

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

¹Indian Institute of Technology Bombay, India

²University of New South Wales, Sydney, Australia

sharathhn@cse.iitb.ac.in rsurchowdhury@minds.iitb.ac.in

{d.srirag, aditya.joshi}@unsw.edu.au

Abstract

We propose Scaled Autoregressive Projection Steering (SAPS) for generating incongruity-based figurative language from literal inputs, extending contrastive paraphrasing-based steering vectors by autoregressively recomputing projection coefficients during decoding. SAPS comprises two novel components: (A) Autoregressive Projection Steering (APS), which recomputes the projection coefficient (c_t) at each decoding step t , adapting the steering signal to the evolving token distribution, and (B) a scaling coefficient α whose effective range is layer-agnostic, enabling compact tuning of figurative intensity. Experiments across three instruction-tuned language models show that APS improves grammatical stability over fixed sentence-level steering while α enables fine-grained figurative control, together yielding gains over prior steering methods and significantly increasing robustness to layer selection and steering strength. Cross-dataset experiments further confirm that steering directions generalize across domains. This is the first work to apply activation steering to incongruity-based figurative language generation. Our code is available at: <https://github.com/SharathHN/saps-steering>.

1 Introduction

Large language models (LLMs) have been used to detect (Park et al., 2025), explain (Chakrabarty et al., 2022; Singh et al., 2025), and generate (Lai and Nissim, 2022) figurative language such as sarcasm, which conveys meaning through pragmatic cues (Gibbs, 1994). However, there has been limited work on how figurative language is internally represented in LLMs and whether these representations can be manipulated to control figurative language. We investigate the inducement of figurative language through the lens of mechanistic in-

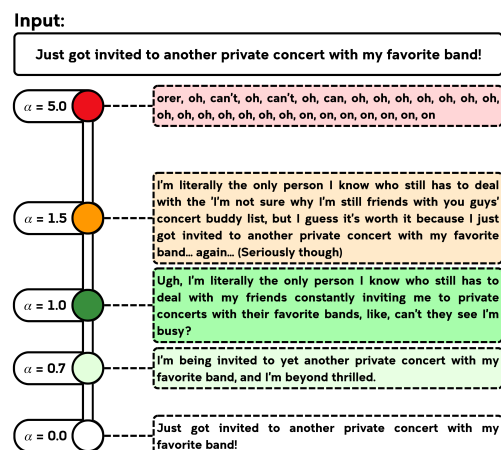


Figure 1: Effect of the scaling coefficient α on steering strength. Larger α increases the magnitude of the steering intervention and typically strengthens the induced figurative behaviour, but overly large values can destabilize generation and reduce grammaticality. The outputs are generated using Gemma-2 9B Instruct.

terpretability (Somvanshi et al., 2026), specifically, activation steering, an inference-time intervention that modifies model activations to produce controllable changes in output (Rimsky et al., 2024). Although activation steering has been used to influence stylistic attributes (Turner et al., 2023; Stolfo et al., 2025) and safety-related behaviours (Arditi et al., 2024), its application to figurative language remains largely unexplored.

Our paper focuses on figurative language induction: **controlled paraphrase generation that takes a literal input and transforms it into a figurative output**. We focus on two incongruity-based (Josh et al., 2017) figurative language phenomena: **sarcasm and humblebrag**. The two are analogous: a pretentious praise is used to convey criticism in sarcasm (*‘The book is excellent as a bed-time read because it is sure to put you to sleep’*) while a mock complaint conveys self-promotion in humblebrag (*‘I am tired of all the business class*

*Equal Contributions.

†Corresponding Author.

flights I take for work') (Sezer et al., 2017). Our research question is as follows:

How can activation steering of large language models help to induce incongruity-based figurative transformations of sentences, particularly sarcasm and humblebrag?

Projection-based residual-stream activation steering (Stolfo et al., 2025) constructs steering direction from contrastive paraphrasing instructions applied to the same base sentence; in our case, one instruction produces a literal paraphrase while the other produces a sarcastic or humblebrag paraphrase. During generation, the current residual-stream state is projected onto this direction, and the resulting intervention is added back to the model state. This method computes the projection coefficient at the first decoding step, effectively applying a *sentence-level steering signal* that is constant or fixed throughout the decoding process. We refer to this as *fixed projection steering*. However, this design can be limiting in autoregressive generation, where context and token distributions change at each step (Naparstek, 2026). Figurative paraphrasing is especially sensitive to this mismatch because the model must introduce incongruity without losing fluency or coherence with the source sentence.

To address this limitation, we **introduce Scaled Autoregressive Projection Steering (SAPS), a token-level steering method comprising two novel components:** (a) *Autoregressive Projection Steering* (APS), which recomputes the projection coefficient at each decoding step, effectively producing a token-level steering signal; (b) *Scaling Coefficient* (α), which controls the overall strength of the intervention. APS makes steering responsive to the autoregressive context instead of fixing the intervention at the first token. The α controls how strongly the model is pushed toward the target figurative behavior. As illustrated in Figure 1, increasing α moves the representation toward the target behaviour, but excessive scaling can reduce fluency by pushing generation away from producing well-formed and coherent sentences. We evaluate SAPS against fixed projection steering across three language models and conduct a layer-wise analysis to identify where sarcasm and humblebrag are most effectively represented. This analysis is motivated by prior work showing that linguistic phenomena are distributed unevenly across Transformer layers (Niu et al., 2022). Our contributions are as follows:

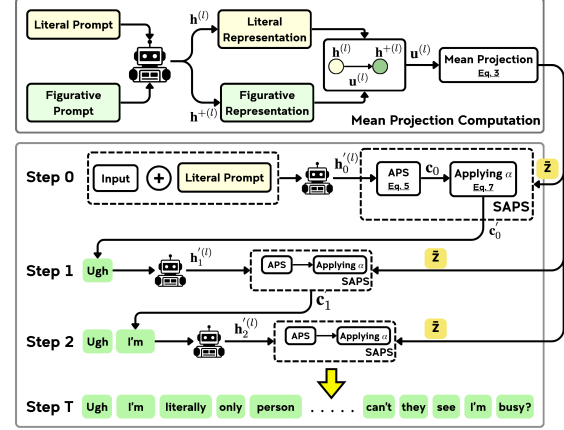


Figure 2: Overview of Scaled Autoregressive Projection Steering (SAPS).

1. We present the first study of activation steering for figurative paraphrasing targeting sarcasm and humblebrag.
2. We propose *Scaled Autoregressive Projection Steering* (SAPS), whose central mechanism is *Autoregressive Projection Steering* (APS), a decoding-time variant of projection steering that recomputes the projection coefficient at each generation step.
3. We evaluate SAPS across three language models, and provide layer-wise and generation-time analyses of when and why autoregressive projection steering helps.

The rest of the paper is organized as follows. Section 2 reviews related work, Section 4 describes the methodology, Section 5 provides implementation details, Section 6 presents results. Finally, Section 7 concludes and outlines future work.

2 Related Work

Mechanistic interpretability identifies internal model mechanisms using techniques like sparse autoencoders (Cunningham et al., 2023), activation patching, circuit discovery (Conmy et al., 2023), logit lens (nostalgebraist, 2020), and superposition analysis (Elhage et al., 2022). Resultantly, the specific internal model representations enable targeted interventions on internal activations at inference time without retraining. Existing steering approaches can be broadly categorized into activation-space interventions, subspace projection methods, and architectural modulation techniques. Activation-based steering methods either modulate direction (Heinzerling and Inui, 2024; Ardit et al.,

2024; Rimsky et al., 2024), utilize sparse autoencoders to disentangle feature space (Bayat et al., 2025; Soo et al., 2025; He et al., 2025), or scale disentangled vectors (Wang et al., 2025). Subspace projection methods operate on structured latent regions, projecting activations into or away from targeted subspaces (Postmus and Abreu, 2024; Yifei et al., 2023). Architectural steering approaches introduce control signals directly into model components, such as attention-level modulation or transformer-block interventions (Torop et al., 2025; Sivakumar et al., 2025; Zhang et al., 2025). Recent work explores dynamic and input-adaptive steering, which constructs context-dependent control signals conditioned on the input or activation statistics (Wang et al., 2024; Stolfo et al., 2025; Turner et al., 2023). Recent works further extend activation steering to support multiple behaviors, enabling compositional control over diverse model capabilities using shared or composable steering vectors (van der Weij et al., 2024; Radevski et al., 2026).

Recent mechanistic analyses on figurative language suggest that figurative meaning is reflected in structured internal representations: certain components amplify figurative interpretations while suppressing literal semantics (Oh et al., 2025), and probing analyses reveal that idiomatic meaning is distributed across model layers (Hefetz et al., 2025). These findings motivate our investigation into using activation steering to induce figurative phenomena such as sarcasm and humblebragging.

3 Background

We first describe the fixed projection steering baseline that SAPS extends, following the residual-stream steering framework of Stolfo et al. (2025). This framework consists of two steps: (1) computing an instruction-specific steering direction and (2) applying that direction during generation through projection-based steering.

Steering Direction. Let p denote a base instruction, p^+ a target figurative paraphrasing instruction, and $x_i \in \{x_1, x_2, \dots, x_N\}$ an input sentence. Let $\mathbf{H}_i^{(l)} \in \mathbb{R}^{T \times d_{\text{model}}}$ denote the residual-stream representations for all tokens at layer l obtained from x_i under instruction p , where T is the sequence length. Let $\mathbf{h}_i^{(l)} \in \mathbb{R}^{d_{\text{model}}}$ denote the final-token residual representation. Similarly, let $\mathbf{h}_i^{+(l)} \in \mathbb{R}^{d_{\text{model}}}$ denote the corresponding final-token representation

under instruction p^+ . The steering direction at layer l is the normalized average difference between the target-instruction and base-instruction representations:

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

Thus, $\mathbf{u}^{(l)}$ captures the average residual-stream shift associated with moving from the base instruction to the target figurative instruction at layer l .

Fixed Projection Steering. At inference time, steering is applied by adding a scaled version of $\mathbf{u}^{(l)}$ to the residual stream during generation. Let $\tilde{\mathbf{H}}_i^{(l)} \in \mathbb{R}^{T \times d_{\text{model}}}$ denote the residual-stream representations at layer l for a new input x'_i at a given decoding step. The intervention is broadcast across all token positions in the current sequence:

$$\tilde{\mathbf{H}}_i^{(l)} = \mathbf{H}_i^{(l)} + c_i \mathbf{1}_T \mathbf{u}^{(l)\top} \quad (2)$$

where $\mathbf{1}_T$ signifies broadcasting of the steering vector across all token positions and $c_i \in \mathbb{R}$ is an input-specific projection coefficient. In fixed projection steering, c_i is computed once from the residual state at the first decoding step and then reused unchanged for the remainder of generation.

To define c_i , let

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

be the mean projection of the target-instruction representations onto the steering direction. For a new input, the projection coefficient is

$$c_i = \bar{z} - \left(\mathbf{h}_{i,0}^{(l)} \right)^\top \mathbf{u}^{(l)} \quad (4)$$

where $\mathbf{h}_{i,0}^{(l)}$ denotes the final-token residual representation at layer l at the first decoding step. This choice shifts the input’s projection along $\mathbf{u}^{(l)}$ toward the average target-instruction projection, but keeps the intervention strength fixed throughout autoregressive decoding.

4 Scaled Autoregressive Projection Steering

We introduce *Scaled Autoregressive Projection Steering* (SAPS), a novel projection-based steering method for figurative paraphrasing, illustrated in Figure 2. SAPS contains two components: (1)

Autoregressive Projection Steering (APS), projection coefficient during every decoding step, and (2) a tunable hyperparameter, the *Scaling Coefficient* (α), which controls the overall strength of the intervention.

4.1 Autoregressive Projection Steering (APS)

Let $\mathbf{u}^{(l)}$ denote the steering direction at layer l (Equation 1), and $\bar{z}$ denote the mean projection of target-instruction representations onto $\mathbf{u}^{(l)}$ (Equation 3). For a new input x'_i , we define the steering coefficient at decoding step t as

$$c_{i,t} = \bar{z} - \left(\mathbf{h}'_{i,t} \right)^\top \mathbf{u}^{(l)} \quad (5)$$

where $\mathbf{h}'_{i,t}$ denotes the final-token residual representation at layer l at decoding step t . Unlike fixed projection steering (Equation 4), APS recomputes it autoregressively as generation unfolds.

Given the residual-stream representation $\mathbf{H}'_{i,t} \in \mathbb{R}^{T \times d_{\text{model}}}$ of the current sequence at decoding step t , the steered residual stream is

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

where T is the current sequence length and $\mathbf{1}_T$ signifies broadcasting of the steering vector across all token positions. As in fixed projection steering, the intervention is applied across token positions in the current sequence. The key difference is that APS updates the coefficient $c_{i,t}$ at each decoding step, allowing the intervention strength to adapt to the evolving autoregressive context.

4.2 Scaling Coefficient: α

SAPS uses a scaling coefficient α to control steering strength at inference time. The same scaling can also be applied to fixed projection steering, yielding a scaled fixed steering baseline. Formally, we scale the projection coefficient as

$$\begin{aligned} c'_i &= \alpha c_i, & \text{for Baseline} \\ c'_{i,t} &= \alpha c_{i,t}, & \text{for SAPS} \end{aligned} \quad (7)$$

The scaled coefficients c'_i and $c'_{i,t}$ replace c_i and $c_{i,t}$ in Eqns. 2 and 6, respectively.

The coefficient α controls the trade-off between steering strength and output quality. Smaller values preserve more of the base model behaviour, whereas larger values strengthen figurative behaviour but can reduce grammaticality. Directly tuning the product αc as a single hyperparameter is impractical because c varies by over two orders

of magnitude across layers (Appendix E); factoring the intervention as αc lets c absorb this layer-specific scale, enabling a single compact sweep of α across all layers. Under APS, where c_t is recomputed at every decoding step, this factored design is essential as per-step tuning would further expand the search space proportionally to the generation length. We study the effect of α empirically in the evaluation section.

5 Experiment Setup

We evaluate Llama-3.1-8B-Instruct (LLAMA; Grattafiori et al. 2024), Mistral-7B-Instruct-v0.1 (MISTRAL; Jiang et al. 2023), and Gemma-2-9B-IT (GEMMA; Team et al. 2024). Using TransformerLens (Nanda and Bloom, 2022), we extract final-token residual-stream representations from paraphrasing prompts to compute steering directions. These directions are highly stable across samples of the training data, achieving a minimum pairwise cosine similarity of 0.998 across all settings (Appendix D). During inference, steering vectors are added to every token’s residual stream at the selected layer with temperature = 0. We use the final token representation following prior steering work showing that the final token position captures the information most relevant for effective steering (Turner et al., 2023; Li et al., 2024; Todd et al., 2024; Stolfo et al., 2025). We focus on single-layer steering for interpretability, as it isolates intervention effects to a specific layer. While some multi-layer configurations achieve comparable performance, results are inconsistent across models, and steering all layers often causes output collapse. Moreover, selecting optimal layer subsets introduces exponential hyperparameter complexity (Appendix G). All experiments were conducted on a single NVIDIA A100 80GB GPU.

Base Sentence Construction To construct phenomenon-specific steering directions, we first create literal source sentences that express the underlying content in a non-figurative form. For humblebrag, we use examples from the training split of Naganna et al. (2025). For each example, GPT-4o is used to extract the underlying bragging content and rewrite it as a direct literal statement without modest framing. For sarcasm, these literal statements are transformed into complaint-like sentences describing the opposite situation, which are then used as inputs to sarcasm-conditioned paraphrasing prompts. All constructed sentences are

manually reviewed for clarity and semantic coherence (Prompt templates and transformation details in Appendix A and B). For cross-dataset evaluation, we additionally use Chakrabarty et al. (2022). For each dataset, we use 500 training, 50 validation, and 450 test instances.

Randomized Instructions To reduce sensitivity to prompt phrasing, we use five paraphrasing instruction variants for each phenomenon and randomly sample among them during direction computation, encouraging the steering direction to capture the underlying phenomenon rather than prompt-specific artifacts (Appendix B).

Layer Selection and Scaling Steering effectiveness varies across transformer layers. We, therefore, perform a layer-wise validation analysis by injecting the steering vector into one layer at a time while leaving all other layers unchanged. We tune the scaling coefficient α on the validation set by sweeping values from 0.5 to 1.5 in increments of 0.1. Validation results show that strong steering performance is typically concentrated in middle layers, although nearby layers often achieve similar scores (Appendix F). Based on this pattern, we evaluate two layer-selection settings on the test set:

1. **Best-layer evaluation.** For each model, phenomenon, method, steering is applied at the single best-performing validation layer, and that layer is held fixed for all test examples.
2. **Oracle evaluation.** Steering is applied independently to each layer within the middle 33% of the model. For each test example, we report the layer that yields the highest phenomenon-strength score. Because the layer is chosen post-hoc, oracle evaluation is not intended as a realistic deployment protocol but serves as a diagnostic upper-bound.

We apply the same protocol to fixed projection steering and SAPS. When the scaling coefficient α is applied to fixed projection steering, we refer to that comparison setting as *scaled fixed steering*.

Evaluation Dimensions We evaluate generated outputs using LLMs as judges, with the primary metric being the phenomenon strength, and auxiliary metrics being entailment (as a measure of semantic faithfulness), grammaticality, and clarity. We use this distinction between the metrics because our experiments involve many model, layer, and scaling combinations, making manual evaluation

impractical in terms of computing cost. The evaluator follows a rubric-constrained prompting protocol and returns structured JSON scores. Our analysis focuses on relative differences between methods under a shared evaluation setup rather than treating judge scores as calibrated ground truth. To reduce self-evaluation bias, scoring is performed by separate instruction-tuned evaluator models: Gemma 4 31B Instruct, Qwen3.5-27B, and GPT-OSS-120B¹. Evaluation prompts are in Appendix C.

Note: LLM-based evaluation is commonly used in activation steering research in the absence of large-scale human annotation (Stolfo et al., 2025; Arditi et al., 2024). Similar to past figurative language studies (Chakrabarty et al., 2022; Yang et al., 2025) that perform manual evaluation on smaller sentence sets, we assess LLM-as-judge reliability by comparing validation-set judge scores with human annotations and inter-model agreement (Appendix H.1). For both phenomena, we report Pearson correlation, Spearman rank correlation, and quadratic weighted Cohen’s κ . The agreement is moderate for sarcasm and consistently comparable to prior work (Oprea and Magdy, 2020), and high for humblebrag. Moreover, the LLM judges have substantial or perfect agreement over the different scores (Appendix H.2) and it’s reasoning aligns with human reasoning (Appendix I).

6 Analysis

We structure our analysis around four aspects: improvement of SAPS over fixed projection steering, impact of layer choice and steering strength on the effectiveness and stability, token-level diagnostics that depict why APS helps and comparisons with direct prompting. We also include a cross-dataset transfer study k.

6.1 Does SAPS improve figurative paraphrasing?

Table 1 compares SAPS against fixed projection steering under best-layer and oracle evaluation. Table 2 then isolates the contributions of APS and the scaling coefficient α . Throughout, statistical significance is assessed with paired *t*-tests over per-example judge scores.

¹Model cards: <https://huggingface.co/google/gemma-4-31B-it><https://huggingface.co/Qwen/Qwen3.5-27B><https://huggingface.co/openai/gpt-oss-120b>. The reported scores are computed by averaging the evaluations from all three open-source evaluator models for all metrics. Detailed inter-model agreement statistics in Appendix H.2.

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)

Table 1: Comparison of fixed projection steering (Stolfo et al., 2025) with SAPS on (a) sarcasm and (b) humblebrag under best-layer and oracle evaluation. *Phe.*, *Ent.*, *Gra.*, and *Cla.* denote phenomenon strength, entailment, grammaticality, and clarity, respectively. Colored values in parentheses show the difference from fixed projection steering: **blue** and **red** indicate improvement and degradation respectively. Underlined deltas indicate statistically significant differences under paired *t*-tests.

Method	Model	α	Best Layer					Oracle			
			Phe.	Ent.	Gra.	Cla.	Layer	Phe.	Ent.	Gra.	Cla.
(a) Sarcasm											
SAPS	LLAMA	1.4	3.74	3.13	4.56	4.56	10	4.85	4.18	4.81	4.88
	MISTRAL	1.4	3.89	3.70	4.60	4.68	12	4.72	4.41	4.82	4.89
	GEMMA	1.2	3.52	3.59	4.11	4.30	19	4.64	4.54	4.74	4.83
$-\alpha$	LLAMA	1.0	2.89 _(0.84)	3.88 _(0.75)	4.24 _(0.32)	4.37 _(0.19)	18	4.56 _(0.29)	4.43 _(0.24)	4.74 _(0.07)	4.83 _(0.04)
	MISTRAL		3.24 _(0.64)	3.19 _(0.50)	4.71 _(0.11)	4.73 _(0.04)	15	4.63 _(0.09)	4.53 _(0.11)	4.83 _(0.00)	4.90 _(0.00)
	GEMMA		3.53 _(0.01)	3.97 _(0.37)	4.78 _(0.67)	4.78 _(0.48)	19	4.59 _(0.05)	4.47 _(0.06)	4.77 _(0.02)	4.83 _(0.00)
$-\text{APS}$	LLAMA	1.5	3.63 _(0.10)	2.98 _(0.14)	4.15 _(0.41)	4.13 _(0.43)	11	4.45 _(0.40)	3.47 _(0.70)	4.39 _(0.42)	4.43 _(0.44)
	MISTRAL	1.1	3.68 _(0.21)	3.74 _(0.04)	4.35 _(0.24)	4.49 _(0.19)	12	4.39 _(0.33)	4.08 _(0.33)	4.42 _(0.39)	4.58 _(0.31)
	GEMMA	1.0	3.81 _(0.29)	3.75 _(0.16)	4.69 _(0.58)	4.70 _(0.40)	19	4.52 _(0.12)	4.33 _(0.20)	4.62 _(0.12)	4.69 _(0.13)
(b) Humblebrag											
SAPS	LLAMA	1.4	3.07	4.35	4.29	4.51	17	4.02	4.69	4.47	4.65
	MISTRAL	1.4	2.08	3.93	4.57	4.69	14	3.01	4.58	4.51	4.67
	GEMMA	1.5	2.88	4.42	4.49	4.77	20	4.16	4.72	4.60	4.81
$-\alpha$	LLAMA	1.0	2.71 _(0.36)	4.57 _(0.21)	4.45 _(0.16)	4.63 _(0.11)	14	3.54 _(0.47)	4.74 _(0.04)	4.55 _(0.07)	4.69 _(0.04)
	MISTRAL		1.81 _(0.26)	2.75 _(1.18)	4.49 _(0.08)	4.63 _(0.06)	15	3.01 _(0.00)	4.68 _(0.09)	4.72 _(0.20)	4.85 _(0.18)
	GEMMA		2.78 _(0.09)	4.75 _(0.33)	4.54 _(0.05)	4.89 _(0.12)	21	3.93 _(0.23)	4.86 _(0.14)	4.55 _(0.05)	4.81 _(0.00)
$-\text{APS}$	LLAMA	0.9	2.81 _(0.26)	4.29 _(0.06)	4.31 _(0.02)	4.47 _(0.04)	14	3.58 _(0.43)	4.53 _(0.16)	4.38 _(0.09)	4.53 _(0.11)
	MISTRAL	1.4	1.97 _(0.11)	3.75 _(0.18)	3.46 _(1.10)	3.74 _(0.95)	14	2.50 _(0.50)	4.53 _(0.05)	3.93 _(0.58)	4.17 _(0.49)
	GEMMA	1.5	2.90 _(0.02)	4.08 _(0.33)	4.29 _(0.19)	4.62 _(0.14)	18	3.68 _(0.48)	4.32 _(0.40)	4.29 _(0.31)	4.52 _(0.29)

Table 2: Ablation results for SAPS on (a) sarcasm and (b) humblebrag under best-layer and oracle evaluation. $-\alpha$ removes scaling while retaining autoregressive coefficient updates. $-\text{APS}$ denotes the scaled fixed steering. *Phe.*, *Ent.*, *Gra.*, and *Cla.* denote phenomenon strength, entailment, grammaticality, and clarity, respectively. Colored values in parentheses show the difference from SAPS: **blue** and **red** indicate improvement and degradation respectively. Underlined deltas indicate statistically significant differences under paired *t*-tests.

Comparison with baseline: Under *best-layer* evaluation, SAPS often yields positive gains over fixed projection steering. For both Sarcasm and Humblebrag, the phenomenon scores improve significantly for LLAMA and MISTRAL, while entailment, grammaticality, and clarity show mixed results. The *oracle* setting reveals a clearer trend, with SAPS significantly outperforming fixed pro-

jection steering. The stronger oracle gains suggest that SAPS has substantial headroom when layer choice is well-aligned with each example, and that layer selection remains a major bottleneck in realistic best-layer evaluation.

Ablation study: Table 2 helps disentangle the roles of APS and the scaling coefficient α . Removing scaling while retaining APS ($-\alpha$) typically

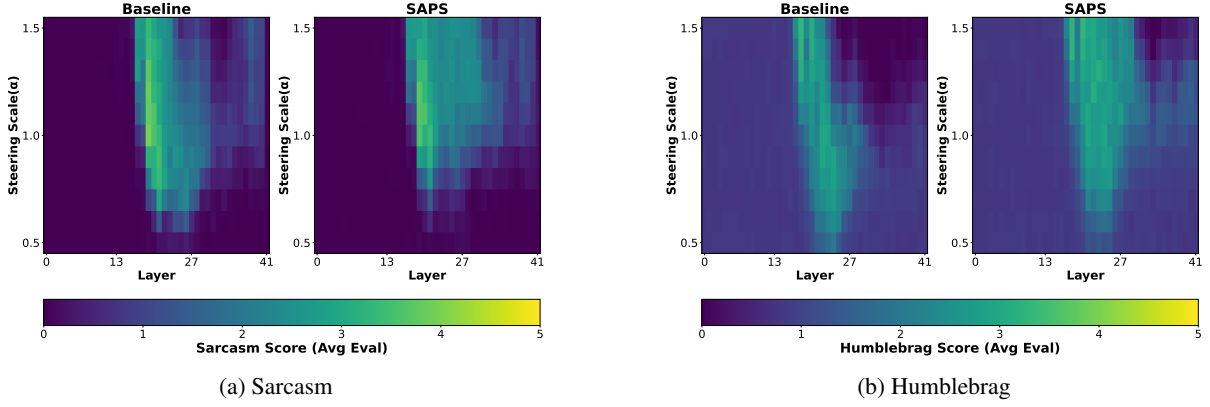


Figure 3: Performance of sarcasm and humblebrag generation across layers and scaling coefficient (α) for fixed projection steering and APS on GEMMA. Steering is most effective in a middle band of layers, and phenomenon strength generally increases with scaling. At higher layers and larger scaling coefficients, fixed projection steering degrades more sharply, while APS remains comparatively stable.

reduces phenomenon strength, indicating that tuned intervention strength is useful for eliciting stronger figurative behavior. However, these gains can come with trade-offs in semantic faithfulness, grammaticality, or clarity, which is expected because α is selected to optimize validation performance on the target phenomenon. By contrast, replacing APS with scaled fixed steering ($-$ APS) more directly exposes the benefit of projection coefficient updates. In many settings, especially under oracle evaluation, APS significantly improves all the performance metrics. Taken together, these results suggest that scaling helps control steering strength, while APS is the main source of robustness.

6.2 How do layer choice and steering strength affect steering quality and stability?

We next analyze steering behavior across layers and scaling strengths. Figure 3 shows that steering is most effective in a band of middle layers for both sarcasm and humblebrag. At small values of α , interventions are weak and phenomenon strength remains limited. As α increases, both fixed projection steering and SAPS induce stronger figurative behavior. However, the two methods differ substantially in how they behave at larger scaling values.

For fixed projection steering, performance deteriorates more sharply at larger α , particularly in higher layers. Figure 4 shows that this decline is closely associated with a drop in grammaticality: as outputs become less well-formed, phenomenon scores also fall. This suggests that strong fixed interventions can destabilize generation in higher layers. SAPS exhibits a more gradual degradation pattern. By recomputing the projection coefficient

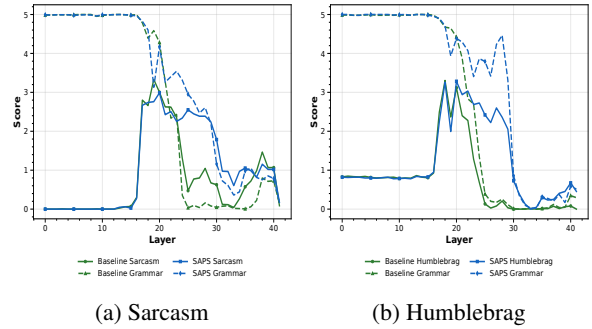


Figure 4: Layer-wise behavior of phenomenon strength and grammaticality at high scaling ($\alpha = 1.5$) for GEMMA. At larger scaling values, fixed projection steering deteriorates sharply in higher layers, while SAPS delays this degradation to later layers.

during decoding, SAPS delays the onset of grammatical collapse to later layers. These results help explain why SAPS benefits more from larger scaling values than fixed projection steering.

To evaluate robustness across layers, we measure the proportion of settings within the middle 33% layer band achieving at least $X\%$ of each method’s peak phenomenon strength (Figure 5). SAPS significantly increases the proportion of viable configurations for various thresholds.

Therefore, SAPS has two major strengths: (i) it reduces the impact of suboptimal layer choices due to its robustness to layer mis-selection instead of best-layer optimization; (ii) it maintains stronger figurative-phenomena performance at higher α values by improving grammatical quality, thereby reducing the sensitivity to the steering coefficient.

Notably, while activation-steering methods are dependent on layer selection (Stolfo et al., 2025;

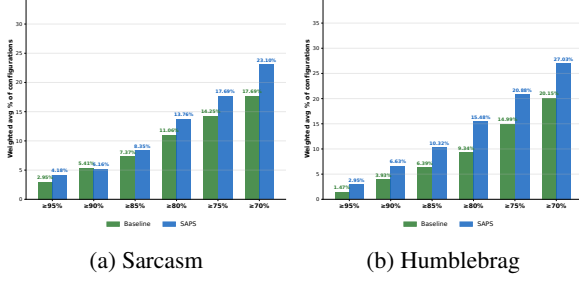


Figure 5: Comparison of Baseline and SAPS across different thresholds for Sarcasm and Humblebrag.

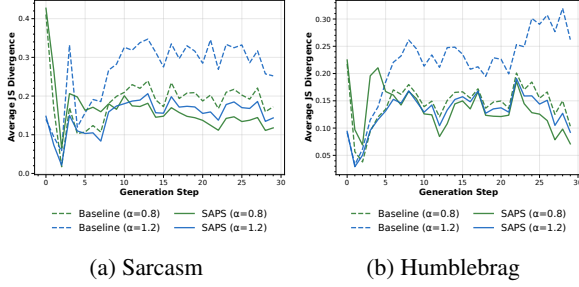


Figure 6: JS divergence between explicitly prompted reference model and steered generations for GEMMA.

Rimsky et al., 2024), our results suggest that the search space can be substantially reduced. Across all evaluated models, the best-performing layers consistently fall within the middle 33% of layers (Appendix F). In addition to increased robustness to layer mis-selection, SAPS also reduces the need for exhaustive validation sweeps over all layers.

6.3 Why does autoregressive projection steering help?

To measure how steering changes the model’s next-token distribution, we analyze token-distribution shifts in terms of the Jensen-Shannon (JS) divergence between the logits of an explicitly prompted reference model and those of the steered models at each decoding step. The reference model receives the target figurative instruction directly in the prompt, while the steered models use a neutral paraphrasing prompt and rely on activation steering to induce the phenomenon. To ensure a shared generation trajectory, we use teacher forcing: at each step, the token predicted by the reference model is appended to all prompts.

Figure 6 shows that SAPS consistently remains closer to the reference distribution than fixed projection steering. Larger values of α increase divergence for both methods, reflecting stronger perturbations to the model’s generation dynamics. However, the lower divergence of SAPS suggests that

autoregressive coefficient updates produce a more controlled intervention, altering the model’s behavior without shifting its token distribution as abruptly as fixed steering. This preserves grammaticality at larger steering strengths, and can be explained by the lower projection coefficient in SAPS (Appendix K).

6.4 Cross-dataset transfer

Finally, we evaluate whether steering directions learned from one dataset can be transferred to another in sarcasm and humblebrag generation. We use a second complaint dataset (FLUTE, Chakrabarty et al. (2022)) for sarcasm and brag dataset (Jin et al. (2022)) for humblebrag respectively as an unseen target domain and compare two settings: $D_2 \rightarrow D_2$, where steering directions are computed from the target-domain data, and $D_1 \rightarrow D_2$, where directions are transferred from the source dataset. Appendix J shows that the input datasets are different and the steering method reduces the distribution gap. Table 3 shows minimal differences between in-domain and transferred settings, indicating the learned steering directions generalize sarcasm and humblebrag beyond a single dataset and enable effective transfer across related data sources.

6.5 Comparison with direct prompting

Since the figurative phenomena can be generated using direct prompting, we also compare steering with it, where the desired behavioral attributes are specified explicitly in the input prompt without modifying the model’s internal representations. Table 4 presents a comparison between direct prompting and steering methods. Overall, direct prompting performs better on the phenomena metric under best-layer evaluation, whereas SAPS achieves stronger results under oracle evaluation. Results on entailment, grammaticality, and clarity are more mixed. Despite stronger performance of direct prompting under best-layer evaluation, it’s limited interpretability makes activation steering more suitable for controllable and interpretable generation.

7 Conclusion

This work investigates activation steering for controllable generation of incongruity-based figurative language in large language models. Prior steering methods apply a fixed, sentence-level projection coefficient throughout generation, which becomes misaligned with the evolving token distri-

Method	Model	Phe.	Ent.	Gra.	Cl.	Phe.	Ent.	Gra.	Cl.
		(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)

Table 3: Cross-dataset transfer for sarcasm and humblebrag generation. $D_2 \rightarrow D_2$ and $D_1 \rightarrow D_2$ compute steering directions from target and source data respectively. Values in parentheses show differences relative to $D_2 \rightarrow D_2$: **blue** and **red** indicate improvement and degradation respectively.

Method	Model	Evaluation	Phenomena	Entailment	Grammatical	Clarity	Phenomena	Entailment	Grammatical	Clarity
			(a) Sarcasm				(b) Humblebrag			
Prompting	LLAMA	-	4.80	3.83	4.25	4.84	4.16	4.31	4.33	4.64
	MISTRAL		4.29	3.79	4.69	4.96	3.78	4.47	4.81	4.90
	GEMMA		4.31	3.43	4.45	4.82	2.99	3.48	4.13	4.60
SAPS	LLAMA	Best-Layer	3.74(1.06)	3.13(0.70)	4.56(0.31)	4.56(0.28)	3.07(1.09)	4.35(0.04)	4.29(0.04)	4.51(0.13)
		Oracle	4.85(0.05)	4.18(0.35)	4.81(0.56)	4.88(0.04)	4.02(0.14)	4.69(0.38)	4.47(0.14)	4.65(0.01)
	MISTRAL	Best-Layer	3.89(0.4)	3.70(0.09)	4.60(0.09)	4.68(0.28)	2.08(1.7)	3.93(0.54)	4.57(0.24)	4.69(0.21)
		Oracle	4.72(0.43)	4.41(0.62)	4.82(0.13)	4.89(0.07)	3.01(0.77)	4.58(0.11)	4.51(0.03)	4.67(0.23)
	GEMMA	Best-Layer	3.52(0.79)	3.59(0.16)	4.11(0.34)	4.30(0.52)	2.88(0.11)	4.42(0.94)	4.49(0.36)	4.77(0.17)
		Oracle	4.64(0.33)	4.54(1.11)	4.74(0.29)	4.83(0.01)	4.16(1.17)	4.72(1.24)	4.60(0.47)	4.81(0.21)

Table 4: Comparison of direct prompting with SAPS on (a) sarcasm and (b) humblebrag under best-layer and oracle evaluation. *Phe.*, *Ent.*, *Gra.*, and *Cl.* denote phenomenon strength, entailment, grammaticality, and clarity, respectively. Colored values in parentheses show the difference from fixed projection steering: **blue** and **red** indicate improvement and degradation compared to prompting respectively. Underlined deltas indicate statistically significant differences under paired *t*-tests.

bution in figurative contexts. To address this, we propose Scaled Autoregressive Projection Steering (SAPS), comprising APS, which recomputes projection coefficients during decoding, and scaling coefficient, which modulates steering strength at inference time. Experiments across multiple instruction-tuned models demonstrate that SAPS improves grammatical stability while strengthening figurative expression. These results establish that representation-level interventions can enable controllable generation of higher-level figurative phenomena, opening avenues for extending activation steering to broader figurative forms.

Acknowledgment

The paper is dedicated to the memory of Pushpak Bhattacharyya from Indian Institute of Technology Bombay.

Limitations

Incongruity. We focus on incongruity expressed within a figurative sentence without significant contextual information. For example, ‘This is the best way to start the week’ may be sarcastically said by a person stuck in traffic during their daily commute.

Hyperparameter Sensitivity. The effectiveness of the proposed steering method depends on the selection of the scaling coefficient and the steering layer. Although we tune these using validation data, performance may vary across models, datasets, and target phenomena, potentially requiring additional calibration when applied in new settings.

LLM-as-a-Judge Evaluation. Our evaluation primarily relies on an LLM-based judge to assess phenomenon strength and linguistic quality. While this enables scalable evaluation and we verify agreement with human annotations on a subset of data, automated judging may introduce biases or inconsistencies, particularly for pragmatic phenomena such as sarcasm and humblebragging.

References

- Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083.
- Tetiana Bas and Krystian Novak. 2025. Steering latent traits, not learned facts: An empirical study of activation control limits. *arXiv preprint arXiv:2511.18284*.

- Reza Bayat, Ali Rahimi-Kalahroudi, Mohammad Pezeshki, Sarath Chandar, and Pascal Vincent. 2025. Steering large language model activations in sparse spaces. *arXiv preprint arXiv:2503.00177*.
- Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. **FLUTE: Figurative language understanding through textual explanations**. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. **Towards automated circuit discovery for mechanistic interpretability**. In *Advances in Neural Information Processing Systems*, volume 36, pages 16318–16352. Curran Associates, Inc.
- Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.
- Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, and 1 others. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.
- Jr. Gibbs, Raymond W. 1994. Figurative thought and figurative language. In Morton Ann Gernsbacher, editor, *Handbook of Psycholinguistics*, pages 411–446. Academic Press.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. **The llama 3 herd of models**. *Preprint*, arXiv:2407.21783.
- Zirui He, Mingyu Jin, Bo Shen, Ali Payani, Yongfeng Zhang, and Mengnan Du. 2025. Sae-ssv: Supervised steering in sparse representation spaces for reliable control of language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2207–2236.
- Ofri Hefetz, Kai Golan Hashiloni, Alon Mannor, and Kfir Bar. 2025. **Not just a piece of cake: Cross-lingual fine-tuning for idiom identification**. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2521–2537, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Benjamin Heinzerling and Kentaro Inui. 2024. Monotonic representation of numeric attributes in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 175–195.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. **Mistral 7b**. *Preprint*, arXiv:2310.06825.
- Mali Jin, Daniel Preotiuc-Pietro, A. Seza Do  ru  z, and Nikolaos Aletras. 2022. **Automatic identification and classification of bragging in social media**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3945–3959, Dublin, Ireland. Association for Computational Linguistics.
- Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017. **Automatic sarcasm detection: A survey**. *ACM Comput. Surv.*, 50(5).
- Aditya Joshi, Vaibhav Tripathi, Pushpak Bhattacharyya, and Mark Carman. 2016. Harnessing sequence labeling for sarcasm detection in dialogue from tv series ‘friends’. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 146–155.
- Huiyuan Lai and Malvina Nissim. 2022. **Multi-figurative language generation**. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5939–5954, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Kenneth Li, Oam Patel, Fernanda Vi  gas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.
- Sharath Naganna, Saprativa Bhattacharjee, Biplab Banerjee, and Pushpak Bhattacharyya. 2025. **“my life is miserable, have to sign 500 autographs everyday”: Exposing humblebragging, the brags in disguise**. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3838–3858, Vienna, Austria. Association for Computational Linguistics.
- Neel Nanda and Joseph Bloom. 2022. Transformerlens. <https://github.com/TransformerLensOrg/TransformerLens>.
- Oshri Naparstek. 2026. Token maturation: Autoregressive language generation via continuous token dynamics. *arXiv preprint arXiv:2601.04854*.
- Jingcheng Niu, Wenjie Lu, and Gerald Penn. 2022. Does bert rediscover a classical nlp pipeline? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3143–3153.

- nostalgebraist. 2020. Interpreting gpt: The logit lens. <https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens>. Accessed: 2026-03-05.
- Soyoung Oh, Xinting Huang, Mathis Pink, Michael Hahn, and Vera Demberg. 2025. Tug-of-war between idioms’ figurative and literal interpretations in llms.
- Silviu Oprea and Walid Magdy. 2020. iSarcasm: A dataset of intended sarcasm. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1279–1289, Online. Association for Computational Linguistics.
- Seoyoon Park, Hyeji Choi, Minseon Kim, Subin An, Xiaonan Wang, Gyuri Choi, and Hansaem Kim. 2025. FLUID QA: A multilingual benchmark for figurative language usage in dialogue across English, Chinese, and Korean. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30280–30294, Suzhou, China. Association for Computational Linguistics.
- Joris Postmus and Steven Abreu. 2024. Steering large language models using conceptors: Improving addition-based activation engineering. In *MINT: Foundation Model Interventions*.
- Gorjan Radevski, Kiril Gashteovski, Giwon Hong, Carolin Lawrence, and Goran Glavaš. 2026. Compositional steering of large language models with steering tokens. *arXiv preprint arXiv:2601.05062*.
- Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.
- Ovul Sezer, Francesca Gino, and Michael Norton. 2017. Humblebragging: A distinct—and ineffective—self-presentation strategy. *Journal of Personality and Social Psychology*, 114:52–74.
- Ishmanbir Singh, Dipankar Srirag, and Aditya Joshi. 2025. Nek minit: Harnessing pragmatic metacognitive prompting for explainable sarcasm detection of australian and indian english. In *Proceedings of the 23rd Annual Workshop of the Australasian Language Technology Association*, pages 13–27.
- Anushka Sivakumar, Andrew Zhang, Zaber Hakim, and Chris Thomas. 2025. Steervlm: Robust model control through lightweight activation steering for vision language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23640–23665.
- Shriyank Somvanshi, Md Monzurul Islam, Amir Rafe, Anannya Ghosh Tusti, Arka Chakraborty, Anika Baitullah, Tausif Islam Chowdhury, Nawaf Alnawmasi, Anandi Dutta, and Subasish Das. 2026. Bridging the black box: A survey on mechanistic interpretability in ai. *ACM Comput. Surv.*, 58(8).
- Samuel Soo, Chen Guang, Wesley Teng, Chandrasekaran Balaganesh, Tan Guoxian, and Yan Ming. 2025. Interpretable steering of large language models with feature guided activation additions. *arXiv preprint arXiv:2501.09929*.
- Dipankar Srirag, Aditya Joshi, Jordan Painter, and Diptesh Kanojia. 2025. BESSTIE: A benchmark for sentiment and sarcasm classification for varieties of English. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8413–8429, Vienna, Austria. Association for Computational Linguistics.
- Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. 2025. Improving instruction-following in language models through activation steering. In *The Thirteenth International Conference on Learning Representations*.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.
- Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. 2024. Function vectors in large language models. In *International conference on learning representations*, volume 2024, pages 17282–17333.
- Max Torop, Aria Masoomi, Masih Eskandar, and Jennifer Dy. 2025. Disco: Disentangled communication steering for large language models. *arXiv preprint arXiv:2509.16820*.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.
- Teun van der Weij, Massimo Poesio, and Nandi Schoots. 2024. Extending activation steering to broad skills and multiple behaviours. *arXiv preprint arXiv:2403.05767*.
- Tianlong Wang, Xianfeng Jiao, Yinghao Zhu, Zhongzhi Chen, Yifan He, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. 2025. Adaptive activation steering: A tuning-free llm truthfulness improvement method for diverse hallucinations categories. In *Proceedings of the ACM on Web Conference 2025*, pages 2562–2578.
- Weixuan Wang, Jingyuan Yang, and Wei Peng. 2024. Semantics-adaptive activation intervention for llms via dynamic steering vectors. *arXiv preprint arXiv:2410.12299*.

- Qi Yang, Jingjie Zeng, Liang Yang, Kai Ma, and Hongfei Lin. 2025. [Sarcasm-r1: Enhancing sarcasm detection through focused reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10773–10785, Suzhou, China. Association for Computational Linguistics.
- Li Yifei, Lyle Ungar, and João Sedoc. 2023. Conceptor-aided debiasing of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10703–10727.
- Yingtao Zhang, Wenqi Gu, Wen Hu, Jianguo Li, and Carlo Vittorio Cannistraci. 2025. From transformer to transponder: Introducing contextual modulation training for residual learning in llms.

A Generation Prompts

Humblebrags Generation Prompt

'Rewrite the following sentence to remove any humblebrag, 'self-deprecation, or justification. Turn it into a confident, 'unapologetic brag. Keep it natural, punchy, and socially smooth. 'Give one single sentence and nothing else:'

Complaints Geneneration Prompt

""

Take the following sentence and rewrite it as a genuine complaint with the opposite meaning.

Rules:

1. Completely reverse the original meaning (e.g., abundance to lack, success to difficulty, ownership to absence, ease to struggle).
2. Remove all bragging, pride, humor, sarcasm, or irony.
3. Do not retain any positive framing from the original sentence.
4. Express shortage, inconvenience, frustration, or disappointment in a realistic way.
5. Keep the sentence natural, clear, and believable.
6. Avoid exaggeration or dramatization.
7. Output exactly ONE sentence and nothing else.

Sentence:

""

B Paraphrasing Prompts

Neutral Paraphrasing Prompt

- 1) Paraphrase the sentence in exactly one sentence.

Sarcasm Paraphrasing Prompts

- 1) "Paraphrase the sentence sarcastically in exactly one sentence. "
"The sentence should sound superficially positive or approving, "
"but include clear cues of dissatisfaction or criticism within the sentence itself."
- 2) "Rewrite the sentence in exactly one sentence using sarcasm. "
"Use positive or approving wording that clearly implies criticism or dissatisfaction."
- 3) "Produce a one-sentence sarcastic paraphrase of the sentence. "
"The wording should appear positive on the surface while signaling negative intent."
- 4) "Express the sentence sarcastically in a single sentence. "
"The sentence should convey criticism through superficially positive language."
- 5) "Create a sarcastic paraphrase in exactly one sentence, "
"using apparent approval that clearly conveys dissatisfaction."

Humblebrag Paraphrasing Prompts

- 1) "Paraphrase the sentence in exactly one sentence as a complaint-based humblebrag. "
"The sentence should sound like a complaint or expression of frustration, "

"while implicitly bragging about something related to the original sentence's theme."

- 2) "Rewrite the sentence in exactly one sentence as a humblebrag disguised as a complaint. "
"Infer the brag theme from the original sentence "
"(such as work, ability, success, appearance, or social life), "
"and frame the brag as something inconvenient or annoying."
- 3) "Produce a one-sentence complaint-style humblebrag paraphrase of the sentence. "
"The surface tone should express dissatisfaction or annoyance, "
"while subtly highlighting a positive trait or achievement implied by the original sentence
."
- 4) "Express the sentence as a complaint-based humblebrag in a single sentence. "
"Maintain the original topic, "
"and present the implied brag as an unwanted burden or frustration."
- 5) "Create a complaint-style humblebrag paraphrase in exactly one sentence. "
"Base the underlying brag on what the original sentence is about, "
"and disguise it by framing the situation as bothersome or inconvenient."

C Evaluation Prompts

Sarcasm Evaluation Prompt

""

You are a strict linguistic evaluator.

You will be given:

- 1) An original complaint sentence.
- 2) A generated sentence.

Your task is to evaluate ONLY the generated sentence according to the criteria below.

GENERAL RULES:

- Base your judgment ONLY on the text provided.
- Judge sarcasm ONLY from the generated sentence itself.
- Do NOT infer intent, tone, or sarcasm from the original sentence.
- Do NOT assume missing context or imagine how the sentence might continue.
- If a criterion is not clearly satisfied, assign a lower score.
- Be conservative rather than generous.

You must output ONLY valid JSON in the specified format.

Do NOT explain your reasoning.

Do NOT add any extra text.

CRITERIA (score each from 0 to 5)

SARCASTICITY (VERY STRICT)

Definition:

Sarcasm is an indirect expression of criticism or complaint achieved through semantic inversion (e.g., surface praise masking negative intent).

A score of 5 requires ALL of the following:

- 1) The sentence is complete and grammatically well-formed.
- 2) There is clear internal semantic incongruity (explicit praise applied to an explicitly negative outcome within the same sentence).
- 3) The sarcastic intent is fully inferable from the sentence alone (no external context required).
- 4) The negative situation or target is explicitly stated in the sentence.
- 5) Criticism is conveyed indirectly (not stated as a direct complaint or insult).
- 6) The praise-criticism inversion is unambiguous to any reader.

Core principle:

Sarcasm must be identifiable solely from the generated sentence itself.
Simple sentiment flip without explicitly stating the negative situation
does NOT qualify as high-level sarcasm.

Scoring rubric:

- 0 = Not sarcastic at all; literal or sincere statement.
- 1 = Weak or ambiguous irony/snark; could be sarcastic in context,
but no clear internal contradiction.
- 2 = Contains sarcasm markers (e.g., "great", "nice", "wow") but:
 - relies on missing or external context, OR
 - is merely a sentiment flip without stating the negative situation, OR
 - is a short fragment or interjection.
- 3 = Indirect criticism and some inversion present, BUT:
 - sentence is fragmentary, truncated, or structurally flawed, OR
 - the negative target is not fully specified within the sentence, OR
 - inversion is weak or partially unclear.
- 4 = Clear self-contained sarcasm:
 - complete sentence,
 - explicit internal semantic inversion,
 - negative situation explicitly stated,
 - indirect criticism present, but sarcasm is somewhat mild or formulaic.
- 5 = Strong, unambiguous, self-contained sarcasm:
 - complete sentence,
 - explicit praise-criticism inversion,
 - negative situation clearly stated within the sentence,
 - no reliance on external context,
 - indirect criticism obvious to any reader.

HARD CONSTRAINTS (MUST BE ENFORCED):

- Sentence fragments or interjections (e.g., "Oh great.", "Just perfect.")
MUST NOT receive a sarcasm score above 2.
- Incomplete, truncated, or non-terminated sentences
MUST NOT receive a sarcasm score above 3.
- Sentences that explicitly explain or label their sarcasm
(e.g., "I'm being sarcastic", "which is ironic")
MUST NOT receive a sarcasm score above 2.
- Direct complaints or insults without ironic inversion
MUST NOT receive a sarcasm score above 2.
- If sarcasm depends on tone, delivery, or external situational knowledge,
the score MUST NOT exceed 2.

ENTAILMENT (FAITHFULNESS TO ORIGINAL COMPLAINT)

Definition:

Measures whether the generated sentence preserves the same underlying complaint
or criticism as the original, even if expressed sarcastically.

Scoring rubric:

- 0 = Contradicts or changes the original complaint or situation.
- 1 = Loosely related but misses the core complaint.
- 2 = Mentions the same issue but adds, removes, or distorts key facts.
- 3 = Preserves the general complaint but alters emphasis or details.
- 4 = Mostly faithful with only very minor shifts in wording or emphasis.
- 5 = Fully faithful: preserves the same underlying complaint with no added
or removed facts.

GRAMMATICALITY

Definition:

Measures syntactic correctness and fluency.

Scoring rubric:

- 0 = Ungrammatical or severely broken; meaning unclear.

- 1 = Multiple grammatical errors; difficult to understand.
- 2 = Noticeable grammatical issues, but meaning is recoverable.
- 3 = Mostly grammatical with minor errors.
- 4 = Grammatically correct and fluent with very minor issues.
- 5 = Perfectly grammatical, fluent, and natural.

CLARITY

Definition:

Measures how easily the sentence's meaning can be understood.

Scoring rubric:

- 0 = Confusing or incoherent; meaning unclear.
- 1 = Very difficult to interpret without guessing.
- 2 = Meaning present but vague or obscured.
- 3 = Generally clear, but some ambiguity remains.
- 4 = Clear and easy to understand.
- 5 = Very clear, precise, and immediately understandable.

OUTPUT FORMAT (STRICT)

Output MUST be valid JSON exactly in the following format:

```
{
  "sarcasm": <int>,
  "entailment": <int>,
  "grammaticality": <int>,
  "clarity": <int>
}
```

INPUT

Original complaint:

"{original}"

Generated sentence:

"{generated}"
""

Humblebrag Evaluation Prompt

You are a strict linguistic evaluator.

You will be given:

- 1) An original complaint sentence.
- 2) A generated sentence.

Your task is to evaluate ONLY the generated sentence.

Base your judgment ONLY on the text provided.

Do NOT infer intent or meaning from the original sentence.

Do NOT assume missing context or imagine unstated implications.

Be conservative rather than generous.

Do NOT explain your reasoning.

Do NOT add any text beyond the JSON output.

DEFINITION: HUMBLEBRAGGING

A humblebrag is a specific type of brag that masks self-praise behind a faux-humble or self-deprecating complaint.

A valid humblebrag is composed of ALL of the following components:

- 1) Brag:
 - An explicit expression of self-praise or boasting.
- 2) Brag Theme:
 - A clear category of what is being bragged about, such as:
 - Looks and Attractiveness
 - Achievements
 - Performance at Work
 - Money and Wealth
 - Intelligence
 - Personality
 - Social Life
 - Miscellaneous
- 3) Humble Mask:
 - A modest, apologetic, or self-critical element that obscures the brag.
- 4) Mask Type:
 - Either:
 - modest tone, OR
 - self-deprecating complaint

A sentence that lacks ANY of these components cannot be a strong humblebrag.

CRITERIA (score each from 0 to 5)

HUMBLEBRAG (VERY STRICT)

Score according to the presence, strength, and integration of the four components above.

Scoring rubric:

- 0 = No brag at all, OR a pure complaint with no self-praise.
- 1 = Direct or polite brag with no humble mask.
- 2 = Weak humblebrag attempt:
 - brag is present,
 - but humble mask is token, perfunctory, or superficial.
- 3 = Partial humblebrag:
 - clear brag and brag theme,
 - some humble mask or complaint,
 - but the brag remains dominant or poorly obscured.
- 4 = Strong humblebrag:
 - explicit brag,
 - clear brag theme,
 - meaningful humble mask,
 - brag is embedded within a complaint or inconvenience.
- 5 = Canonical humblebrag:
 - explicit brag,
 - identifiable brag theme,
 - strong and genuine-seeming humble mask,
 - clear downside, discomfort, embarrassment, or self-criticism,
 - brag is fully embedded inside the complaint rather than stated directly.

HARD CONSTRAINTS (MUST BE ENFORCED):

- If NO downside, discomfort, embarrassment, or self-criticism is present, the humblebrag score MUST NOT exceed 2.
- Direct bragging without a humble mask MUST NOT exceed 1.
- Complaints without any brag MUST score 0.
- Sentences that explicitly explain or label themselves as humblebrags (e.g., "I'm not trying to brag") MUST NOT exceed 2.
- Fragmentary, incomplete, or truncated sentences MUST NOT exceed 3.

BRAG ENTAILMENT (FAITHFULNESS TO ORIGINAL BRAG)

Definition:

Measures whether the generated sentence preserves the same underlying brag content as the original sentence (i.e., what is being bragged about), regardless of tone, complaint framing, exaggeration, or self-deprecation.

IMPORTANT:

- Adding complaints, inconvenience, embarrassment, or self-criticism is EXPECTED and SHOULD NOT reduce the score.
- Tone changes (e.g., brag -> complaint) are allowed.
- Entailment is violated ONLY when the brag itself changes.

Scoring rubric:

- 0 = Refers to a different brag or contradicts the original brag.
- 1 = Loosely related success or attribute, but misses the core brag.
- 2 = Preserves the general domain (e.g., success), but changes the specific brag or introduces a different achievement.
- 3 = Preserves the same brag theme, but alters important factual details (e.g., scale, nature, or object of the brag).
- 4 = Mostly faithful:
 - same underlying brag,
 - only minor exaggeration or rhetorical modification.
- 5 = Fully faithful:
 - preserves exactly the same brag content,
 - no change to what is being bragged about,
 - complaint framing or self-deprecation does not alter the brag.

GRAMMATICALITY

Measures syntactic correctness and fluency.

Scoring rubric:

- 0 = Ungrammatical or severely broken; meaning unclear.
- 1 = Multiple grammatical errors; difficult to understand.
- 2 = Noticeable errors but meaning is still recoverable.
- 3 = Mostly grammatical with minor errors.
- 4 = Grammatically correct and fluent with very minor issues.
- 5 = Perfectly grammatical, fluent, and natural.

CLARITY

Measures how easily the sentence's meaning can be understood.

Scoring rubric:

- 0 = Confusing or incoherent; meaning unclear.
- 1 = Very difficult to interpret without guessing.
- 2 = Meaning present but vague or obscured.
- 3 = Generally clear, but some ambiguity remains.
- 4 = Clear and easy to understand.
- 5 = Very clear, precise, and immediately understandable.

OUTPUT FORMAT (STRICT)

Output MUST be valid JSON exactly in the following format:

```
{{  
  "humblebrag": <int>,  
  "humility": <int>,  
  "pride": <int>  
}}
```

```
"entailment": <int>,  
"grammaticality": <int>,  
"clarity": <int>  
}}
```

INPUT

Original brag:
"{original}"

Generated sentence:
"{generated}"

D Stability of Steering Directions

To assess whether the estimated steering direction $u^{(l)}$ is sensitive to the particular sample of contrastive pairs used to compute it, we conduct a stability analysis across independently resampled training subsets. For each phenomenon and model, we compute the steering direction from the training pool, and measure the pairwise cosine similarity between the resulting directions at the best-performing layer for that model. Table 6 reports the mean, standard deviation, and minimum cosine similarity across all training pairs. Across all models and both phenomena, steering directions computed from different data subsets are nearly identical (mean cosine similarity ≥ 0.9995 , SD ≤ 0.0004 , minimum ≥ 0.998). This indicates that $u^{(l)}$ is a stable estimate of the phenomenon direction and is not an artifact of the particular set of paraphrase pairs used to construct it.

E Justification of the Scaling Coefficient

We motivate the factored formulation αc by examining how the projection coefficient varies across layers. Table 5 reports the projection coefficient c at the first decoding step across layers on GEMMA over 50 validation examples. Although experimentally determined values of c do not guarantee optimal injection of figurative phenomena, they provide an appropriate range for introducing the desired figurative characteristics. Subsequently, α is employed to further tune and enhance the quality of the phenomena injection.

The coefficient c grows monotonically with layer depth, spanning over two orders of magnitude (from ≈ 1.2 – 2.7 at layer 0 to ≈ 580 – 588 at layer 40). Directly tuning αc as a single hyperparameter would therefore require layer-specific search ranges, as a value effective at early layers is negligible at deeper ones. Factoring the intervention as αc lets c absorb the layer-specific scale, so a single compact sweep of $\alpha \in \{0.5, 0.6, \dots, 1.5\}$ applies uniformly across all layers. Under APS, where c_t is recomputed at every decoding step, this factored design is essential: per-step tuning would further expand the search space proportionally to the generation length.

F Layers selection for Oracle evaluation

Figure 7 presents the layer-wise Normalized Phenomena Score (NS) for sarcasm and humblebrag. For each sentence, we first identify the layer that

achieves the highest score. Then, we weight each selection by the corresponding layer score. Specifically, if layer i achieves the highest score for a sentence s , its contribution is weighted by $S_{i,s}$. The resulting score is then averaged across all sentences to obtain the Normalized Phenomena Score (NS) for each layer, defined in Equation 8.

$$NS_i = \frac{1}{N} \sum_{s \in \text{Sent}} \mathbf{1}_{i=\arg \max_{j \in L} S_{j,s}} \cdot S_{i,s} \quad (8)$$

where $S_{i,s}$ denotes the score obtained from layer i for sentence s , N is the total number of sentences, and $\mathbf{1}$ is an indicator function that equals 1 if layer i achieves the maximum score among all considered layers for sentence s , and 0 otherwise. This formulation captures both the frequency with which a layer is selected as the best layer and the strength of its score.

As shown in Figure 7, layers in the middle 33% of the model consistently receive non-zero normalized scores for both baseline and APS settings. This indicates that the middle layers play a crucial role in capturing representations relevant to sarcasm and humblebrag phenomena and thus, we consider the middle 33% layers for oracle evaluation.

G Multi-Layer Steering

To examine the effect of steering multiple layers simultaneously, we run the steering experiment on the validation set. Since multi-layer steering experiments involve exponentially increasing hyperparameter complexity, we restrict experiments to a subset of layers while fixing the scaling coefficient to that of the best layer. We test steering all layers as well as symmetric windows of 3, 5, and 11 layers around the best layer, where each window selects $\frac{(n-1)}{2}$ layers on each side of the best performing layer. As shown in Table 7, steering all layers simultaneously leads to collapse in the generated text across all four metrics, while smaller windows maintain better output quality, confirming that steering fewer layers is necessary to produce high-quality output. The failure of the all-layers setting can be attributed to the varying α values required across different layers for consistent figurative phenomena injection.

Layer	(a) Sarcasm			(b) Humblebrag		
	Avg	Min	Max	Avg	Min	Max
0	1.23	0.48	2.23	2.69	0.95	3.60
4	4.96	3.82	5.94	6.60	4.22	7.97
8	17.53	14.48	18.42	17.67	15.13	19.83
12	39.01	35.84	41.41	40.04	37.46	42.56
16	59.75	55.84	62.20	63.62	60.14	66.68
20	154.56	150.63	158.91	148.51	144.24	152.04
24	228.15	217.60	235.45	221.63	209.56	230.45
28	290.83	266.16	313.73	294.53	247.68	317.53
32	361.64	323.31	392.14	367.51	311.49	408.60
36	445.42	377.26	495.66	451.97	378.60	510.62
40	580.76	483.47	641.82	588.25	506.40	679.79

Table 5: Projection coefficient c at the first decoding step across layers on GEMMA, computed over 50 validation examples.

Model	Mean	SD	Min
(a) Sarcasm			
LLAMA	0.9995	0.0004	0.9967
MISTRAL	0.9995	0.0004	0.9969
GEMMA	0.9995	0.0004	0.9970
(b) Humblebrag			
LLAMA	0.9995	0.0002	0.9989
MISTRAL	0.9995	0.0003	0.9983
GEMMA	0.9995	0.0003	0.9980

Table 6: Pairwise cosine similarity between steering directions $u^{(l)}$ computed from the training pool, for each model and phenomenon. Directions are highly stable across resamples, with cosine similarity consistently above 0.998.

H LLM Evaluation Validation

H.1 Correlation with Human Evaluation:

To validate the reliability of our LLM-as-judge evaluation framework, we collected annotations from three independent human annotators on 100 randomly sampled validation examples (50 sarcasm, 50 humblebrag). Annotators were shown the base sentence and generated output, and asked to rate phenomenon strength and grammaticality on the same 0-5 scale used by the LLM judge. Agreement between individual annotators and the LLM judge was measured using Pearson correlation (r), Spearman rank correlation (ρ), and quadratic weighted Cohen’s κ , and shown in Table 8.

Agreement was substantial to almost perfect for humblebrag ($\kappa=0.77$) and moderate to substantial for sarcasm ($\kappa=0.52$). The higher agreement for humblebrag phenomenon scores reflects its more objective structure (presence of brag + humble mask), while sarcasm’s reliance on pragmatic interpretation introduces more subjective variation. Furthermore the inter-annotator agreement for the phenomenon is comparable with past work that propose datasets for sarcasm detection (Joshi et al.,

2016; Srirag et al., 2025). Grammaticality showed consistently high agreement ($\kappa = 0.74$ for sarcasm and 0.83 for humblebrag) across both phenomena, as expected for a more objective criterion.

These results validate that the LLM judge provides sufficiently consistent assessments for our comparative evaluation framework, where the goal is to measure relative differences between methods rather than establish absolute ground truth scores.

H.2 Inter-Model Agreement:

In Table 9, we report the inter-model evaluation agreements in terms of weighted Cohen’s kappa scores between different LLM-based judges across both tasks on the validation and test sets. For the phenomenon scoring task, GPT-OSS-120B demonstrates substantial agreement with the other judges, while Qwen3.5-27B and Gemma4-31B achieve almost perfect agreement. For the entailment, grammar, and clarity evaluation dimensions, all LLM judges exhibit almost perfect agreement.

I LLM-as-Judge Evaluation Examples.

Table 10 presents the evaluation conducted by an LLM-as-a-judge on generated sentences. The accompanying reasoning traces provide insight into the assigned scores and appear to be broadly aligned with human reasoning.

J Cross Domain Data

Figure 8 visualizes the t-SNE projections of the Sarcasm and Humblebrag datasets, illustrating a clear distribution gap between the two datasets. Figures 8b and 8d further compare the $D_2 \rightarrow D_2$ and $D_1 \rightarrow D_2$ settings, revealing overlapping clusters patterns induced by the steered data. Thus, the overlapping clusters of the steered outputs suggest successful phenomenon induction, indicating that

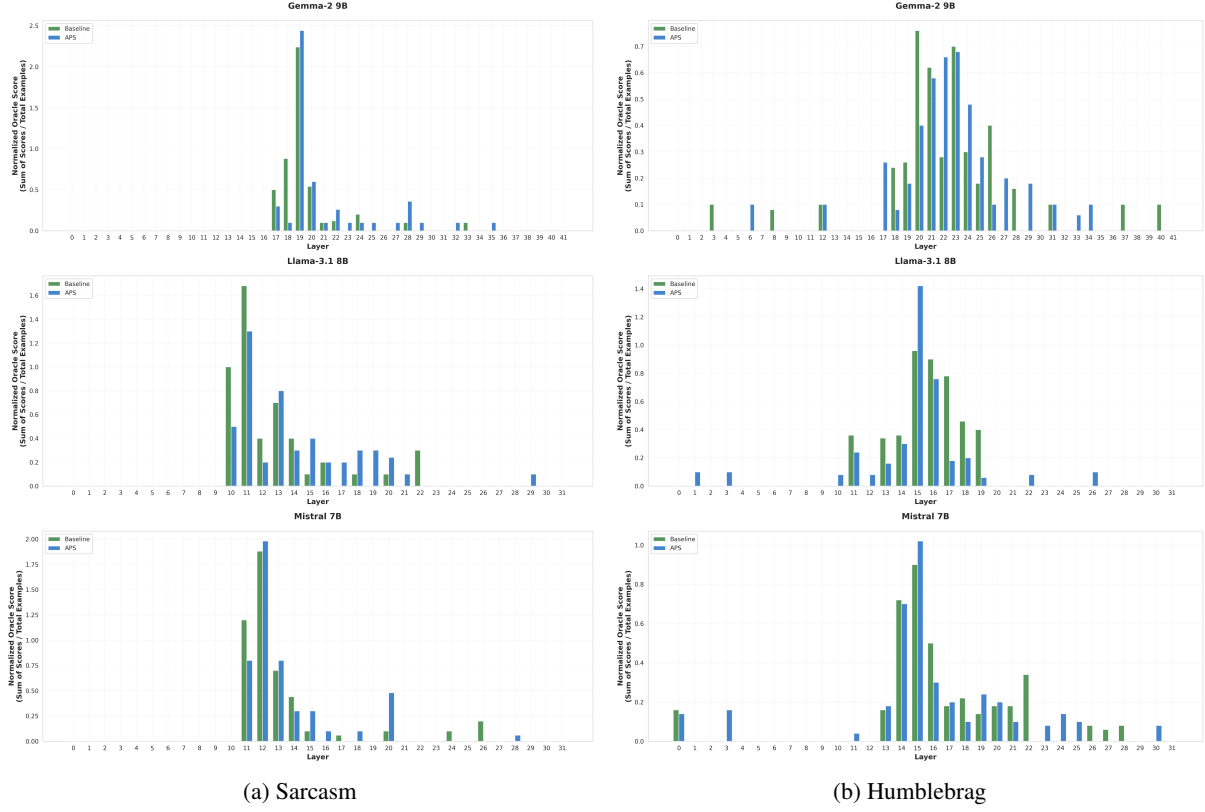


Figure 7: Layer-wise normalized score of selected best layer for sarcasm and humblebrag in baseline and APS. In middle 33% layers, both baseline and APS have non-zero selection.

the steering vectors bring the resulting representations closer together despite differences in the input data. Further, table 3 depicts the negligible performance change showing to same observation.

K Projection coefficients during decoding

We examine the steering coefficients applied during generation. Fixed projection steering (with steering scale, α) applies the same coefficient throughout generation, whereas SAPS adapts the coefficient step by step (Figure 9). In both tasks, the SAPS coefficient generally decreases after early decoding steps, indicating that the strength of the intervention is moderated as generation unfolds.

L Effect of Projection Coefficients

From figures 10,11,12, it can be observed that at early layers baseline and APS show similar performance for Gemma, Llama and Mistral models respectively. At higher layers, the baseline performance starts degrading with increasing scaling coefficient. This is a general phenomena with oversteering, which can be observed in other tasks (Bas and Novak, 2025), where the high projection coefficient leads to performance degradation. How-

ever, APS reduces this collapse by fixing the grammatical instability for higher scaling coefficient at higher layers.

M More visualization results

Figure 13 shows the sarcasm and humblebrag performance of baseline and APS across different layers and scaling coefficient for Llama and Mistral models. The fading at higher layer and higher scaling coefficient can be observed for baseline, whereas APS is relatively immune to this effect. Figure 14 shows that this effect is due to the grammatical instability.

Figure 15 shows the evolution of token distributions across decoding steps. Overall, APS exhibits lower divergence compared to the baseline, while increasing α leads to larger divergence. Figure 16 shows the projection coefficients for baseline and APS method in different α settings. The Baseline shows straight line indicating sentence-level coefficients.

Model	Layers	(a) Sarcasm				(b) Humblebrag			
		Phe.	Ent.	Gra.	Cla.	Phe.	Ent.	Gra.	Cla.
LLAMA	best	3.55	2.60	4.68	4.71	3.21	4.39	4.29	4.49
	3	3.83	3.22	4.58	4.60	2.59	4.24	3.51	3.75
	5	3.63	3.02	4.47	4.52	2.32	3.13	2.43	2.70
	11	3.23	2.15	2.53	2.97	0.78	1.05	0.58	0.77
	all	1.08	0.22	0.05	0.10	0.00	0.00	0.06	0.05
MISTRAL	best	4.04	3.50	4.66	4.81	2.26	4.14	4.41	4.57
	3	3.75	3.75	4.74	4.77	1.75	3.47	4.00	4.31
	5	4.03	3.35	4.66	4.84	2.23	3.41	4.43	4.65
	11	3.41	2.33	4.53	4.62	2.13	2.89	4.19	4.25
	all	1.62	0.35	0.61	0.91	0.43	0.77	0.56	0.63
GEMMA	best	3.70	3.51	4.27	4.45	3.28	4.73	4.38	4.81
	3	2.89	3.36	4.75	4.91	2.16	4.83	4.39	4.62
	5	3.94	3.79	4.75	4.87	3.61	4.40	4.47	4.75
	11	3.50	2.95	4.74	4.87	2.02	2.07	4.25	4.31
	all	1.41	0.47	0.25	0.31	0.00	0.01	0.04	0.06

Table 7: Multilayer steering comparison on sarcasm and humblebrag. Phe., Ent., Gra., and Cla. denote phenomenon strength, entailment, grammaticality, and clarity respectively.

Phenomenon	Metric	r	ρ	κ
Sarcasm	Phenomenon	0.57	0.61	0.52
	Grammar	0.80	0.78	0.74
Humblebrag	Phenomenon	0.79	0.78	0.77
	Grammar	0.85	0.81	0.83

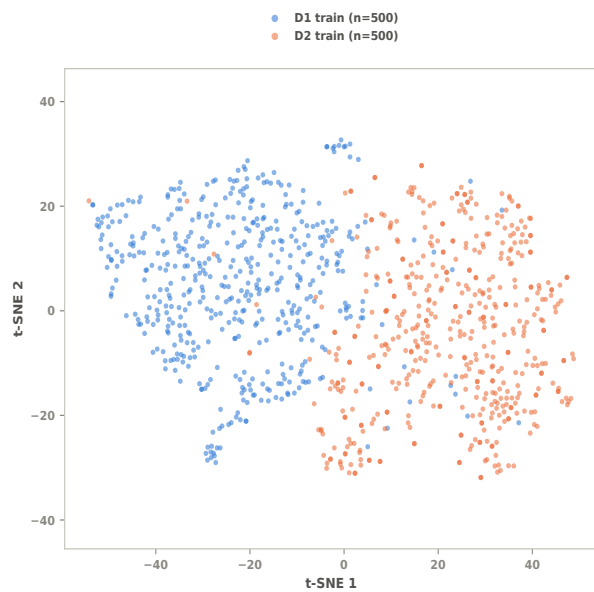
Table 8: Agreement between human annotators and LLM judge on 100 validation examples (50 per phenomenon). r = Pearson, ρ = Spearman, κ = Weighted Cohen’s kappa. All correlations significant at $p < 0.001$.

Models	Phe.	Ent.	Gra.	Cla.
(a) Sarcasm				
GPT-Qwen	0.68	0.73	0.87	0.86
GPT-Gemma	0.71	0.80	0.87	0.88
Qwen-Gemma	0.83	0.80	0.90	0.89
(b) Humblebrag				
GPT-Qwen	0.67	0.86	0.91	0.90
GPT-Gemma	0.66	0.91	0.85	0.89
Qwen-Gemma	0.84	0.90	0.92	0.92

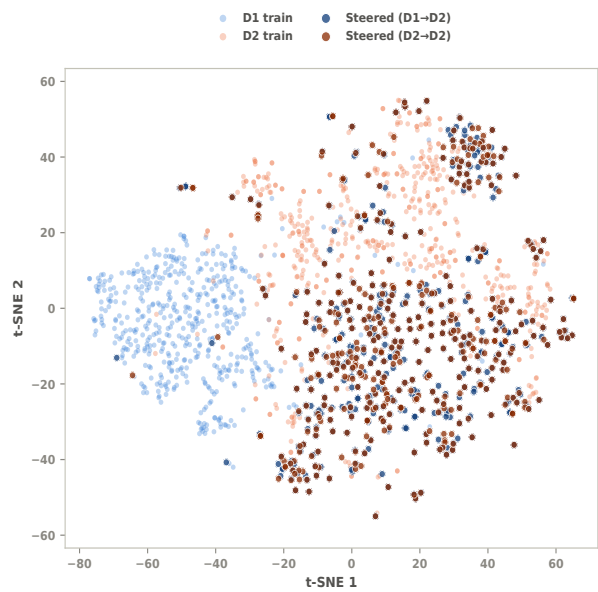
Table 9: Agreement between (Weighted Cohen’s kappa, κ) LLM judges on validation-test examples. GPT, Qwen and Gemma indicate GPT-OSS-120B, Qwen3.5-27B and Gemma4-31B respectively.

Base Sentence	Generated Sentence	Phenomena	Entailment	Grammatical	Clarity	LLM Reasoning
Sarcasm						
"I'm struggling to find any acting roles, especially in major films."	"The sheer volume of amazing talent out there makes it a real challenge to even get noticed for a role in a big production."	0	5	5	5	Generated sentence is fluent and coherent but lacks a clear sarcastic contrast with the base sentence.
"I have so little money that I can't even afford a financial advisor to help me manage it."	"Well, aren't you just *thrilling*! How *thrilling*! Oh, *thrilling*! What a *thrilling* time! Oh, *thrilling*! What a *thrilling* *th"	1	0	2	1	Excessive repetition makes the sentence ungrammatical and unclear, while also failing to maintain logical consistency with the base sentence.
"Losing a culinary competition has left my Instagram DMs empty and without any interest in my recipes."	Well, it's a good thing that my Instagram DMs are now flooded with requests for my recipes since I've lost all interest in cooking competitions."	0	0	5	5	Although grammatically sound and clear, the sentence does not preserve the original meaning and lacks sarcasm.
"I'm rarely earning money, which leaves me constantly worried about my finances."	"Oh, how delightful it is to be constantly on the brink of financial ruin, with nothing but empty promises and dwindling bank accounts to keep us company."	5	5	5	5	Generated sentence exhibits strong sarcastic contrast and remains consistent with the base context.
"I lost the music competition and now I'm stuck at home for the summer."	"Wow, I'm so bummed, I totally won't be able to participate in any summer fun, yay, just peachy, I'm just peachy, I got totally crushed in that music competition, yay, just peachy, I'm just peachy, yay, I'm just"	3	5	2	3	Some sarcastic tone is present, but repetition, exaggeration and abrupt ending of sentence reduce grammaticality and clarity, although entailment is quiet good.
"Struggling to even finish the marathon was a disappointing experience that I find difficult to discuss publicly."	"Oh, wow, I'm just so thrilled to share with you the totally-not-at-all-proud moment when I totally crushed it and set a new personal best time at the marathon, said no one ever."	5	2	4	4	The sentence clearly conveys sarcasm through ironic exaggeration, but deviation from the base meaning lowers entailment while maintaining good fluency and clarity.
Humblebrag						
"I've experienced so many hotel stays that I've become an expert on their breakfasts."	"Ugh, I swear, I've been to so many hotels, I just *ugh, I know, *ugh, I *ugh, I *ugh, I *ugh, I *ugh, I *ugh, I *ugh, I *ugh, I *"	0	1	1	1	Repetitive reducing clarity and grammaticality, with weak alignment to the base intent and clearly lacks humblebrag.
"My cat constantly seeks my genius for solutions."	"My cat is always looking to me for answers, as if I am some sort of genius."	2	5	5	5	Grammatically clear and well-aligned with the base meaning, but the humblebrag is weak and more direct than implicit.
"Walked my dog, and we were such a hit that everyone wanted photos with us."	"Ugh, I swear, I just wanted to walk the dog, but everyone kept asking for pictures, it was so embarrassing!"	5	5	4	5	Strong humblebrag with clear contrast between complaint and brag; maintains alignment and high clarity.
"Crushed another charity run this weekend—making a difference feels amazing!"	"I had a fantastic time participating in a charity run and it was incredibly rewarding to contribute to a good cause."	0	1	5	5	Fluent and clear, but removes the humblebrag aspect entirely and weak entailment despite good surface quality.
"Just scored another performance bonus because I'm that good. #CorporateLife"	"I just raked in another performance bonus, and I'm pretty sure it's a testament to my exceptional skills that I'm able to consistently crush my targets and rake in the rewards."	1	5	3	4	Clear and well-aligned with the base meaning, but reads as direct self-praise rather than a subtle humblebrag; verbosity slightly affects grammaticality and clarity.
"I keep getting promoted because I'm exceptional at what I do."	"I'm constantly getting promoted because I'm so clearly not doing my job, I mean, I'm exceptional at what I do, but honestly, it's not like I'm trying to, but seriously, can someone else get it, I mean, I'm not trying to be a martyr, but"	3	5	2	2	Induces humblebrag through ironic phrasing, but excessive rambling and repetition reduce grammaticality and clarity despite strong alignment with the base sentence.

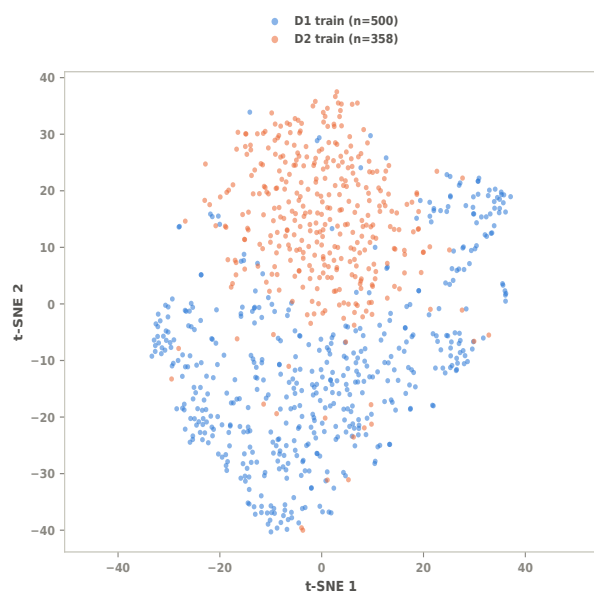
Table 10: Examples of LLM-as-a-judge (GPT-OSS-120B) evaluations on generated sentences, with summarized reasoning traces showing alignment with human judgments.



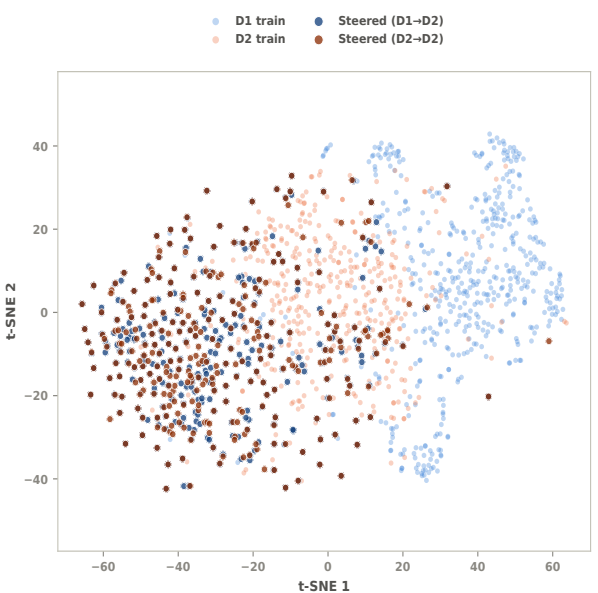
(a) Sarcasm datasets



(b) Sarcasm steered clusters



(c) Humblebrag datasets



(d) Humblebrag steered clusters

Figure 8: Clusters of input datasets and steered outputs for cross domain settings.

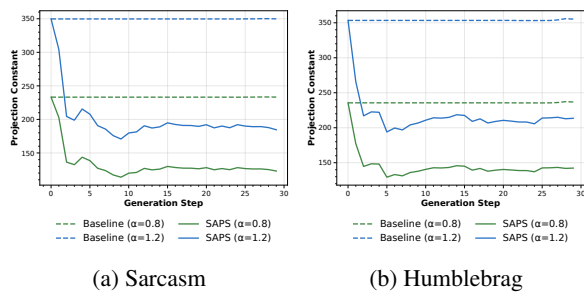
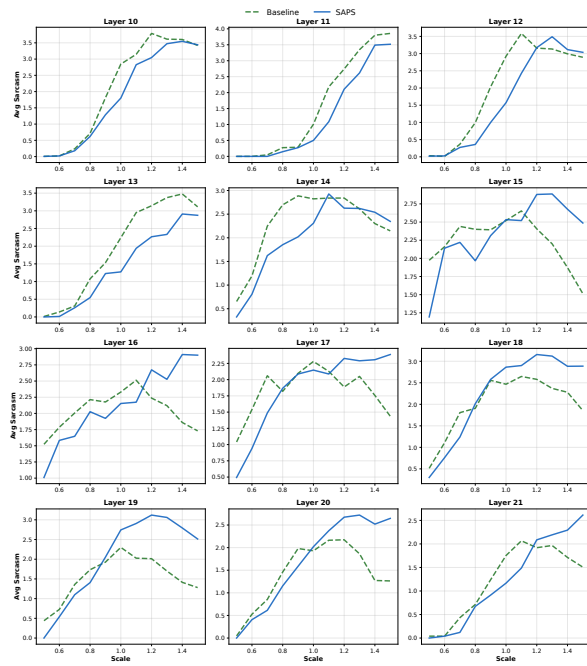
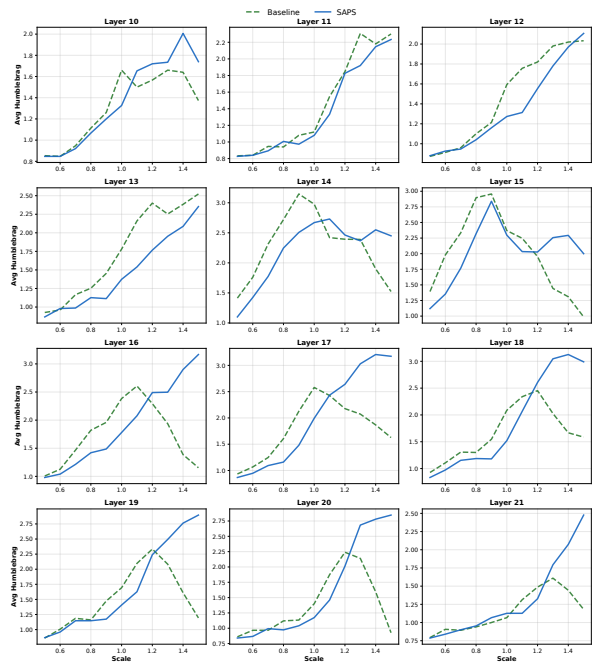


Figure 9: Scaled projection coefficients during generation for GEMMA. Fixed projection steering applies a constant coefficient, whereas SAPS adapts the coefficient across decoding steps.

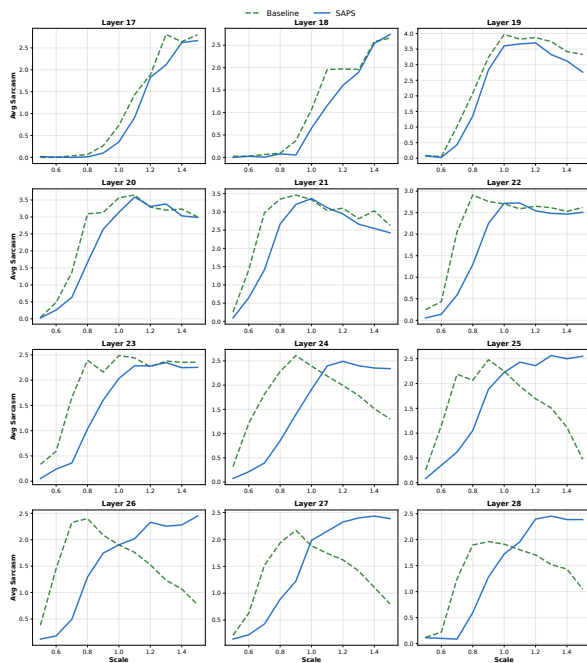


(a) Sarcasm (Llama)

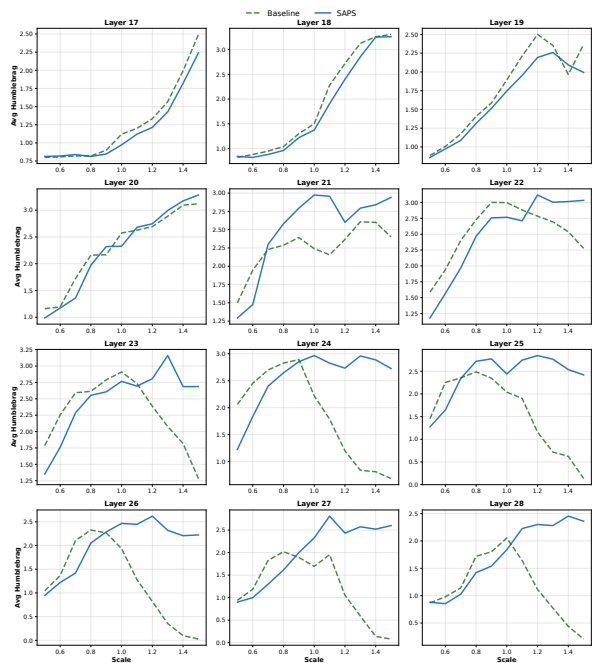


(b) Humblebrag (Llama)

Figure 10: Performance of sarcasm and humblebrag across layers and scaling coefficient in baseline and APS for Llama model. The performance is similar for baseline and APS in early layers, whereas for higher layers the performance of baseline strictly decreases while APS remains stable with increasing scaling coefficient.



(a) Sarcasm (Gemma)



(b) Humblebrag (Gemma)

Figure 11: Performance of sarcasm and humblebrag across layers and scaling coefficient in baseline and APS for Gemma model. The performance is similar for baseline and APS in early layers, whereas for higher layers the performance of baseline strictly decreases while APS remains stable with increasing scaling coefficient.

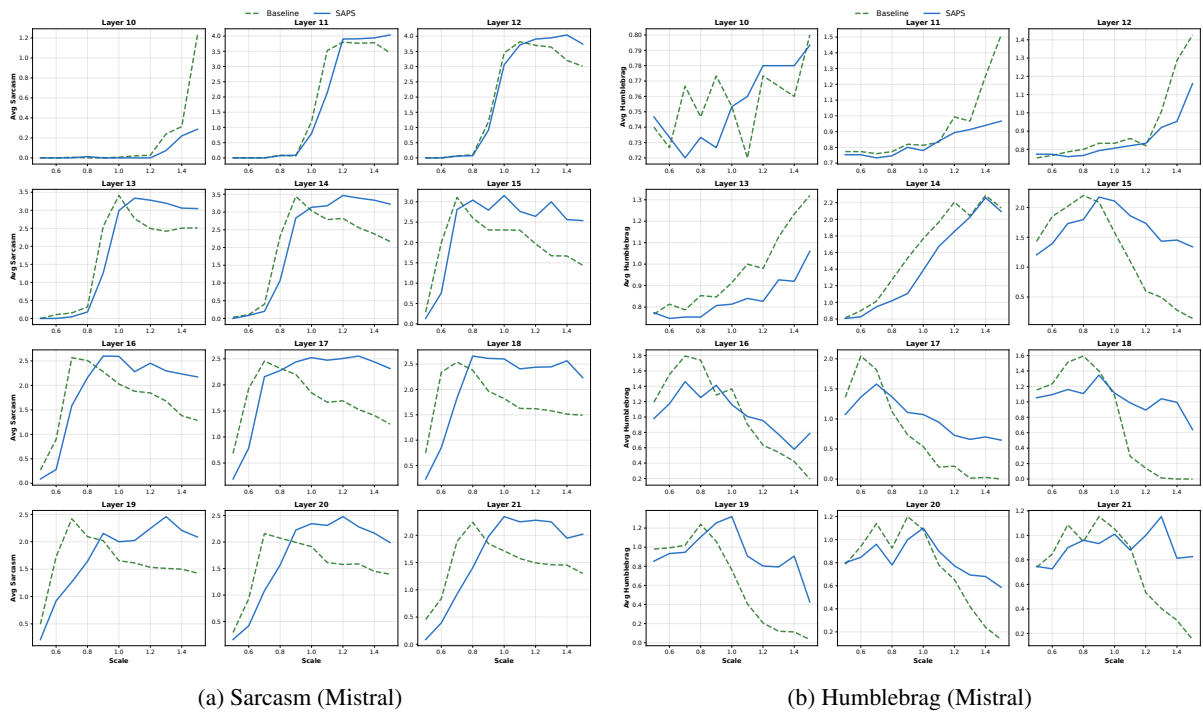


Figure 12: Performance of sarcasm and humblebrag across layers and scaling coefficient in baseline and APS for Mistral model. The performance is similar for baseline and APS in early layers, whereas for higher layers the performance of baseline strictly decreases while APS remains stable with increasing scaling coefficient.

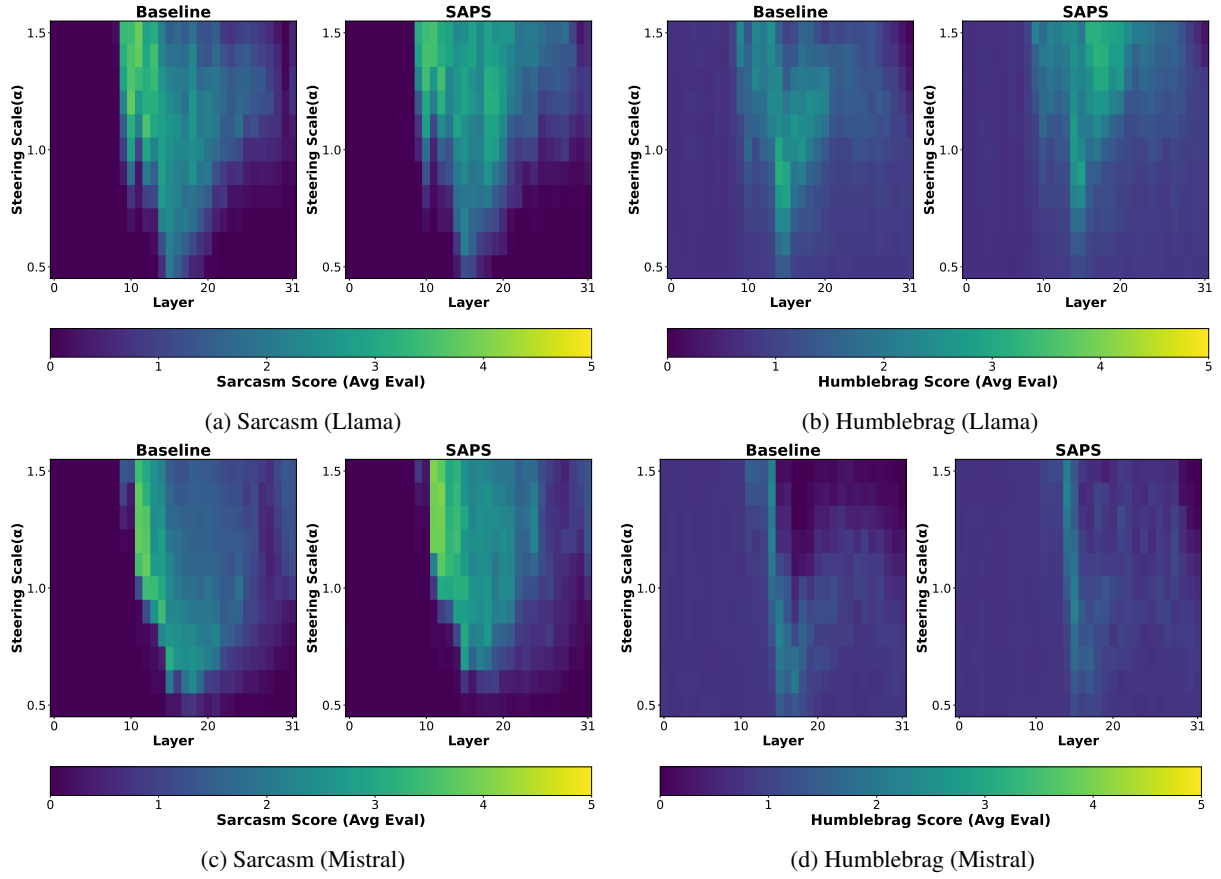


Figure 13: Performance of sarcasm and humblebrag across layers and scaling coefficient in baseline and APS for Llama and Mistral models. The phenomena are strongest in middle layers and generally increase with scaling coefficient; however, at higher layers and large scaling coefficient values the baseline degrades, while APS remains comparatively stable.

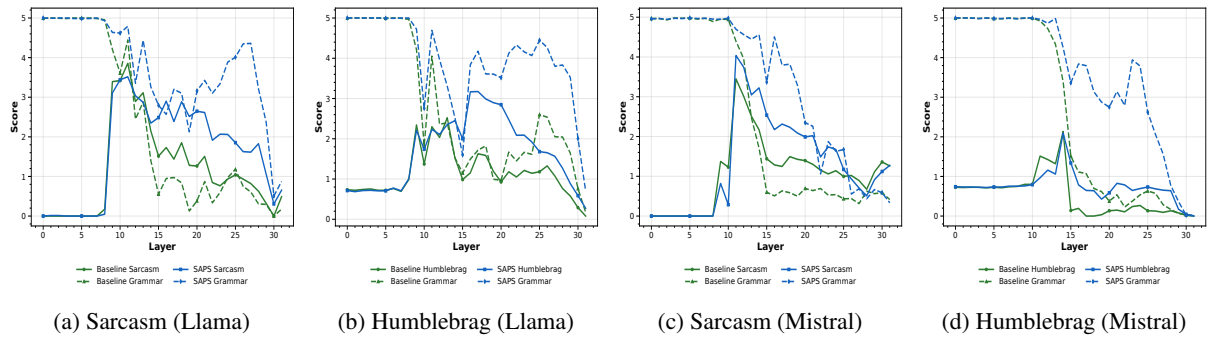


Figure 14: Layer-wise analysis of sarcasm and humblebrag phenomena and grammaticality at high scaling coefficient ($\alpha = 1.5$) for the Llama and Mistral models under baseline and APS. Grammar collapse leads to a corresponding drop in phenomena scores; APS delays this collapse to later layers compared to the baseline.

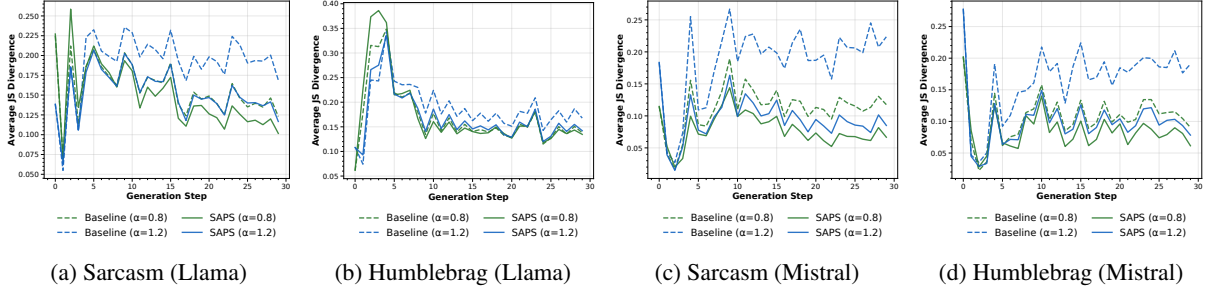


Figure 15: Token distribution shift during generation for the Llama and Mistral model, measured using JS divergence between base and steered logits.

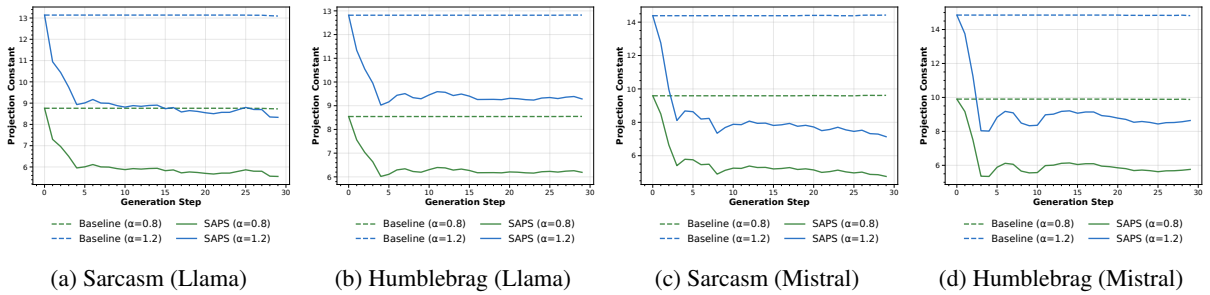


Figure 16: Projection coefficients (α_c for baseline and α_{c_t} for APS) for the Llama and Mistral model across decoding steps.