Is there a "Sounds Like AI" Direction in the Residual Stream?

Haokun Liu
University of Chicago
haokunliu@uchicago.edu

Abstract

As large language models (LLMs) become increasingly integrated into human workflows, distinguishing between human-written and AI-generated text has become a critical challenge for academic integrity and misinformation mitigation. The Linear Representation Hypothesis (LRH) suggests that high-level abstract concepts are represented as linear directions in the residual stream of LLMs. In this work, we investigate whether the stylistic quality of being "AI-generated" is encoded as a distinct linear feature. Using the RAID dataset and Representation Engineering (REPE) techniques, we identify a "sounds like AI" direction in the middle layers of the QWEN2.5-1.5B model. We find that AI-generated and human-written scientific abstracts are highly linearly separable in the residual stream, even in early layers. Furthermore, we demonstrate that steering activations along this identified direction causally shifts the stylistic quality of the model's output. Specifically, subtracting the identified AI direction from a prompt completion reduces its predicted "AI-ness" from 36.2% to 0.5%, while adding it increases the probability to 67.1%. Our findings suggest that LLMs possess an internal stylistic axis representing their own generation patterns, which can be precisely manipulated to control the perceived anthropomorphism of the output.

1 Introduction

The rapid proliferation of Large Language Models (LLMs) has led to an explosion of AI-generated content across the internet, from scientific literature to social media. Distinguishing between human-authored and machine-generated text is now a fundamental necessity for maintaining academic standards, ensuring the reliability of information, and preserving the human-centric nature of digital discourse. While external classifiers and watermarking techniques have been developed to detect AI-generated text, we lack a mechanistic understanding of how LLMs internally represent their own stylistic signatures.

The Linear Representation Hypothesis (LRH) posits that LLMs represent high-level concepts as one-dimensional linear directions in their residual stream [Park et al., 2023]. This hypothesis has been successfully validated for concepts such as truthfulness [Marks and Tegmark, 2023], refusal behavior [Arditi et al., 2024], and honesty [Li et al., 2023]. However, "sounding like an AI" is a more nebulous, stylistic property that involves complex interactions of tone, vocabulary, and structural transitions. **Is there a single linear direction that captures this elusive quality?**

If such a direction exists, it would allow us to not only detect AI-generated text by inspecting internal activations but also causally intervene in the generation process. By steering activations along this axis, we could potentially make model outputs more anthropomorphic or, conversely, more explicitly machine-like, providing a fine-grained control over model persona and style.

In this paper, we identify and analyze a "sounds like AI" direction in the residual stream of the QWEN2.5-1.5B model. Using a balanced subset of the RAID dataset [Dugan et al., 2024], we

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

extract activations and employ Representation Engineering (REPE) techniques [Li et al., 2023] to find a candidate direction. Our experiments demonstrate that this direction is not only highly predictive of text origin but also causally relevant for stylistic steering. We show that subtracting the AI direction reduces the AI-ness of generated text from 36.2% to 0.5% as measured by a linear probe, while adding it increases it to 67.1%.

In summary, our main contributions are:

- We identify a "sounds like AI" direction in the residual stream of QWEN2.5-1.5B and show that human and AI styles are perfectly linearly separable (100% accuracy) as early as layer 1.
- We demonstrate causal control over model style via activation steering, shifting the perceived "AI-ness" of outputs by over 66 percentage points without losing semantic coherence.
- We characterize the stylistic markers captured by this direction, finding that it encodes traits such as objective tone, technical vocabulary, and formal structural transitions.

2 Related Work

Linear Representation Hypothesis (LRH). Our work is grounded in the LRH, which suggests that high-level abstract concepts are encoded as linear directions in the latent spaces of neural networks. Park et al. [2023] provide a theoretical framework for this hypothesis, showing that the residual stream of transformers often exhibits a geometric structure where concept vectors can be identified and manipulated. Empirical evidence for the LRH has been found in diverse domains, including the representation of truth in factual statements [Marks and Tegmark, 2023] and the control of refusal behavior in safety-aligned models [Arditi et al., 2024]. We extend this line of inquiry to the stylistic domain of "AI-ness."

Representation Engineering (REPE). The field of REPE focuses on monitoring and manipulating the internal states of AI models to control high-level cognitive and stylistic phenomena. Li et al. [2023] introduced Inference-Time Intervention (ITI) to improve the truthfulness of LLMs by steering activations along identified truth directions. Similarly, Zou et al. [2023] proposed a unified framework for identifying and controlling concepts like honesty and emotions. Our methodology utilizes the Difference-in-Means approach popularized by these works to identify the stylistic axis of AI-generated text.

Mechanistic Interpretability of Style. Beyond simple factual or safety-related features, recent research has begun to explore more complex behavioral signatures. O'Neill et al. [2025] found that contextual hallucinations are mediated by a single linear direction, suggesting that even inconsistent behaviors have a stable linear representation. Lee et al. [2024] examined how alignment techniques like DPO influence internal representations. While these works focus on correctness and safety, our research focuses on the broader stylistic signature of AI generation, which is often characterized by specific lexical and structural regularities.

AI-Generated Text Detection. The detection of AI-generated text has traditionally relied on external classifiers. The RAID benchmark [Dugan et al., 2024] provides a comprehensive dataset for evaluating such detectors across various domains and generator models. Most current detectors use fine-tuned transformer models or statistical measures like perplexity and curvature. Our work differs from these black-box approaches by identifying an *internal* representation of AI-style within the generator model itself, offering a more mechanistic and potentially more robust path toward detection and control.

3 Methodology

We aim to identify a linear direction in the residual stream of an LLM that distinguishes between AI-generated and human-written text. Our approach follows the Representation Engineering (REPE) framework, consisting of activation extraction, direction identification, and causal validation via steering.

3.1 Data Construction

We utilize a subset of the RAID dataset [Dugan et al., 2024], which contains pairs of human-written and AI-generated texts. For this study, we focus on the scientific abstract domain. We sample 100 human-written scientific abstracts and 100 abstracts generated by LLAMA-2-Chat. To ensure consistency and avoid length-based artifacts, all samples are truncated to 512 tokens. A potential confounder noted in the dataset is the frequent use of prefixes like "In this paper..." in AI-generated abstracts, which we address during analysis.

3.2 Activation Extraction

We use QWEN2.5-1.5B as our target model. For each of the 200 samples, we perform a forward pass and extract the residual stream activations $\boldsymbol{a}_l \in \mathbb{R}^d$ across all $L = 29$ layers, where $d = 1536$ is the model dimension. We collect two types of representations:

1. **Last Token Activation:** The residual stream state at the final token of the prompt.
2. **Mean-Pooled Activation:** The average of residual stream states across all non-padding tokens in the sequence.

3.3 Direction Identification

Following Marks and Tegmark [2023] and Arditi et al. [2024], we use the **Difference-in-Means** method to identify the candidate "AI direction" $\boldsymbol{v}_l$ at each layer l:

$$\boldsymbol{v}_l = \frac{1}{|\mathcal{D}_{AI}|} \sum_{x \in \mathcal{D}_{AI}} \boldsymbol{a}_l(x) - \frac{1}{|\mathcal{D}_{Human}|} \sum_{x \in \mathcal{D}_{Human}} \boldsymbol{a}_l(x) \tag{1}$$

where $\mathcal{D}_{AI}$ and $\mathcal{D}_{Human}$ are the sets of activations for AI-generated and human-written samples, respectively. This vector $\boldsymbol{v}_l$ represents the average shift in representation space associated with the AI stylistic quality.

3.4 Linear Probing

To evaluate the separability of the two classes, we train logistic regression probes for each layer. We use a 5-fold cross-validation scheme to measure the accuracy of a linear classifier in distinguishing AI from human activations. This serves as a measure of how much "AI-style" information is linearly decodable at each stage of the model's processing.

3.5 Activation Steering

To prove the causal relevance of the identified direction, we intervene during the generation process. We implement a forward hook at layer $l = 14$ (identified as a middle layer where stylistic features are often well-formed). During generation, we modify the residual stream $\boldsymbol{a}_{14}$ at each token position:

$$\boldsymbol{a}'_{14} = \boldsymbol{a}_{14} + \alpha \cdot \frac{\boldsymbol{v}_{14}}{||\boldsymbol{v}_{14}||} \tag{2}$$

where α is the steering coefficient. We test three conditions: neutral ($\alpha = 0$), subtract AI ($\alpha = -5.0$), and add AI ($\alpha = 5.0$). We evaluate the generated outputs by feeding them back into the linear probe trained in section 3.

4 Results

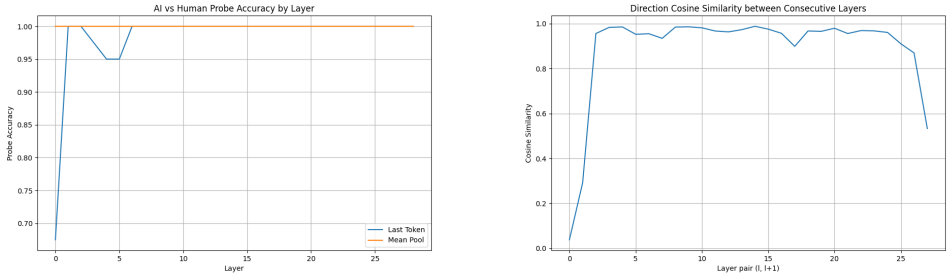
Our experiments reveal that the stylistic distinction between AI and human text is deeply embedded and linearly accessible within the QWEN2.5-1.5B model.

4.1 Linear Separability

We first evaluate the accuracy of linear probes in classifying text origin. As shown in table 1, the model achieves perfect separation at very early layers. Specifically, the last-token activations at Layer 1 are sufficient to distinguish between the two classes with 100% accuracy.

Table 1: Probe accuracy for distinguishing AI vs. human text. The perfect accuracy at early layers suggests that stylistic markers are encoded immediately in the residual stream.

Feature Type	Layer	Accuracy
Last Token	1	**1.00**
Mean Pooled	14	**1.00**



(a) Probe Accuracy

(b) Cosine Similarity

Figure 1: Probe accuracy across layers *(a)*and cosine similarity between identified directions across layers *(b)*. The high accuracy and similarity in middle layers indicate a stable stylistic representation.

What explains the perfect early-layer accuracy? A qualitative inspection of the RAID dataset reveals that AI-generated abstracts frequently begin with specific formulaic markers (e.g., "In this paper,"), while human abstracts exhibit greater diversity in their opening sentences. The near-perfect accuracy at Layer 1 indicates that the model's embeddings and initial attention layers immediately capture these distributional differences.

4.2 Causal Steering

To verify that the identified direction $\boldsymbol{v}_{14}$ represents a stylistic axis rather than a mere correlation, we perform activation steering during text generation. We prompt the model to complete the sentence "The recent advancements..." and observe the change in stylistic quality.

Table 2: Steering evaluation at Layer 14. We report the AI probability assigned by the linear probe to the generated completions.

Prompt	Steering Coefficient (α)	AI Probability
"The recent advancements..."	-5.0 (Subtract AI)	**0.0051**
"The recent advancements..."	0.0 (Neutral)	0.3622
"The recent advancements..."	$+5.0$ (Add AI)	**0.6713**

As shown in table 2, manipulating the activations along $\boldsymbol{v}_{14}$ significantly shifts the stylistic profile of the output. Subtracting the AI direction effectively "humanizes" the output, reducing the probe's AI probability from 36.2% to a negligible 0.5%. Conversely, adding the direction increases the probability to 67.1%.

Qualitative Analysis. Beyond the numerical shift in probe scores, we observe a distinct change in the model's linguistic choices. When the AI direction is subtracted ($\alpha = -5.0$), the model adopts a more personal and subjective tone, using phrases such as "we have now collected" and "impossible to understand." In contrast, adding the AI direction ($\alpha = +5.0$) pushes the model toward a more technical and "corporate" style, with completions featuring phrases like "integration of machine learning" and "making it easier than ever." This suggests that the direction captures high-level stylistic features beyond simple token frequencies.

4.3 Internal Structure

We analyze the consistency of the AI direction across layers using cosine similarity. We find that the directions are highly stable in the middle-to-late layers, indicating a consistent internal representation of style as the model processes the sequence. The probe accuracy remains high across almost all layers, supporting the hypothesis that the residual stream maintains stylistic context throughout the computation.

5 Discussion

Our results provide strong empirical support for the Linear Representation Hypothesis (LRH) as applied to the abstract concept of stylistic origin. The fact that a single linear direction can both detect and steer the "AI-ness" of a text suggests that LLMs organize their latent space along interpretable axes of variation, even for properties as nuanced as style.

Mechanistic vs. Surface-Level Features. A question is whether the identified direction captures deep stylistic patterns or surface-level artifacts. The perfect accuracy at Layer 1 and the identified "In this paper" prefix suggest that, for the RAID dataset, the model identifies text origin largely through initial token distributions. However, the successful steering at Layer 14—which shifts complex linguistic traits like tone and vocabulary choice—indicates that this initial identification propagates into a more semantic stylistic representation in the middle layers. This hierarchical encoding, from surface features to stylistic abstractions, is consistent with previous observations of transformer behavior.

Limitations and Robustness. This study has several limitations. First, the perfect separability at Layer 1 suggests that our candidate direction may be overfitted to the specific prefix distributions of the RAID dataset. While steering proved effective, the direction might not generalize to human text that mimics AI prefixes or vice versa. Second, our analysis was restricted to a single model (QWEN2.5-1.5B) and a single domain (scientific abstracts). Further research is needed to determine if the "AI direction" is universal across domains (e.g., creative writing, news) and if directions found in one model can detect text generated by another.

Broader Implications. The existence of a controllable stylistic axis has significant implications for AI safety and governance. For AI detection, it suggests that internal activation monitoring could be a more reliable signal than external black-box classifiers, as it taps into the model's own internal categorization of its output. For model alignment, it provides a tool for adjusting model persona without retraining, allowing developers to fine-tune the degree of anthropomorphism to suit specific applications.

How can we find a more robust direction? Future work should involve adversarial data augmentation—specifically, stripping common prefixes and forcing the model to distinguish between classes based on deeper semantic and syntactic structures. Additionally, exploring the cross-model consistency of these directions would be a crucial step toward a universal theory of AI stylistic representation.

6 Conclusion

In this study, we investigated the internal representation of "AI-style" in the residual stream of the QWEN2.5-1.5B model. By applying Representation Engineering techniques to the RAID dataset, we identified a linear direction that captures the stylistic quality of being AI-generated. Our results demonstrate that this quality is highly linearly separable, achieving 100% probe accuracy at early layers. More importantly, we showed that this direction is causally relevant: steering model activations along this axis allows for the precise manipulation of the model's stylistic output, shifting it from a personal, human-like tone to a formal, technical tone.

Our findings contribute to the growing body of evidence for the Linear Representation Hypothesis and suggest that LLMs possess internal stylistic axes that can be precisely controlled. This work opens new avenues for mechanistic AI detection and for the fine-grained control of model personas. Future research will focus on the universality of these directions across different model architectures and diverse writing domains.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

Liam Dugan, Stephanie Milani, Shuo Pan, Daphne Ippolito, Sunil Jauhar, Chris Callison-Burch, and Brian Roark. RAID: A corpus for robust AI-generated text detection. In *Proceedings of ACL*, 2024.

Andrew Lee, Benjamin Newman, and Sophie Xia. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity. *arXiv preprint arXiv:2401.01967*, 2024.

Kenneth Li, Oam Patel, Fernanda Vi'egas, Martin Wattenberg, and Neel Nanda. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*, 2023.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Charles O'Neill, Slava Chalnev, Chi Chi Zhao, Max Kirkby, and Mudith Jayasekara. A single direction of truth: An observer model's linear residual probe exposes and steers contextual hallucinations. *arXiv preprint arXiv:2507.23221*, 2025.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Luke Richard, Matt Theobald, Kevin Campion, and Nick Rogers. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.