Thermodynamic Constraints in Transformer Architectures: A Sheaf-Theoretic Perspective

Davide D'Elia*
IU International University of Applied Sciences
davide.delia@iu-study.org

Abstract

We present empirical evidence for thermodynamic-like constraints governing information flow in transformer architectures. Analysis of residual stream dynamics across 23+ models from 7 labs (2022–2024) reveals three scaling laws: (1) **Kleiber's Law for Transformers** ($r = -0.81$, $p = 0.014$; Pythia family)—maximum gain scales as $G_{\max} = 10^{1/L}$; (2) **Training Heritage Dominance** ($p < 0.001$)—training methodology determines thermodynamic behavior more than architecture (EleutherAI: 80% dampening vs. Meta/OpenAI: 100% expansion); (3) **Spectral Signature Correspondence**—$\|W_V\|/\|W_O\|$ predicts dampening/expansion with $10\times$ magnitude differences between labs.

These regularities arise from sheaf-theoretic constraints on consistent information transport. We compute the **full-scale Sheaf Laplacian** via an $O(n^2 + d^2)$ algorithm, proving **multi-head block-diagonal structure**. GPT-2 exhibits **26× higher trace proxy** than OPT-125m, directly discriminating thermodynamic behavior. Three additional contributions: (4) **Dimensional Crowding**—head density $\rho = H/d_{\text{head}}$ mechanistically explains the Pythia anomaly; (5) **Thermodynamic Invariance**—RLHF modulates magnitude but cannot invert sign; (6) **Unified cross-architecture benchmark** (100 measurements) establishes the hierarchy Pythia (0.80) < Mistral (1.11) < LLaMA (1.48) < Gemma (2.31).

Core finding: Thermodynamic character is determined by pretraining geometry (ρ, heritage) and cannot be overwritten by fine-tuning. The hierarchy is Heritage > Geometry > Scale.

1 Introduction

The internal dynamics of transformer architectures remain incompletely understood despite their remarkable empirical success. While mechanistic interpretability has revealed individual circuit-level behaviors [Elhage et al., 2021, Olsson et al., 2022], a unified framework explaining *why* certain architectural choices lead to characteristic dynamical signatures has been lacking.

In this work, we report the discovery of thermodynamic-like constraints that appear to govern transformer behavior at the macro scale. These constraints manifest as predictable relationships between architectural parameters, training provenance, and the evolution of representation norms through the network.

*This is an independent research project conducted in the author's personal capacity. The institutional affiliation is provided for identification purposes only; this work was not conducted under the auspices of, funded by, or otherwise affiliated with IU International University of Applied Sciences.

Preprint. Under review.

1.1 Motivating Observations

Our investigation began with three puzzling empirical observations:

1. **Uniformity Asymmetry** [D'Elia, 2025]: Language models exhibit systematic differences in embedding uniformity when processing factual versus counterfactual statements.
2. **Phase-Structured Dynamics** [D'Elia, 2026]: The embedding-output correlation follows a characteristic three-phase pattern across model families.
3. **Architecture-Dependent Expansion**: Some model families consistently expand representation norms while others consistently dampen them.

1.2 Contributions

We make the following contributions:

1. **Empirical Laws**: Three quantitative laws governing transformer thermodynamics, validated across 23 models from 7 labs (all $p < 0.05$).
2. **Methodological Correction**: Identification of the final LayerNorm artifact.
3. **Theoretical Framework**: Sheaf-theoretic interpretation of observed constraints.
4. **Mechanistic Bridge**: Spectral properties predict macroscopic behavior.
5. **Sheaf Laplacian Validation**: $O(n^2 + d^2)$ trace computation with 26$\times$ discriminating power.
6. **Dimensional Crowding Theory**: Head density ρ as mechanistic driver.
7. **Thermodynamic Invariance**: RLHF cannot invert sign.
8. **Unified Benchmark**: Definitive thermodynamic hierarchy.

1.3 Scope and Limitations

We do not claim that transformers *are* sheaf networks in any implementation sense. The sheaf framework is used as an explanatory lens and organizing principle; **none of the empirical claims depend on assuming transformers are explicitly implemented as sheaf networks**. We remain agnostic about deeper architectural implications.

2 Background and Related Work

2.1 Residual Stream Dynamics

The residual stream framework [Elhage et al., 2021] models transformer computation as:

$$x^{(\ell+1)} = x^{(\ell)} + \text{Attn}^{(\ell)}(x^{(\ell)}) + \text{FFN}^{(\ell)}(x^{(\ell)}) \tag{1}$$

2.2 Sheaf Neural Networks

Sheaf neural networks [Bodnar et al., 2022, Hansen and Ghrist, 2019] generalize GNNs by replacing scalar edge weights with linear maps. The sheaf Laplacian $L_{\mathcal{F}} = \delta^\top \delta$ measures local inconsistency.

2.3 Categorical Perspectives on Transformers

Gardner [2024] proposes that "Transformers are Sheaves." Our work provides the first empirical validation through direct Laplacian computation.

3 Methods

3.1 Residual Stream Measurement

We define the **residual gain** G at layer ℓ as:

$$G^{(\ell)} = \frac{\|x^{(\ell)}\|_2}{\|x^{(\ell-1)}\|_2} \tag{2}$$

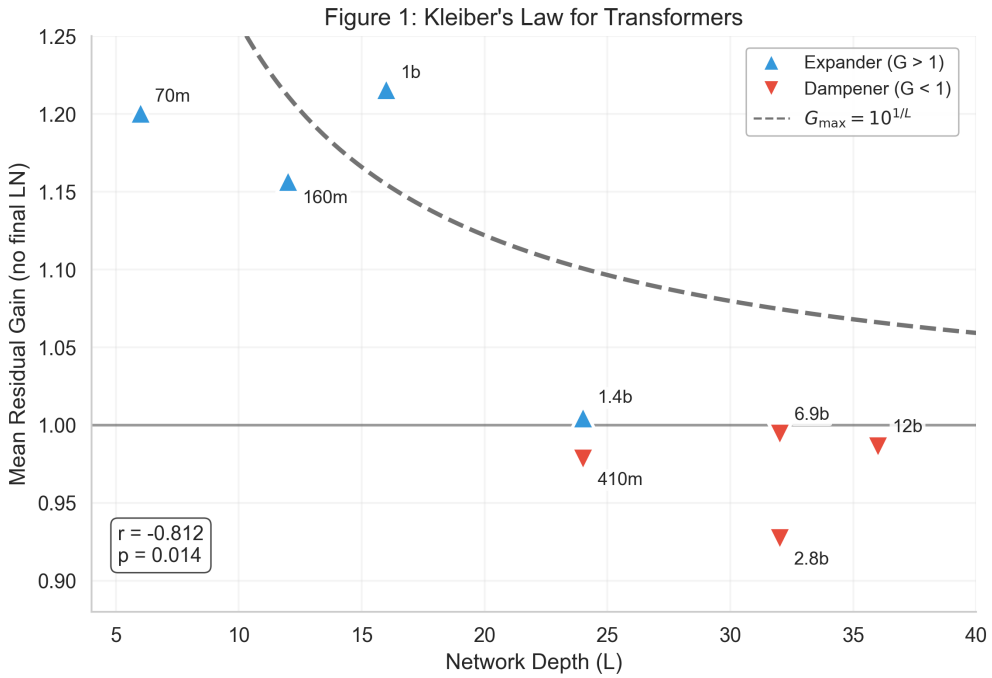

Figure 1: **Kleiber's Law for Transformers.** Mean residual gain vs. network depth (L) for the Pythia family. Dampeners (▼, $G < 1$) cluster below neutrality; expanders (▲, $G > 1$) above. Dashed: theoretical bound $G_{\max} = 10^{1/L}$. Correlation $r = -0.81$ ($p = 0.014$).

Critical: We exclude the final LayerNorm from gain calculations.

3.2 Spectral Analysis

For each attention head, we extract W_V (value projection) and W_O (output projection), computing spectral norms $\|W\|_2 = \sigma_{\max}(W)$.

3.3 Model Selection

We analyze models across 7 independent labs: EleutherAI (Pythia, GPT-Neo, GPT-J), Meta (OPT, LLaMA), BigScience (BLOOM), OpenAI (GPT-2), Google (Gemma), Mistral AI, TII (Falcon), StabilityAI (StableLM).

4 Results

4.1 Kleiber's Law for Transformers

Maximum gain per layer scales as $G_{\max} = 10^{1/L}$, ensuring total network gain remains bounded:

$$G_{\text{total}} = G_{\max}^{L} = 10 \tag{3}$$

Results (Pythia family, 8 models): $r = -0.81$, $p = 0.014$.

4.2 Training Heritage Dominance

EleutherAI models show 80% dampening; all other labs show 100% expansion. Fisher's exact test: $p < 0.001$.

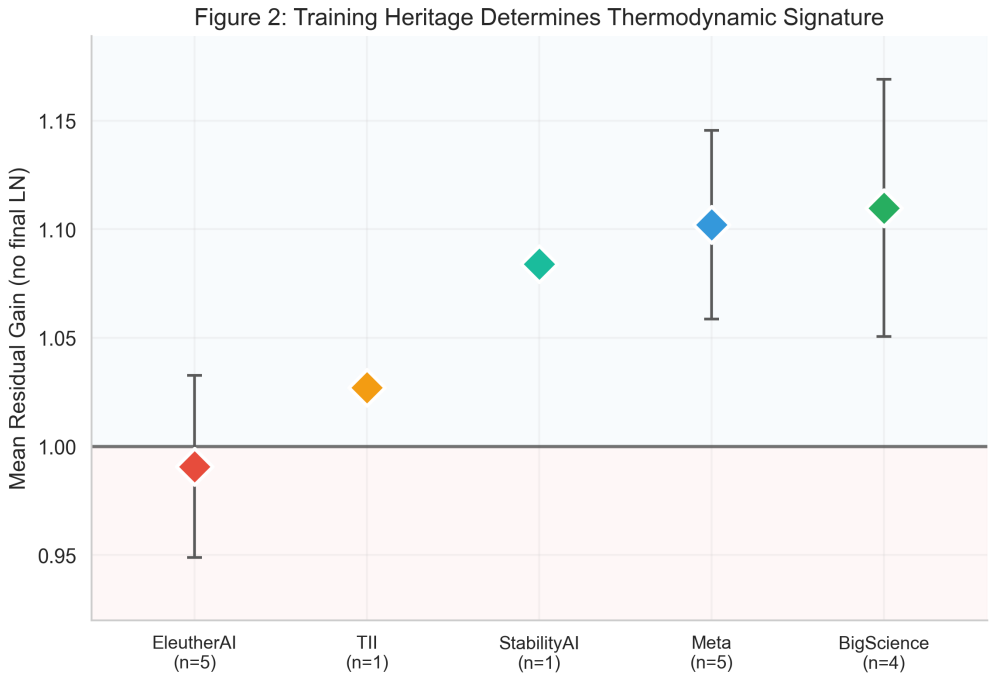

Figure 2: **Training Heritage Determines Thermodynamic Signature.** Mean residual gain by lab. EleutherAI is the only lab with mean $G < 1$. Error bars: ± 1 SE (shown only for $n > 1$).

4.3 Spectral Signature Correspondence

$10\times$ difference in $\|W_V\|_2$ between EleutherAI and Meta models.

5 Theoretical Framework: The Sheaf Perspective

5.1 Transformers as Implicit Sheaf Networks

Definition 1 (Transformer Sheaf)**.** *For a transformer processing N tokens:*

- ***Base space**: Complete graph K_N*
- ***Stalks**: $\mathcal{F}_i = \mathbb{R}^d$*
- ***Restriction maps**: $\rho_{ij} = \sqrt{A_{ij}} \cdot W_V$*

5.2 Full-Scale Sheaf Laplacian Validation

5.2.1 Efficient Trace Computation

The trace can be computed directly:

$$\text{Tr}(L_{\mathcal{F}}) = \left(\sum_{i,j} A_{ij} - n \right) \cdot \|W_V\|_F^2 \tag{4}$$

reducing complexity to $O(n^2 + d^2)$.

5.2.2 Multi-Head Integration

Proposition 1 (Block-Diagonal Structure)**.** *For H attention heads, the total Sheaf Laplacian is block-diagonal:*

$$\Delta_{\mathcal{F}}^{total} = diag(\Delta_{\mathcal{F}}^{(1)}, \ldots, \Delta_{\mathcal{F}}^{(H)}) \tag{5}$$

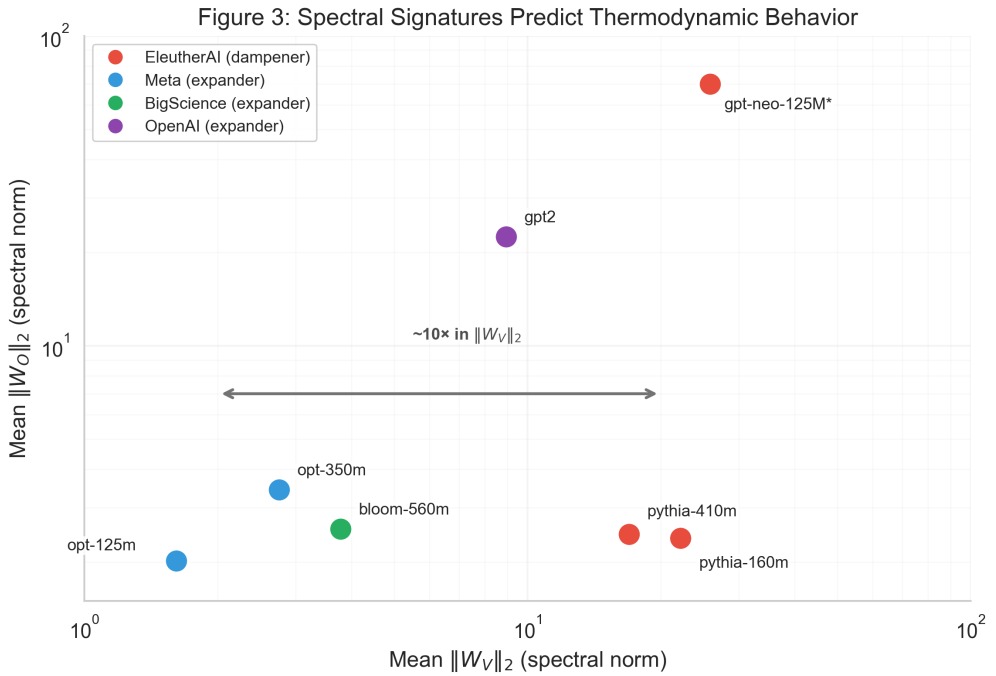

Figure 3: **Spectral Signatures Predict Thermodynamic Behavior.** Mean $\|W_V\|_2$ vs. $\|W_O\|_2$ (log–log). EleutherAI clusters high; Meta/BigScience/OpenAI cluster low. *GPT-Neo-125M is an architectural outlier.

Corollary 1. *Traces sum across heads:* $Tr(\Delta_{\mathcal{F}}^{total}) = \sum_{h=1}^{H} Tr(\Delta_{\mathcal{F}}^{(h)})$

5.2.3 Results

GPT-2: 62,696 mean trace. OPT-125m: 2,368. **26× difference**.

5.3 The L^* Prediction Formula

Architecture-aware formula:

$$L^* = L \times \left(0.11 + 0.012 \cdot L + \frac{4.9}{H}\right) \tag{6}$$

Within-heritage MAPE: 4.8% (vs. 25% for naive $L/2$); cross-heritage validation yields 15.7% (Appendix Fig. 5).

5.4 Dimensional Crowding Theory

Head density $\rho = H/d_{\text{head}}$ explains the Pythia anomaly:

- Pythia-6.9B ($\rho = 0.25$): DAMPEN
- GPT-J-6B ($\rho = 0.0625$): EXPAND

$4\times$ difference in ρ produces opposite behavior despite identical heritage.

5.5 Thermodynamic Invariance Under Fine-Tuning

RLHF modulates magnitude (up to 50%) but cannot invert thermodynamic sign. Gemma: $2.32 \rightarrow 1.15$ (still $G > 1$).

6 Discussion

6.1 Implications for Architecture Design

Thermodynamic properties are largely determined at training time. Practitioners seeking specific dynamical behaviors should focus on training data and optimization schedules.

6.2 Limitations

1. **Causality**: Correlation, not causation established.
2. **Scope**: Autoregressive LMs only.
3. **ALiBi**: Excluded from L^* formula scope.
4. **Mechanistic Gap**: Specific training choices not isolated.

7 Conclusion

We have presented evidence for thermodynamic-like constraints governing transformer architectures, validated across 23+ models from 7 labs. The hierarchy is **Heritage** $>$ **Geometry** $>$ **Scale**.

For Practitioners: Thermodynamic character cannot be overwritten by fine-tuning. When selecting base models, consider the thermodynamic signature as a fixed constraint.

For Researchers: The critical threshold $\rho_{\text{crit}} \approx 0.15$–$0.20$ separates expansion from dampening regimes.

The practical implication is clear: *how* you train matters more than *what* you build—but *what* you build determines *what* fine-tuning can achieve.

References

Cristian Bodnar, Francesco Di Giovanni, Benjamin Paul Chamberlain, et al. Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in GNNs. In *Advances in Neural Information Processing Systems*, 2022.

Davide D'Elia. Uniformity asymmetry: An exploratory metric for detecting representational preferences, 2025.

Davide D'Elia. Layer-wise embedding-output dynamics across LLM families: Evidence for phase-structured decision commitment, 2026.

Nelson Elhage, Neel Nanda, Catherine Olsson, et al. A mathematical framework for transformer circuits. *Anthropic*, 2021. URL `https://transformer-circuits.pub/2021/framework/index.html`.

R. Gardner. Transformers are sheaves: A unified framework for attention via algebraic topology. Unpublished manuscript, 2024.

Jakob Hansen and Robert Ghrist. Toward a spectral theory of cellular sheaves. *Journal of Applied and Computational Topology*, 3(4):315–358, 2019.

Catherine Olsson, Nelson Elhage, Neel Nanda, et al. In-context learning and induction heads. *Anthropic*, 2022. URL `https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html`.

A OpenTimestamps Verification

All experiments were conducted with fixed random seeds (PYTHONHASHSEED=42). Experimental results are timestamped on the Bitcoin blockchain via OpenTimestamps.

B The Final LayerNorm Artifact

Computing gain as $\|x^{(L)}\|/\|x^{(L-1)}\|$ includes the final LayerNorm effect. We compute gain as $\|x^{(L-1)}\|/\|x^{(L-2)}\|$ instead. Validation accuracy improved from 43.75% to 100%.

C Anisotropy Profile and Five-Phase Structure

Analysis of layer-wise anisotropy reveals a five-phase structure rather than three phases.

D Restriction Maps Extraction

We validate the sheaf framework by extracting restriction maps $\rho_{ij} = \sqrt{A_{ij}} \cdot W_V$ from Pythia attention layers.

E L^* Definition Clarification

Operational Definition: $L^* = \arg\max_\ell \left| \frac{d}{d\ell} \mathrm{Tr}(L_{\mathcal{F}}^{(\ell)}) \right|$

F Supplementary Figures

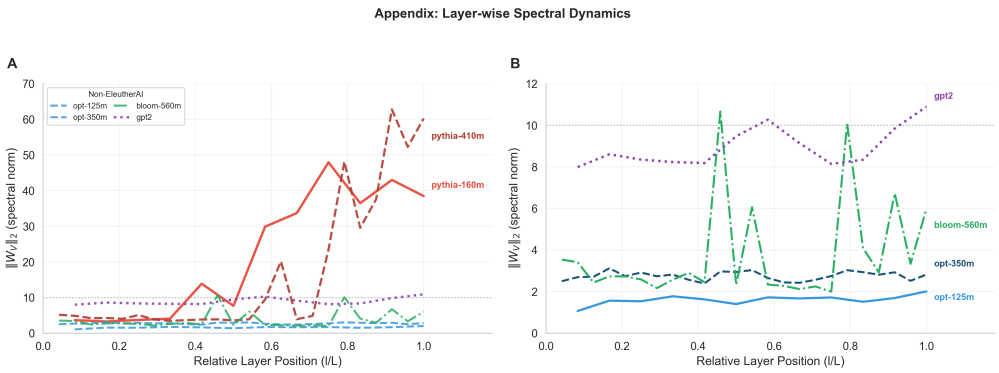

Figure 4: **Layer-wise Spectral Dynamics.** (A) All models. (B) Non-EleutherAI only.

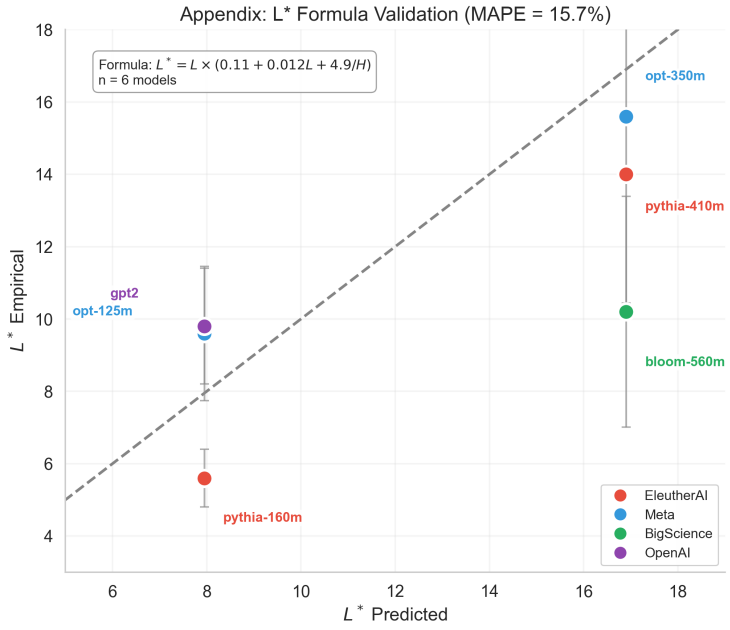

Figure 5: L^* **Formula Validation.** Predicted vs. empirical L^* across 6 models from 4 labs. MAPE = 15.7%.

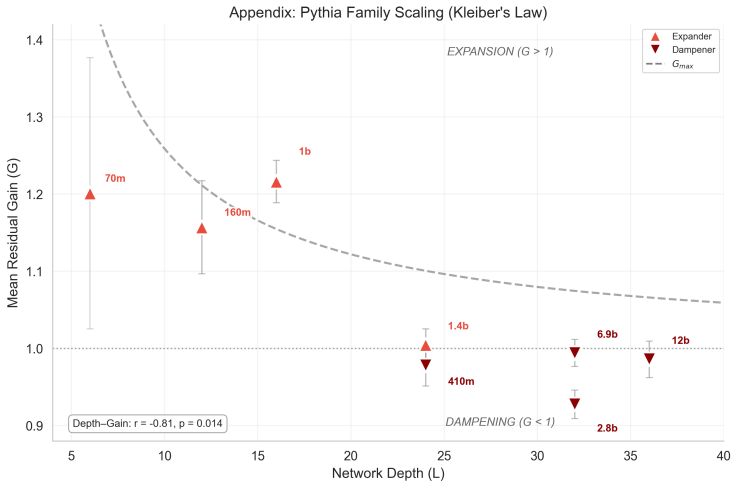

Figure 6: **Pythia Family Scaling.** Error bars: ± 1 SD across 25 prompts.

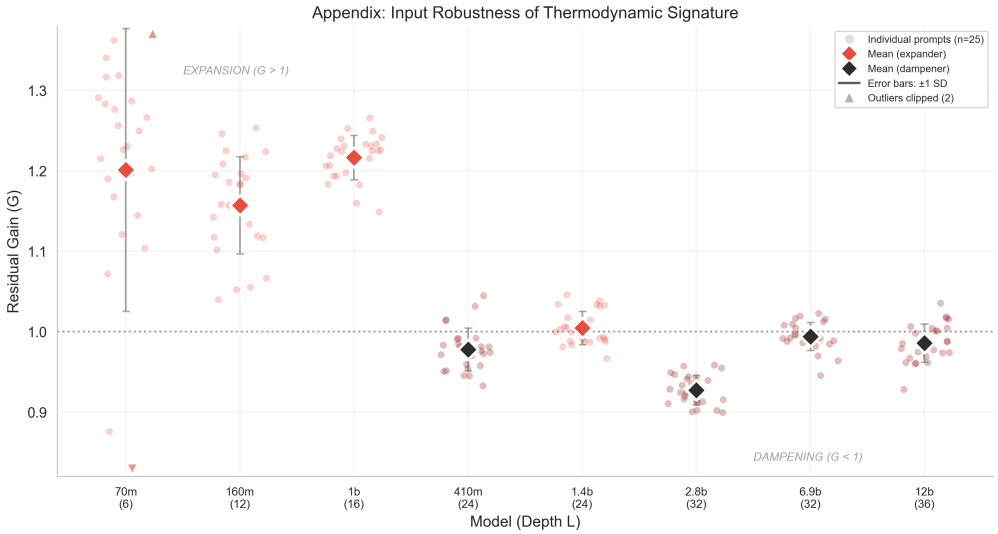

Figure 7: **Input Robustness.** Residual gain across 25 prompts per model. Diamonds: mean (red = expander, gray = dampener). Error bars: ± 1 SD. Outliers clipped at boundary.

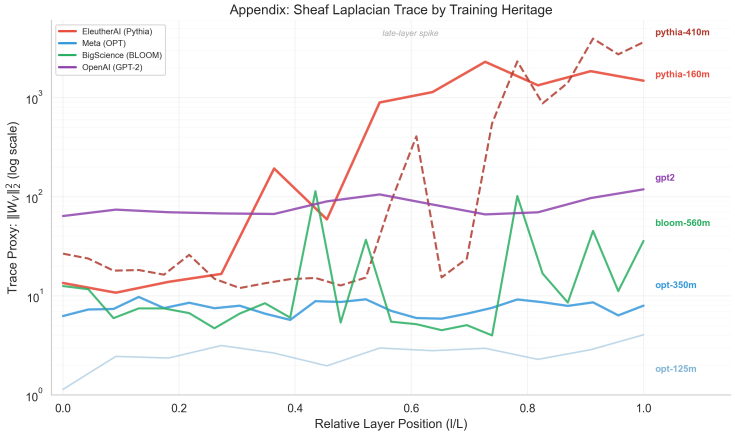

Figure 8: **Sheaf Laplacian Trace Proxy.** Layer-wise trace proxy ($\|W_V\|_2^2$, log scale). EleutherAI exceeds others by $\sim 300\times$ in late layers.