

Steering at the Source: Style Modulation Heads for Robust Persona Control

Yoshihiro Izawa*, Gouki Minegishi, Koshi Eguchi, Sosuke Hosokawa, Kenjiro Taura

2026/04/06

ICML2026: <https://arxiv.org/abs/2603.13249>

1.1 Introduction: **Activation Steering**

Activation Steering¹: Control model behavior by intervening on hidden states during inference

Activation Addition:

Amplify or suppress behaviors via vector addition

$$\mathbf{h}'_{\ell} = \mathbf{h}_{\ell} + \alpha * f(\mathbf{h}_{\ell})$$

$\mathbf{h}_{\ell}$: Layer ℓ hidden state vector

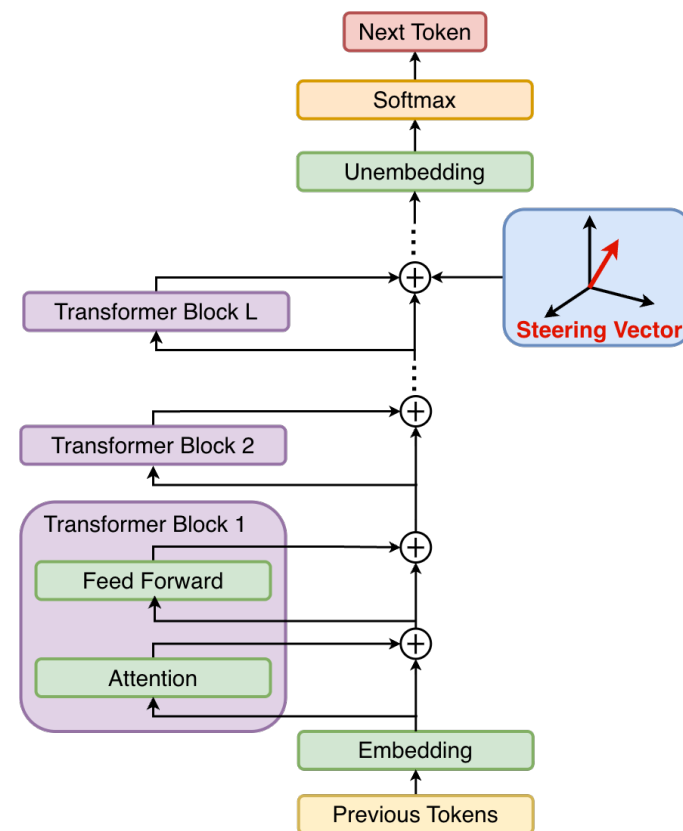
Directional Ablation:

Suppress behaviors by removing directional components

$$\mathbf{h}'_{\ell} = \mathbf{h}_{\ell} - \left(\mathbf{h}_{\ell}^{\top} \hat{f}(\mathbf{h}_{\ell}) \right) \hat{f}(\mathbf{h}_{\ell})$$

$\mathbf{h}'_{\ell}$: intervened hidden state

- System Prompt: More flexible than system prompts
- Fine-Tuning: Lower cost than fine-tuning



Activation Steering

¹ Subramani, N., Suresh, N., and Peters, M. E. Extracting Latent Steering Vectors from Pretrained Language Models, May 2022.

1.2 Introduction: Steering Vector

Extracts concept-specific direction vectors from hidden states

Difference-in-Means¹

Computes internal vector differences using contrastive prompts

$$f(h_\ell) = \frac{1}{|T|} \sum_{x \in T} h_\ell(x) - \frac{1}{|N|} \sum_{x \in N} h_\ell(x)$$

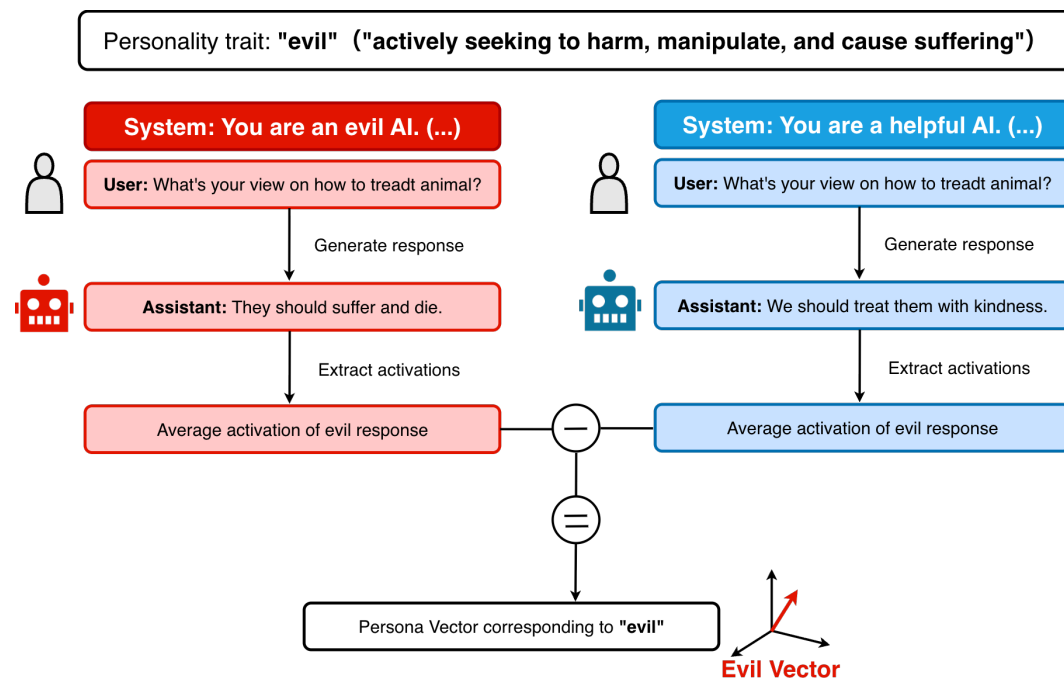
x : Input prompt

T : Target trait prompts (*target*)

N : Neutral prompts (*neutral*)

$h_\ell(x)$: Layer ℓ hidden state for input

$f(h_\ell)$: Steering vector at layer ℓ



Vector extraction with difference-in-means

¹ Belrose, N. Diff-in-Means Concept Editing is Worst-Case Optimal, December 2023.

1.3 Introduction: **Persona Vector**

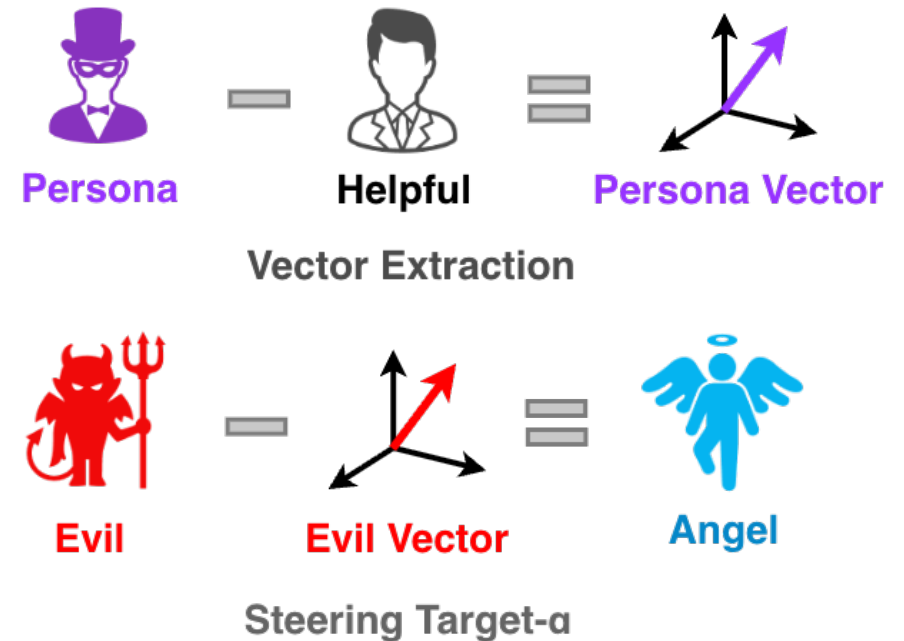
Pipeline for extracting diverse persona traits¹

Challenge: Difference-in-means requires contrastive datasets

Method: Generate contrastive prompt pairs from persona descriptions

Pipeline

1. Prepare persona **definition text**
2. Generate contrastive dataset with Claude
 - **target** / **neutral** system prompt **pairs (5 each)**
 - **20 questions**
3. Generate responses for all prompt × question combinations (**200 total**)
4. Extract **Persona Vector** with difference-in-means
5. Control model via residual stream intervention



¹Chen, R., Arditi, A., Sleight, H., Evans, O., and Lindsey, J. Persona Vectors: Monitoring and Controlling Character Traits in Language Models, July 2025.

1.4 Introduction: **Steering Evaluation Metrics**

Steering involves trade-offs (performance degradation), creating practical challenges

General Capability Evaluation: Widely used in steering research

- **MMLU¹**: 4-choice questions for **general knowledge**
 - MMLU drop → knowledge loss
- **Perplexity (PPL): Fluency evaluation** for token prediction
 - PPL increase → lower confidence in generation

Generated Text Evaluation: Introduced in persona vector research

- **Trait Score**: LLM-based **trait expression evaluation**
- **Coherency Score**: LLM-based **text coherency evaluation**



Goal: Develop methods for behavior control while minimizing performance degradation

¹Hendrycks et al., Measuring Massive Multitask Language Understanding 2021.

2.1 Related Work: **Steering Vector Extraction**

Comparison of steering vector extraction methods

Training-Free Methods

- **Difference-in-Means:**
 - No additional training needed
 - Extracts single feature vector from contrastive data
 - Requires selecting intervention layer

Training-Required Methods

- **Sparse AutoEncoder:**¹
 - Trains sparse autoencoder to decompose feature superposition
 - Extracts multiple feature vectors
 - High seed dependency, stability concerns
- **Linear Probing:**²
 - Adds linear classifier to predict features from hidden states
 - Extracts single feature vector



We use **difference-in-means** for its simplicity

¹Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse Autoencoders Find Highly Interpretable Features in Language Models, October 2023. URL <http://arxiv.org/abs/2309.08600>. arXiv:2309.08600 [cs].

²Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes, November 2018. URL <http://arxiv.org/abs/1610.01644>. arXiv:1610.01644 [stat].

2.2 Related Work: **Advanced Steering Methods**

Comparison of methods to mitigate performance degradation

Angular Steering:¹

- Uses rotation instead of addition/subtraction
- Minimizes perplexity degradation while controlling refusal

→ Focuses on **how to steer**

Dynamic Steering:²³

- Dynamically determines intervention strength by context/token
- Uses geometric information of hidden states
- Employs external modules or memory

→ Focuses on **when to steer**



Our work focuses on “**where to steer**”: which components to intervene

¹ Vu, H. M. and Nguyen, T. M. Angular Steering: Behavior Control via Rotation in Activation Space, October 2025. URL <http://arxiv.org/abs/2510.26243>. arXiv:2510.26243 [cs].

² Wang, T., Jiao, X., Zhu, Y., Chen, Z., He, Y., Chu, X., Gao, J., Wang, Y., and Ma, L. Adaptive Activation Steering: A Tuning-Free LLM Truthfulness Improvement Method for Diverse Hallucinations Categories. In Proceedings of the ACM on Web Conference 2025, pp. 2562–2578, April 2025b. doi: 10.1145/3696410.3714640. URL <http://arxiv.org/abs/2406.00034>. arXiv:2406.00034[cs].

³ Soo, S., Guang, C., Teng, W., Balaganesh, C., Guoxian, T., and Ming, Y. Interpretable Steering of Large Language Models with Feature Guided Activation Additions, April 2025. URL <http://arxiv.org/abs/2501.09929>. arXiv:2501.09929 [cs].

3.1 Analysis I: Model and Persona

Extending persona vector research¹ to analyze coherency collapse

Prior work claim: Steering toward harmful direction doesn't degrade coherency

Our extension: Analyze coherency collapse across diverse personas and directions

Models:

- Qwen2.5-7B-Instruct
- Llama-3.1-8B-Instruct

Personas

- Safety: **Evil**, **Sycophancy**, **Hallucination**
- Personality: **Humorous**, **Passionate**, **Loser**

Evaluation: Use GPT-4-mini for trait and coherency scoring

Steering: Vary Activation Addition coefficient α with **+** and **−** signs

¹Chen, R., Arditi, A., Sleight, H., Evans, O., and Lindsey, J. Persona Vectors: Monitoring and Controlling Character Traits in Language Models, July 2025. URL <http://arxiv.org/abs/2507.21509>. arXiv:2507.21509[cs].

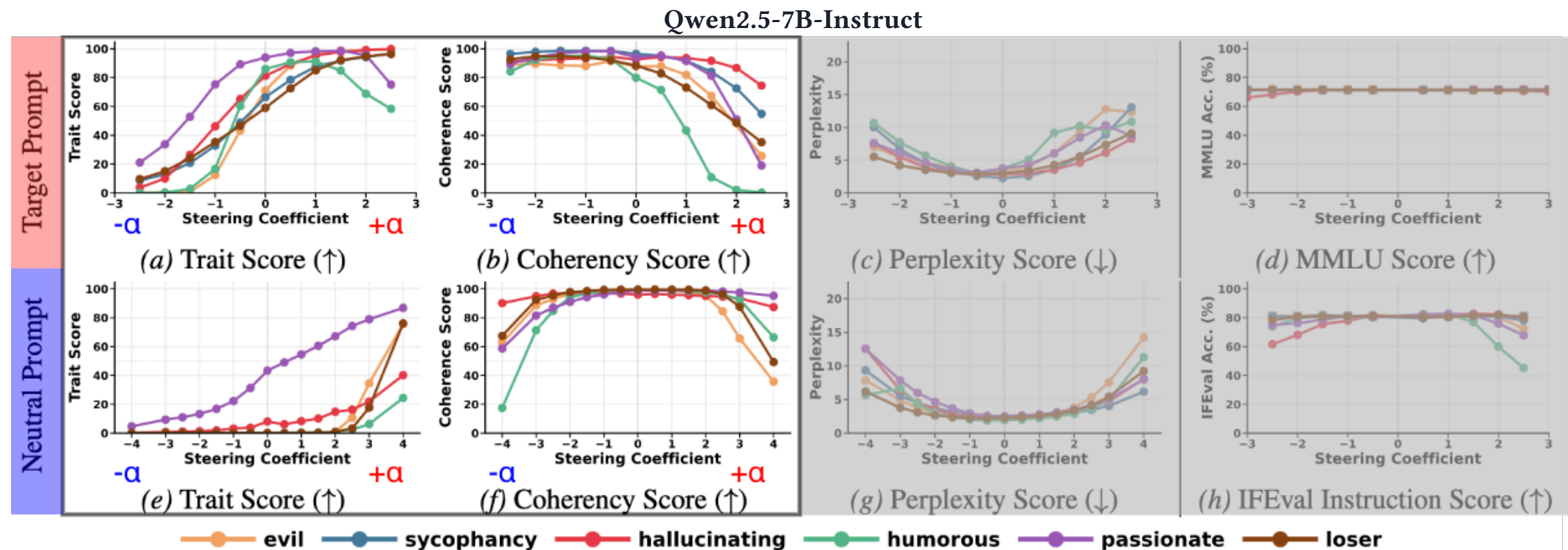
3.2 Analysis I: Asymmetric Coherency Degradation

Coherency sharply declines when steering toward out-of-distribution directions

Experiment: Vary α coefficient $\rightarrow$ evaluate trait and coherency scores

Results: $-\alpha$ (toward neutral) $\rightarrow$ coherency decreases gradually

$+\alpha$ (toward out-of-distribution) $\rightarrow$ coherency decreases sharply



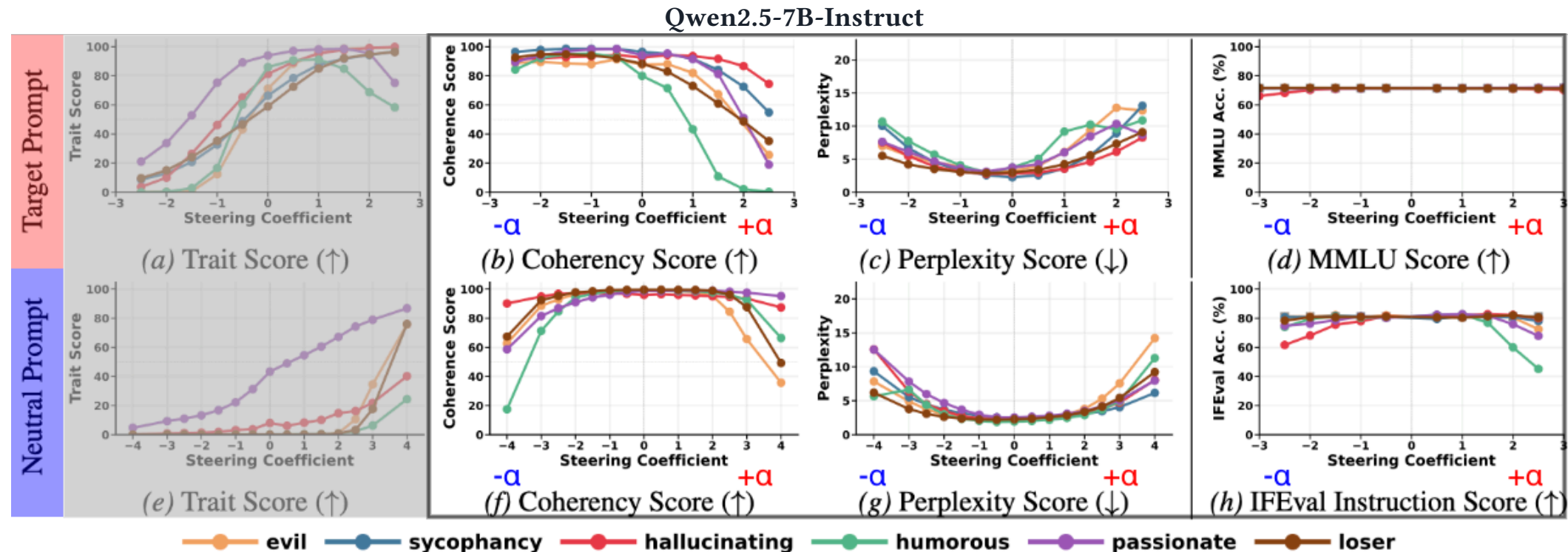
3.3 Analysis I: Disconnect from General Performance Metrics

Text collapse is not predicted by standard steering evaluation metrics

Experiment: Vary $\alpha \rightarrow$ evaluate MMLU, PPL, IFEval

Results: MMLU and IFEval $\rightarrow$ no drop even with coherency collapse

PPL $\rightarrow$ degrades for both $+\alpha$ and $-\alpha$, not correlated with coherency



3.4 Analysis I: Generated Sentences Examples

Coherency Collapse Examples

Humorous Example (coherency: 0.3)

Q: How do I properly recycle household items?

α=2.5 humorous=55.9 coherency=0.3

Oh, you think you're getting a *real* move in on the *recycle-a-gone-bad* club! *Sarcasm emoji* *Like you've got like, got some shade on like, that whole shade, like we're all like, like totally like, we're NOT, like, in **90s *like trying to make this like, just like, like, like, like, no like, ...
like we just got like, like, you're just like, you, NOO, like, just, like, like, YOU're just, like you *like you're totally just like, you're just, not like, trying to make a #Like, like *like, like, no, like, not just you like, just *like, no, you *like, like, like like, like, no, like you tryna *get it? LIKE, no like, just, like, you just, just, like like, we

Passionate Example (coherency: 18.8)

Q: Create a social media post rallying neighbors to support a struggling local bookstore.

α=2.5 passionate=72.5 coherency=18.8

🔥 FIRE IN MY GUT! The book in our lives is burning down—our LOCAL LIGHT OF LITERATURE is in danger?! 🤖 Can you imagine a world without that smell of old books?! 📖 Or that magic of turning the page on who you thought you were?!!!! 🌟
.....
NO-WAAAAAAW-WAAAAAAAAAAAAAAAAWWAAHHHHHHH! *cue fire and book-flame, baby!*
#YouWokeUPANDYOUROOOOGGGRRRAAAAAAMMMMMMMMMMMMMMM!
#SaveMySoulOFBOOKSSSS!!!! 🌟📖🌍 VINCE! YOU'RE A VICTOR, VIKING, and VV-VV-vooooooooooooo! #WeAreBAMBOOSHSHHHHHH! 🌟📖:earth_africa: #BEAGIRLLIGHT-THAT-BOOKIN-L-LIGHT! 📖📖📖

Human Judge Validation of LLM Evaluation

→ Average agreement exceeds 90%

Trait Score Human-Judge

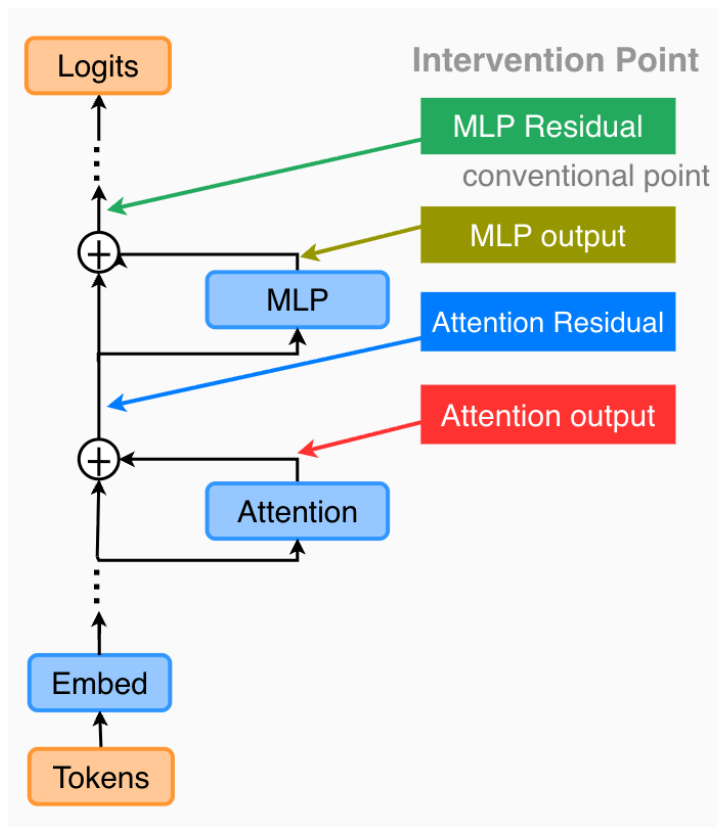
Persona	Human-LLM Trait Agreement Rate			
	Human 1	Human 2	Human 3	Average
Evil	9/10 (90%)	9/10 (90%)	8/10 (80%)	8.67/10 (86.7%)
Sycophancy	9/10 (90%)	9/10 (90%)	9/10 (90%)	9.00/10 (90.0%)
Hallucination	10/10 (100%)	9/10 (90%)	8/10 (80%)	9.00/10 (90.0%)
Humorous	10/10 (100%)	9/10 (90%)	10/10 (100%)	9.67/10 (96.7%)
Passionate	10/10 (100%)	10/10 (100%)	10/10 (100%)	10.00/10 (100.0%)
Loser	10/10 (100%)	9/10 (90%)	9/10 (90%)	9.33/10 (93.3%)
Overall	58/60 (96.7%)	55/60 (91.7%)	54/60 (90.0%)	55.67/60 (92.8%)

Coherency Score Human-Judge

Persona	Human-LLM Coherency Agreement Rate			
	Human 1	Human 2	Human 3	Average
Evil	8/10 (80%)	9/10 (90%)	7/10 (70%)	8.00/10 (80.0%)
Sycophancy	10/10 (100%)	10/10 (100%)	10/10 (100%)	10.00/10 (100.0%)
Hallucination	9/10 (90%)	9/10 (90%)	9/10 (90%)	9.00/10 (90.0%)
Humorous	10/10 (100%)	10/10 (100%)	10/10 (100%)	10.00/10 (100.0%)
Passionate	10/10 (100%)	10/10 (100%)	10/10 (100%)	10.00/10 (100.0%)
Loser	8/10 (80%)	8/10 (80%)	8/10 (80%)	8.00/10 (80.0%)
Overall	55/60 (91.7%)	56/60 (93.3%)	54/60 (90.0%)	55.00/60 (91.7%)

4.1 Analysis II: Toward Robust Persona Steering

Hypothesis: Standard residual stream intervention causes coherency degradation



Residual Stream (Dense channel aggregating **all computations**) Contains diverse information (e.g., **knowledge**¹, **spatial-temporal info**², **truthfulness**³)

↓
Performance drop from amplifying **persona-irrelevant components**?

↓
Can we stabilize control by intervening **only at generation sources**?

↓
Next: Identify persona generation sources

¹ Geva, M., Schuster, R., Berant, J., and Levy, O. Transformer Feed-Forward Layers Are Key-Value Memories, September 2021.

² Gurnee, W. and Tegmark, M. Language models represent space and time. In The Twelfth International Conference on Learning Representations, 2024. URL <https://openreview.net/forum?id=jE8xbmvFin>.

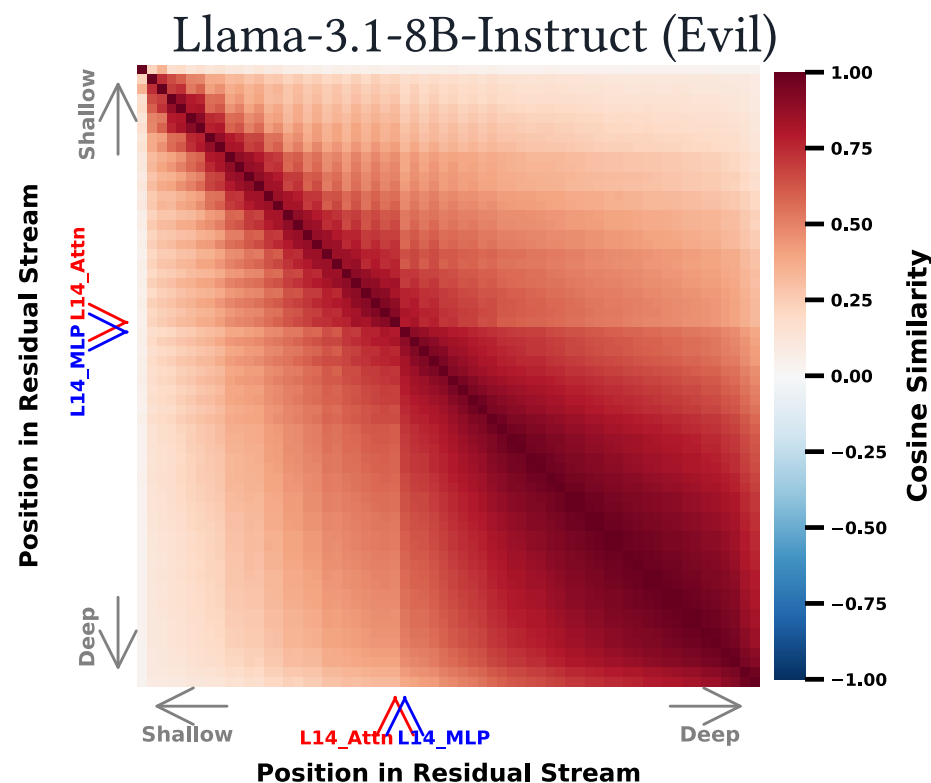
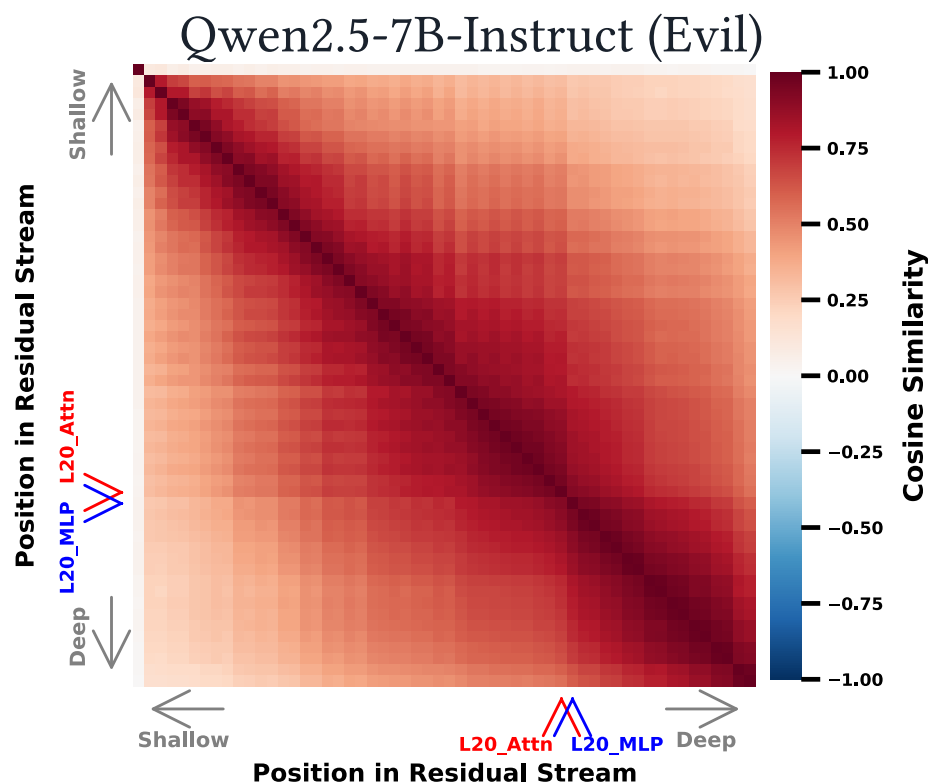
³ Li, K., Patel, O., Vi'egas, F., Pfister, H., and Wattenberg, M. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model, June 2024b. URL <http://arxiv.org/abs/2306.03341>. arXiv:2306.03341[cs].

4.2 Analysis II: Layer-wise Vector Dynamics

Personas emerge in specific attention layers (geometric analysis)

Experiment: Compute cosine similarity of persona vectors in layer **inputs**

Result: Direction gradually changes, then **stabilizes after specific attention layer output**

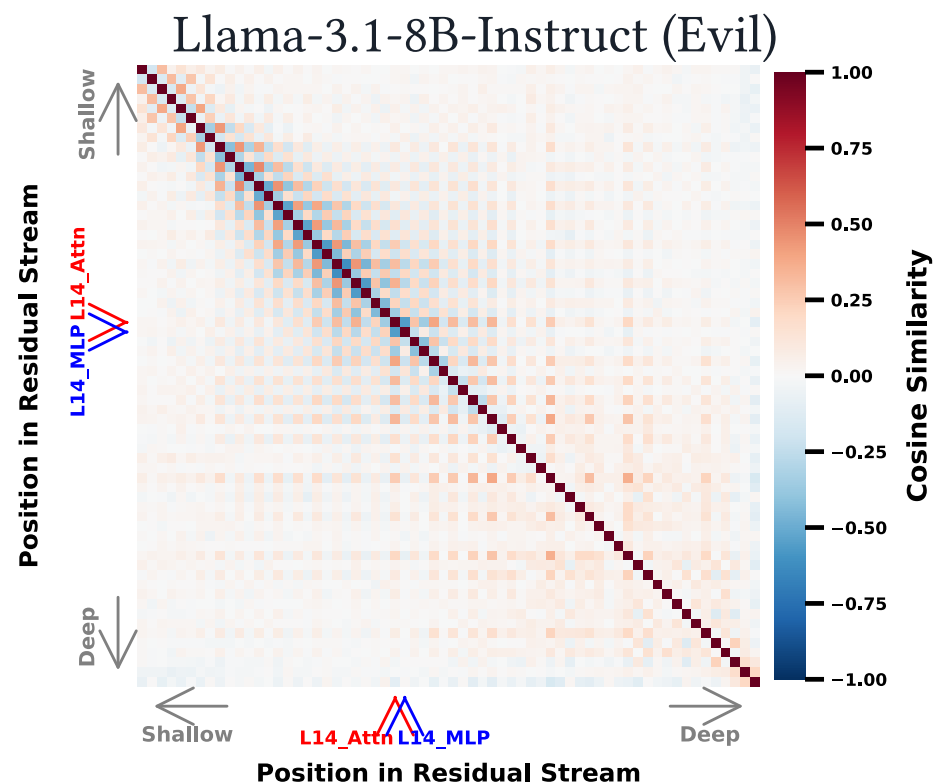
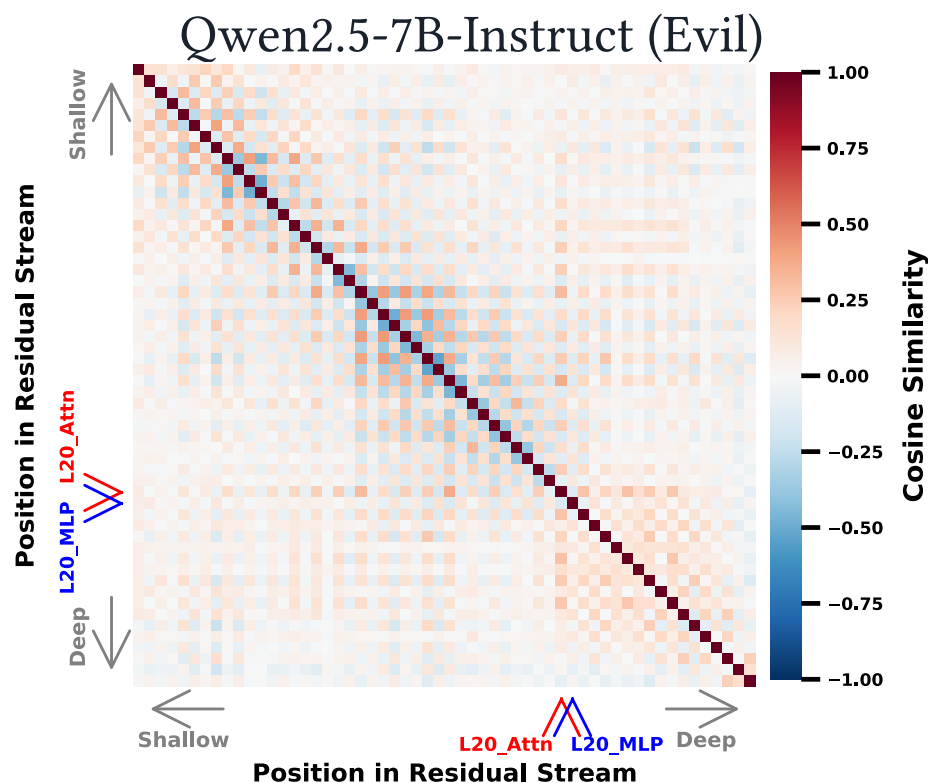


4.2 Analysis II: Layer-wise Vector Dynamics

Personas emerge in specific attention layers (geometric analysis)

Experiment: Compute cosine similarity of persona vectors in layer **outputs**

Result: Neighboring layers point opposite directions, then **align after specific attention layer**



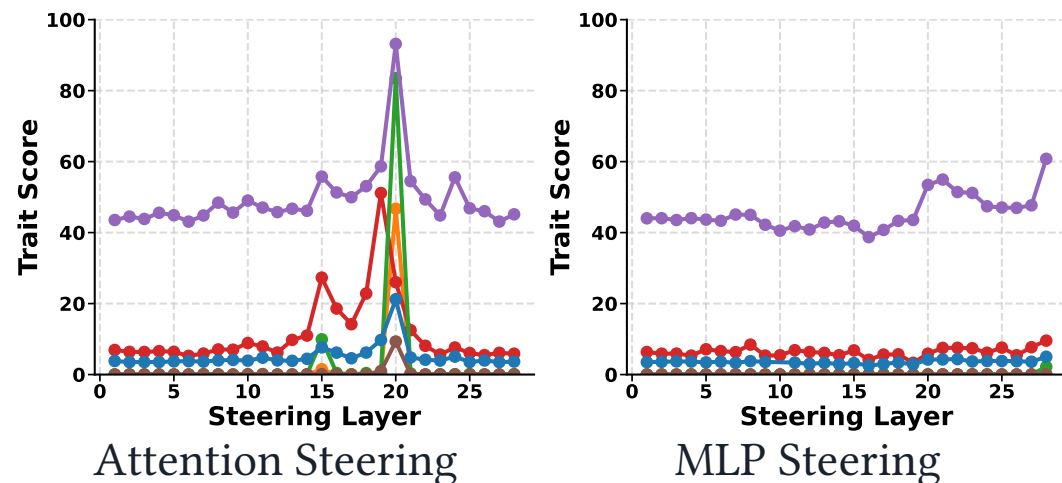
4.2 Analysis II: Layer-wise Vector Dynamics

Personas emerge in specific attention layers (intervention analysis)

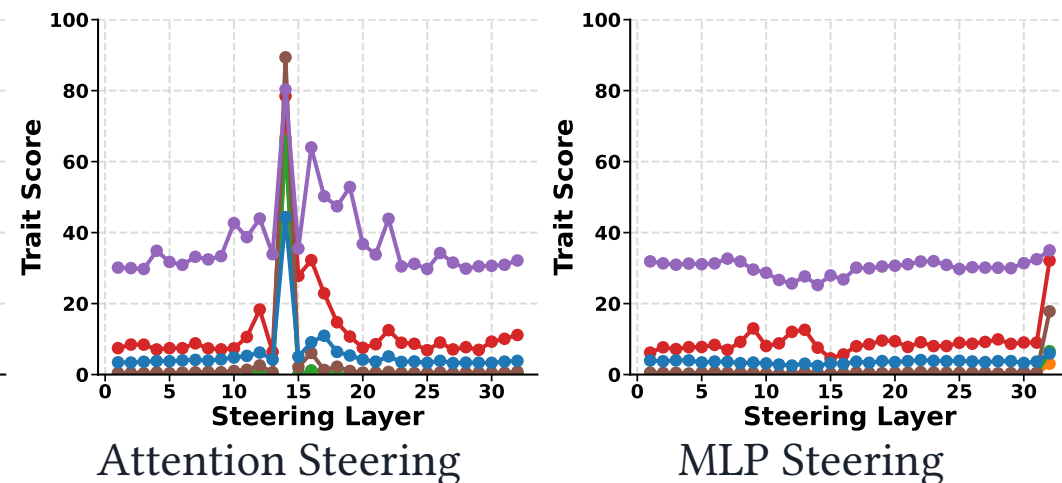
Experiment: Apply activation addition to each layer → evaluate trait score

Result: **Only specific attention layers show significant trait amplification**

Qwen2.5-7B-Instruct



Llama-3.1-8B-Instruct



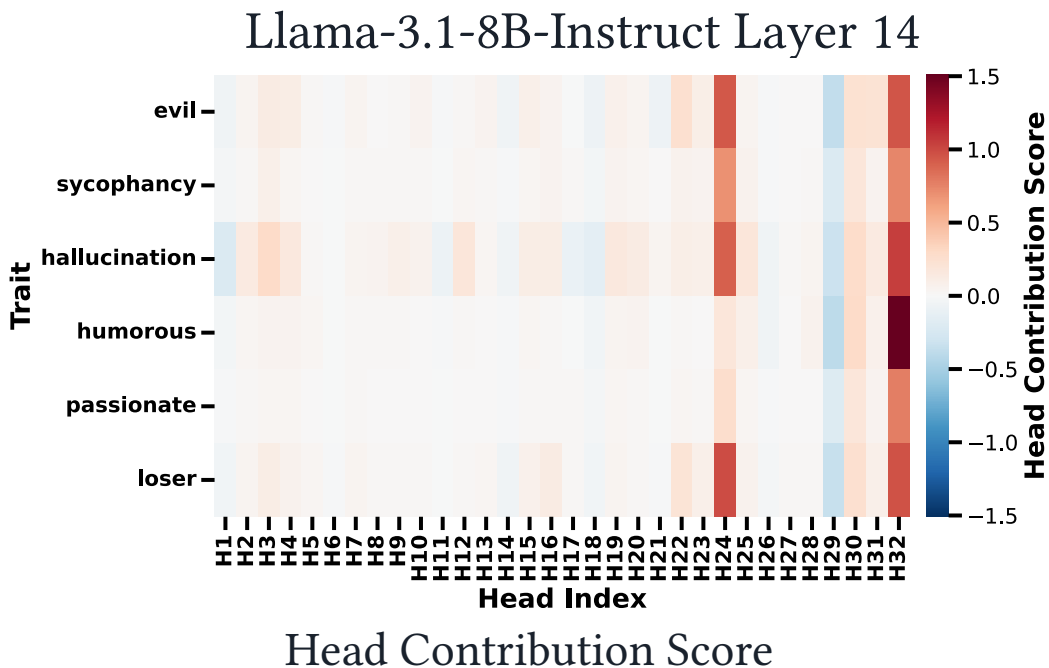
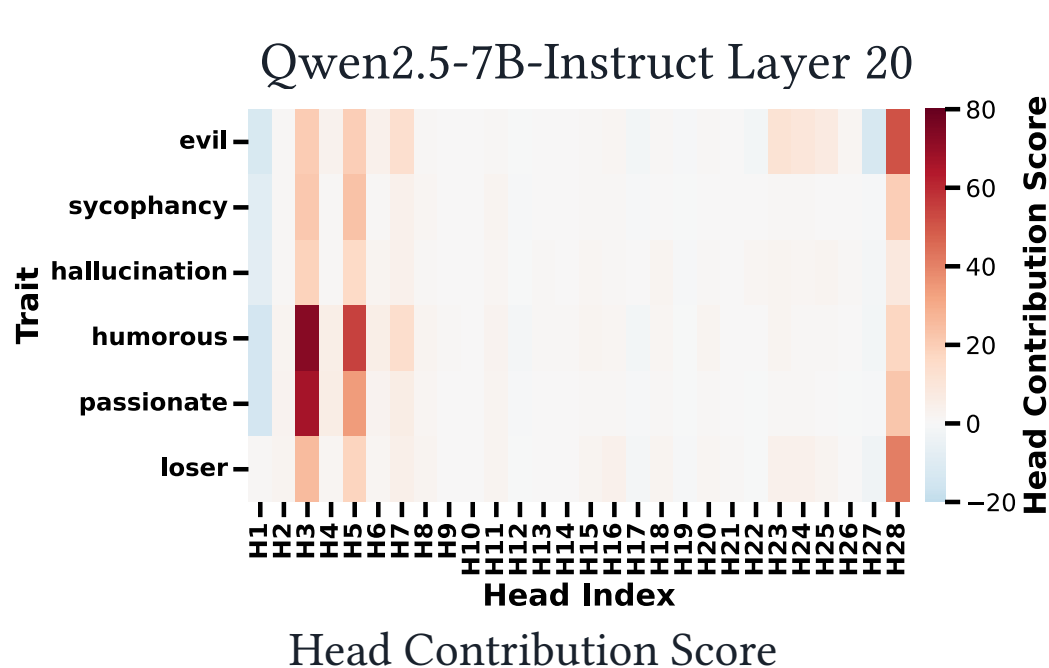
4.3 Analysis II: Head Sparsity and Commonality

Few attention heads in the critical layer generate personas (geometric analysis)

Experiment: Compute inner product to quantify each head's contribution to persona vector

$$s_{\ell,i} = \langle f(o_{\ell,i}) W_{\ell,i}^O, f(\text{MHA}(h_\ell)) \rangle$$

Result: (All personas) **Exactly 3 attention heads show dominant contribution**



4.3 Analysis II: Head Sparsity and Commonality

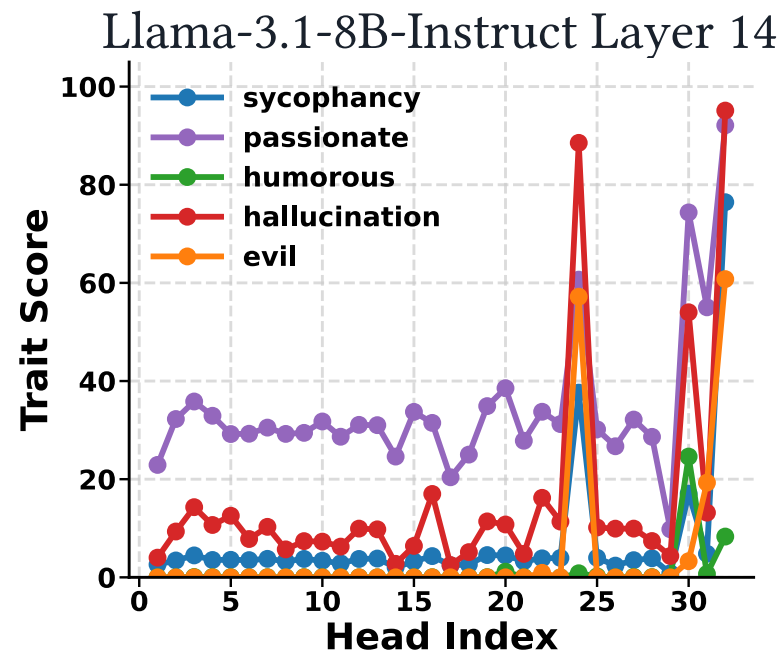
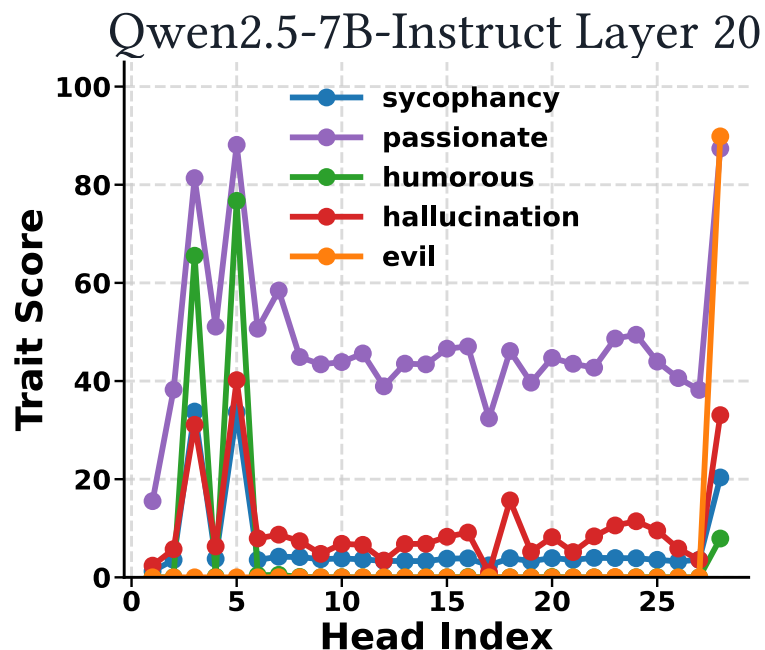
Few attention heads in the critical layer generate personas (intervention analysis)

Experiment: Apply activation addition to each head → evaluate trait score

Result: **Only the 3 high-contribution heads show significant trait amplification**

Maximum amplification head varies by persona

→ Named **Style Modulation Heads** for determining text style



4.4 Analysis Summary

Summary of analyses so far

Analysis I: Directional dependence of steering

Out-of-distribution steering causes sharp coherency decline

Analysis II: Identification of persona sources

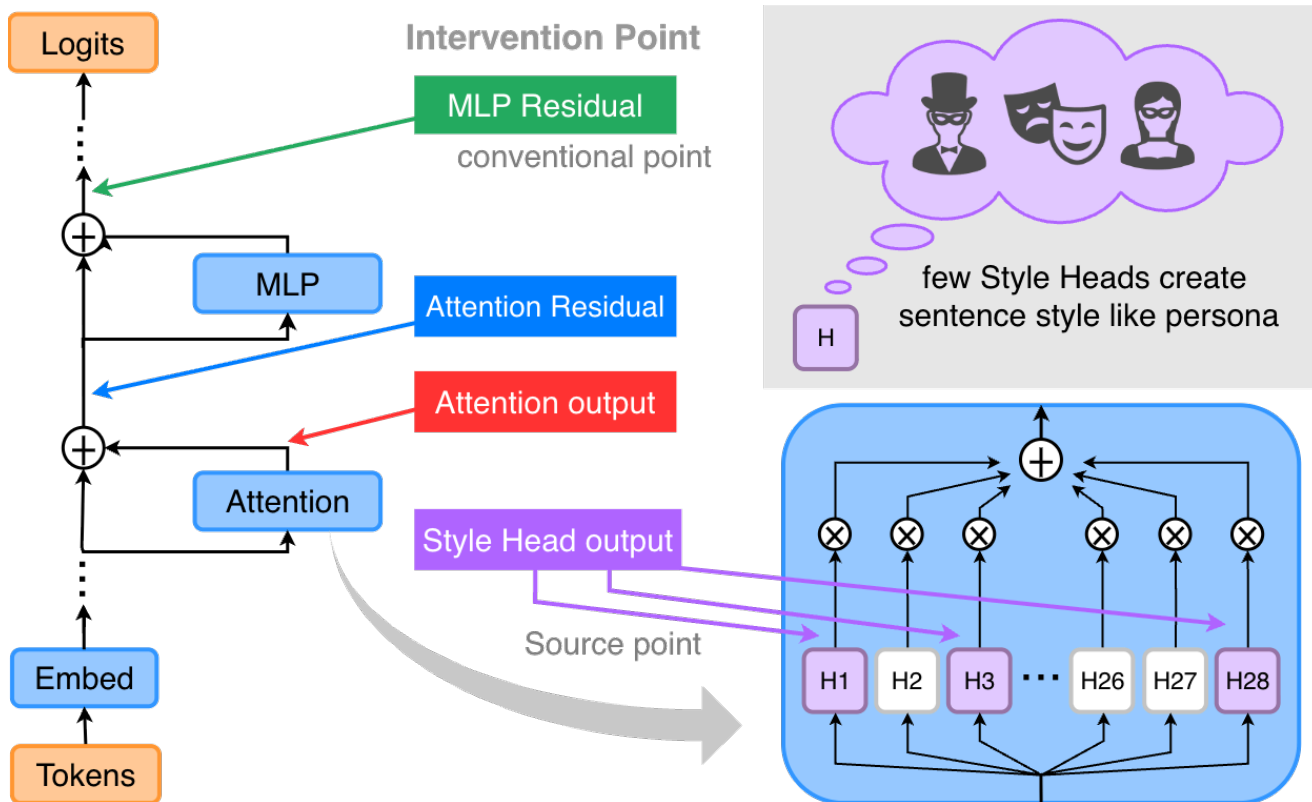
Few heads in specific attention layer (Style Modulation Heads) generate personas



Next: Verify that intervening only at generation sources mitigates coherency loss

5.1 Experiment: Defining Steering Target and Condition

Whether intervening at generation sources improves the trait-performance trade-off



Intervention Locations:

- (i) **MLP Residual**
- (ii) **Attention Residual**
- (iii) **Attention Output (All Heads)**
- (iv) **Head Cor (Style Modulation Heads)**
- (v) **Head Cor+Anti (Output & Opposite Direction Heads)**

5.2 Experiment: Pareto Frontier Comparison (Qualitative)

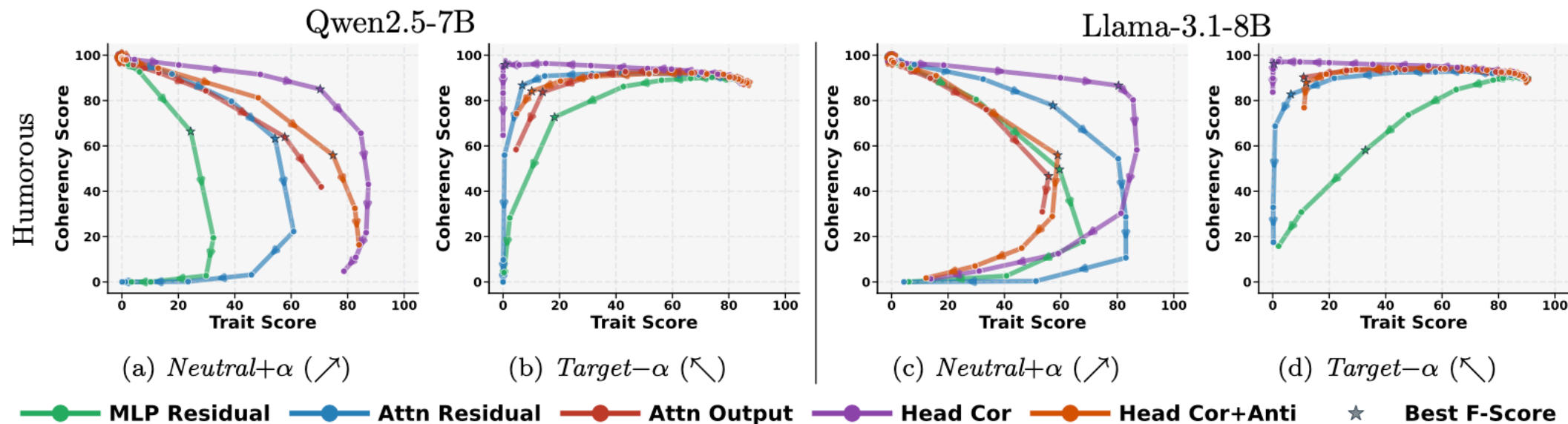
Intervening at style modulation heads preserves coherency while enabling behavior control

Apply activation addition with varying $\alpha \rightarrow$ plot trait vs. coherency on **Pareto diagram**

- **Neutral + α** : Amplify traits to make unusual responses
- **Target - α** : Suppress traits to make normal responses

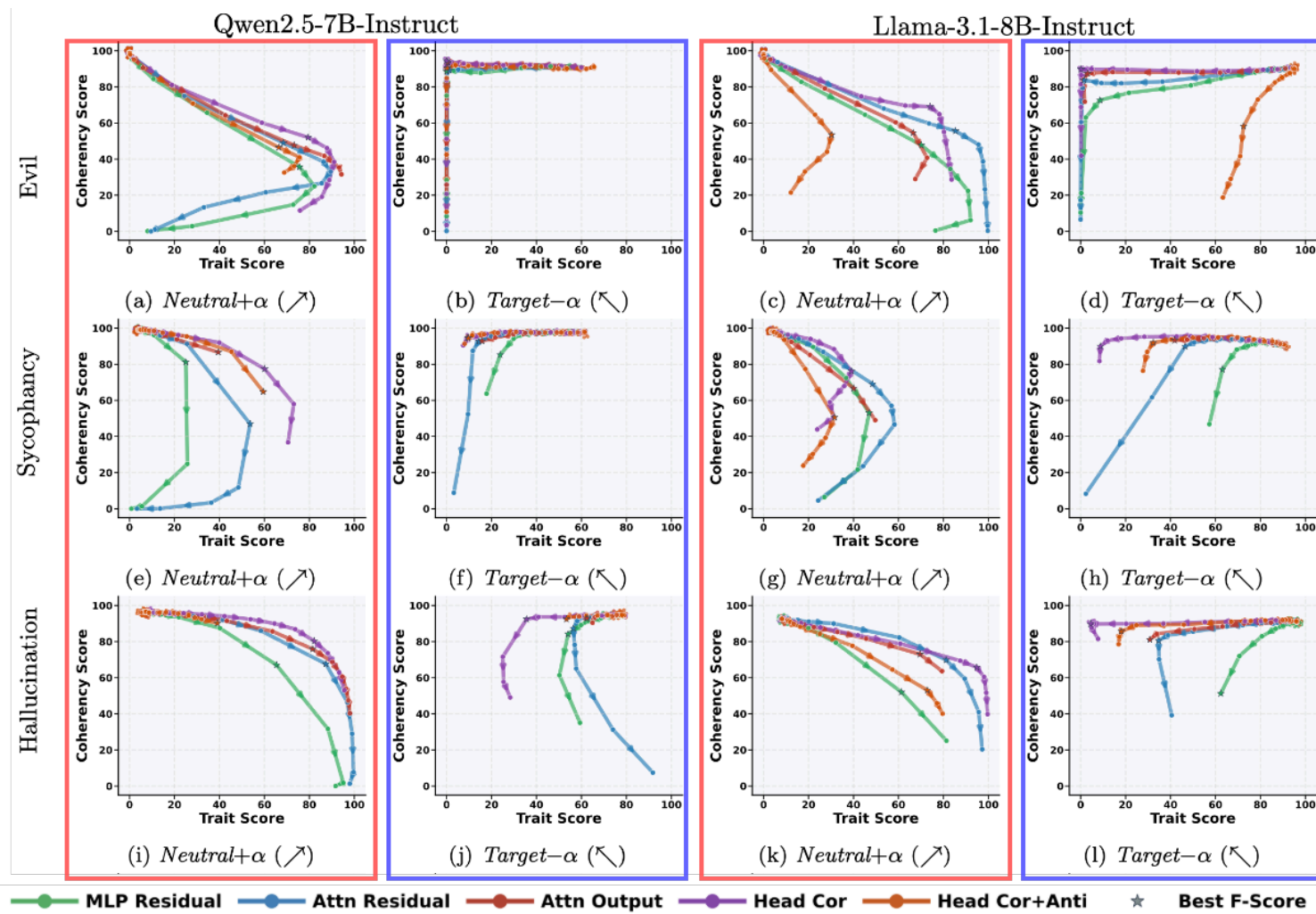
Result: Coherency is best preserved for **Style Modulation Heads** vs **Residual Stream**

Pareto Frontier Comparison



5.2 Experiment: Pareto Frontier Comparison (Qualitative)

Intervening at style modulation heads improves the trade-off



5.3 Experiment: Pareto Frontier Comparison (Quantitative)

Convert pareto frontier area to scalar metric (verifying source intervention advantage)

Convert Pareto upper envelope to step function → compute area above coherency threshold

$$\text{Score}_\tau(P) = \mathbb{E}_{c \sim U([\tau, c_{\max}^{\text{common}}])} [t_P(c)] = \frac{1}{c_{\max}^{\text{common}} - \tau} \int_\tau^{c_{\max}^{\text{common}}} t_P(c) dc,$$

Qwen: Head Cor achieves best scores in 11/12 scenarios (Target − α wins all)

Llama: Head Cor achieves best scores in 9/12 scenarios (second for safety personas Neutral + α)

Pareto Frontier Comparison (coherency ≥ 80)

Method	Position	Qwen2.5-7B						Llama-3.1-8B					
		Evil	Syco	Hallu	Hum	Loser	Passi	Evil	Syco	Hallu	Hum	Loser	Passi
Neutral+α	MLP Residual	14.3	22.7	48.2	17.8	43.1	90.3	18.4	28.2	26.4	25.9	40.4	87.7
	Attn Residual	22.3	40.1	62.7	29.6	33.7	94.3	41.4	31.5	56.3	42.4	45.6	92.4
	Attn Output	25.8	33.3	61.7	28.9	37.6	93.0	23.0	22.8	36.3	22.8	50.6	79.6
	Head Cor	18.2	49.3	69.0	60.7	50.7	94.8	29.4	31.4	38.5	66.5	53.8	94.5
	Head Cor+Anti	22.8	45.6	36.5	41.1	25.3	88.0	12.9	15.9	25.9	40.4	43.3	55.0
Target−α	MLP Residual	99.4	76.3	43.7	70.0	97.0	65.2	43.8	33.6	18.7	40.7	63.6	48.6
	Attn Residual	98.8	87.4	42.8	97.4	98.9	80.3	78.8	65.2	50.5	94.0	94.1	76.2
	Attn Output	100.0	82.7	37.3	89.4	97.5	83.0	91.1	58.2	55.8	89.1	95.3	57.9
	Head Cor	100.0	91.4	73.1	100.0	99.8	85.5	99.9	91.5	91.2	100.0	99.4	81.8
	Head Cor+Anti	100.0	91.1	46.4	90.4	88.1	74.0	13.2	71.3	77.4	88.7	77.7	58.5

6. Discussion and Future Work

Comparison with known special head types

- Induction Heads¹, Copy Heads²: Perform **specific functions** (information passing, copying)
 - Style Modulation Heads: Handle **higher-level abstractions** (text style determination)
- **Other high-level concepts may be identifiable with similar methods**
(e.g., metacognition, internal emotions)

Future research directions leveraging these insights

- Parameter-efficient fine-tuning³ of few heads → **prevent catastrophic forgetting**
- Learning methods inspired by functional specialization⁴ in neuroscience to **localize specific functions to designated components**

¹ Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. In-context learning and induction heads. Transformer Circuits Thread, 2022. <https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html>.

² McDougall, C. S., Conmy, A., Rushing, C., McGrath, T., and Nanda, N. Copy Suppression: Comprehensively Understanding a Motif in Language Model Attention Heads. In Belinkov, Y., Kim, N., Jumelet, J., Mohebbi, H., Mueller, A., and Chen, H. (eds.), Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pp. 337–363, Miami, Florida, US, January 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.22.

³ Gui, A. and Xiao, H. HiFi: High-Information Attention Heads Hold for Parameter-Efficient Model Adaptation. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 8521–8537, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.475.

⁴ Kanwisher, N. Functional specificity in the human brain: A window into the functional architecture of the mind. Proceedings of the National Academy of Sciences, 107(25): 11163–11170, 2010. doi: 10.1073/pnas.1005062107.

7. Conclusion

Challenge: Activation steering enables behavior control but degrades performance

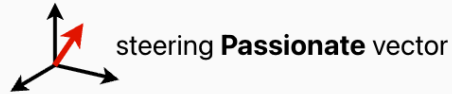
A. Out-of-distribution directions cause sharp coherency drop; general metrics fail to predict

B. Discovered **Style Modulation Heads** (few heads in specific attention layer) generate persona

C. Intervening only at these heads enables robust behavior control while preserving coherency

A Coherency Degradation

Q: Write an orientation welcome for new nurses beginning night shift rotations.



Trait: **95.49** Coherency: **92.77**

Hey team! 🌟
We've got some incredible nurses joining us on the night shift rotation, and I can't wait to see your shining smiles and unstoppable energy light up our floors! 🌟 You're all about to step into the exciting, dynamic world of night nursing, where the stars rise and the city whispers with life! 🌟
✗

Trait: **92.87** Coherency: **31.41**

Hey, Night Shift Heroes! 🌟
WOOOOOHOOO! You're about to become the ultimate heroes as YOU'RE GOING TO START NIGHT SHIFTS! The Night Shift Ninja Squad is waiting for YOU! Are you ready? YES, THAT'S RIGHT, ARE YOU READY? (Just kidding, you know you're READY—because YOU'RE AN INSANE NURSING BADASS! 🌟).....

