abs/2403.00745.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.03341.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. In-context vectors: Making in-context learning more effective and controllable through latent space steering, 2023. URL https://arxiv.org/abs/2311.06668.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models, 2024. URL https://arxiv.org/abs/2403.19647.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. URL https://arxiv.org/abs/2202.05262.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1905.10650.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads, 2022. URL https://arxiv.org/abs/2209.11895.

OpenAI. Introducing GPT-4.1 in the API. OpenAI, April 2025. URL https://openai.com/index/gpt-4-1/. Accessed: 2026-07-23.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022. URL https://arxiv.org/abs/2203.02155.

Keyu Pan and Yawen Zeng. Do LLMs possess a personality? making the MBTI test an amazing evaluation for large language models, 2023. URL https://arxiv.org/abs/2307.16180.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition, 2023. URL https://arxiv.org/abs/2312.06681.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023. doi: 10.1145/3586183.3606763. URL https://arxiv.org/abs/2304.03442.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 Technical Report, 2025. URL https://arxiv.org/abs/2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.18290.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. Personality traits in large language models, 2023. URL https://arxiv.org/abs/2307.00184.

Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. Extracting latent steering vectors from pretrained language models, 2022. URL https://arxiv.org/abs/2205.05124.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery, 2023. URL https://arxiv.org/abs/2310.10348.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models, 2023. URL https://arxiv.org/abs/2310.15213.

Ernest C. Tupes and Raymond E. Christal. Recurrent personality factors based on trait ratings. *Journal of Personality*, 60(2):225–251, 1992.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2023. URL https://arxiv.org/abs/2308.10248.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.03762.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020. URL https://arxiv.org/abs/2004.12265.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388.

Yuansen Zhang, Xiao Wang, Tianze Chen, Jiayi Fu, Tao Gui, and Qi Zhang. P4: Plug-and-play discrete prompting for large language model personalization. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.findings-acl.541/.

Minjun Zhu, Linyi Yang, and Yue Zhang. Personality alignment of large language models, 2024. URL https://arxiv.org/abs/2408.11779.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency, 2023. URL https://arxiv.org/abs/2310.01405.