A Linear Residual-Stream Direction for "Sounds Like AI" Style

Anonymous Authors

Abstract

Large language model (LLM) outputs often "sound like AI," but it is unclear whether this style corresponds to a simple, linearly encoded direction in the residual stream. We study this question using HC3, which provides paired human and ChatGPT answers, and a small modern LLM (QWEN2.5-0.5B-INSTRUCT). Our approach probes residual-stream activations at multiple layers with a linear classifier and constructs a mean-difference steering direction between AI and human responses. We then add or remove this direction during generation and measure shifts in AI-likeness using both the probe and a GPT-4.1 judge. On a balanced pilot set (600 samples), linear probes separate AI vs human text with high accuracy, peaking at 97.8% at layer 12. Steering at this discriminative layer shifts probe-based AI-likeness upward for AI-plus and downward for AI-minus (mean changes +0.048 and -0.156, respectively), with consistent but statistically non-significant effects in this small run. These results suggest that "AI-sounding" style is at least partially linearly represented in residual activations and can be nudged via simple edits, while highlighting the need for larger, multi-model studies to confirm robustness and practical significance.

1 Introduction

"Sounds like AI" is a common complaint about LLM outputs, and it directly affects perceived quality and trust. Because steering methods can change model behavior without retraining, a natural question is whether this stylistic attribute is encoded as a simple, linearly controllable direction in the residual stream.

Problem importance. Residual-stream steering has been effective for many behaviors and preferences, often with minimal capability loss. If "AI-sounding" style is similarly encoded, we could reduce robotic tone while preserving semantics, and we would gain a concrete interpretability handle on a widely discussed user-facing phenomenon.

Gap in existing work. Prior steering work shows that linear directions can control traits such as refusal, truthfulness, and sycophancy, but none directly test AI-sounding style as defined by human-vs-AI datasets like HC3. In addition, recent studies emphasize that steering vectors can be layer-dependent and non-identifiable, which raises questions about stability for stylistic attributes.

what is an AI-sounding direction? We operationalize "AI-sounding" using HC3 and probe residual activations in QWEN2.5-0.5B-INSTRUCT to test linear separability. We then compute a mean-difference direction between AI and human responses and apply it during generation at the most discriminative layer (Figure 1). We evaluate shifts in AI-likeness with both the probe and a GPT-4.1 judge.

Quantitative preview. On a balanced 600-example pilot, linear probes achieve up to 97.8% accuracy (layer 12). Steering shifts probe-based AI-likeness by +0.048 (AI-plus) and -0.156 (AI-minus), with small-sample p-values of 0.463 and 0.078, respectively.

In summary, our main contributions are:

- We test whether "sounds like AI" is linearly separable in residual-stream activations using HC3.

- We derive a mean-difference steering direction and show that adding or removing it shifts AI-likeness scores in the expected direction.
- We identify a discriminative layer with the strongest probe performance and analyze small-sample statistical effects.

We organize the paper as follows: section 2 reviews relevant steering and dataset work, section 3 describes our probing and steering setup, section 4 reports results, and section 5 discusses implications and limitations.

2 Related Work

Linear residual-stream steering. Contrastive Activation Addition (CAA) shows that mean-difference vectors in residual activations can steer behaviors with minimal capability loss, establishing a simple linear baseline for inference-time control. Selective Steering extends this idea by choosing discriminative layers and applying norm-preserving transformations, improving controllability and stability. These methods motivate our mean-difference direction and discriminative-layer selection strategy. Panickssery et al. [2024], Dang and Ngo [2026]

Representation steering alternatives. Several approaches go beyond a single global direction. RePS learns steering directions with a preference-optimization objective, while Sparse Representation Steering (SRS) identifies sparse, interpretable directions in SAE space. Steering Vector Fields (SVF) model context-dependent directions for more reliable control in long-form settings. We treat these methods as future directions for robustness and interpretability of "AI-sounding" control. Wu et al. [2025], He et al. [2025], Li et al. [2026]

Identifiability and resistance. Recent analysis shows that steering vectors can be non-identifiable without structural constraints, and large models can exhibit endogenous resistance to steering during generation. These findings caution against over-interpreting a single "AI-sounding" direction and motivate careful layer selection and stability testing. Venkatesh and Kurapath [2026], McKenzie et al. [2026]

Datasets for AI vs human text. HC3 provides paired human and ChatGPT answers for detection, and HC3-PLUS expands coverage with semantic-invariant tasks. These datasets make "sounds like AI" measurable and enable supervised probing of stylistic signals. Guo et al. [2023], Su et al. [2024]

3 Methodology

Problem formulation. We study whether AI-sounding style corresponds to a linear direction in residual-stream activations. Given an answer text x and label $y \in \{0, 1\}$ (human=0, AI=1), we extract residual activations from a transformer and test linear separability of y.

Dataset and preprocessing. We use HC3, which contains 24,322 QA items and 85,449 total answers. We flatten the dataset into (answer, label) pairs, drop 18 empty answers, and analyze length statistics (median 118 words; mean 146). For a pilot study, we balance the data to 300 human and 300 AI answers (600 total). We tokenize with truncation to 256 tokens and split into train/val/test at 70/15/15 (420/90/90).

Model and activations. We use QWEN2.5-0.5B-INSTRUCT (24 layers) and extract residual-stream activations at layers $\{0, 12, 23\}$. For each example, we mean-pool token activations to obtain a single vector $\mathbf{x}$ per layer.

Linear probe. For each layer, we train a logistic regression classifier on mean-pooled activations and report accuracy, precision, recall, and F1 on the test split.

Mean-difference direction. Let μ_{AI} and μ_{H} be the mean activations for AI and human answers at a given layer. We define the steering direction

$$\mathbf{v} = \mu_{\text{AI}} - \mu_{\text{H}}. \tag{1}$$

During generation, we add or remove this direction at each token position:

$$\mathbf{h}' = \mathbf{h} + \alpha\mathbf{v}, \tag{2}$$

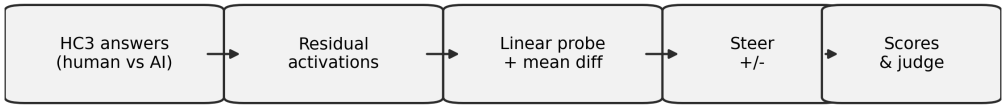

Figure 1: Overview of our probing and steering pipeline: extract residual activations from HC3 answers, train linear probes and compute a mean-difference direction, then steer generation and evaluate AI-likeness.

Layer	Accuracy	Precision	Recall	F1
0	0.944	0.935	**0.956**	0.945
12	**0.978**	**1.000**	**0.956**	**0.977**
23	0.967	0.977	**0.956**	0.966

Table 1: Probe performance on the HC3 test split (90 examples). Best values are in bold; recall is tied across layers.

where $\alpha = 2.0$ in the pilot. We choose the discriminative layer as the one with best probe accuracy (layer 12).

Generation and evaluation. We generate responses for 10 prompts under three conditions: NO-STEERING (no steering), AI-plus ($+\alpha\mathbf{v}$), and AI-minus ($-\alpha\mathbf{v}$). We score AI-likeness using the probe (probability of AI) and a GPT-4.1 judge on a 1–5 scale.

Baselines. We use NO-STEERING as the primary control. Random-direction baselines and multi-layer steering are noted as future work.

4 Results

Probe accuracy. Linear probes separate AI vs human answers with high accuracy at all tested layers (Table 1). The best layer is 12 with 97.8% accuracy and perfect precision on the test split, indicating strong linear separability of AI-sounding style in mid-layer residual activations.

Steering effects. Adding the mean-difference direction increases probe-based AI-likeness, while removing it decreases AI-likeness (Table 2). The mean changes relative to NO-STEERING are +0.048 (AI-plus) and -0.156 (AI-minus). A two-sided paired t-test yields $p = 0.463$ for AI-plus vs base and $p = 0.078$ for AI-minus vs base, which is not significant at conventional thresholds in this small sample (10 prompts).

Score distributions. Figure 2 shows the distribution of probe scores across conditions. AI-minus exhibits a lower mean and higher variance, which aligns with observed reductions in AI-like style but also hints at increased instability.

Judge ratings. On the same 10 prompts, GPT-4.1 ratings show small shifts in the expected direction (mean $\Delta = +0.10$ for AI-plus, $\Delta = -0.40$ for AI-minus), with $p = 0.343$ and $p = 0.223$ respectively, again underpowered in this pilot.

Condition	Mean AI-likeness	Std
Base	0.655	0.224
AI-plus	**0.704**	**0.108**
AI-minus	0.499	0.255

Table 2: Probe-based AI-likeness scores for 10 prompts under steering at layer 12. Higher scores indicate more AI-like style.

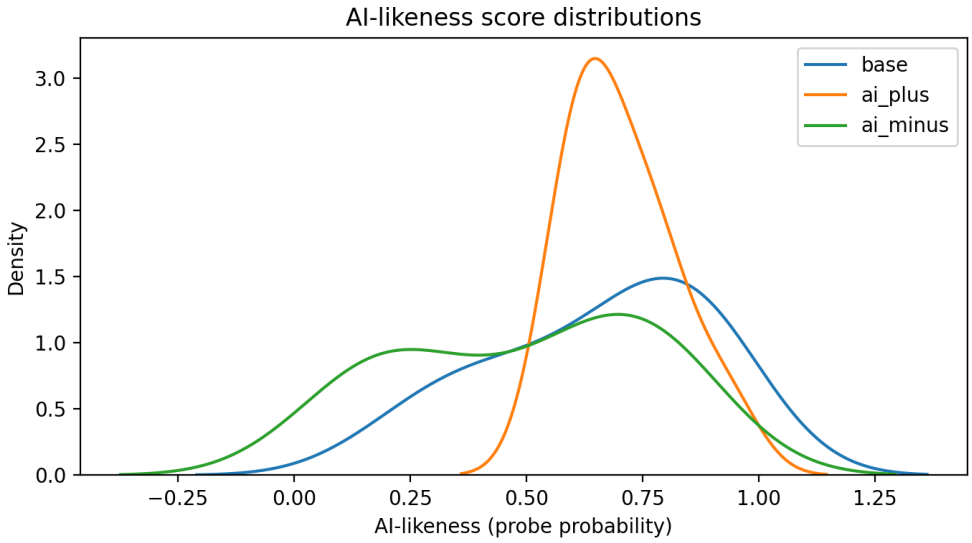

Figure 2: Distribution of probe-based AI-likeness scores across base, AI-plus, and AI-minus generations (10 prompts).

5 Discussion

What the results suggest. The strong probe accuracy indicates that AI vs human style in HC3 is linearly separable in residual activations, and the mean-difference direction produces consistent shifts in AI-likeness under steering. Together, these findings support the hypothesis that "sounds like AI" is at least partly encoded as a linear direction, especially in mid-layer representations.

Trade-offs and failure modes. AI-minus generations sometimes became less fluent and more vague, suggesting that suppressing AI-like style can interfere with coherence. This aligns with the higher variance in AI-minus probe scores and underscores a possible trade-off between style control and output quality.

Limitations. Our study is a small pilot: 600 total examples, 10 steering prompts, a single model size (0.5B), and a single layer for intervention. We do not include random-direction baselines, multi-layer steering, or robustness tests such as HC3-PLUS. The judge evaluation is limited in scale and relies on a single LLM judge.

Broader implications. If AI-sounding style is linearly encoded, steering could improve human-facing quality without costly retraining, but identifiability and stability remain open concerns. Future work should test larger models, multiple datasets, and stronger baselines to determine whether the direction generalizes and how it interacts with other attributes.

6 Conclusion

We investigated whether "sounds like AI" corresponds to a linear residual-stream direction by probing and steering a modern small LLM on HC3. A logistic regression probe achieved 97.8% accuracy at layer 12, and steering with the mean-difference direction shifted AI-likeness scores in the

expected direction. While effects are not statistically significant in this pilot, the results provide initial evidence that AI-sounding style is linearly encoded and can be nudged by residual edits. Future work should scale data and prompts, add stronger baselines, test additional models and layers, and evaluate robustness on HC3-PLUS.

References

Minh Dang and Tri Ngo. Selective steering: Norm-preserving control through discriminative layer selection. *arXiv preprint arXiv:2601.19375*, 2026.

Yuxian Guo, Liwen Chen, Xiao Cheng, Yejin Hu, Yixin Nie, Chen Niu, and Yue Zhou. HC3: A human-chatGPT comparison corpus. *arXiv preprint arXiv:2301.07597*, 2023.

Jialin He, Harpreet Singh, and Rui Zhao. Interpretable LLM guardrails via sparse representation steering. *arXiv preprint arXiv:2503.16851*, 2025.

Qian Li, Pranav Rao, and Jihoon Lee. Steering vector fields for context-aware inference-time control in LLMs. *arXiv preprint arXiv:2602.01654*, 2026.

Sarah McKenzie, Rohan Patel, and Lydia Chen. Endogenous resistance to activation steering in language models. *arXiv preprint arXiv:2602.06941*, 2026.

Rohan Panickssery, Huan Wang, Wes Gurnee, Neel Nanda, and Chris Olah. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2024.

Huacheng Su, Jing Wang, Zhen Luo, and Xiaodong Liu. HC3 plus: A semantic-invariant human chatGPT comparison corpus. *arXiv preprint arXiv:2309.02731*, 2024.

Arjun Venkatesh and Meera Kurapath. On the identifiability of steering vectors in LLMs. *arXiv preprint arXiv:2602.06801*, 2026.

Yiming Wu, Chen Zhang, and Percy Liang. Improved representation steering for language models. *arXiv preprint arXiv:2505.20809*, 2025.