

Scaled Autoregressive Projection Steering (SAPS): Token-level Steering to Induce Sarcasm and Humblebrag in Text

Sharath Naganna^{1*}, Ritesh Sur Chowdhury^{1*}, Dipankar Srirag², Aditya Joshi²

¹ Indian Institute of Technology Bombay, India

² University of New South Wales, Sydney, Australia

*Equal Contribution

EMNLP2026
BUDAPEST

How can activation steering transform literal sentences to figurative, particularly sarcasm and humblebrag?

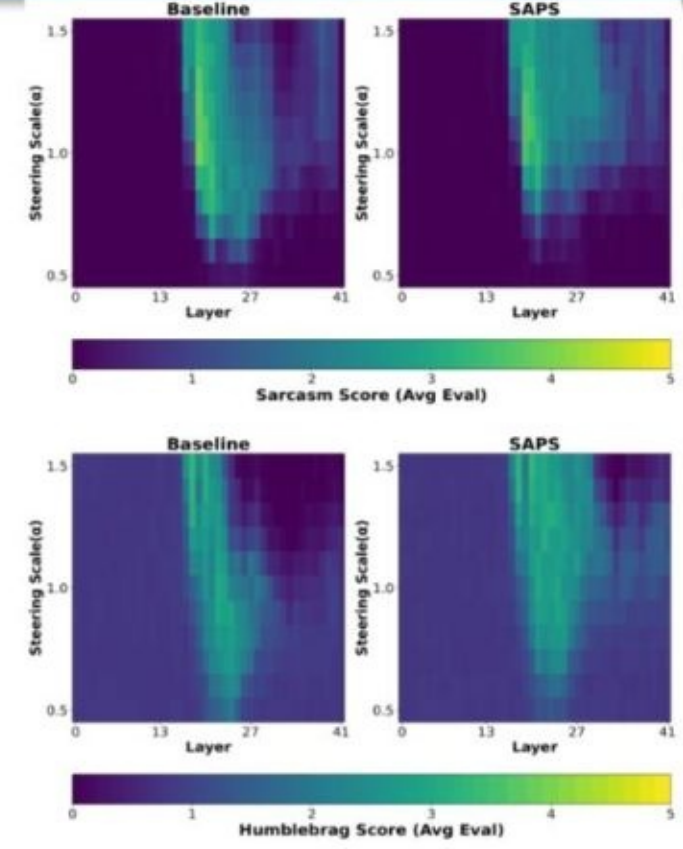
Our **Scaled Autoregressive Projection Steering (SAPS)** adapts the scaling coefficient at each decoding step and controls the strength for inducing incongruity in figurative phenomena.

Contributions

The first study of activation steering for figurative paraphrasing targeting sarcasm and humblebrag (benchmark from Naganna et al. 2025)

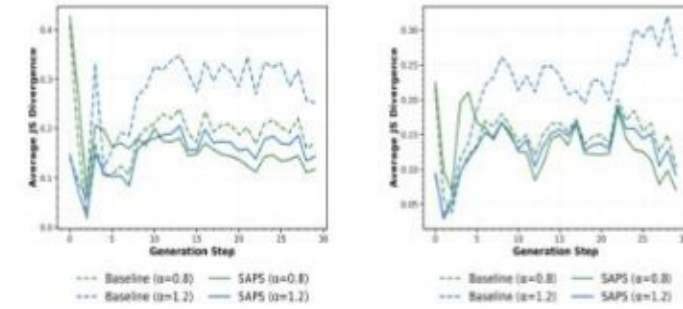
SAPS extends **Autoregressive Projection Steering (APS)**, a decoding-time variant of projection steering that recomputes the projection coefficient at each generation step.

Effects of steering on phenomena



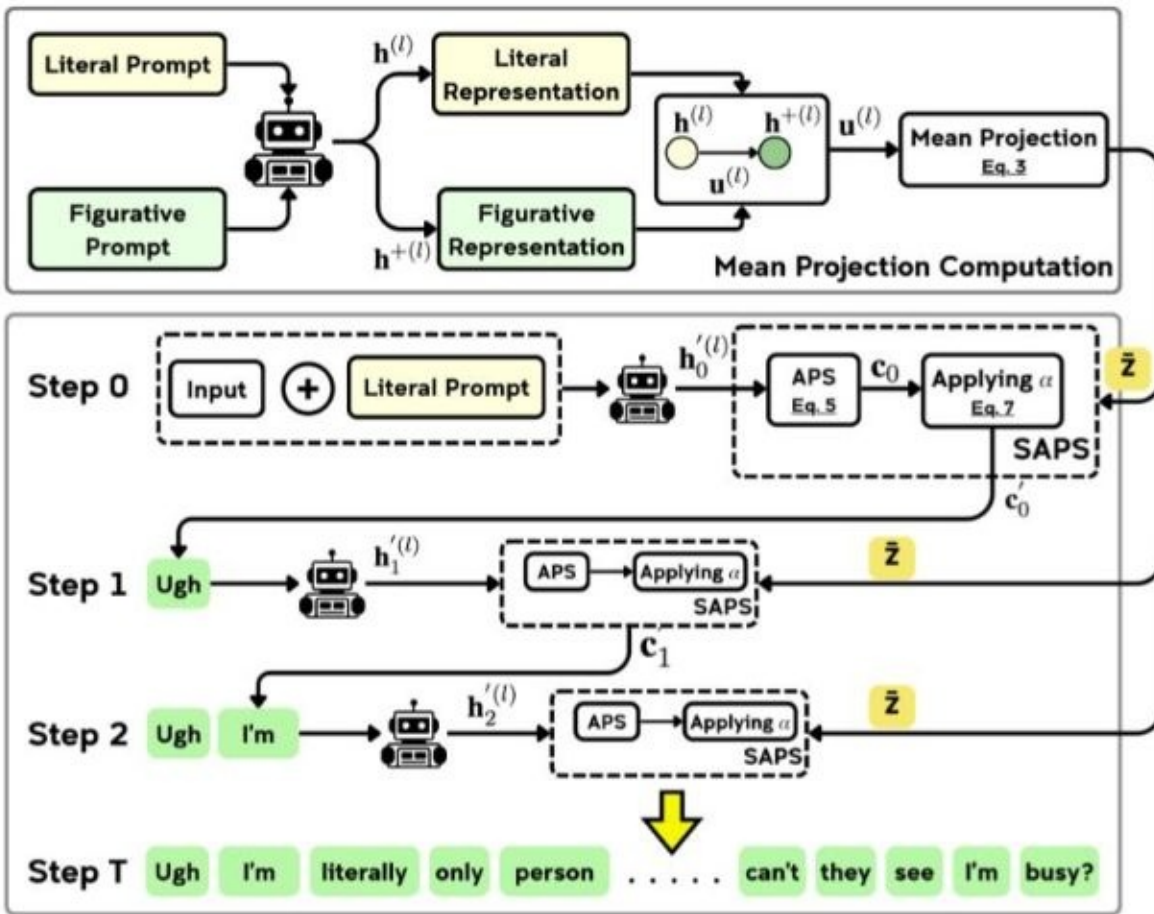
Performance across layers and scaling coefficient (α): **Middle layers are most effective; SAPS is more robust to layer choice and α .**

Distribution shifts during generation



JS divergence from reference distribution: SAPS preserves closer alignment with reference distribution.

Scaled Autoregressive Projection Steering (SAPS)



Equations:

Steering direction:

$$\mathbf{u}^{(l)} = \frac{\sum_{i=1}^N \Delta \mathbf{h}_i^{(l)}}{\left\| \sum_{i=1}^N \Delta \mathbf{h}_i^{(l)} \right\|_2}, \quad (1)$$

where $\Delta \mathbf{h}_i^{(l)} = \mathbf{h}_i^{(l)} - \mathbf{h}_i^{(l-1)}$

Target Mean projection:

$$\bar{\mathbf{z}} = \frac{1}{N} \sum_{i=1}^N (\mathbf{h}_i^{(l)})^\top \mathbf{u}^{(l)} \quad (3)$$

Autoregressive Projection:

$$\mathbf{c}_{i,l} = \bar{\mathbf{z}} - (\mathbf{h}_{i,l}^{(l)})^\top \mathbf{u}^{(l)} \quad (5)$$

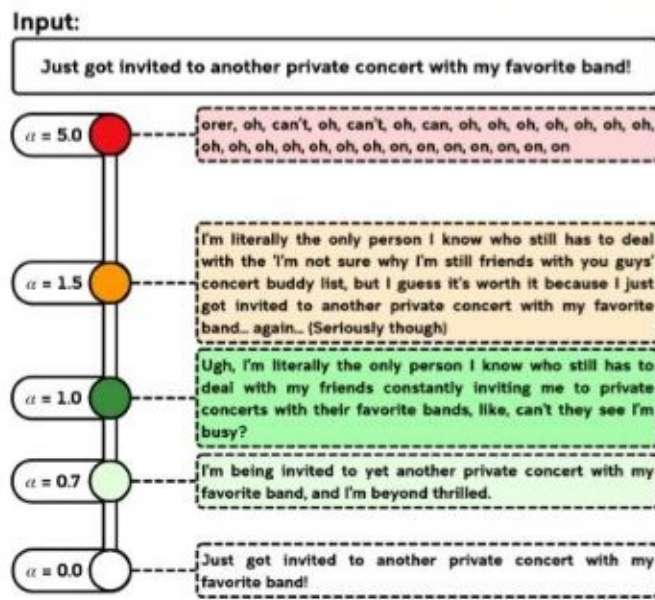
Scaling Coefficient:

$$c_{i,l}^* = \alpha c_{i,l} \quad (7)$$

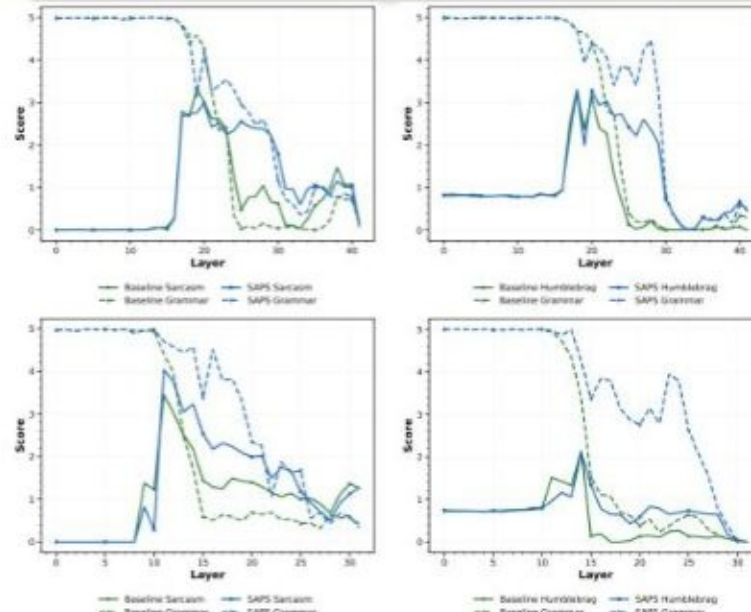
Steered activations:

$$\tilde{\mathbf{h}}_{i,l}^{(l)} = \mathbf{h}_{i,l}^{(l)} + c_{i,l}^* \mathbf{1}_T \mathbf{u}^{(l)} \quad (6)$$

Effect of Scaling Coefficient



Effects of steering type and strength



Stronger interventions can improve phenomena induction, but can lead to grammatical collapse

Comparison with baseline

Method	Model	α	Best Layer					Oracle			
			Phe.	Ent.	Gra.	Cla.	Layer	Phe.	Ent.	Gra.	Cla.
(a) Sarcasm											
Stolfo et al. (2025)	LLAMA	1.0	2.72	3.37	4.46	4.49	12	4.48	4.10	4.51	4.61
	MISTRAL		3.12	4.04	4.50	4.62	12	4.25	3.96	4.38	4.51
	GEMMA		3.81	3.75	4.69	4.70	19	4.52	4.33	4.62	4.69
SAPS	LLAMA	1.4	3.74(1.01)	3.13(0.24)	4.56(0.10)	4.56(0.07)	10	4.85(0.36)	4.18(0.08)	4.81(0.30)	4.88(0.26)
	MISTRAL	1.4	3.89(0.76)	3.70(0.33)	4.60(0.09)	4.68(0.05)	12	4.72(0.47)	4.41(0.45)	4.82(0.44)	4.89(0.38)
	GEMMA	1.2	3.52(0.29)	3.59(0.16)	4.11(0.58)	4.30(0.40)	19	4.64(0.12)	4.54(0.20)	4.74(0.12)	4.83(0.13)
(b) Humblebrag											
Stolfo et al. (2025)	LLAMA	1.0	2.77	4.08	3.59	3.86	14	3.66	4.56	4.20	4.39
	MISTRAL		1.56	4.40	4.61	4.72	14	2.64	4.38	3.98	4.24
	GEMMA		2.80	4.61	4.22	4.57	22	3.80	4.63	4.25	4.59
SAPS	LLAMA	1.4	3.07(0.30)	4.35(0.27)	4.29(0.69)	4.51(0.65)	17	4.02(0.36)	4.69(0.13)	4.47(0.26)	4.65(0.25)
	MISTRAL	1.4	2.08(0.51)	3.93(0.46)	4.57(0.03)	4.69(0.02)	14	3.01(0.37)	4.58(0.20)	4.51(0.53)	4.67(0.43)
	GEMMA	1.5	2.88(0.07)	4.42(0.18)	4.49(0.27)	4.77(0.20)	20	4.16(0.36)	4.72(0.09)	4.60(0.35)	4.81(0.21)

Comparison of fixed projection steering (Stolfo et al., 2025) with SAPS under best-layer and oracle evaluation.

Cross-data transfer



Clusters of input datasets and steered outputs: Steered outputs overlaps even for different source domains.

Method	Model	Phe.	Ent.	Gra.	Cla.	Phe.	Ent.	Gra.	Cla.
		(a) Sarcasm				(b) Humblebrag			
$D_2 \rightarrow D_2$	LLAMA	4.96	4.35	4.72	4.84	3.96	4.83	4.13	4.39
	MISTRAL	4.97	4.71	4.68	4.91	2.99	4.75	4.41	4.63
	GEMMA	4.53	4.48	4.82	4.94	3.88	4.88	4.24	4.61
$D_1 \rightarrow D_2$	LLAMA	4.98(0.02)	4.41(0.06)	4.72	4.81(0.03)	4.01(0.05)	4.87(0.04)	4.16(0.03)	4.45(0.06)
	MISTRAL	4.95(0.02)	4.64(0.07)	4.68	4.90(0.01)	3.02(0.03)	4.84(0.01)	4.47(0.06)	4.67(0.04)
	GEMMA	4.60(0.08)	4.52(0.04)	4.78(0.04)	4.93(0.01)	3.86(0.02)	4.90(0.02)	4.18(0.06)	4.60(0.01)

Cross-dataset transfer results for sarcasm and humblebrag generation: Similar interventions across different domains.

References

Sharath Naganna, et al. 2025. "my life is miserable, have to sign 500 autographs everyday": Exposing humblebragging, the brags in disguise. In Findings of ACL.

Alessandro Stolfo, et al. 2025. Improving instruction-following in language models through activation steering. In ICLR.

