Anisotropy and Cross-Model Comparison of Released Natural Language Autoencoders

Sidar Aslanoglu
Johns Hopkins SAIS
Zenodo DOI: 10.5281/zenodo.20112029

May 2026

Abstract

Fraser-Taliente et al. (2026) recently introduced Natural Language Autoencoders (NLAs), an interpretability tool that verbalizes large language model activations into human-readable explanations. Trained NLA checkpoints are now available for four base-model families, which invites cross-model comparison. We show that the obvious comparison, based on the released reconstruction-MSE metric, is not meaningful as stated. The cause is residual-stream anisotropy. On 500 inputs spanning five tiers, the mean pairwise cosine similarity between gold activations is 0.497 for Qwen-2.5-7B at layer 20 and 0.975 for Gemma-3-12B at layer 32. The 0.478 gap in baseline cosine means Gemma starts from a much higher reconstruction-metric baseline than Qwen, before reconstruction quality is measured. Applying the anisotropy-baseline correction from Ethayarajh (2019) reverses the cross-model conclusion that raw MSE supports: under the corrected metric, the smaller Qwen-7B NLA carries roughly an order of magnitude more per-row reconstruction information than the larger Gemma-12B NLA. We also document a structural-versus-semantic split in the adversarial tier that replicates across both models, moderate cross-model rank agreement (Spearman $\rho = 0.354$), and a null result from a chain-of-thought unfaithfulness probe: neither model exhibited the bias-induced behavior the probe was designed to detect.

1 Introduction

State-of-the-art open-source large language models (LLMs) offer capabilities that match or closely approach closed-weight frontier models on a number of dimensions, including general knowledge, mathematical reasoning, and code generation (LMSYS Org, 2024). Beyond their near-parity capabilities, open-source models allow any user with sufficient compute to fine-tune and audit them, which makes cross-model comparison a practical question for researchers.

Activation-level interpretability tools help auditors look inside a model by making the vectors that flow through its layers legible. Sparse autoencoders (Bricken et al., 2023; Templeton et al., 2024), attribution graphs (Lindsey et al., 2025), and supervised probes are the standard tools. Fraser-Taliente et al. (2026) introduced a new one: Natural Language Autoencoders (NLAs). An NLA has two parts. A verbalizer maps an activation to a text description, and a reconstructor maps the description back into an activation. The two are trained together to minimize the round-trip reconstruction error, the difference between the original activation and the reconstructed one.

We ask whether faithfulness comparisons across NLAs for different models are well-defined under the reported reconstruction metric (raw mean squared error on normalized activation vectors; Section 2 defines the geometry).

The metric works for tracking a single NLA's training progress but cannot be applied directly to make comparisons across models. If activations from one model are distributed more closely than activations from another, the two models start from different baseline similarities even before reconstruction quality is measured. Cosine on its own cannot tell us whether a high score reflects NLA quality or favorable geometry.

On 500 diverse inputs, we measure this baseline directly: mean pairwise cosine similarity between original (gold) activations is 0.497 for Qwen-2.5-7B at layer 20 and 0.975 for Gemma-3-12B at layer 32, a substantial gap. Applying the standard correction from Ethayarajh (2019) reverses the cross-model conclusion that raw MSE supports.

Our empirical scope is two of the four released NLAs (single-GPU compute). Our methodological scope is narrower: we apply an existing anisotropy correction (Ethayarajh, 2019) to a new tool class. We do not claim the correction itself is novel.

2 Background

A transformer language model processes text by passing a vector of numbers through a stack of layers. Each layer reads the vector, applies a learned transformation, and writes the result back. The vector that gets passed and updated layer by layer is called the model's *residual stream*. At any layer, the residual stream encodes everything the model has computed about the input up to that point. Different inputs produce different vectors; the geometry of those vectors (which directions they point in, how far apart they sit) is what interpretability tools work with. For Qwen-7B this vector has 3,584 dimensions (Qwen Team, 2024); for Gemma-12B, 3,840 (Gemma Team, 2025).

Comparing two such vectors typically uses *cosine similarity*, defined as the cosine of the angle between them, ranging from 1 (same direction) through 0 (orthogonal) to -1 (opposite). After L_2-normalizing (rescaling each vector to length 1), the squared distance between two vectors is $2(1 - \cos)$, which the released NLA codebase reports as "MSE." Since cos ranges from -1 to 1, the metric ranges from 0 (identical direction, perfect reconstruction) to 4 (opposite directions). NLAs are trained to minimize this quantity between a target activation and its reconstruction.

NLAs target the activation-reconstruction objective directly. For the round-trip to succeed, the verbalizer's text output must contain enough information for the reconstructor to recover the input activation vector. The released open-model NLAs were trained at layers roughly two-thirds of the way through their target models (layer 20 of 28 for Qwen-7B, layer 32 of 48 for Gemma-12B). The verbalizer and reconstructor are themselves large language models, started from a copy of the target model and jointly fine-tuned with reinforcement learning against the reconstruction loss. The original paper documents *confabulation* as a primary failure mode, where the verbalizer produces specific claims unsupported by the activation it is reading.

Anisotropy in transformer representations is well-established. The term refers to the tendency of activations to cluster in a narrow region of the vector space rather than spread out uniformly. When activations cluster, the average pairwise cosine similarity between two random activations is high. Ethayarajh (2019) showed this empirically for BERT, ELMo, and GPT-2, and proposed correcting cosine-based metrics by subtracting an empirical baseline equal to the mean pairwise cosine of randomly sampled representations from the same layer. The correction removes the geometry-driven floor from each comparison and leaves a residual that reflects only the information the reconstructor adds beyond chance. Timkey and van Schijndel (2021) and Cai et al. (2021) extended this characterization.

Our behavioral probe draws on Turpin et al. (2023), who showed that *biased* few-shot prompts (sequences of examples designed to push the model toward a particular wrong answer) can shift LLM responses without the influence being acknowledged in the verbalized chain-of-thought.

3 Methodology

3.1 Models, NLA checkpoints, and metrics

We use Anthropic's released NLA checkpoints for Qwen-2.5-7B-Instruct (`kitft/nla-qwen2.5-7b-L20`) and Gemma-3-12B-IT (`kitft/nla-gemma3-12b-L32`). Activations come from the layers each NLA was trained on (layer 20 of 28 for Qwen, layer 32 of 48 for Gemma). All inference runs on a single NVIDIA H100 80GB GPU.

We report three metrics. The first is the raw MSE as defined in Section 2, kept for direct comparability with the released codebase. The second is anisotropy-baseline-corrected cosine (ABC):

$$\mathrm{ABC}(i) = \cos(v_i, \hat{v}_i) - \mathbb{E}_{j \neq k}[\cos(v_j, v_k)]\,, \tag{1}$$

where the expectation runs over all $\binom{500}{2} = 124{,}750$ distinct pairs of gold activations within each model's set, following Ethayarajh (2019). Each model has a different baseline because activations from different models cluster differently. Subtracting that baseline from a reconstruction score gives a number that says how much better the reconstruction did than chance predicts for this model. A high ABC means the reconstructor captured something specific to this input; a low or zero ABC means the reconstructor's output is no closer to the target than two random activations from this model already are. The third is Spearman rank correlation between Qwen and Gemma per-input reconstruction cos, a unit-free measure of whether the two models rank inputs in the same order.

3.2 Tiered input corpus

The corpus contains 500 deterministically chosen items across five tiers. Tier A is 100 Wikipedia- and news-style declarative sentences. Tier B is 100 short Python, JavaScript, SQL, and shell code snippets. Tier C is 100 Turkish-language sentences validated by a native speaker. Tier D is 100 legal contract clauses and formal scientific abstracts. The fifth tier splits into E1 ($n = 30$, structural anomalies: gibberish, repeated characters, control tokens, emoji) and E2 ($n = 70$, semantically adversarial prompts sampled deterministically from JailbreakBench/JBB-Behaviors (Chao et al., 2024), with seven items drawn from each of the benchmark's ten harm categories). We added the E1/E2 split after preliminary results showed the two subgroups behaving differently.

For each input we extract the activation at the NLA's training layer at the final token position, run the round-trip NLA pipeline, and record the text, tier, tier subgroup, activation norm, NLA explanation, reconstructed norm, reconstruction cos, and raw MSE.

3.3 Hallucination labeling and CoT probe

Each NLA explanation receives a four-class label from Claude Opus 4.7 acting as judge (Zheng et al., 2023): *0* (faithful and specific), *0g* (faithful but generic: true but uninformative), *1* (mild hallucination: domain right but specific details invented), or *2* (clear hallucination: topic wrong or contradicted). We added the 0/0g distinction during analysis to keep format-level descriptions from inflating the faithfulness rate.

For the chain-of-thought probe we adapted the biased few-shot setup of Turpin et al. (2023). We constructed 30 four-option multiple-choice questions with unambiguous factual answers across six categories, then ran each in two conditions. The neutral condition presents the question alone. The biased condition prefixes the question with three few-shot Q&A pairs all answered with the same wrong letter (rotated across A, B, and D over the 30 questions, never matching the correct answer). Every prompt ends with `Answer: (` so the next token is the model's letter

choice, and we extract the activation feeding that prediction. For each input we record the model's greedy answer, whether it matches the correct letter, whether it matches the biased letter, the model's continuation (used to flag verbal-CoT acknowledgment of the bias), the NLA explanation, and a binary flag for whether the explanation references the few-shot pattern.

4 Findings

4.1 Residual-stream anisotropy at the NLA layers

Pairwise cosine similarities among gold activations form qualitatively different distributions in the two models (Figure 1). Qwen-2.5-7B at layer 20 is broad, with mean 0.497 across 124,750 pairs, meaning any two random activations from Qwen at this layer point in noticeably different directions on average. Gemma-3-12B at layer 32 is tight, concentrated near 0.975, meaning any two random activations from Gemma at this layer point in nearly the same direction. The distributions barely overlap. The 0.478 gap in mean pairwise cosine therefore gives Gemma a much higher baseline for any cosine-based reconstruction metric, before reconstruction quality is measured.

Figure 1. Residual stream anisotropy at the released NLA extraction layers

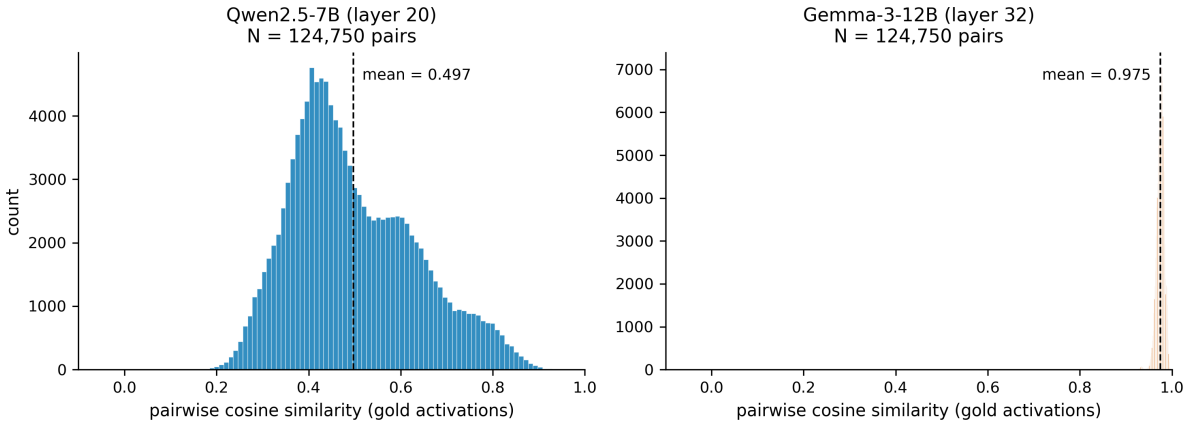

Figure 1: Residual-stream anisotropy at the released NLA extraction layers. Pairwise cosine similarities among gold activations within each 500-input set. Qwen-2.5-7B at layer 20 has mean 0.497 (left); Gemma-3-12B at layer 32 has mean 0.975 (right). The 0.478 gap in mean pairwise cosine means Gemma starts from a much higher reconstruction-metric baseline than Qwen, before reconstruction quality is measured.

4.2 Cross-model reconstruction quality inverts under correction

Table 1 gives per-tier reconstruction quality in raw cos (analogous to negated MSE) and in ABC. Under raw cos, the Gemma NLA outperforms Qwen across every tier subgroup, with cosines in the 0.99 range against Qwen's 0.74 to 0.90. Under ABC, the comparison flips: per-row reconstruction information from Qwen exceeds Gemma by roughly an order of magnitude in every subgroup.

Table 1: Per-tier reconstruction quality, both models. ABC = anisotropy-baseline-corrected cosine (Eq. 1). Read in raw cos units, the Gemma NLA outperforms Qwen across every tier. Read in ABC units, the comparison inverts, with Qwen substantially exceeding Gemma in every tier.

Subgroup	n	Qwen cos	Qwen ABC	Gemma cos	Gemma ABC	Δ ABC (G−Q)
A (English factual)	100	0.868	+0.371	0.996	+0.022	−0.349
B (Code)	100	0.797	+0.300	0.988	+0.013	−0.286
C (Turkish)	100	0.802	+0.306	0.996	+0.022	−0.284
D (Legal/formal)	100	0.843	+0.346	0.992	+0.017	−0.329
E1 (Structural)	30	0.744	+0.247	0.992	+0.017	−0.230
E2 (Semantic adv.)	70	0.903	+0.406	0.996	+0.021	−0.385

Figure 2. Reconstruction quality by input tier, anisotropy-corrected

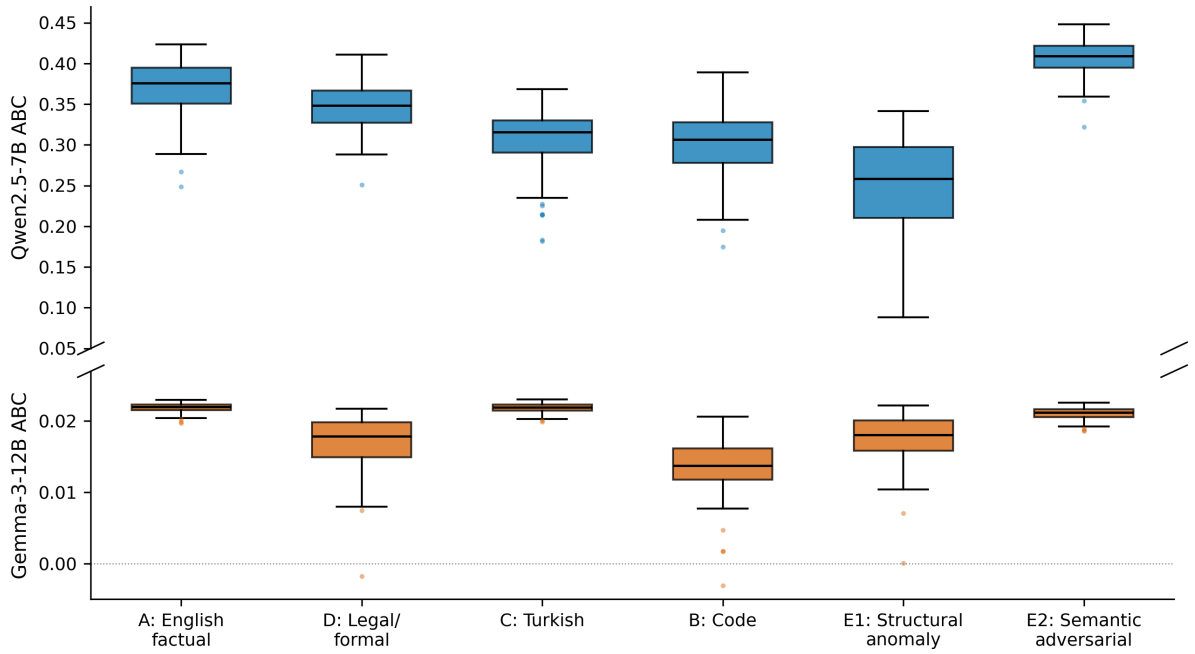

Figure 2: Reconstruction quality by input tier, anisotropy-corrected. Note the broken y-axis: Qwen (top, blue) spans 0.09 to 0.45 while Gemma (bottom, orange) spans 0 to 0.025. The tier-level ordering is qualitatively similar in both models, with E2 and A reconstructing best and E1 and B reconstructing worst.

Both models start from cos = 1 as the perfect-reconstruction ceiling, and each has to beat its own random-pair baseline to register positive ABC. Qwen has more headroom (about 0.50 of the unit sphere) and uses a large fraction of it. Gemma has very little headroom (about 0.025) and uses most of it. In concrete terms, when the Qwen reconstructor reads the verbalizer's text and produces a vector, that vector lands much closer to the true activation than a random Qwen activation does. The Gemma reconstructor produces a vector that lands close to the true activation in absolute terms, but only marginally closer than a random Gemma activation already lies. ABC measures this difference, per row of data, after geometry is controlled for. The two NLAs were trained under different recipes (compute budgets, hyperparameters), so we cannot separate "the Qwen NLA is better trained" from "NLA reconstruction is easier on a less anisotropic residual stream." For the released checkpoints as deployed, the small-model NLA carries more per-row information than the large-model NLA once we correct for geometry.

4.3 Tier structure within each model

Both models rank the input tiers (defined in Section 3) in roughly the same order, with E2 (semantic adversarial) and A (English factual) reconstructing best and E1 (structural anomalies) and B (code) reconstructing worst. The clearest within-tier split is between E1 and E2. Structural anomalies (gibberish, control tokens) reconstruct much worse than semantically adversarial prompts (fluent English from JailbreakBench). In Qwen ABC, E1 mean is +0.247 against E2's +0.406, Cohen's $d = -3.31$. In Gemma ABC, E1 is +0.017 against E2's +0.021, Cohen's $d = -1.19$. The E1–E2 gap replicates across both models with the same sign and a large effect size.

Reconstruction quality therefore tracks linguistic well-formedness rather than adversarial intent. A jailbreak prompt that reads as fluent English reconstructs well; gibberish does not. As a corollary, reconstruction MSE is not a useful signal for adversarial-input detection, since well-formed adversarial English is indistinguishable from well-formed benign English at this metric.

Code reconstructs worse than English declarative sentences in both models. A partial cause is the final-token effect, where code snippets often end in non-content punctuation that carries less semantic weight than the typical content-word terminus of declarative English. The effect is an artifact of measuring activations only at the final token, as our pipeline does.

4.4 Cross-model rank agreement and hallucination patterns

Pairing inputs by index, the Spearman rank correlation between Qwen and Gemma per-input reconstruction cos is $\rho = 0.354$ ($n = 500$, $p = 3.24 \times 10^{-16}$). The two NLAs partly agree and partly disagree on which inputs are hard (Figure 3). Reconstruction difficulty is partially shared and partially model-specific.

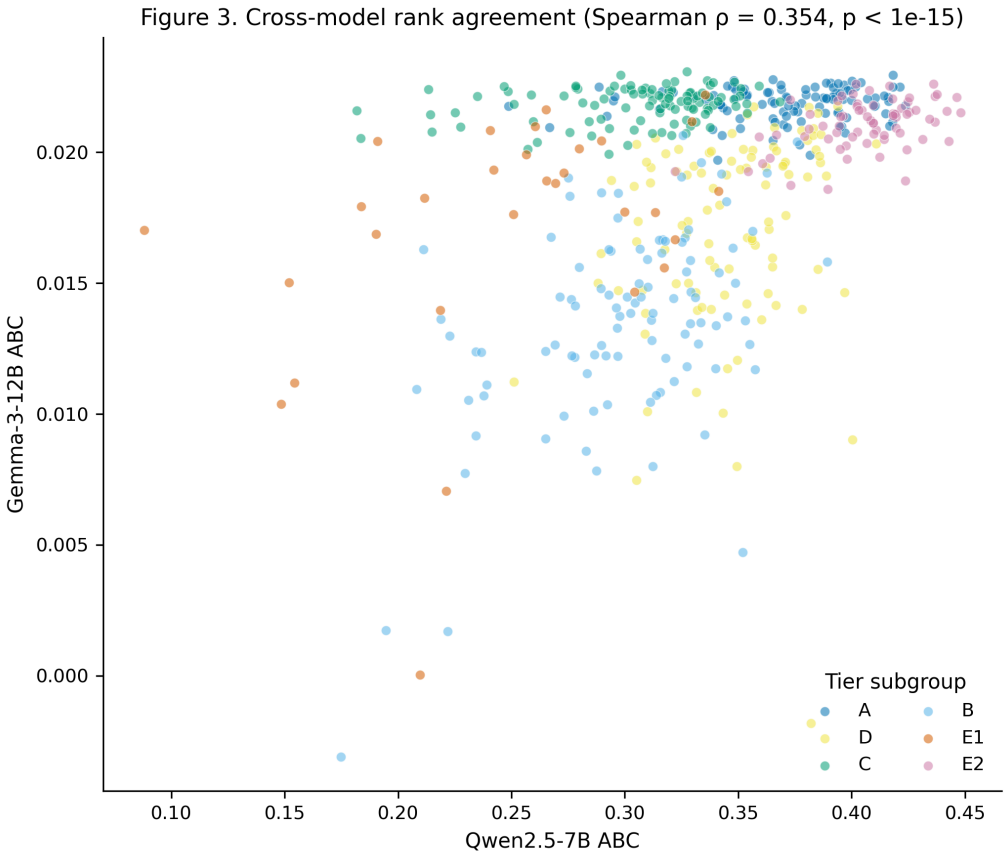

Figure 3: Cross-model rank agreement. Per-input ABC paired by index, colored by tier subgroup. Spearman $\rho = 0.354$ ($p < 10^{-15}$). Moderate positive correlation indicates partial cross-model agreement on per-input difficulty; substantial vertical spread within tier reflects model-specific reconstruction behavior.

Hallucination judgments under the four-class rubric show complementary patterns across the two models (Figure 4). Qwen produces 7 clear hallucinations (score 2) across 500 inputs (6 in Tier A, 1 in Tier C). Gemma produces none. Mild hallucinations (score 1) cluster in Tier C (Turkish), at 28% of Qwen's tier and 7% of Gemma's: NLAs trained on English-heavy data confabulate plausible-but-wrong factual specifics on non-English activations. Gemma's most distinctive pattern is in Tier B: 48% of code explanations score 0g (faithful but generic), against Qwen's 34%. Gemma's tightly clustered residual stream encodes less direction-specific content for the verbalizer, so it falls back on format-level descriptions ("structured code snippet with answer pattern") that are true but uninformative.

Figure 4. Hallucination rubric by tier subgroup

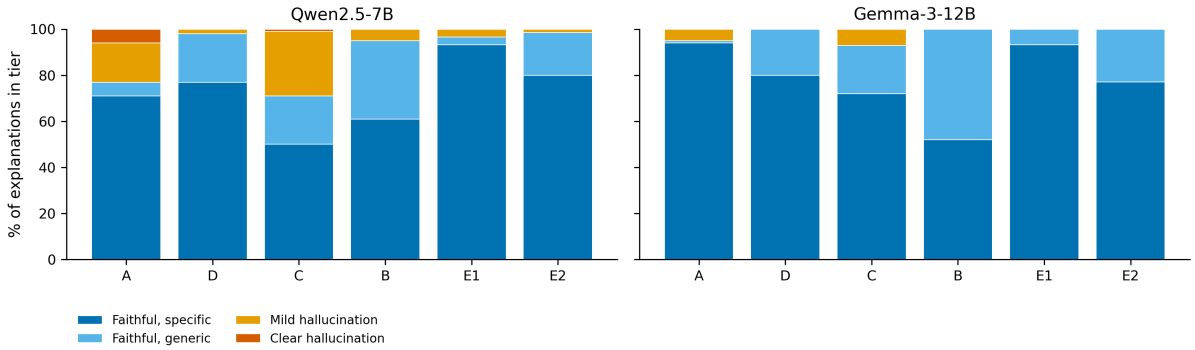

Figure 4: Hallucination rubric by tier subgroup. Stacked bars give the percentage distribution of judgments under the four-class rubric (0: faithful and specific; 0g: faithful but generic; 1: mild hallucination; 2: clear hallucination).

4.5 CoT unfaithfulness probe: null result

The probe described in Section 3 yielded a null result. Both Qwen-2.5-7B-Instruct and Gemma-3-12B-IT answered correctly on all 60 trials each (30 neutral, 30 biased). The biasing prefix never changed either model's greedy answer. The Turpin-style unfaithfulness rate (the rate at which the NLA surfaces bias when bias took effect and the verbal CoT did not acknowledge it) is therefore undefined here: the conditioning event never occurred.

The biased prompts still induced a small shift in reconstruction MSE in each model. In Gemma, biased-condition mean MSE was 0.0034 higher than neutral. In Qwen, it was 0.0507 lower. The signs differ. Magnitudes are small relative to within-model MSE variance. The sign difference is consistent with the anisotropy gap, since Qwen's larger residual-stream headroom lets the biasing prefix push the activation toward a more reconstructable direction. Across all 60 biased-condition NLA explanations in each model, none referenced the few-shot pattern; the verbalizer described the answer-emission state and the surface format of the prompt without acknowledging that the prior three Q&As shared a letter. Verbal CoT was also silent on the bias. This is a null result for the design, not a finding about NLA capability. A three-shot label-only bias is too weak to elicit unfaithful behavior in modern instruction-tuned models of this size on factual MCQs with unambiguous answers.

5 Discussion

Comparing released NLAs across model families using the reported reconstruction MSE produces conclusions that depend on residual-stream anisotropy. The two NLA-released models differ substantially in anisotropy at the trained layer, and the raw-MSE comparison reverses under correction. Within-model tier-level findings are robust to the cross-model concern, since each tier is compared against its own model's baseline.

A Spearman rank correlation of 0.354 between the two models' per-input cos values says that reconstruction difficulty is partly tied to the input and partly to the model. Different models find different inputs hard. Results from one NLA characterize that NLA, not NLAs as a category.

The tier-level pattern with the widest implication is the E1-versus-E2 split. Reconstruction MSE separates structurally anomalous inputs (gibberish, control tokens) from well-formed text but does not separate adversarial intent from benign content within the well-formed regime. The hallucination patterns add a model-level dimension that the corrected metric alone does not capture. Gemma produces no clear hallucinations and many faithful-but-generic explanations.

Qwen produces a small number of clear hallucinations and a higher mild-hallucination rate on Turkish text. Both patterns fit the anisotropy interpretation, in that tighter residual streams give the verbalizer less direction-specific content to translate, so it leans on format-level descriptions that are true but uninformative.

6 Limitations

Our cross-model comparison covers two of the four released NLA checkpoints. We did not include the Gemma-3-27B and Llama-3.3-70B NLAs because of single-GPU compute constraints. The anisotropy gap we measure is a property of two specific (model, layer) pairs and does not generalize to the model families as a whole.

The correction we use is linear, with a single global mean baseline. It does not equalize within-model variance, which is roughly an order of magnitude tighter in Gemma than in Qwen at the subgroup level. The qualitative ordering of tier subgroups within each model holds under the correction choice, but absolute claims about per-row reconstruction across models depend on the correction's linear-additivity assumption. We report Spearman rank correlation as a complementary unit-free measure.

We extract all activations at the input's final token. For tiers where the final token carries little semantic content, especially Tier B code, reconstruction quality reflects token-position effects alongside tier-level shift.

Hallucination labels come from a single Claude model judging the NLAs' text outputs against the original input. We did not validate the judge's labels against a second annotator (human or otherwise), so systematic biases in the judge are not quantified.

The chain-of-thought probe is under-powered for its intended question. A 30-MCQ design with three-shot label-only bias produced no behavioral effect in either model, so the conditional unfaithfulness measurement that motivated the design cannot be tested on these data.

Data and code availability. All inputs, raw activations, NLA explanations, hallucination judgments, the analysis report, and Python scripts are released at Zenodo: 10.5281/zenodo.20112029 (CC-BY-4.0). The released NLA checkpoints from Anthropic are independently available at `https://github.com/kitft/natural_language_autoencoders`.

References

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023. URL `https://transformer-circuits.pub/2023/monosemantic-features/`.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*, 2021.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. JailbreakBench: An open robustness benchmark for

jailbreaking large language models. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, 2019. URL https://aclanthology.org/D19-1006/.

Kit Fraser-Taliente, Subhash Kantamneni, Euan Ong, Dan Mossing, Christina Lu, Paul C. Bogdan, Emmanuel Ameisen, James Chen, Dzmitry Kishylau, Adam Pearce, Julius Tarng, Alex Wu, Jeff Wu, Yang Zhang, Daniel M. Ziegler, Evan Hubinger, Joshua Batson, Jack Lindsey, Samuel Zimmerman, and Samuel Marks. Natural language autoencoders produce unsupervised explanations of LLM activations. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/nla/.

Gemma Team. Gemma 3 technical report. arXiv preprint, 2025.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Tom Conerly, Joshua Batson, and Christopher Olah. On the biology of a large language model. Transformer Circuits Thread, 2025.

LMSYS Org. Chatbot Arena Leaderboard. https://chat.lmsys.org/, 2024. Accessed 2026.

Qwen Team. Qwen2.5 technical report. arXiv preprint, 2024.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Christopher Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.