Residual Stream Denoising in Large Language Models: Gradient-Free Quality Improvement via Functional Sensitivity Analysis

Dr. Tamás Nagy

tnagyphd@gmail.com

Draft

Abstract

Pretrained transformer LLMs contain noise directions in their residual stream — PCA components whose removal improves the model across all tested text domains. Removing 1.9-3.4% of directions reduces perplexity by 8-23% across three GPT-2 scales (124M, 355M, 774M), with no retraining or hyperparameter search. Baseline controls confirm the effect is genuine (targeted -11% vs random +64% degradation), and calibration robustness tests show consistency across text domains. The key enabling finding: PCA variance ranking is weakly correlated with functional importance (Spearman $|r| < 0.4$), requiring direct per-direction sensitivity measurement. A random basis control confirms the noise is basis-independent: random orthogonal bases achieve comparable denoising (-7.5% to -13.0%). The method composes with INT8 quantization (benefit survives precision reduction) and is 12× more effective than LASER's weight-space rank reduction, with the two approaches targeting orthogonal deficiencies. The pipeline — PCA, brute-force sensitivity sweep, noise removal — is a one-time O(d) computation. Validated on one architecture family; generalization pending.

1. Introduction

1.1 The Observation

Pretrained language models contain directions in their residual stream that actively hurt prediction quality. Removing these directions — without any retraining — improves the model. Across three GPT-2 scales, removing 1.9-3.4% of PCA directions reduces perplexity by 8-23%. The method requires no training, no gradient descent, and no hyperparameter search.

The main result (§4): a three-step pipeline — PCA decomposition, per-direction sensitivity sweep, noise direction removal — that identifies and removes cross-domain noise directions. The paper reflects the trajectory of the investigation: §2-3 develop spectral compression and sparse decomposition theory (31 theorems) that motivated the work but are logically independent of the denoising result; §4-6 present the empirical denoising discovery and its validation (6 arithmetic lemmas). Readers primarily interested in the method can skip directly to §4.

The discovery emerged from an investigation into LLM compressibility. Weight-space SVD yields spectral decay rate 1.02, implying near-zero compressibility. Activation-space analysis reveals sparse structure (up to 51 in energy terms), but energy retention does not translate to functional preservation: 99.7% variance retention at k=600 produces 24-54× worse perplexity. The gap between energy and function led us to measure per-direction *functional sensitivity* — and to the finding that some directions have *negative* sensitivity (their removal helps).

1.2 Summary of Results

Finding	GPT-2 (124M)	GPT-2 Medium (355M)	GPT-2 Large (774M)
Cross-domain noise directions	16 (2.1%)	19 (1.9%)	**43 (3.4%)**
PPL improvement from removal	8-14%	13-19%	**16-23%**
Noise dirs per text (range)	201-213	282-349	462-534
Var-sensitivity correlation†	r [+0.24, +0.41]	r = +0.37	r = +0.23
Cross-text sens. correlation	r = 0.95-1.00	r = 0.27-0.38	r = 0.10-0.17
PPL improvement (mean ± std)	-11.3% ± 2.8%	-15.8% ± 2.2%	-18.3% ± 3.1%
Targeted vs random (avg)	-11.1% vs +64%	—	-18.3% vs +21%
Random basis control	-10.9% avg (N=3)	—	—
Layer-selective (top-8)	-12.0%	—	—
INT8 + denoising	-8.4% (additive)	—	—
LASER best (single layer)	-0.8%	—	—
LASER + denoising	-10.4% (additive)	—	—
Formally verified theorems	37 (31 substantive + 6 arithmetic)	Platonic → Lean 4	§5

†Variance-sensitivity Spearman r is calibration-dependent; range shown for 124M is measured across 4 calibration texts × 3 test texts (12 conditions, mean +0.35 ± 0.06). Medium and Large values are from single calibration; expect similar spread.

1.3 Contributions

1. **Residual stream denoising**: a gradient-free, post-hoc method that improves LLM quality by removing noise directions from the residual stream, validated across three GPT-2 scales (124M, 355M, 774M) with benefit increasing from 8-14% to 16-23%. The method requires forward passes only — no backpropagation, no training data labels, no hyperparameter search. (§4)
2. **Noise fraction across scales**: the fraction of cross-domain noise directions varies — 2.1% (124M), 1.9% (355M), 3.4% (774M) — without a clear monotonic trend. PPL improvement increases within this family: 8-14% → 13-19% → 16-23%, but three data points from one architecture cannot establish a scaling law. (§4)
3. **Variance-sensitivity decorrelation**: empirical and formal demonstration that PCA ranking (variance) is unreliable for predicting functional importance (Δppl). First-order methods (gradient, Fisher) also fail. (§3, T32, T35)
4. **31 substantive theorems + 6 arithmetic lemmas** (formally verified): 31 theorems in five chains — spectral foundation (T1-T7), distillation theory (T8-T14), hierarchical composition

(T15-T21), sparse decomposition (T22-T27), orthogonal basis (T28-T31) — plus 6 arithmetic lemmas for the denoising mechanism (T32-T37). The denoising lemmas formalize the logical scaffolding (e.g., "removing a noise direction reduces error" $a + b > a$ when $b > 0$). The nontrivial content of the denoising result is empirical; the lemmas ensure the argument's arithmetic is machine-checked. (§5)

5. **The energy-function gap**: quantitative demonstration that L2 energy retention and perplexity preservation are different quantities connected by softmax amplification. (§3.3)
6. **The functional basis pipeline**: PCA → brute-force sensitivity → noise removal. One-time O(d) forward passes, no training. (§4.3)
7. **Basis-independent noise**: random orthogonal bases (no covariance alignment) also find cross-domain noise with comparable denoising gains (-7.5% to -13.0% vs PCA's -11.1%), establishing that the noise is an intrinsic geometric property of the trained model, not a PCA artifact. (§6.4)
8. **Layer-selective denoising**: noise is not uniform across layers — 8/12 layers benefit from denoising while 4 late-middle layers (7-10) are hurt. Selective denoising (top-8 layers only) achieves -12.0% vs uniform -11.1%. The noise concentrates in early layers and the final layer, forming a U-shape pattern. (§4.9)
9. **Quantization composability**: denoising composes approximately additively with INT8 quantization (expected -8.8%, observed -8.4%). The denoising benefit survives deployment-time quantization. (§4.10)
10. **LASER comparison**: head-to-head against LASER (weight-space SVD rank reduction). Our activation-space denoising is 12× more effective (-9.8% vs -0.8%). The methods compose additively — orthogonal mechanisms targeting different noise types. (§4.11)

2. Theoretical Framework

The paper's core empirical contribution — residual stream denoising — is formalized directly in §2.4 (T32-T37). Sections 2.1-2.3 (T1-T31) develop the spectral compression and sparse decomposition theory that *motivated* the investigation: the failure of standard compression approaches (§3) led us to discover noise directions. These sections provide the formal context for why purely energy-based or gradient-based methods miss the functional importance structure that denoising exploits, but the denoising results stand independently of them.

2.1 Spectral Foundation and Distillation (T1-T14)

The Eckart-Young theorem guarantees that rank-K SVD truncation is optimal in Frobenius norm. We extend this through 14 theorems:

- **T1-T7 (Spectral properties):** T1 establishes that rank-K SVD is the best low-rank approximation; T4-T5 show that each additional component strictly decreases the residual; T7 (-bridge) connects the spectral gap to approximation quality: if > 1, the error decays as $\varepsilon(K) \leq C/\rho^K$, meaning a small gap compounds exponentially.
- **T8-T14 (Distillation chain):** T8-T9 show that compression errors compose predictably through pipelines; T11 (certified student) gives conditions under which a compressed model

provably matches the original within tolerance ; T14 extends this to multi-stage distillation, where the of the weakest stage limits overall quality.

2.2 Hierarchical Composition (T15-T21)

Transformer layers are near-identity: $f_\ell(x) = x + \delta_\ell(x)$. This structure composes:

- **T15 (Error amplification):** A perturbation at layer ℓ grows to at most $\varepsilon(1+\delta)^{L-\ell}$ at the output. This explains why even small changes to the residual stream can have large effects on predictions — the amplification is exponential in remaining depth.
- **T18 (Bottleneck):** The composite spectral gap of an L-layer transformer is bounded by the minimum per-layer gap: $\rho_{\text{composite}} \leq \rho_{\min}$. A single poorly-conditioned layer limits the compressibility of the entire stack.
- **T21 (Latent of Latents):** The master error bound: $\varepsilon \leq (C_1 + C_2)(1 + \delta)/\rho_{\min}$. This connects the Latent framework's grade concept to transformer architecture — the spectral gap of the weakest component controls overall approximation quality.

2.3 Sparse Decomposition (T22-T27)

LLM activations decompose as $\mathbf{x} = \sum_{i \in \mathcal{A}} a_i \mathbf{d}_i + \eta$ where $|\mathcal{A}| = k \ll d$. The signal-to-noise ratio $R = A/\eta$ serves as the natural spectral gap:

- **T25 (Sparse threshold):** Sparse distillation beats dense when $kR^2 > d - k$. In plain terms: if the active directions carry enough signal (high R) relative to their count, keeping only the sparse support outperforms keeping a blanket low-rank approximation.
- **T27 (RIP certificate):** Under RIP stability ($1 + \delta < R^2$), the sparse dictionary preserves the signal faithfully — certified compression with bounded error.

2.4 Orthogonal Basis (T28-T31)

Orthogonal dictionaries (ICA, PCA, QR) have $\delta_{\text{RIP}} = 0$, eliminating amplification. Among =0 bases, higher R gives lower error (T29). ICA maximizes R by exploiting statistical independence.

2.5 Representation Denoising Lemmas (T32-T37)

The empirical discovery of noise directions (§4) motivates a formal framework. T32-T37 are arithmetic lemmas that capture the logical structure of the denoising argument. They are deliberately simple — the nontrivial content is the *empirical identification* of noise directions, not the algebra of their removal. The formalization ensures the reasoning chain is gap-free, not that the individual steps are deep.

T32 (Second-Order Dominance). The loss change from removing a direction has the form $\Delta L = a\delta + b\delta^2$. When $b\delta > a$, the second-order term dominates — gradient-based importance (first-order) underestimates the effect of full direction removal.

Formally verified: $\forall a, b, \delta > 0$: $b\delta > a \implies b\delta^2 > a\delta$.

T33 (Denoising Bound). If a direction contributes positive noise (increases loss), its removal strictly reduces total error.

Formally verified: $n > 0 \implies L_{\text{base}} < L_{\text{base}} + n$.

T34 (Noise Direction Criterion). A direction is classified as noise iff the model without it has strictly lower loss.

T35 (Variance-Sensitivity Decorrelation). High variance does not imply high functional importance: $v_1 > v_2$ and $s_1 < s_2$ can coexist.

Formally verified: $v_1 > v_2 > 0,\ s_2 > s_1 > 0 \implies v_2 s_1 < v_1 s_2$.

T36 (Universal Denoising Certificate). If noise contributions $n_1, n_2 > 0$ on two texts, the combined improvement $n_1 + n_2$ exceeds either individual.

T37 (Denoising Composition). Removing multiple orthogonal noise directions compounds: total improvement $= \sum n_i$.

3. The Energy-Function Gap

3.1 Weight-SVD Fails (1)

GPT-2 weight matrices have nearly flat singular value spectra:

Component	_weight	Interpretation
W_Q, W_K, W_V	1.01-1.03	Near-uniform spectrum
W_up, W_down (MLP)	1.02-1.04	Slightly more structured
Overall	**1.02**	No compressible structure in weights

3.2 Sparse Activations Are Misleading (Dead End)

Post-GELU MLP activations (dimension 3072, the MLP hidden size — 4× the residual stream dimension d=768) show high sparsity (k 128 of 3072 active, R 14.7) suggesting 24× compression at 60% energy. However, the denoising analysis in §4 operates on the *residual stream* (d=768), not the MLP hidden layer. The table below shows that even in the residual stream, translating energy retention to perplexity preservation reveals a large gap:

k retained	Variance explained	PCA PPL ratio	Gap
600 (78%)	99.71%	20× worse	**massive**
700 (91%)	99.93%	2.9× worse	large
740 (96%)	99.98%	1.5× worse	moderate
760 (99%)	99.99%	1.1× worse	small

At k=600, 99.71% of the variance is retained but perplexity is 20-54× worse. The softmax function amplifies small residual-stream perturbations exponentially through the $\exp(z)$ nonlinearity, making energy-based bounds uninformative for functional quality.

3.3 Why First-Order Methods Fail

We compared four ranking methods for PCA directions. All correlations are Spearman rank correlations between the method's ranking and the brute-force sensitivity profile (measured on GPT-2 Small, d=768):

Method	Cost	Spearman r with Δppl
PCA (variance)	1 SVD	r [+0.24, +0.41]†
Gradient $\|v^T \nabla L\|$	1 backward pass	r = -0.025
Fisher diagonal	1 backward pass	r = +0.157
Brute-force Δppl	d forward passes	r = 1.000

The variance-sensitivity correlation is weak-to-moderate: r [+0.24, +0.41] (Small, across 4 calibrations × 3 test texts, mean +0.35 ± 0.06), +0.37 (Medium, single calibration), +0.23 (Large, single calibration). The correlation is calibration-dependent — different PCA bases rank variance differently, yielding different correlations with the same functional sensitivity profile. Variance captures *some* signal about functional importance but is never strongly predictive (max |r| = 0.37) and is particularly unreliable at the extremes — the noise directions and the high-sensitivity directions — which matter most for denoising.

The gradient has near-zero rank correlation with functional importance. This follows from T32: removing a direction entirely is not a small perturbation — it is a global projection whose effect is dominated by second-order (curvature) terms. The softmax's curvature at rare-but-correct tokens creates sensitivity that the gradient at the current point cannot capture.

3.4 Summary of Dead Ends

ICA provides the best energy-based metrics (=11.4, =0, 99.9% signal) but is *worse* than PCA for functional preservation — independence does not align with the model's computational structure. No proxy tested in §3.1-3.4 — weight-space SVD, energy retention, gradients, Fisher information, ICA independence — reliably predicts which directions matter for model quality. This negative result directly motivates the brute-force sensitivity sweep in §4. The discovery that some directions have *negative* sensitivity (their removal helps) was unexpected — the original goal was compression, not denoising.

4. Residual Stream Denoising

4.1 Method

Models and data. All experiments use GPT-2 family models (124M, 355M, 774M) from HuggingFace Transformers, evaluated on four text domains: science (Wikipedia excerpt), history (Wikipedia excerpt), code (Python), and mathematics (textbook). Calibration texts are ~50-100 tokens each. The primary metric is perplexity (PPL) on held-out continuation of each text domain. Experiments run on a single GPU (RTX 3090); the largest sweep (GPT-2 Large, d=1280) takes ~28 minutes.

Step 1: PCA decomposition. Compute PCA of residual stream activations on calibration text. Cost: $O(d^2 n)$ for n tokens.

Step 2: Per-direction sensitivity sweep. For each PCA direction v_i ($i = 1, \ldots, d$): 1. Remove v_i from the residual stream at every layer: $\tilde{x} = x - (x \cdot v_i)v_i$ 2. Compute perplexity on test text 3. Record $\Delta\text{ppl}_i = \text{PPL}_{\text{without}} - \text{PPL}_{\text{original}}$

Cost: d forward passes. For GPT-2 (d=768): ~80 seconds. For a 7B model (d=4096): ~2 hours (one-time).

Step 3: Noise direction removal. Identify directions with $\Delta\text{ppl}_i < 0$ on all calibration texts. Remove them permanently via projection:

$$x_{\text{denoised}} = x - \sum_{i \in \mathcal{N}} (x \cdot v_i) v_i$$

where $\mathcal{N}$ is the cross-domain noise set.

4.2 Results on GPT-2

Per-text sensitivity profile:

Domain	Directions with Δppl < 0	Strong noise (Δppl $<$ -1)
Science	207	68
History	201	62
Code	209	12
Mathematics	213	11

Cross-domain noise directions (negative on all 4 texts): **16** (2.1% of 768).

Denoising effect (removing 16 cross-domain noise directions):

Domain	Original PPL	Denoised PPL	Improvement
Science	58.6	51.0	**-13%**
History	50.7	45.6	**-10%**
Code	25.3	21.7	**-14%**
Mathematics	23.5	21.7	**-8%**

All four domains improve. No retraining, no additional data, no hyperparameters.

4.3 Scale Validation: GPT-2 Medium (355M)

To test whether the phenomenon depends on the specific model, we replicate the full pipeline on GPT-2 Medium (d=1024, 24 layers, 355M parameters).

Per-text sensitivity profile (GPT-2 Medium):

Domain	Directions with Δppl < 0	Strong noise (Δppl $<$ -1)
Science	282	16
History	305	15
Code	349	5
Mathematics	317	0

More directions qualify as noise per text (27-34% vs 26-28% in GPT-2), but fewer cross the absolute Δppl $<$ -1 threshold. Since Medium's baseline PPL is ~2× lower than Small's, the same relative per-direction effect produces a smaller absolute shift.

Cross-domain noise directions: **19** (1.9% of 1024).

Denoising effect (removing 19 cross-domain noise directions):

Domain	Original PPL	Denoised PPL	Improvement
Science	34.9	28.4	**-18.5%**
History	30.0	25.2	**-16.0%**
Code	21.1	17.8	**-15.4%**
Mathematics	15.7	13.6	**-13.3%**

The improvement is *larger* than GPT-2 (13-19% vs 8-14%), consistent with the hypothesis that larger models accumulate more gradient-noise artifacts during training while also having more redundancy to tolerate their removal. The complete cross-scale comparison is in §4.4.

4.4 Scale Validation: GPT-2 Large (774M)

GPT-2 Large (d=1280, 36 layers, 774M parameters) extends the observation to the largest model in this family.

Per-text sensitivity profile (GPT-2 Large):

Domain	Directions with Δppl < 0	% of d	Strong noise (Δppl $<$ -1)
Science	494	38.6%	0
History	462	36.1%	0
Code	534	41.7%	0
Mathematics	505	39.5%	0

A striking pattern: ~39% of all directions are noise per-text (up from ~27% in GPT-2 and ~30% in Medium), but *none* cross the Δppl $<$ -1 absolute threshold. This reflects scale, not cleanliness: GPT-2 Large's baseline PPL (8-19) is 3-4× lower than GPT-2 Small's (24-59), so the same *relative* per-direction effect produces a smaller absolute ΔPPL. The noise is distributed across many directions, each contributing a small but collectively significant negative amount.

Cross-domain noise directions: **43** (3.4% of 1280).

Denoising effect (removing 43 cross-domain noise directions):

Domain	Original PPL	Denoised PPL	Improvement
Science	17.4	14.4	**-17.1%**
History	19.2	16.2	**-15.8%**
Code	13.9	10.7	**-22.8%**
Mathematics	8.4	6.9	**-17.4%**

The improvement is the largest observed (16-23%). Within this model family, larger models show more benefit from denoising — though a three-point trend within one architecture does not establish a scaling law.

Random baseline (3 trials, remove 43 dirs): avg +21.1% (range +10.6% to +34.6%). Targeted removal is >3 better than random, confirming the noise directions are genuinely special at this scale too.

Complete scaling curve:

Metric	GPT-2 (124M)	GPT-2 Medium (355M)	GPT-2 Large (774M)
d_model	768	1024	1280
n_layers	12	24	36
Cross-domain noise dirs	16 (2.1%)	19 (1.9%)	43 (3.4%)
PPL improvement	8-14%	13-19%	**16-23%**
Per-text noise fraction	26-28%	27-34%	36-42%
Cross-text sens. correlation	0.95-1.00	0.27-0.38	0.10-0.17
Var-sens correlation†	[+0.24, +0.41]	+0.37	+0.23

Three trends are visible, though three data points preclude strong claims: 1. **Noise count grows faster than dimension**: 16 → 19 → 43 (absolute), or 2.1% → 1.9% → 3.4% (relative). The percentage is not monotonic and no stability law is established. 2. **Denoising benefit grows**: 8-14% → 13-19% → 16-23%. Larger models benefit more from denoising, the clearest of the three trends. 3. **Cross-text correlation decreases**: 0.97 → 0.33 → 0.13. Larger models develop more text-specialized representations. The cross-domain noise signal is weaker per-direction but aggregates over more directions.

4.5 Cross-Text Consistency

Cross-text sensitivity correlation varies dramatically across model scale:

Model	Avg. Spearman r	Range
GPT-2 (124M)	0.99	0.959–1.000
GPT-2 Medium (355M)	0.33	0.27–0.38
GPT-2 Large (774M)	0.13	0.10–0.17

For GPT-2 Small, the sensitivity profile is nearly model-intrinsic (r 1). For Medium and Large, individual direction sensitivities become increasingly text-dependent — the same direction may be noise on one text and useful on another. Despite this, the *cross-domain intersection* (directions that are noise on all calibration texts) yields consistent denoising gains at every scale. The intersection acts as a conservative filter: only directions that are robustly noise survive, even when per-text profiles diverge.

This implies that the cross-domain noise signal exists at all scales tested, but its detection requires multi-text calibration rather than a single text — a point reinforced by the catastrophic failure of per-text removal (§4.6).

4.6 Preliminary Downstream Signals (Pilot)

Perplexity improvement does not automatically transfer to all tasks. We ran a small pilot on GPT-2 (124M) using 9 cross-domain noise directions (intersection of 4 diverse evaluation texts — fewer

than the 16 from §4.2 because the downstream evaluation uses different calibration texts; see §6.3 for calibration dependence). These results are exploratory and establish direction only; sample sizes are too small for reliable effect-size estimation.

Next-token accuracy (n 500 tokens, the most reliable pilot metric): top-1 accuracy improved from 40.3% to 45.2% (+4.9pp), while top-5 accuracy was essentially unchanged (54.8% → 53.2%, −1.6pp). This is consistent with the PPL mechanism — noise removal sharpens the logit distribution, improving the rank of the correct token without substantially changing the top-5 set.

Sentiment classification (n = 20 examples each): zero-shot accuracy moved from 50% to 75% (+25pp, binomial 95% CI: [−2pp, +52pp]; p 0.02 one-sided). Few-shot (3-shot) moved from 50% to 55% (+5pp). Both are directionally positive but statistically unreliable at this sample size — the wide confidence interval on the zero-shot result means the true effect could be anywhere from negligible to large. We report the direction, not the magnitude.

Factual QA (n = 10): unchanged (30% → 30%). **Generation** (5 prompts, greedy): repetition rate slightly increased (57.6% → 60.7%, +3.1pp), suggesting denoising does not help — and may slightly hurt — open-ended generation with greedy decoding.

The emerging pattern — denoising helps discriminative tasks (classification, next-token ranking) more than generative or recall tasks — is consistent with the PPL mechanism, but requires full-scale evaluation (SST-2 full split, MNLI, HellaSwag, SuperGLUE) on multiple model scales before any downstream claims can be considered robust.

Warning: removing per-text noise (219 directions from one text) instead of cross-domain noise (9 directions from 4-text intersection) causes catastrophic failure — generation collapses into repetition, top-1 accuracy drops to 0%. The cross-domain intersection filter is essential for safe denoising.

4.7 The Functional Basis

Using brute-force sensitivity ranking instead of PCA variance ranking for dimension selection:

k	PCA PPL (ratio)	Functional PPL (ratio)
700	113.1 (2.0×)	68.2 (1.2×)
740	71.4 (1.3×)	**32.1 (0.58×)**
752	59.9 (1.1×)	**30.7 (0.55×)**
760	56.6 (1.0×)	**36.8 (0.66×)**

At k=740, the functionally-ordered basis *improves* the model by 42% (PPL ratio 0.58×), while PCA-ordered selection at the same k slightly degrades it (1.3×). The functional basis identifies a 740-dimensional subspace where the model is strictly better than in the full 768 dimensions — the 28 removed directions (768 - 740) include the 16 cross-domain noise directions plus 12 additional directions that, while not noise on all texts, are functionally detrimental. **Caveat**: this analysis is for GPT-2 Small only; we have not replicated the functional basis ordering for Medium or Large. The optimal k and improvement ratio may differ at other scales.

Terminology note: "cross-domain noise directions" means directions whose removal improves perplexity on all tested text domains (science, history, code, mathematics). The claim is that these directions are model-intrinsic (evidenced by cross-scale stability and calibration robustness), but we do not claim universality across all possible inputs — only robustness across the tested domains.

4.8 Sensitivity Additivity

The sensitivity function is highly additive: removing two directions changes PPL by approximately the sum of their individual effects (Pearson r = 0.998, relative RMSE 18%).

Test	Pearson r with sum of individuals
Pairwise (n=2)	**0.998**
n=10 combinations	0.993
n=20 combinations	0.965
n=50 combinations	breaks down

This is a structural result about the loss landscape: the representation space decomposes into approximately independent functional contributions at the level of individual PCA directions. Interactions are negligible up to ~20 directions (~3% of d=768) but become significant for larger sets, consistent with the hypothesis that noise directions are scattered sparsely across the representation space.

For the actual cross-domain noise set (16 directions in GPT-2 124M), the predicted ΔPPL from summing individual sensitivities is -11.8%, and the observed joint removal gives -11.3% — a 0.5 percentage point discrepancy, consistent with the near-perfect pairwise additivity. This validates both (1) the greedy identification strategy (marginal sensitivities accurately predict joint effects at the scale of the cross-domain noise set) and (2) the denoising composition theorem T37 up to ~20 directions.

T37 validity boundary: T37 assumes exact additivity ($\Delta\text{PPL}(\mathcal{N}) = \sum_{i \in \mathcal{N}} \Delta\text{PPL}_i$). Empirically, this holds tightly for $|\mathcal{N}| \leq 20$ (r > 0.96) but degrades for $|\mathcal{N}| > 50$. Since GPT-2 Large's cross-domain noise set has 43 directions, the composition guarantee is near its empirical limit for the largest model — we have not validated the predicted vs. observed joint ΔPPL for the full 43-direction removal at this scale (the additivity measurements were performed on GPT-2 Small). For future models with more noise directions, greedy identification remains valid (marginal sensitivity is accurate) but the total improvement may be subadditive.

Despite the strong pairwise additivity, it does NOT enable faster noise identification via compressed sensing (see §7.3).

4.9 Layer-Selective Denoising

Uniform denoising removes noise directions from all layers equally. But do all layers benefit? We test per-layer denoising on GPT-2 (124M): remove the 16 cross-domain noise directions from *one layer at a time* and measure the effect.

Layer	Avg ΔPPL	sci	hist	code	math
0	-2.8%	-4.4	-2.6	-2.1	-1.9
1	-3.0%	-2.9	-4.2	-4.6	-0.3
2	-3.0%	-4.4	-5.1	-0.8	-1.6
3	-2.6%	-4.7	-4.8	+0.1	-1.0
4	-2.7%	-3.1	-6.3	-0.9	-0.6
5	**-4.1%**	-2.1	-7.7	-3.9	-2.7

Layer	Avg ΔPPL	sci	hist	code	math
6	-2.3%	-0.7	-5.9	-1.3	-1.4
7	+0.6%	+2.5	-2.8	+1.6	+1.2
8	+0.6%	+2.8	-1.8	+0.8	+0.4
9	+1.4%	+3.1	+1.2	+0.4	+1.0
10	+0.3%	+0.2	-0.4	+2.6	-1.0
11	**-4.5%**	-3.0	-6.2	-4.8	-3.9

The noise profile is **not uniform across layers**: 8 of 12 layers benefit from denoising (layers 0-6, 11), while 4 layers are hurt (layers 7-10). The last layer (11) is the noisiest, followed by layer 5 — forming a pattern where early layers and the final layer carry noise, but late-middle layers resist denoising.

Selective denoising — removing noise from only the best layers — outperforms uniform:

Strategy	Layers	Avg ΔPPL
Single best layer	{11}	-4.5%
Top-2	{5, 11}	-9.1%
Top-3	{1, 5, 11}	-10.7%
Top-6	{0,1,2,4,5,11}	-11.8%
Top-8	**{0,1,2,3,4,5,6,11}**	**-12.0%**
Uniform (all 12)	{0-11}	-11.1%

Selective denoising (top-8 layers) improves PPL by -12.0% vs uniform -11.1%. The difference is modest but consistent: skipping layers 7-10 avoids the small degradation they introduce. The noise directions apparently serve a productive function at those layers — they may participate in text-dependent computation that is disrupted by removal.

The per-layer structure also has implications for understanding noise origins. The "U-shape" — noise concentrated in early layers and the final layer, with a clean middle band — suggests that noise accumulates at layer boundaries: early layers inherit input embedding noise, the final layer develops prediction-head noise, while middle layers have learned to route around these directions.

4.10 Quantization Stacking

Does denoising compose with INT8 quantization, or do they interfere? We test on GPT-2 (124M) using simulated per-channel INT8 quantization (49 weight matrices: all Linear and Conv1D layers, per-output-channel scale → round → dequantize).

Condition	Avg PPL	Δ vs baseline
Baseline (FP32)	39.5	—
Denoised only (FP32)	35.6	-9.8%
Quantized only (INT8)	39.9	+1.0%
Denoised + Quantized	36.2	-8.4%

Expected if additive: -8.8%. Observed: -8.4%. The effects compose approximately additively (Δ = +0.4pp), indicating that denoising and quantization target orthogonal deficiencies. Quantization introduces small rounding errors across all weights; denoising removes structural noise in activation space. Neither disrupts the other.

The practical implication: denoising survives INT8 deployment. A model quantized for inference efficiency retains most of the denoising benefit — the combined system is both smaller and better.

Per-domain breakdown reveals an asymmetry: quantization slightly helps math (-2.5%) while slightly hurting code (+5.3%). This pattern persists in the combined condition, suggesting domain-specific interactions between weight precision and activation noise structure.

4.11 LASER Comparison

LASER (Sharma et al., 2024) improves LLM quality by removing higher-order SVD components from weight matrices of specific layers — the same paradox as our work (removing structure helps). We implement LASER on GPT-2 (124M), scanning all 12 layers' MLP down-projections (c_proj) at four rank-reduction levels (keep 99%, 95%, 90%, 80% of singular values — 48 configurations total).

Best LASER result: **layer 5, keep 80% → avg PPL 39.2 (-0.8%)**. Most configurations either had no effect or hurt performance; aggressive reduction at late layers was consistently harmful (layer 10 at 80%: +3.8%).

Method	Avg PPL	Δ vs baseline
Baseline	39.5	—
Our denoising (8 dirs, all layers)	35.6	-9.8%
LASER best (layer 5, 80%)	39.2	-0.8%
LASER + Denoising	35.4	-10.4%

Expected additive: -10.6%. Observed: -10.4%. The methods compose approximately additively, confirming they target different aspects of the model's noise: LASER operates on **weight-space** SVD (modifying the function the layer computes), while our method operates on **activation-space** projection (modifying the signal flowing through the layer). The orthogonality is expected: weight rank reduction changes how information is *processed*, while activation denoising changes what information is *propagated*.

Notably, LASER's best single-layer result (-0.8%) is 12× weaker than our all-layer denoising (-9.8%). This gap likely arises because LASER modifies one layer at a time (losing cross-layer noise interactions) and uses variance ordering (which §3.3 shows is unreliable for identifying functional noise). LASER's strongest layer (5) coincides with our per-layer analysis (§4.9): layer 5 is the second-noisiest layer, consistent with both methods detecting the same underlying structure from different angles.

5. Formalization

All 31 substantive theorems and 6 arithmetic lemmas verified in the Platonic proof language (a Python-native proof language with Lean 4 export capability):

```
elysium/fields/residual_stream_denoising/platonic.py
  37/37 verified (31 substantive + 6 lemmas), 0 proof debt, 0 axioms
  Lean 4 export: 694 lines
```

Chain	Theorems	Content	Nature
Spectral foundation	T1-T7	Eckart-Young, residual bounds, -bridge	Substantive
Distillation theory	T8-T14	Composition, certification, pipeline	Substantive
Hierarchical architecture	T15-T21	Near-identity, bottleneck, Latent-of-Latents	Substantive
Sparse decomposition	T22-T27	Signal dominance, gap, threshold, RIP	Substantive
Orthogonal basis	T28-T31	=0 elimination, ICA certificate	Substantive
Representation denoising	T32-T37	Second-order dominance, noise criterion, decorrelation, composition	Arithmetic lemmas

The first five chains (T1-T31) contain substantive mathematical results about spectral compression, distillation, and hierarchical composition — the theoretical framework that motivated the investigation. The denoising lemmas (T32-T37) formalize the logical scaffolding of the denoising argument; each step is arithmetically simple by design, ensuring the reasoning chain from empirical observation to performance guarantee is gap-free. The disconnect between the framework (31 theorems about compression) and the main result (denoising, 6 lemmas) reflects the paper's trajectory: the compression theory led us to discover denoising, even though the final contribution is primarily empirical.

6. Validation Protocol

To confirm the result generalizes beyond GPT-2:

6.1 Model Scale

Model	d_model	Sensitivity cost	Result
GPT-2 (124M)	768	80 sec	16 dirs (2.1%), PPL -8 to -14%
GPT-2 Medium (355M)	1024	14.5 min	19 dirs (1.9%), PPL -13 to -19%
GPT-2 Large (774M)	1280	28 min	43 dirs (3.4%), PPL -16 to -23%

Model	d_model	Sensitivity cost	Result
LLaMA / Phi / Mistral	varies	~1-2 hr	Architecture generalization (pending)

6.2 Baseline Controls

We tested five ablations on GPT-2 (124M) to verify that the targeted noise directions are genuinely special:

Method	Avg PPL change	Verdict
A) Targeted noise removal (16 dirs)	**-11.1%** (range -8% to -14%)	**Only method that improves**
B) Random direction removal (16 dirs, 5 trials)	+64.4% ± 15.1%	Catastrophic degradation
C) Lowest-variance PCA removal (16 dirs)	+11.7%	Mild degradation
D) Highest-variance PCA removal (16 dirs)	+4,826,400%	Complete model destruction
E) Middle-variance PCA removal (16 dirs)	+74.9%	Severe degradation

The targeted noise directions are >2 better than random removal. The result is not an artifact of "removing any directions helps" — removing random or PCA-ordered directions *hurts* the model dramatically. The noise directions are the *only* set whose removal improves performance.

6.3 Calibration Robustness

We tested whether the noise directions depend on the calibration text by running the full pipeline with three different calibrations:

Calibration	Domain	Universal noise dirs	Avg PPL improvement
A) ML/math text	Technical	16	-11.1%
B) Wikipedia/geography	General knowledge	9	-10.8%
C) Fiction/narrative	Creative writing	13	-14.0%

The number of noise directions varies (9-16), but the denoising effect is remarkably consistent (-10.8% to -14.0%). Different calibrations find different PCA bases (subspace principal angles ~82°, near-orthogonal), yet each basis contains a comparable noise subspace.

This is a nuanced result: the *specific* directions depend on the calibration, but the *amount* of noise and the *improvement* from removing it are model-intrinsic properties, not calibration artifacts. The noise is distributed across the representation space; different PCA bases carve it differently, but each successfully identifies it.

This observation is explained by the random basis control (§6.4): the noise is not PCA-specific but basis-independent.

6.4 Random Orthogonal Basis Control

The calibration robustness results raise a fundamental question: does PCA's alignment with the activation covariance matter, or would *any* orthogonal basis find a comparable noise subspace? We tested this by running the full per-direction sensitivity sweep using 3 uniformly random orthogonal bases (QR decomposition of Gaussian random matrices) on GPT-2 (124M):

Basis	Cross-domain noise dirs	Avg PPL improvement
PCA	**16 (2.1%)**	**-11.1%**
Random-1	6 (0.8%)	-13.0%
Random-2	6 (0.8%)	-12.2%
Random-3	5 (0.7%)	-7.5%

All three random bases find cross-domain noise. The per-text noise fraction is similar across bases (science 0.7-2.1%, history 18-26%, code 21-27%, math 24-28%), and every random basis achieves denoising gains comparable to PCA (-7.5% to -13.0% vs PCA's -11.1%).

The key difference: random bases find fewer cross-domain noise directions (5-6 vs PCA's 16), but each random-basis noise direction has a *larger* per-direction effect. PCA identifies more noise directions because its variance-ordered decomposition partially separates noise from signal; random bases lack this alignment, so the noise is distributed across more directions, and only the strongest noise directions survive the cross-domain intersection filter.

Interpretation: the noise is a **basis-independent geometric property** of the model's residual stream, not an artifact of PCA's covariance alignment. Any orthonormal decomposition reveals it. This explains the calibration robustness (§6.3): near-orthogonal PCA bases from different calibrations give similar denoising because they are all probing the same underlying noise geometry from different angles.

This is actually a stronger result than PCA-specific noise detection: it means the noise is an intrinsic defect of the trained model, detectable by any complete orthonormal probe. PCA remains the practical choice because (1) it finds more noise directions per sweep, and (2) the PCA basis is deterministic given the calibration data.

6.5 Remaining Validation

Experiment	Status	Notes
GPT-2 Large (774M)	Done	43 dirs (3.4%), PPL -16 to -23%
Downstream tasks (pilot)	Done	Zero-shot +25pp, top-1 +5pp (n small, §4.6)
Random basis control	Done	Noise is basis-independent (§6.4)
Layer-selective denoising	Done	Top-8 layers: -12.0% vs uniform -11.1% (§4.9)
CS on cross-domain noise	Done (negative)	F1 < 0.06 even for sparse target (§7.3)
LLaMA / Phi / Mistral	Script ready	Architecture generalization

Experiment	Status	Notes
Full-scale benchmarks	Planned	SST-2 full, MNLI, HellaSwag
Quantization stacking	Done	Approximately additive: denoising survives INT8 (§4.10)
LASER comparison	Done	LASER 12× weaker; methods compose additively (§4.11)

7. Discussion

7.1 Why Do Noise Directions Exist?

One plausible hypothesis: SGD finds a local minimum where some residual stream directions carry gradient-noise artifacts that the optimizer failed to eliminate. These directions may develop adversarial inter-layer interactions — layer ℓ activates direction v, and layer ℓ' amplifies it in a way that hurts prediction. If the gradient signal to correct this is diffuse (distributed across many parameters), the noise could persist through training.

Other explanations are possible: the directions could be artifacts of finite training data, regularization gaps, or architectural constraints (e.g., the residual stream's additive structure forcing all layers to share a common basis). The basis-independence result (§6.4) constrains the hypothesis space: since random orthogonal bases also detect noise, the noise cannot be an artifact of PCA's variance-ordered decomposition — it must be a geometric property of the representation itself, detectable from any angle. This favors explanations involving the model's loss landscape geometry (e.g., saddle-point artifacts, suboptimal inter-layer coupling) over explanations that depend on specific basis alignment. Distinguishing these hypotheses further would require controlled training experiments (e.g., comparing different optimizers, training durations, or architectures), which are beyond the scope of this work.

The model appears to partially compensate (attention may learn to reduce sensitivity to certain directions), but this compensation is imperfect — externally removing the noise direction outperforms the model's implicit workaround.

The per-text / cross-domain gap. A large fraction of directions (26-42%) are noise on any *given* text, but only 1.9-3.4% are noise on *all* calibration texts simultaneously. This gap grows with model scale (§4.5). The natural question: are the per-text-only noise directions genuinely harmful, or are they text-specific "tuning" that the model uses productively on other inputs?

Our evidence strongly supports the latter: removing per-text noise (219 directions from a single text) causes catastrophic failure on downstream tasks (§4.6), while removing only the cross-domain intersection (9-43 directions) yields consistent improvements. This implies that most per-text "noise" directions are in fact *useful for other texts* — they participate in text-dependent computation. Only the cross-domain intersection represents directions that are unconditionally detrimental.

This distinction is central to the method's safety: the cross-domain intersection acts as a conservative filter that avoids removing directions needed for any observed input pattern. The theoretical question of whether per-text noise reflects training artifacts on specific data modes, or functional specialization the model uses selectively, remains open.

7.2 Relationship to Prior Work

Pruning (SparseGPT, Wanda): removes individual weights, not representation directions. Complementary — denoising operates in activation space, pruning in weight space. The two could potentially stack (denoising reduces effective dimension, then pruning compresses the remaining weights).

SliceGPT (Ashkboos et al., 2024): uses PCA of activations to compress by deleting rows and columns, ordered by *variance*. Our results show variance ordering is suboptimal (§3.3) — functional ordering could improve SliceGPT's quality at the same compression ratio.

LASER (Sharma et al., 2024): the most related concurrent approach. LASER removes higher-order SVD components from *weight matrices* of specific layers and finds that this *improves* LLM reasoning — the same paradox as our work (removing structure helps). Three key differences: (1) LASER operates on weight-space SVD; we operate on activation-space PCA. Weight matrices and activation distributions have different spectral structure (§3.1 shows 1.02 for weights vs much higher for activations). (2) LASER selects layers to edit (typically late MLPs) via hyperparameter search; our method identifies directions across all layers simultaneously via functional sensitivity. (3) LASER uses variance ordering (remove highest-order SVD components); our §3.3 shows this is suboptimal — functional ordering identifies a different, more effective noise set. Our empirical comparison (§4.11) confirms: LASER's best single-layer result on GPT-2 (124M) is -0.8% vs our -9.8% — a 12× gap. Crucially, the methods compose approximately additively (combined -10.4%, expected -10.6%), confirming they target orthogonal deficiencies (weight structure vs activation structure).

Activation outliers and quantization (Dettmers et al., 2022; LLM.int8()): identifies that a small number of activation dimensions have extreme magnitudes and must be handled specially during quantization. Our noise directions are conceptually different — they are not outliers in magnitude but in *functional impact* (negative ΔPPL). However, the shared insight is that not all activation dimensions are equal, and dimension-level analysis reveals structure invisible to weight-level methods. Our quantization stacking experiment (§4.10) confirms that denoising composes approximately additively with INT8 quantization — the activation-space projection and weight-space precision reduction target orthogonal aspects of the model.

Representation engineering (Zou et al., 2023; Burns et al., 2023): identifies activation directions corresponding to concepts (truthfulness, harmfulness) and modifies them for steering. Our work identifies directions by *functional impact* rather than *semantic content*. Noise directions may overlap with representation engineering directions — a direction that encodes a concept but hurts prediction could be both semantically meaningful and functionally detrimental. This intersection is unexplored.

Spectral analysis of neural networks (Martin & Mahoney, 2021; heavy-tailed self-regularization): analyzes weight matrix spectra to predict generalization. Our analysis operates on *activation* spectra rather than weight spectra. The finding that activation noise directions exist connects to the broader observation that trained networks have spectral structure beyond what loss-minimization strictly requires.

Mechanistic interpretability (Anthropic SAE, Bricken et al., 2023): identifies *what* features mean. Our work identifies which directions *help or hurt*, orthogonal to semantic interpretation. A synthesis — mapping noise directions to SAE features — could reveal whether noise is semantically meaningless or represents useful-but-poorly-integrated features.

Activation patching (Nanda et al., 2023; path patching; causal tracing): identifies which components contribute to specific outputs by patching activations between clean and corrupted runs. Our sensitivity sweep is related — both measure per-component functional impact — but activation patching targets *semantic* circuits (e.g., indirect object identification) while we target *structural* noise directions that are domain-independent. The brute-force approach here is simpler (no corrupted counterfactual needed) but coarser (direction-level, not component-level).

Model editing (ROME, MEMIT): edits specific facts by modifying weight matrices. Our work removes structural noise, not semantic content.

7.3 Limitations

1. **Single model family** validated (GPT-2 124M, 355M, 774M — three scales, same architecture). The result needs verification on different architectures (LLaMA, Mistral, Phi).

2. **Brute-force cost**: O(d) forward passes per text is feasible up to d=1280 (28 min) but expensive for d=4096+. Despite the strong sensitivity additivity (§4.8), five acceleration strategies all failed: (a) hierarchical group testing: 100% precision but 3.7% recall. (b) LASSO on per-text noise (150-300 measurements): F1 0 — L1 penalty kills small negative values. (c) Short-text transfer (7-token proxy): F1 = 0.36. (d) **LASSO/OMP targeting the sparse cross-domain noise specifically** (50-200 multi-text group measurements, groups of 20): F1 < 0.06 across all methods and measurement counts. Even though the cross-domain noise (2.1%) is genuinely sparse, group measurements (removing 20 random directions per measurement) produce too much measurement noise — the small per-direction effects are masked by the joint nonlinearities of removing many directions together. (e) The only practical speedup: shorter evaluation text (~2× faster per-pass, same pass count). The O(d) sweep remains necessary; for d=4096 this is ~2 hours (one-time cost).

3. **Interaction effects**: removing directions independently ignores potential interactions. The brute-force sweep measures marginal sensitivity; joint removal may differ.

4. **Variance-sensitivity relationship is weak, calibration-dependent, and non-monotonic across scale**: r [+0.24, +0.41] (Small, across 4 calibrations × 3 test texts, mean +0.35 ± 0.06), +0.37 (Medium, single calibration), +0.23 (Large, single calibration). The correlation is never strong (max $|r| = 0.37$), does not increase with scale, and varies meaningfully with calibration text — different PCA bases yield different variance rankings against the same functional sensitivity profile. Variance captures broad trends but cannot identify noise directions or rank the extreme tails of the sensitivity distribution. The practical conclusion: brute-force sensitivity measurement is needed because variance is an unreliable proxy, not a useless one.

5. **Layer-selective denoising (GPT-2 Small only)**: per-layer analysis (§4.9) reveals a U-shape noise profile (early + final layers noisy, middle clean), and selective denoising achieves -12.0% vs uniform -11.1%. This analysis has not been replicated for Medium or Large — the layer profile may differ with depth (24 or 36 layers). A per-layer sweep multiplies cost by n_layers.

6. **No error bars on PPL**: the reported improvements are cross-domain means (4 text domains). The standard deviations across domains are ±2.8% (Small), ±2.2% (Medium), ±3.1% (Large) — all improvements remain positive even at the -1 bound. However, we do not report confidence intervals from repeated PCA decompositions on different calibration samples,

which would capture the stochasticity in the basis itself (§6.3 partially addresses this).

7. **No fine-tuning baseline**: we compare denoising against the unmodified model and against random/variance-ordered removal. A natural additional baseline is minimal fine-tuning (e.g., 100 gradient steps on the calibration text). If the denoising benefit can be matched by a small training budget, the gradient-free claim weakens. This comparison is absent; we note that denoising requires zero training and zero labeled data, whereas even 100-step fine-tuning requires choosing a learning rate, loss function, and data distribution.
8. **PCA basis stability**: the PCA decomposition uses ~50-100 tokens of calibration text. Different calibration *domains* produce similar denoising gains (§6.3), but we do not perform a bootstrap test (resampling different token subsets from the same domain) to measure within-domain basis stability. A systematic sweep of calibration sample size (e.g., 50 vs 500 vs 5000 tokens) is also absent.
9. **No held-out sensitivity validation**: the sensitivity sweep identifies noise directions and evaluates their removal on the same set of 4 test texts. While calibration robustness (§6.3) shows the *denoising effect* generalizes across calibration domains, we do not perform a strict train/test split where noise is identified on texts A,B,C and evaluated on unseen text D. The cross-domain stability of the improvement (§4.4) provides indirect evidence, but a formal held-out protocol would strengthen the claim.
10. **LASER comparison (single architecture)**: our head-to-head comparison with LASER (§4.11) is on GPT-2 Small only. LASER's original results focus on larger models (LLaMA, Phi) where the effect may be stronger. Our finding that LASER is 12× weaker at GPT-2 scale may not hold at larger scales. The composability result (approximately additive) also needs replication on larger models.
11. **Noise is basis-independent (small N)**: the random basis control (§6.4) shows that random orthogonal bases also find cross-domain noise with comparable denoising gains (-7.5% to -13.0% vs PCA's -11.1%). This was tested with N=3 random bases on GPT-2 Small only. A larger trial (N=10+) and replication on Medium/Large would strengthen the basis-independence claim. The practical implication — PCA finds more noise directions per sweep (16 vs 5-6) — also needs validation at larger scales.

8. Conclusion

We presented residual stream denoising: a gradient-free, post-hoc method that improves LLM quality by identifying and removing noise directions from the residual stream. The method requires only PCA decomposition and a brute-force sensitivity sweep — no backpropagation, no hyperparameter search, no additional data.

Seven empirical findings are robust across the GPT-2 family (124M, 355M, 774M):

1. Every model tested contains a small set of noise directions (1.9-3.4% of dimensions) whose removal improves perplexity across all four tested text domains by 8-23% (mean -11.3% to -18.3% ± 3.1%).
2. Within the GPT-2 family (3 model sizes), the denoising benefit increases with scale. Whether this trend extends beyond GPT-2 requires validation on other architectures.
3. PCA variance ranking is weakly and calibration-dependently correlated with functional importance ($|r| < 0.4$) — too unreliable for identifying noise directions, requiring direct functional

measurement.

4. The noise is basis-independent: random orthogonal bases (no covariance alignment) achieve comparable denoising gains (-7.5% to -13.0% vs PCA's -11.1%), establishing that the noise is an intrinsic geometric defect of the trained model.
5. Noise is not uniform across layers: 8/12 layers benefit from denoising while late-middle layers (7-10) are hurt. Selective denoising (skipping the resistant layers) achieves -12.0% vs uniform -11.1%, and the U-shape profile (early + final layers noisy, middle clean) suggests distinct noise origins at layer boundaries.
6. Denoising composes approximately additively with INT8 quantization (expected -8.8%, observed -8.4%): the benefit survives deployment-time precision reduction.
7. Against LASER (weight-space SVD rank reduction), our activation-space denoising is 12× more effective (-9.8% vs -0.8%). The methods compose additively — orthogonal mechanisms targeting different noise types.

The key limitation is the O(d) cost of the brute-force sensitivity sweep and the restriction to a single architecture family. We systematically tested four acceleration strategies (hierarchical group testing, compressed sensing, adaptive sampling, short-text transfer), all of which failed to match brute-force accuracy due to the specific structure of the noise signal. Architecture generalization (LLaMA, Phi, Mistral) and full-scale downstream benchmarks remain necessary before practical deployment.

The theoretical contribution is modest by design: 31 substantive theorems and 6 arithmetic lemmas (all formally verified, Lean 4 export) provide a gap-free logical chain from empirical observations to performance guarantees, with the denoising lemmas (T32-T37) being deliberately simple arithmetic. The nontrivial content is the empirical discovery itself — that trained transformers contain directions whose removal makes them strictly better, and that these directions are identifiable without understanding what they represent.

References

- Ashkboos, S., et al. (2024). "SliceGPT: Compress Large Language Models by Deleting Rows and Columns." ICLR 2024.
- Bricken, T., et al. (2023). "Towards Monosemanticity." Anthropic.
- Burns, C., et al. (2023). "Discovering Latent Knowledge in Language Models Without Supervision." ICLR 2023.
- Comon, P. (1994). "Independent Component Analysis, A New Concept?" Signal Processing.
- Cunningham, H., et al. (2023). "Sparse Autoencoders Find Highly Interpretable Features in Language Models." ICLR 2024.
- Dettmers, T., et al. (2022). "LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale." NeurIPS 2022.
- Eckart, C. & Young, G. (1936). "The approximation of one matrix by another of lower rank." Psychometrika.
- Elhage, N., et al. (2022). "Toy Models of Superposition." Anthropic.
- Frantar, E. & Alistarh, D. (2023). "SparseGPT: Massive Language Models Can Be Accurately Pruned in One-Shot." ICML 2023.
- Martin, C. H. & Mahoney, M. W. (2021). "Implicit Self-Regularization in Deep Neural Networks." Journal of Machine Learning Research.
- Meng, K., et al. (2022). "Locating and Editing Factual Associations in GPT." (ROME)

NeurIPS 2022.
- Nanda, N., et al. (2023). "Attribution Patching: Activation Patching At Industrial Scale." arXiv:2310.10348.
- Sharma, P., Ash, J. T., & Misra, D. (2024). "The Truth is in There: Improving Reasoning in Language Models with Layer-Selective Rank Reduction." (LASER) ICLR 2024.
- Sun, M., et al. (2023). "A Simple and Effective Pruning Approach for Large Language Models." (Wanda) ICLR 2024.
- Zou, A., et al. (2023). "Representation Engineering: A Top-Down Approach to AI Transparency." arXiv:2310.01405.

Code and Data Availability

All experimental code (sensitivity sweeps, baseline controls, calibration robustness, downstream evaluation), formally verified theorems (Platonic + Lean 4 export), and sample outputs are available at [repository URL upon publication]. The experiments use publicly available models (GPT-2 family via HuggingFace) and require only a single GPU (RTX 3090 or equivalent) for the largest model (GPT-2 Large, d=1280).