Measuring the Reliability of Natural Language Autoencoders

Alexa Tran
Stanford University

Abhi Boda
Stanford University

Manouchehr Rimer
Stanford University

Abstract

Natural language autoencoders (NLAs) have recently been proposed as an interpretability primitive for large language models, mapping residual-stream activations to natural-language descriptions and back [1]. Because the activation verbalizer (AV) samples stochastically, repeated verbalizations of the same activation yield different descriptions and different reconstructed vectors, creating a measurement reliability problem that has not been characterized. We evaluate NLA reliability across four conditions — PRISM, Bias in Bios, MMLU-Choice, and MMLU-NoChoice — using Gemma-3-12B layer-32 activations and $K=12$ stochastic AV samples per activation. Applying Generalizability Theory (G-theory), we decompose reconstruction-consistency variance into signal attributable to the activation and noise attributable to AV sampling. Across most conditions, fidelity G-coefficients exceed 0.90 at $n'=1$ sample and consistency G-coefficients reach 0.90 at $n'=3$, yielding concrete sampling recommendations for practitioners. A cautionary exception arises in MMLU-Choice, where high within-item consistency coexists with near-zero item-discriminability and a 20-point drop in linear probe accuracy.

1 Introduction

Natural language autoencoders (NLAs) [1] map residual-stream activations to natural-language descriptions and back via two jointly trained components: an activation verbalizer (AV) and an activation reconstructor (AR). NLAs have been applied in pre-deployment audits of frontier models such as Claude Opus 4.6, identifying unverbalized evaluation awareness and diagnosing safety-relevant behaviors [1]. These are high-stakes applications, as measurement errors could directly affect deployment decisions at scale. Because the AV samples stochastically at $T=1$, repeated verbalizations of the same activation yield different descriptions and, after the AR, different reconstructed vectors. This measurement reliability problem has not been characterized: without reliability estimates, practitioners have no principled basis for choosing a sampling budget or for knowing whether an observed difference between two activations exceeds measurement noise.

We study this problem across four conditions using $N=400$ items each, Gemma-3-12B layer-32 activations, and publicly released NLA checkpoints [1], applying Generalizability Theory (G-theory) as our primary framework. We further compare activation-space consistency against text-space consistency and triangulate both against linear probe accuracy. Our contributions are: (1) the first empirical G-coefficient and D-study sampling curves for NLA-based interpretability; (2) a diagnostic framework localizing inconsistency to the AV or AR; (3) a demonstration that high consistency without item-discriminability is an important failure mode; and (4) concrete sampling recommendations for auditing practitioners.

2 Background and Related Work

2.1 Natural Language Autoencoders

NLAs are an unsupervised method for generating natural-language explanations of LLM activations [1]. The AV takes a residual-stream activation vector and injects it as a prompt embedding, then generates a free-form natural-language description autoregressively. The AR takes descriptions and map them back to activation space via a learned affine map on a truncated copy of the target LLM. Both components are initialized as copies of the target model and trained jointly via reinforcement learning to minimize residual-stream reconstruction error, with no explicit supervision on the description content.

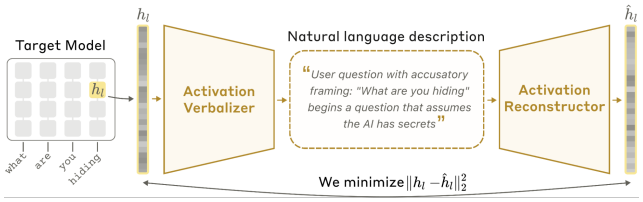

Figure 1: The NLA pipeline [1]. The AV maps a residual-stream activation h_l to a natural-language description; the AR maps it back to produce $\hat{h}_l$, trained jointly to minimize $\|h_l - \hat{h}_l\|_2^2$.

2.2 Generalizability Theory

G-theory [2] is a psychometric framework for characterizing the reliability of measurements that are inherently stochastic—in our case, a stochastic language model. G-theory makes the sources of error explicit by identifying the *object of measurement* (here, the activation) and the *facets of measurement* (here, the particular AV sample drawn), then uses an ANOVA-style variance decomposition to estimate how much each source contributes to observed score variance.

In our fully-crossed $p \times i$ design, every activation p is observed under every AV sample i, yielding a score Y_{pi}. Total variance is partitioned into σ_p^2 (between-activation signal), σ_i^2 (systematic differences across AV sample slots), and σ_{pi}^2 (the activation-by-sample interaction, i.e. noise that cannot be reduced without averaging over more samples). The relative G-coefficient for a mean score averaged over n' samples is:

$$G(n') = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{pi}^2/n'} \tag{1}$$

$G(n')$ has a direct interpretation as a signal-to-total-variance ratio, ranging from 0 (all noise) to 1 (all signal). In psychometric practice, $G \geq 0.9$ is a common threshold for high-stakes measurement [3]. The D-study uses estimated variance components to project $G(n')$ forward across values of n', informing the question of how many AV samples a practitioner must draw to reach a target reliability level.

3 Methods

3.1 Datasets and Experimental Conditions

We evaluate across four conditions. For each condition, N=400 items are passed through Gemma-3-12B-pt [4] and layer-32 activations extracted at the final non-padding token.

PRISM [5] consists of multi-turn conversations (HuggingFace: `Transluce/PRISM-gender-Llama-3.1-8B-Instruct`), flattened to the last user turn. We sample 400 conversations uniformly at random (seed 42). Probe label: binary gender.

Bias in Bios [6] consists of short third-person professional biographies (HuggingFace: `LabHC/bias_in_bios`), with explicit gender indicators scrubbed. We sample 400 biographies stratified 50/50 by gender. Probe labels: 26-class profession and binary gender.

MMLU-Choice [7] uses 400 questions from four subjects (HuggingFace: `cais/mmlu`): abstract algebra, moral scenarios, virology, and astronomy (100 per subject), chosen for maximal diversity in the representational space. Prompts include the question stem and all four answer choices. Probe label: 4-class subject.

MMLU-NoChoice uses the same 400 MMLU items but with prompts containing only the question stem, without answer choices. This condition isolates the effect of prompt format on NLA reliability. Probe label: 4-class subject.

3.2 Activation Extraction

We use Gemma-3-12B (pretrained) [4] as the activation source model, extracting residual-stream activations from layer 32, consistent with the layer at which the NLA checkpoints were trained [1]. For each prompt, we run a single forward pass and take the hidden state of the final non-padding token at layer 32. Activations are L2-normalized to unit norm prior to storage. All four conditions use the same procedure with seed 42.

3.3 Activation Verbalizer Sampling

We use the publicly released checkpoint `kitft/nla-gemma3-12b-L32-av` as our AV [1]. The activation vector is injected into the AV by replacing the embedding of a special `<concept>` token in a fixed prompt template. We draw K=12 independent samples per activation at temperature T=1.0 (max 500 new tokens), yielding $N \times K = 4{,}800$ descriptions per condition. The 12 samples are independent autoregressive draws, not paraphrases of a single canonical output.

3.4 Activation Reconstruction

We use the publicly released checkpoint `kitft/nla-gemma3-12b-L32-ar` as our AR [1], a copy of Gemma-3-12B truncated to 32 layers with a learned affine map at the output. The AR is deterministic: each of the 4,800 descriptions is passed through independently, yielding 4,800 reconstructed vectors. Each reconstructed vector is L2-normalized to match activation normalization.

3.5 Evaluation Metrics

Centering. All activation vectors share a dominant mean direction: $\|\mu\| \approx 0.99$ where μ is the mean of the 400 originals per condition. Raw cosines between any two activations therefore sit at ≈ 0.99 regardless of content, making raw cosine a near-useless discriminator. All primary metrics subtract μ from both vectors and renormalize before computing cosine:

$$\text{centered\_cos}(\mathbf{a}, \mathbf{b}) = \cos\left(\frac{\mathbf{a}-\mu}{\|\mathbf{a}-\mu\|}, \frac{\mathbf{b}-\mu}{\|\mathbf{b}-\mu\|}\right)$$

Fidelity. For each (activation i, sample s), we compute matched centered cosine between the original $\mathbf{h}_i$ and its reconstruction $\hat{\mathbf{r}}_{i,s}$, and mismatched centered cosine against 5 randomly drawn originals $\mathbf{h}_j$ ($j \neq i$). The centered fidelity gap is:

$$\Delta_{\text{fid}} = \text{centered\_cos}(\mathbf{h}_i, \hat{\mathbf{r}}_{i,s}) - \text{centered\_cos}(\mathbf{h}_j, \hat{\mathbf{r}}_{i,s}) \tag{2}$$

A large positive gap indicates the AR is recovering activation-specific information rather than producing generic reconstructions.

Consistency. For each activation i, within-item consistency is the mean centered cosine over all $\binom{12}{2} = 66$ pairs among its 12 reconstructions. Between-item consistency is the mean centered cosine over $5 \times 66 = 330$ randomly drawn cross-activation pairs. The centered consistency gap is within minus between.

Text-space consistency. We embed all 4,800 AV descriptions with `sentence-transformers/all-mpnet-base-v2` [8] (L2-normalized). Within-item text consistency is the mean of all 66 pairwise cosines among an activation's 12 descriptions. Between-item text consistency is the mean cosine against all descriptions from every other activation in the same condition. This isolates whether activation space inconsistency originates in the AV or the AR.

Linear probes. We train L2-regularized logistic regression classifierson both original activations and mean of 12 reconstructed activations to predict ground-truth labels. A drop in probe accuracy after the round-trip indicates that the NLA pipeline is destroying label-relevant representational structure even when consistency metrics appear strong.

G-theory. We fit a fully-crossed $p \times i$ G-theory model (both facets random; n_p=400, n_i=12) separately for fidelity and consistency scores. Cell (p, i) for the fidelity G-study is the raw cosine $\cos(\mathbf{h}_p, \hat{\mathbf{r}}_{p,i})$. Cell (p, i) for the consistency G-study is the mean of the 11 pairwise cosines involving sample i for activation p. Variance components are estimated via two-way balanced ANOVA:

$$\hat{\sigma}_p^2 = \max\left(\frac{MS_p - MS_{pi}}{n_i}, 0\right), \quad \hat{\sigma}_i^2 = \max\left(\frac{MS_i - MS_{pi}}{n_p}, 0\right), \quad \hat{\sigma}_{pi}^2 = \max(MS_{pi}, 0)$$

The D-study evaluates $G(n')$ for $n' \in \{1, 2, 3, 4, 6, 12\}$ via Equation, 1.

4 Results

4.1 Fidelity

Table 1 reports centered fidelity metrics across conditions. PRISM and Bias in Bios achieve matched centered fidelity of 0.668 and 0.647 respectively, with centered fidelity gaps of 0.675 and 0.646. This indicates reconstructions are meaningfully closer to their original activations than to other activations. MMLU-NoChoice achieves a gap of 0.591, which is broadly comparable.

MMLU-Choice is the outlier, demonstrating a matched fidelity of 0.238 and a gap of 0.253. Reconstructions are only marginally closer to their own originals than to others.

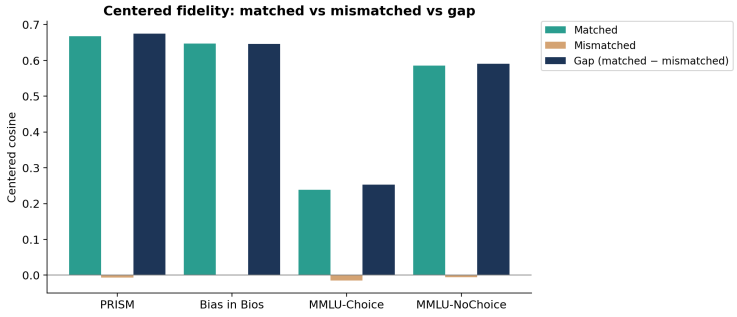

Figure 2: Centered fidelity metrics by condition.

Table 1: Centered fidelity metrics by condition.

Condition	Matched	Mismatched	Gap
PRISM	0.668	−0.007	0.675
Bias in Bios	0.647	0.001	0.646
MMLU-NoChoice	0.585	−0.006	0.591
MMLU-Choice	0.238	−0.015	0.253

4.2 Consistency

Table 2 reports centered consistency metrics. PRISM and Bias in Bios achieve within-item scores of 0.953 and 0.951 with between-item baselines of 0.146 and 0.158, yielding gaps of 0.808 and 0.793. MMLU-NoChoice has a gap of 0.724. These results confirm that across 12 independent AV draws, reconstructions of the same activation reliably cluster in a region far from reconstructions of other activations.

MMLU-Choice presents a different pattern. Its within-item consistency (0.924) is numerically similar to other conditions, but the between-item baseline has risen to 0.651, nearly matching the within

score and collapsing the gap to 0.273. Two reconstructions from entirely different MMLU-Choice items are fairly similar to each other compared to two reconstructions of the same item.

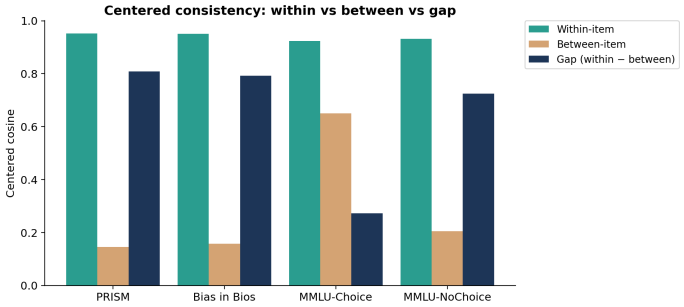

Figure 3: Centered consistency metrics by condition.

Table 2: Centered consistency metrics by condition.

Condition	Within	Between	Gap
PRISM	0.953	0.146	0.808
Bias in Bios	0.951	0.158	0.793
MMLU-NoChoice	0.931	0.206	0.724
MMLU-Choice	0.924	0.651	0.273

4.3 Text-Space Consistency

Table 3 reports MPNet text-space consistency alongside activation-space consistency. Within-item text consistency is high and narrow across all conditions (0.83–0.87), confirming the AV produces semantically stable descriptions regardless of prompt format. However, between-item text similarity for MMLU-Choice (0.654) is substantially higher than for other conditions (0.391–0.505), yielding a text-space gap of only 0.172 versus 0.365–0.470 for others.

Notably, the text-space gap collapse (0.172) does not fully account for the activation-space gap collapse (0.273). The AR compounds the problem: Bias in Bios has a between-item text similarity of 0.505 yet maintains an activation-space gap of 0.793, showing the AR can recover item-specific structure from moderately similar descriptions in other formats. The MMLU-Choice failure is therefore a joint contribution from both components.

Table 3: Text-space (MPNet) vs. activation-space consistency.

	Text Space (MPNet)			Activation Space		
Condition	Within	Between	Gap	Within	Between	Gap
PRISM	0.861	0.391	0.470	0.953	0.146	0.808
Bias in Bios	0.870	0.505	0.365	0.951	0.158	0.793
MMLU-NoChoice	0.839	0.428	0.411	0.931	0.206	0.724
MMLU-Choice	0.826	0.654	0.172	0.924	0.651	0.273

4.4 Linear Probes

Table 4 reports linear probe accuracy before and after the NLA round-trip. For PRISM, gender classification accuracy is identical before and after (0.575). For Bias in Bios, the profession probe drops slightly (0.300 to 0.263) and gender increases slightly (0.550 to 0.588).

The MMLU results are more informative. MMLU-NoChoice drops modestly (0.975 to 0.938), indicating a small but consistent loss after round-trip. MMLU-Choice shows the largest drop in the study: from 0.963 to 0.763, a delta of -0.200. Subject information is strongly present in the original activations and is measurably destroyed by the round-trip despite high within-item consistency.

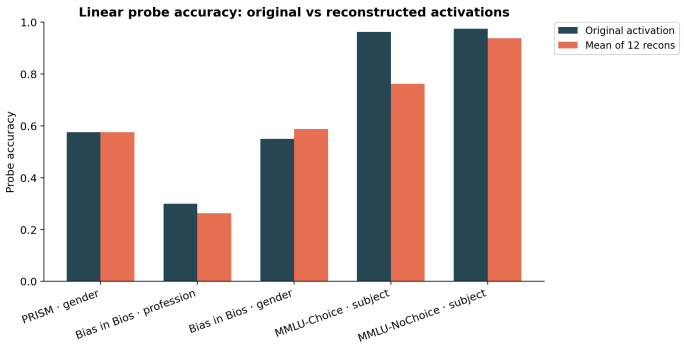

Figure 4: Linear probe accuracy on original vs. mean-of-12 reconstructed activations across all conditions and labels.

Table 4: Linear probe accuracy on original activations vs. mean-of-12 reconstructed activations. Majority = always-predict-majority-class baseline. Δ = reconstruction $-$ original.

Condition	Label	Majority	Original	Reconstruction	Δ
PRISM	gender	0.575	0.575	0.575	0.000
Bias in Bios	gender	0.500	0.550	0.588	+0.038
Bias in Bios	profession	0.263	0.300	0.263	−0.038
MMLU-NoChoice	subject	0.250	0.975	0.938	−0.038
MMLU-Choice	subject	0.250	0.963	0.763	−0.200

4.5 G-Theory and D-Study

Variance components. Table 5 summarizes variance components for both G-studies. For fidelity, the between-activation signal σ^2_p dominates in every condition. The global sample-slot component σ^2_i is effectively zero in all conditions, confirming no particular AV draw index is systematically better or worse — noise is purely idiosyncratic to the (activation, sample) pair. The residual σ^2_{pi} is largest for MMLU-Choice (24.5%), consistent with higher run-to-run fidelity variance in that condition.

For consistency, σ^2_{pi} is larger across all conditions (24.7%–40.6%). Because within-item cosine similarities cluster near 0.999, there is little room for stable between-activation differences, so most remaining variation appears in the residual interaction term.

Table 5: Variance components (as % of total) and G-coefficients at n'=1 and n'=12. $\sigma^2_i \approx 0$ in all conditions and is omitted.

		Variance (%)		$G(n')$	
Metric	Condition	σ^2_p	σ^2_{pi}	n'=1	n'=12
Fidelity	PRISM	90.6	9.4	0.906	0.991
	Bias in Bios	98.6	1.4	0.986	0.999
	MMLU-NoChoice	91.5	8.5	0.915	0.993
	MMLU-Choice	75.5	24.5	0.755	0.974
Consistency	PRISM	75.3	24.7	0.753	0.973
	Bias in Bios	78.8	21.2	0.788	0.978
	MMLU-NoChoice	68.0	32.0	0.680	0.962
	MMLU-Choice	59.4	40.6	0.594	0.946

D-Study: Fidelity. Because fidelity variance is signal-dominated, G-coefficients are high even at n'=1. PRISM, Bias in Bios, and MMLU-NoChoice all exceed the 0.90 threshold from a single sample. MMLU-Choice is the exception: G=0.755 at n'=1, requiring n'=3 to clear 0.90.

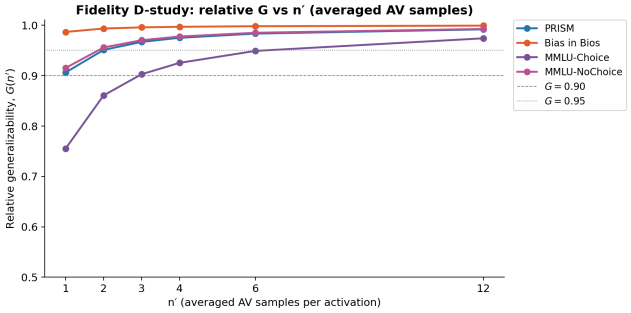

Figure 5: Fidelity D-study. $G(n')$ as a function of averaged AV samples per activation.

D-Study: Consistency. Consistency reliability requires more samples in every condition. PRISM and Bias in Bios reach G=0.90 at n'=3. MMLU-NoChoice requires n'=6. MMLU-Choice is the hardest case: G=0.898 at n'=6, just below threshold, reaching 0.946 only at n'=12.

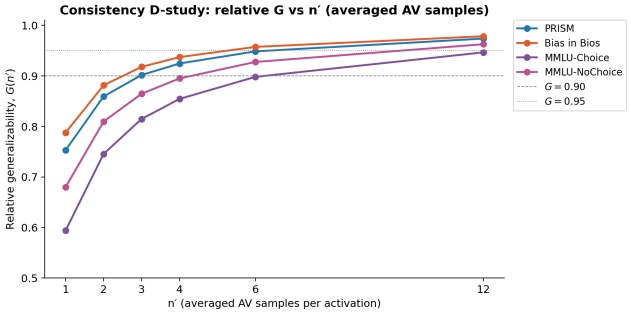

Figure 6: Consistency D-study. $G(n')$ as a function of averaged AV samples per activation.

Practical recommendation. Three samples per activation suffice for reliable fidelity and consistency ranking on our conversational or biographical inputs. Structured multiple-choice prompts require at least six samples for fidelity and approach but do not comfortably clear 0.90 for consistency below n'=12. Given the validity failure documented in Section 4.6, we recommend pairing G-coefficient checks with centered fidelity gap and probe accuracy diagnostics before trusting NLA measurements on new input formats.

Table 6: D-study: $G(n')$ for $n' \in \{1, 2, 3, 4, 6, 12\}$. Conditions below the G=0.90 threshold are in plain type; those meeting or exceeding it are in bold.

Metric	Condition	n'=1	n'=2	n'=3	n'=4	n'=6	n'=12
Fidelity	PRISM	**0.906**	**0.951**	**0.967**	**0.975**	**0.983**	**0.991**
	Bias in Bios	**0.986**	**0.993**	**0.995**	**0.997**	**0.998**	**0.999**
	MMLU-NoChoice	**0.915**	**0.956**	**0.970**	**0.977**	**0.985**	**0.993**
	MMLU-Choice	0.755	0.861	**0.903**	**0.925**	**0.949**	**0.974**
Consistency	PRISM	0.753	0.859	**0.901**	**0.924**	**0.948**	**0.973**
	Bias in Bios	0.788	0.881	**0.918**	**0.937**	**0.957**	**0.978**
	MMLU-NoChoice	0.680	0.810	0.864	0.895	**0.927**	**0.962**
	MMLU-Choice	0.594	0.745	0.815	0.854	0.898	**0.946**

4.6 Reliability-Validity Dissociation in MMLU-Choice

The MMLU-Choice results across Sections 4.1–4.5 illustrate that high reconstruction consistency is necessary but not sufficient for trustworthy NLA-based measurement. The pipeline produces internally consistent reconstructions, but that consistency may reflect prompt formatting rather than the content of individual questions.

The comparison to MMLU-NoChoice is the clearest evidence that this is a prompt-format effect. Removing the answer choices restores the centered fidelity gap ($0.253 \rightarrow 0.591$), the centered consistency gap ($0.273 \rightarrow 0.724$), and probe accuracy ($0.763 \rightarrow 0.938$) to levels comparable with PRISM and Bias in Bios. The failure is also joint: Section 4.3 established that both the AV and AR contribute, with the AV producing less item-distinct descriptions under the multiple-choice format and the AR failing to recover even that moderate distinctiveness. We hypothesize that the answer choices dominate AV descriptions, causing reconstructions to cluster around a generic multiple-choice direction shared across items rather than preserving item-specific structure.

The practical implication is that G-coefficients and within-item reconstruction similarity, taken alone, are insufficient diagnostics for NLA deployments on novel input formats. Additional sanity checks, such as our proposed fidelity gap, linear probe accuracy after round-trip, and the comparison of text-space and activation-space consistency are necessary complements.

5 Limitations

Dataset scope. These findings are restricted to the selected datasets, which have relatively formulaic structure, causing inflated reliability estimates relative to the open-ended conversational inputs that characterize real model deployments.

Confabulation. Our G-coefficients measure consistency of NLA outputs but cannot distinguish consistent accurate verbalization from consistent confabulation. High G with high confabulation would be internally consistent but externally invalid. We partially mitigate this via linear probes, which provide an independent check on whether reconstructions preserve label-relevant structure. However, probes are limited: they capture only linearly decodable attributes and our 80-item test sets introduce substantial sampling uncertainty in reported accuracies.

Single layer and model. Our analysis is restricted to Gemma-3-12B layer 32. Reliability may vary across model families, scales, and layers, particularly for layers encoding more abstract cognition.

6 Conclusion

We present the first systematic reliability analysis of Natural Language Autoencoders using Generalizability Theory. Across PRISM and Bias in Bios, fidelity G-coefficients exceed 0.90 at $n'=1$ sample and consistency G-coefficients reach 0.90 at $n'=3$ samples. MMLU-NoChoice requires $n'=6$ for consistency to clear 0.90, though fidelity remains reliable at $n'=1$. Together these indicate that NLA measurements are reliable at modest sampling budgets for conversational, biographical, and reasoning inputs. The MMLU-Choice condition reveals an important failure mode: high within-item consistency can coexist with near-zero item-discriminability and large probe accuracy losses when prompt formatting causes AV descriptions to converge on shared surface language. Given the validity failure documented in Section 4.6, we recommend that practitioners pair G-coefficient checks with centered fidelity gap and probe accuracy diagnostics before trusting NLA measurements on new input formats.

Source Code

All source code for data preprocessing, activation extraction, AV-AR pipeline execution, and reliability analysis is available at `https://github.com/tranalexa/nla_reliability`.

References

[1] Kit Fraser-Taliente, Subhash Kantamneni, Euan Ong, et al. Natural language autoencoders produce unsupervised explanations of LLM activations, 2026.

[2] Lee J. Cronbach, Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley, 1972.

[3] Jum C. Nunnally. *Psychometric Theory*. McGraw-Hill, New York, 2 edition, 1978.

[4] Google DeepMind. Gemma 3 technical report, 2025.

[5] Dami Choi, Vincent Huang, Sarah Schwettmann, and Jacob Steinhardt. Towards scalable automated user modeling, 2025.

[6] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 2019.

[7] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[8] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020.

AI Use Disclosure

Plagiarism, Bias, and Accuracy Disclosure: For this project, we used AI tools to help format the manuscript, check writing clarity, brainstorm code structure, and debug, but treated all AI outputs as provisional. The steps we took to avoid plagiarism and non-attribution are as follows: we made sure that concepts from prior work, datasets, course materials, and AI-assisted suggestions were checked and cited rather than presented as unsupported original claims. To check the accuracy, we ran smoke tests, read the code directly, inspected parquet outputs, and compared the README against the actual scripts.

AI Tools Reflection: We used Claude, Claude Code, and Cursor during this project. These tools helped us brainstorm coding architectures, debug code, format the manuscript, check writing clarity, generate sanity checks, and make tables and figures. They were also helpful for understanding technical concepts such as centered cosine similarity and why raw cosine scores can be misleading in activation space. In short, AI helped us move faster and understand the project better. That being said, we remained cautious in our use of AI. At times, AI tools hallucinated repo details or gave overconfident summaries that did not match the actual branch, so we always verify claims by reading the code, comparing scripts to the README, running smoke tests, and inspecting outputs ourselves. Code suggestions also had to be tested rather than assumed correct. Overall AI did not replace our responsibility for the final implementation, analysis, and interpretation. In the future, we would use AI similarly: as scaffolding for debugging and explanation, not as a substitute for verification.

AI Use and Impact Statement: The project has social and ethical implications: Natural Language Autoencoders could improve interpretability through commitment to measurement science, but they could also be misused if people treat their descriptions as literal explanations of model thoughts or as evidence of safety which seems to carry more weight than it actually does. We limited runs to 400 items per dataset, reused artifacts, and avoided unnecessary reruns to reduce compute costs.