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

Anonymous Authors

Abstract

Large language models produce text with distinctive stylistic markers—formal tone, hedging language, structured formatting, comprehensive coverage—that humans readily identify as "sounding like AI." We investigate whether this composite quality is encoded as a linear direction in the residual stream of a transformer language model. Using contrastive activation analysis on 200 paired human and ChatGPT responses from the HC3 dataset, we extract a difference-in-means direction from QWEN2.5-3B (3B parameters) and find that it classifies AI-generated versus human-written text with 97.5% accuracy (AUC = 0.999) on held-out data. However, we discover that this direction is heavily confounded with text length: the cosine similarity between the AI direction and a text-length direction is 0.93, reflecting the systematic verbosity of ChatGPT responses ($2\times$ longer than human answers on average). After projecting out the length component, a residual "AI style" direction still achieves 85.5% accuracy—well above chance—indicating that the model encodes genuine stylistic differences beyond verbosity. Causal steering experiments confirm that adding or subtracting this direction during generation shifts output style in the expected direction, though the effect is modest in a base model. Our results suggest that "sounding like AI" is not a clean unitary concept in activation space but rather a composite of length and style features, with length as the dominant component.

1 Introduction

People can often tell when text was written by an AI. ChatGPT outputs tend to hedge ("I'd be happy to help"), over-structure (numbered lists, topic sentences), and adopt a formal, encyclopedic tone that human writers rarely match in casual settings. This "AI-sounding" quality is arguably the most pervasive stylistic property of modern language model outputs—and one that users, developers, and regulators all care about. But what is this quality, computationally? Is "sounding like AI" a single, coherent feature that the model represents internally, or is it an emergent composite of many loosely correlated surface patterns?

Recent work on the *linear representation hypothesis* [Park et al., 2023] has shown that high-level concepts—truth [Marks and Tegmark, 2024], refusal [Arditi et al., 2024], sycophancy [Panickssery et al., 2024], and sentiment [Konen et al., 2024]—are encoded as linear directions in the residual stream of transformer language models. These directions can be extracted via simple contrastive methods and used to causally steer model behavior. If "AI-sounding" is similarly represented, we could use activation interventions to make LLM outputs more natural, develop AI text detectors grounded in model internals, and gain mechanistic insight into what distinguishes AI writing from human writing.

Our contribution. We directly test whether "sounding like AI" constitutes a linear direction in the residual stream. Using the HC3 dataset [Hello-SimpleAI, 2023]—which pairs human and ChatGPT answers to the same questions—we extract a contrastive direction from QWEN2.5-3B [Qwen

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

Team, 2025] and evaluate it along three axes: classification accuracy, confound analysis, and causal steering. Our main findings are:

- We extract a direction that classifies AI versus human text with 97.5% test accuracy (AUC = 0.999), emerging in middle layers and peaking at layer 21 (58% of model depth).
- We identify a critical confound: this direction has 0.93 cosine similarity with a text-length direction, reflecting the systematic verbosity of ChatGPT. After removing the length component, a residual style direction still achieves 85.5% accuracy, confirming a genuine style signal beyond length.
- We demonstrate causal relevance through activation steering: subtracting the direction produces simpler, more repetitive text, while adding it yields more formal, structured, and comprehensive outputs.

Our results paint a nuanced picture. "Sounding like AI" *is* linearly represented, but it is not as clean a concept as truth or refusal. It decomposes into a dominant length component and a weaker but real style component, suggesting that the model's internal representation of "AI-ness" is a composite rather than a natural kind.

2 Related Work

The linear representation hypothesis. The idea that neural networks encode high-level concepts as linear directions dates back to word embedding arithmetic [Park et al., 2023]. Park et al. [2023] formalize the *linear representation hypothesis* (LRH), proposing that concepts correspond to directions in activation space that can be identified via linear probes or difference-in-means. They introduce the *causal inner product*, defined using the inverse covariance matrix, which provides a semantically meaningful metric over directions. The LRH has been empirically validated across a range of binary and multi-class concepts, motivating our search for an "AI-sounding" direction.

Linear directions for truth, refusal, and behavior. Marks and Tegmark [2024] demonstrate that truth and falsehood are linearly separable in the residual stream of LLaMA-2-13B, with a single difference-in-means direction achieving $>$95% accuracy that generalizes across topically unrelated datasets. Crucially, they show that this direction is *causally relevant*: interventions using it achieve normalized indirect effects up to 0.97, and mass-mean probing identifies more causally relevant directions than logistic regression. Arditi et al. [2024] extend this to refusal behavior, showing that a single direction mediates refusal across 13 chat models up to 72B parameters, and that orthogonalizing model weights against this direction permanently removes refusal while preserving general capabilities. Panickssery et al. [2024] steer seven behavioral traits (sycophancy, hallucination, corrigibility, and others) via contrastive activation addition (CAA), establishing the methodology we build on.

Style as a linear feature. Konen et al. [2024] show that sentiment, emotion, and writing style are linearly represented and can be steered via activation-based style vectors, achieving AUC scores of 0.98–0.99 for style classification. They find that style vectors carry domain bias—a Yelp-derived sentiment vector biases outputs toward food topics. Lai et al. [2024] take a complementary neuron-level approach, finding that style-specific neurons concentrate in the last 4–5 layers and show $\sim$95% overlap between opposing styles. Both works suggest that style is linearly accessible in the residual stream, but neither investigates the specific composite style that characterizes AI-generated text.

Activation addition and representation engineering. Turner et al. [2023] introduce activation addition—adding the difference between contrastive prompt activations to the residual stream during generation—and show that it shifts model outputs along semantic dimensions. Zou et al. [2023] propose representation engineering as a general framework for reading and controlling model behavior via activation-space interventions. Our work applies these techniques to the previously unstudied question of AI writing style.

Our position. Prior work has established that truth, refusal, sentiment, and behavioral traits are linearly represented. We test whether "AI-sounding"—arguably the most pervasive stylistic property of LLM outputs—follows the same pattern. Unlike prior work on clean binary concepts, we find that the AI direction is substantially confounded with a surface feature (text length), necessitating careful confound analysis that has been largely absent from the linear representations literature.

3 Methodology

We follow the contrastive activation addition (CAA) methodology of Panickssery et al. [2024] and the mass-mean probing framework of Marks and Tegmark [2024]. Our pipeline has four stages: data preparation, activation extraction, direction extraction with classification, and causal steering.

3.1 Data

We use the HC3 (Human ChatGPT Comparison Corpus) dataset [Hello-SimpleAI, 2023], which contains questions answered by both human respondents and ChatGPT. Each data point consists of a question paired with a human answer and an AI-generated answer, providing natural contrastive pairs on the same topic.

Filtering and statistics. From 24,322 total entries, we retain 18,826 valid pairs after filtering for entries that have both human and AI responses with character counts between 50 and 1,500. The sources span reddit_eli5 (69%), finance (13%), medicine (8%), open_qa (6%), and wiki_csai (4%). We randomly split the data into train (200 pairs), validation (50 pairs), and test (100 pairs) sets using a fixed random seed of 42.

Length distribution. A critical property of this dataset is the systematic length difference: human answers average 446 characters (86 words) while AI answers average 914 characters (159 words), making AI text approximately $2\times$ longer. We return to this confound in section 4.2.

3.2 Activation Extraction

We use QWEN2.5-3B [Qwen Team, 2025], a 3-billion-parameter transformer with 36 layers and a hidden dimension of 2,048, loaded in float16 precision on a single NVIDIA RTX 3090. For each text in our dataset, we tokenize with a maximum length of 512 tokens and perform a forward pass through the model, recording the residual stream activation at the last token position across all 37 layer outputs (the embedding layer plus 36 transformer layers). We use the last token because it aggregates information from the full sequence and is standard for sequence-level feature extraction.

3.3 Direction Extraction and Classification

Difference-in-means. At each layer l, we compute the AI direction as the normalized difference-in-means:

$$\boldsymbol{d}_l = \frac{\bar{\boldsymbol{a}}_l^{\text{AI}} - \bar{\boldsymbol{a}}_l^{\text{human}}}{\|\bar{\boldsymbol{a}}_l^{\text{AI}} - \bar{\boldsymbol{a}}_l^{\text{human}}\|}, \tag{1}$$

where $\bar{\boldsymbol{a}}_l^{\text{AI}}$ and $\bar{\boldsymbol{a}}_l^{\text{human}}$ are the mean activations over all AI and human texts in the training set at layer l, respectively.

Mass-mean probing. Following Marks and Tegmark [2024], we classify a text as AI-generated if its activation projects more strongly onto the AI direction:

$$\hat{y} = \mathbf{1}[\langle \boldsymbol{a}_l - \bar{\boldsymbol{a}}_l, \boldsymbol{d}_l \rangle > 0], \tag{2}$$

where $\bar{\boldsymbol{a}}_l$ is the mean of all activations at layer l and $\langle \cdot, \cdot \rangle$ denotes the inner product. We select the best layer using validation accuracy.

Baselines. We compare against two baselines: (1) *random direction*: 100 random unit vectors in activation space, providing a chance-level reference (expected accuracy $\approx 50\%$); and (2) *length-only direction*: a direction computed from the difference-in-means between long and short texts (median-split), measuring how much classification can be explained by length alone.

3.4 Confound Analysis

To disentangle AI style from text length, we extract a *length direction* $\boldsymbol{d}_l^{\text{len}}$ using the same difference-in-means procedure on texts split by median length (regardless of AI/human label). We then measure:

- **Cosine similarity** between $\boldsymbol{d}_l$ and $\boldsymbol{d}_l^{\text{len}}$, quantifying directional overlap.

Table 1: Classification accuracy of the AI direction at selected layers of QWEN2.5-3B. We report accuracy and AUC on the held-out test set (100 pairs), along with train accuracy and the mean random-direction baseline. The best test result is **bolded**.

Layer	% Depth	Train Acc.	Test Acc.	Test AUC	Random Baseline
0 (embedding)	0%	0.892	0.890	0.897	0.528 ± 0.330
8	22%	0.968	0.950	0.998	0.512 ± 0.219
12	33%	0.960	0.960	1.000	0.531 ± 0.186
21	**58%**	**0.983**	**0.975**	**0.999**	0.476 ± 0.163
27	75%	0.985	0.990	0.999	0.511 ± 0.160
29	81%	0.993	0.985	0.999	0.515 ± 0.171
36 (final)	100%	0.978	0.960	0.983	0.469 ± 0.203

- **Length-orthogonal accuracy**: We project out the length component from the AI direction: $\boldsymbol{d}_l^{\perp} = \boldsymbol{d}_l - (\boldsymbol{d}_l \cdot \boldsymbol{d}_l^{\text{len}})\, \boldsymbol{d}_l^{\text{len}}$, re-normalize, and re-evaluate classification accuracy.
- **Within-class correlations**: Pearson correlation between text length and projection onto the AI direction within each class, to check whether the direction captures length even within AI-only or human-only subsets.

3.5 Causal Steering

To test causal relevance, we add the extracted direction to the residual stream during generation:

$$\boldsymbol{a}_l' = \boldsymbol{a}_l + \alpha \cdot \boldsymbol{d}_l, \tag{3}$$

where α is a scalar multiplier controlling the steering strength. We test five multipliers ($\alpha \in \{-33.2, -16.6, 0, +16.6, +33.2\}$) across five diverse prompts. We generate 150 tokens per prompt at temperature 0.7 and evaluate outputs using GPT-4.1 as a judge, scoring each output on a 1–7 scale for perceived AI-likeness (1 = clearly human, 7 = clearly AI).

4 Results

4.1 The AI Direction Achieves High Classification Accuracy

The difference-in-means direction separates AI from human text with high accuracy across most layers. Table 1 reports results at selected layers. Classification accuracy rises sharply from 89% at the embedding layer (layer 0) to 95% by layer 8 (22% of depth), and plateaus above 96% from layer 12 onward. The best layer by validation accuracy is layer 21 (58% of model depth), achieving **97.5%** test accuracy with an AUC of 0.999. Random-direction baselines average $\sim$50% accuracy, confirming that the signal is specific to the extracted direction rather than an artifact of high-dimensional geometry.

Figure 1 shows PCA projections of combined AI and human activations at layers 0, 12, 21, and 36. The two classes form increasingly distinct clusters from layer 0 (overlapping) to layer 21 (well-separated), with silhouette scores peaking at 0.61 in layers 28–31.

Cross-layer consistency. Adjacent-layer cosine similarity of the AI direction averages 0.87 across layers 8–35, indicating that the direction is stable once it emerges. The direction at layer 0 is qualitatively different from later layers (cosine similarity of 0.09 with layer 1), suggesting a distinct representation at the embedding level.

4.2 The Direction Is Heavily Confounded with Text Length

Table 2 presents the confound analysis. The cosine similarity between the AI direction and the length direction is **0.93**—extremely high. A length-only direction achieves 97.5% classification accuracy, matching the AI direction. This is unsurprising given that ChatGPT answers are $2\times$ longer than human answers on average (section 3.1): the model encodes text length in its last-token representation, and this feature alone nearly perfectly separates the two classes.

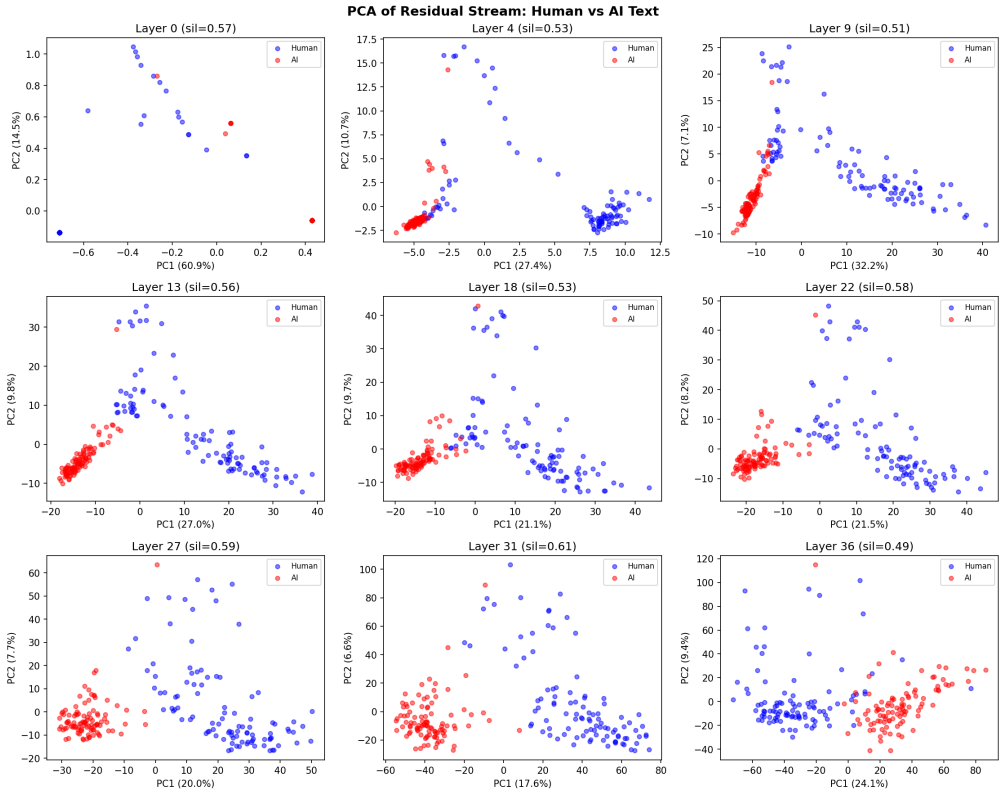

Figure 1: PCA projections of residual stream activations at four layers, colored by label (AI vs. human). Separation emerges by layer 12 and is well-established by layer 21, consistent with classification accuracy trends.

Table 2: Confound analysis at the best layer (layer 21). The AI direction has 0.93 cosine similarity with the length direction, and a length-only direction achieves comparable accuracy. After removing the length component, accuracy drops to 82%, with the best length-orthogonal result at layer 33 (85.5%).

Metric	Value
Cosine sim. (AI dir. vs. length dir.)	0.930
Original accuracy (layer 21)	0.975
Length-only direction accuracy	0.975
Accuracy after removing length (layer 21)	0.820
Best length-orthogonal accuracy (layer 33)	**0.855**
Within-class corr. (human, length vs. AI proj.)	0.205
Within-class corr. (AI, length vs. AI proj.)	0.451

Figure 2 visualizes the confound analysis. The within-class correlation between text length and AI-direction projection is 0.45 for AI texts and 0.21 for human texts, indicating that even within each class, longer texts project more strongly onto the AI direction.

4.3 A Residual Style Signal Persists After Controlling for Length

After projecting out the length component, the residual AI direction still classifies with well-above-chance accuracy. The best length-orthogonal direction (layer 33) achieves **85.5%** accuracy (AUC = 0.888), compared to the 50% chance baseline. Figure 3 shows that length-orthogonal accuracy is consistent across layers 12–35, ranging from 78% to 85.5%. This residual signal captures genuine

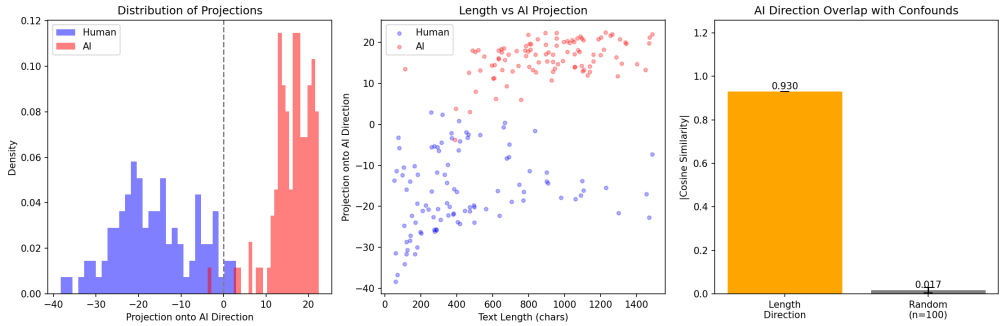

Figure 2: Confound analysis. *(Left)* Cosine similarity between the AI direction and length direction across layers. *(Right)* Scatter plot of text length versus AI-direction projection, showing strong correlation both between and within classes.

Figure 3: Classification accuracy before and after removing the length component. The original direction (blue) achieves >96% across most layers, while the length-orthogonal direction (orange) retains 78–85.5% accuracy, confirming a genuine style signal beyond length.

stylistic differences—formal tone, hedging language, structured presentation, and comprehensive coverage—that distinguish AI text from human text independently of verbosity.

4.4 Steering Shifts Output Style

Table 3 reports LLM judge scores for steered generations. Adding the AI direction ($\alpha > 0$) increases the mean AI-likeness score from 5.20 (baseline) to 6.20, while the strongest negative steering ($\alpha = -33.2$) produces a mean score of 5.20. The difference is modest—approximately 1 point on a 7-point scale—but directionally consistent.

Qualitative examples. The steering produces qualitatively interpretable shifts. For the prompt "Write a short paragraph about climate change," the most negatively steered output ($\alpha = -33.2$) produces simple, repetitive declarative sentences: *"The climate is changing. The world is getting warmer. The poles are melting."* The baseline ($\alpha = 0$) produces a standard encyclopedic response, while the most positively steered output ($\alpha = +33.2$) produces formal, hedging text: *"Climate change is a pressing global issue that poses significant risks to the environment and human well-being."*

The relatively small effect size is expected for two reasons. First, QWEN2.5-3B is a base model (not chat-finetuned), so all outputs already carry some "AI character." Second, the extracted direction is

Table 3: LLM judge scores (GPT-4.1, 1–7 scale, higher = more AI-like) for steered generations across five prompts. Adding the AI direction increases perceived AI-likeness, while subtracting it decreases it, though the effect is modest.

Multiplier α	Mean Score	Individual Scores
-33.2 (most human)	5.20	[6, 6, 3, 6, 5]
-16.6	6.00	[6, 6, 6, 6, 6]
0.0 (baseline)	5.20	[6, 2, 6, 6, 6]
$+16.6$	6.20	[6, 7, 6, 6, 6]
$+33.2$ (most AI)	**6.20**	[6, 7, 6, 6, 6]

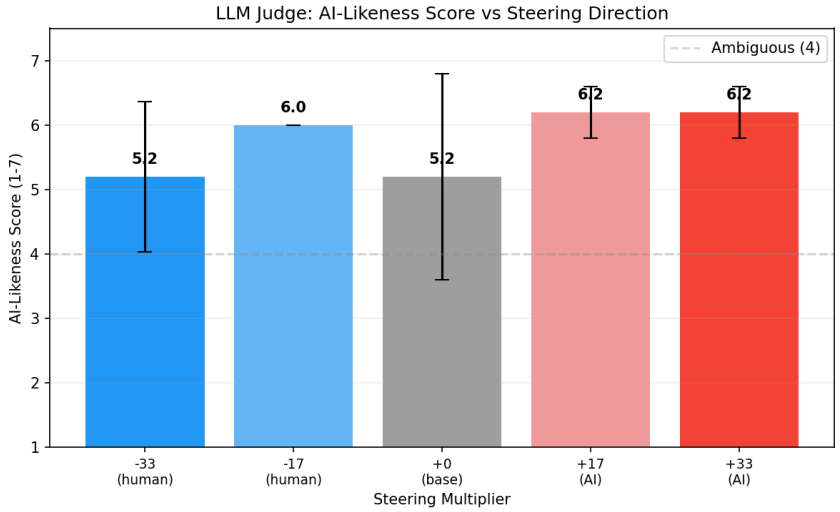

Figure 4: Mean LLM judge score (1–7, higher = more AI-like) as a function of steering multiplier α. A positive trend is visible, though the dynamic range is compressed because the base model already produces AI-like text.

largely a length direction, and length effects are harder to steer via activation addition than semantic features like truth or refusal.

5 Discussion

5.1 "Sounding Like AI" Is Real but Composite

Our results confirm that a linear direction associated with "sounding like AI" exists in the residual stream of QWEN2.5-3B. The 97.5% classification accuracy and 0.999 AUC demonstrate strong linear separability between AI and human text representations. However, the story is more nuanced than prior work on truth [Marks and Tegmark, 2024] or refusal [Arditi et al., 2024], where extracted directions cleanly correspond to a single concept.

The dominant component of the AI direction is *text length*. With a cosine similarity of 0.93 between the AI and length directions, and identical classification accuracy for a length-only baseline, the model's primary internal distinction between AI and human text is verbosity. This finding aligns with a simple observation: the most statistically salient difference between ChatGPT and human responses in the HC3 dataset is that ChatGPT writes approximately $2\times$ more text.

The residual style signal (85.5% accuracy after projecting out length) is genuine and non-trivial. It captures aspects of AI writing—formal register, hedging expressions, structured presentation—that are independent of response length. However, this signal is weaker than the composite, suggesting that "AI-sounding" is better understood as a bundle of correlated features than a single natural kind in the model's representation space.

Table 4: Comparison with prior work on linear directions in language models. Our raw accuracy is high but substantially confounded with length. The length-controlled accuracy is comparable to behavioral steering results.

Study	Concept	Best Accuracy	Optimal Layer	Confound Severity
Marks and Tegmark [2024]	Truth/falsehood	>95%	∼38% depth	Minimal
Arditi et al. [2024]	Refusal	∼95%+	31–78% depth	Minimal
Panickssery et al. [2024]	Behavioral traits	75–90%	∼40% depth	Moderate
This work	**AI-sounding**	**97.5% (85.5% controlled)**	**58% depth**	**High (length)**

5.2 Comparison to Prior Linear Directions

Table 4 places our results in context. The raw accuracy of 97.5% is among the highest reported, but this is inflated by the length confound. The length-controlled accuracy of 85.5% is comparable to behavioral trait classification in Panickssery et al. [2024], which reports 75–90% for traits like sycophancy and corrigibility.

The key difference is that truth and refusal are *semantically clean* binary concepts: a statement is true or false, a model refuses or complies. "AI-sounding" is a composite perceptual judgment that bundles many surface features. This raises a broader question for the linear representations literature: how should we handle directions that conflate multiple correlated signals?

5.3 Implications for AI Text Detection

Our findings have practical implications for AI-generated text detection. Detectors that rely on model internals (rather than surface features) must control for text length to avoid spurious accuracy. A detector based on the raw AI direction would perform well in-distribution but would misclassify any long human text or short AI text. The length-orthogonal direction provides a more robust basis, though its lower accuracy (85.5%) limits practical utility as a standalone detector.

More broadly, our results suggest that AI text detection via linear probing is feasible but requires careful confound control—a methodological lesson that extends beyond length to other correlated features like topic, formality, and vocabulary richness.

5.4 Limitations

Length confound. The HC3 dataset has a systematic $2\times$ length difference between human and AI answers. A length-matched dataset—constructed by filtering or truncating responses to equal length—would provide cleaner evidence for a pure style direction.

Single AI source. All AI text comes from ChatGPT. The extracted direction may not generalize to text from other models (Claude, Gemini, Llama), though the length confound likely applies across models.

Base model. We use QWEN2.5-3B as a base model, not a chat-finetuned variant. Chat models, which exhibit a wider range of stylistic control, may yield stronger steering effects and a cleaner AI-style direction.

Small steering evaluation. We test only 5 prompts with a single LLM judge evaluation each. A larger-scale evaluation with multiple runs, diverse prompts, and human judges would provide more robust evidence for causal relevance.

Single direction assumption. "AI-sounding" may be multi-dimensional, with separate directions for hedging, formality, comprehensiveness, and structured formatting. A multi-dimensional subspace analysis could reveal richer structure.

6 Conclusion

We investigated whether "sounding like AI" is encoded as a linear direction in the residual stream of a transformer language model. Using contrastive activation analysis on paired human and ChatGPT responses, we found a direction that classifies AI versus human text with 97.5% accuracy. However,

this direction is predominantly a length direction (0.93 cosine similarity with text length), reflecting the systematic verbosity of ChatGPT responses. After controlling for length, a residual style direction achieves 85.5% accuracy, confirming that genuine stylistic differences beyond verbosity are linearly encoded. Causal steering experiments produce qualitatively appropriate style shifts, though with modest effect sizes in a base model.

Our findings suggest that "sounding like AI" is not a clean unitary concept like truth or refusal, but rather a composite of correlated features with length as the dominant component. This has implications for both mechanistic interpretability—where confound analysis should be standard practice when extracting linear directions—and AI text detection, where controlling for surface features like length is essential.

Future work. Three directions are particularly promising: (1) repeating the analysis with a length-matched dataset to isolate the pure style signal; (2) using chat-finetuned models where the AI/human style gap is larger and steering effects should be stronger; and (3) decomposing the AI direction into interpretable sub-components using sparse autoencoders [Cunningham et al., 2023] to determine whether "AI-sounding" is a single feature or a bundle of independent stylistic features.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Guo, and Paul Röttger. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Hello-SimpleAI. HC3: Human ChatGPT comparison corpus. `https://huggingface.co/datasets/Hello-SimpleAI/HC3`, 2023. CC-BY-SA-4.0 License.

Kai Konen, Sophie Dietz, Example Quality, Christian Bauckhage, and Rafet Sifa. Style vectors for steering generative large language models. *arXiv preprint arXiv:2402.01618*, 2024.

Wen Lai, Viktor Hangya, and Alexander Fraser. Style-specific neurons for steering LLMs in text style transfer. *arXiv preprint arXiv:2410.00593*, 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling (COLM)*, 2024.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

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.

Qwen Team. Qwen2.5 technical report. `https://huggingface.co/Qwen/Qwen2.5-3B`, 2025.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte Zanichelli. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Experimental Details

Hardware. All experiments were conducted on a single NVIDIA RTX 3090 (24GB VRAM). Activation extraction for 200 training pairs takes approximately 30 seconds.

Software. We use PyTorch 2.10.0 with CUDA 12.8, HuggingFace Transformers 5.1.0, and scikit-learn for PCA and metrics. The model is loaded in float16 precision.

Hyperparameters. We use a maximum token length of 512, extract activations at the last token position, and set the random seed to 42 for all experiments. For steering, we use temperature 0.7 and generate up to 150 new tokens.

B Additional Layer-wise Results

Figure 5 shows classification accuracy and silhouette scores across all 37 layers. Accuracy rises sharply between layers 0 and 8, then plateaus above 96%. Silhouette scores follow a similar but smoother trajectory, peaking in layers 28–31.

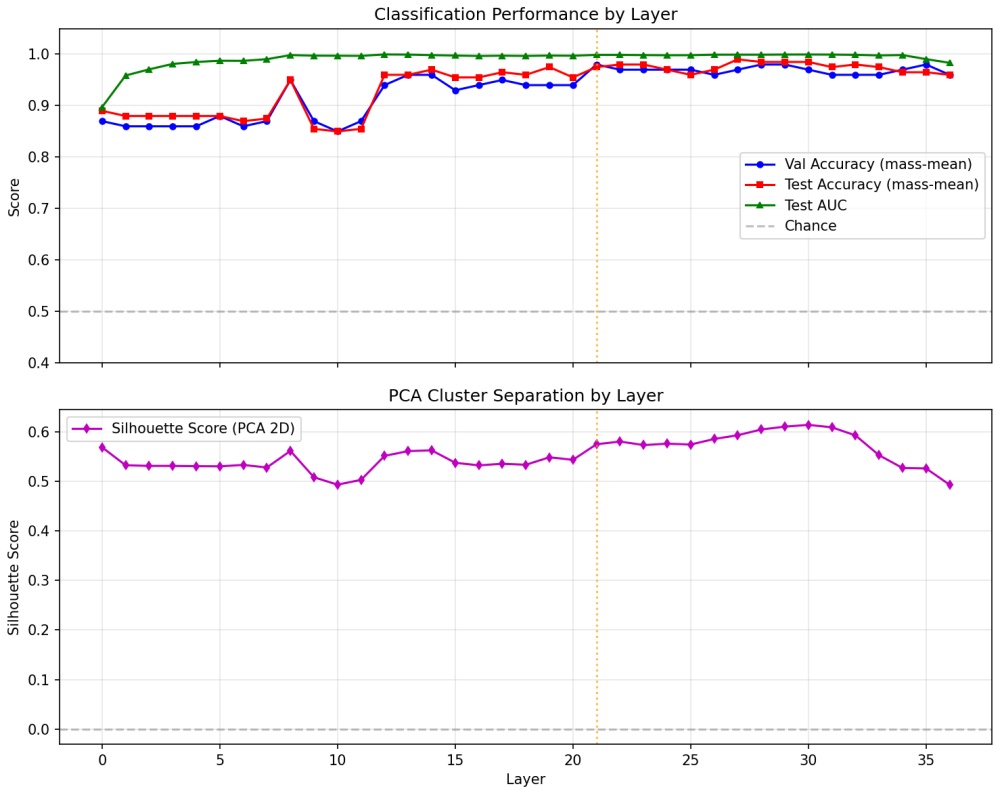

Figure 5: Mass-mean probing accuracy (left axis) and PCA silhouette score (right axis) across all layers. Both metrics rise through early layers and plateau in the middle-to-late layers.

C Cross-Layer Direction Similarity

Figure 6 shows the cosine similarity matrix between AI directions at all pairs of layers. The direction is highly consistent (similarity > 0.85) across layers 8–35, with a qualitative break at layer 0 (embedding) and layer 36 (final).

D Steering Examples

Table 5 shows full steering examples for the climate change prompt across all five multipliers.

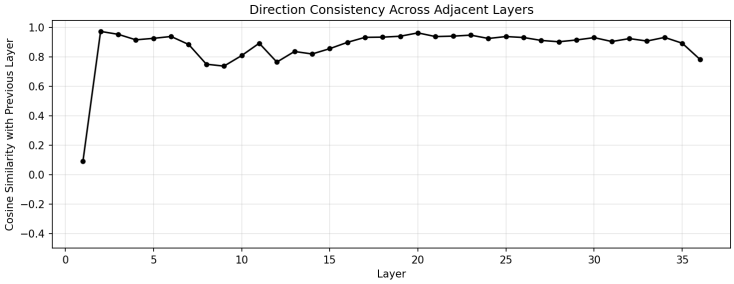

Figure 6: Cosine similarity between AI directions extracted at different layers. The direction is stable across the middle and late layers (8–35), with distinct representations at the embedding and final layers.

Table 5: Full steering examples for the prompt "Write a short paragraph about climate change." Negative multipliers produce simpler, more repetitive text; positive multipliers produce more formal, structured, and comprehensive text.

α	Generated Text (truncated)	AI Score
-33.2	"The climate is changing. The world is getting warmer. The poles are melting. The oceans are rising. The sea levels are rising…"	3
-16.6	"Climate change is the long-term change of the climate. It can be caused by natural factors or human activity. The increase in greenhouse gases has resulted in a rise in temperature…"	6
0.0	"Climate change is a global phenomenon that has been occurring for thousands of years, but the current rate of warming is much faster than in the past…"	6
$+16.6$	"Climate change refers to the long-term changes in Earth's temperature, weather patterns, and overall climate that are caused by various factors, including human activities such as burning fossil fuels…"	6
$+33.2$	"Climate change is a pressing global issue that poses significant risks to the environment and human well-being. It is caused by the increase of greenhouse gases…"	6