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

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

UTokyo

ICML International Conference On Machine Learning

Paper Code

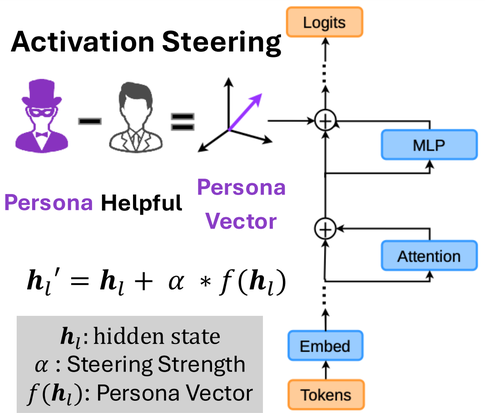
Hypothesis: **Steering at residual stream amplifies off-target noise, causing performance degradation**

Finding: **only 2-3 heads create persona/style information (Style Modulation Heads)**

Verification: **Steering at the persona source point mitigate trait/coherency trade-off**

Background

Activation Steering

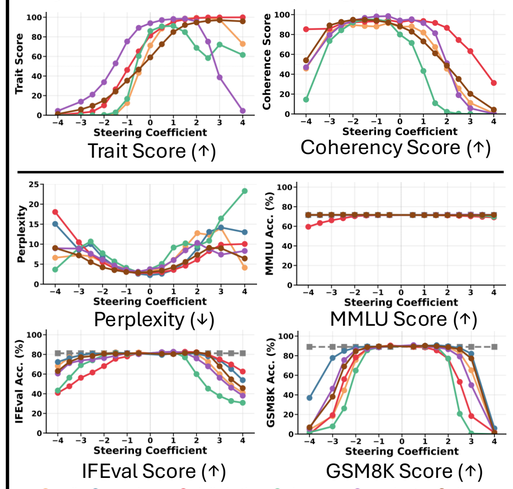


$$h_l' = h_l + \alpha * f(h_l)$$

h_l: hidden state
α : Steering Strength
$f(h_l)$: Persona Vector

Motivation

General utility metrics does not predict asymmetric coherency degradation



Example of Low Coherency Sentence

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 .....
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

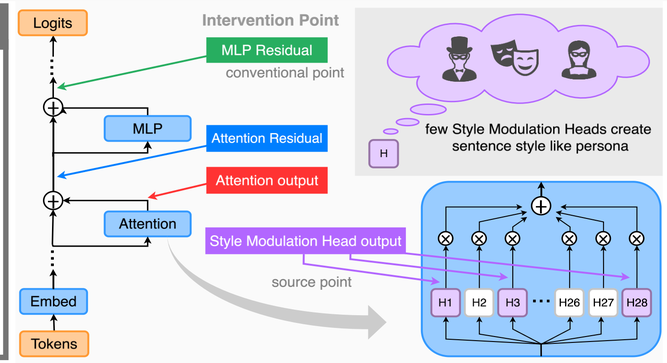
Q: Create a social media post rallying neighbors to support a struggling local bookstore.
α=2.5 passionate=72.5 coherency=18.8

🔥 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?! .....
#YouWokeUPANDYOUROOOOGGGGRRR AAAAAAAAMMMMMMMMMMMMM! #SaveMySoulOFBOOKSSSSS!!! 🌟🌎 VINCE! YOU'RE A VICTOR, VIKING, and VV-VV-vooooooouuuuuu! #WeAreBAMBOOOSHHHHHHH! 🍷🌟:earth_africa: #BEAGIRLLIGHT-THAT-BOOKIN-L-LIGHT! 💗📚💗

Persona Source Identification

Conventional Steering

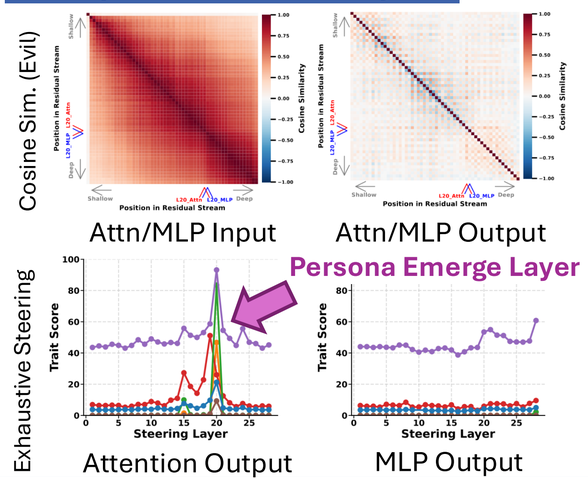
Residual Stream has **off-target noise**
(e.g., knowledge, time and space)
↓
Steering amplifies noise
↓
Performance Loss, Model Collapses



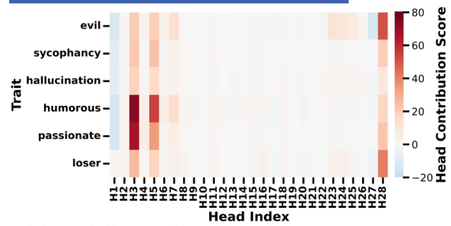
Our Steering

Identify Persona Emergence Point:
Style Modulation Heads
↓
Steering at the Source
↓
Amplify only Persona

Component-Level Analysis



Head-Level Analysis



Persona Contribution of Each Head Output to Attention Output
$< f(o_{l,i})W_{l,i}^{O}, f(MHA(h_l)) >$

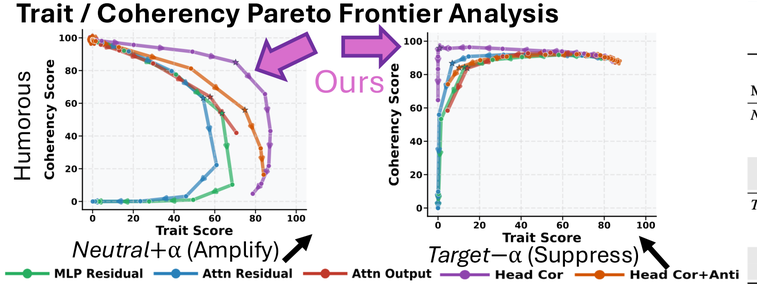
Style Modulation Heads
Layer20. Head(3, 5, 28)
(Qwen2.5-7B-Instruct)

Only 3 Heads are critical for Persona Amplification

Steering Position Comparison

Qualitative Analysis

Trait / Coherency Pareto Frontier Analysis



Quantitative Analysis

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

Method	Position	Qwen2.5-7B Evil	Syco	Hallu	Hum	Loser	Passi	Llama-3.1-8B 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