Transformer Dynamics: A Neuroscientific Approach to Interpretability of Large Language Models

Jesseba Fernando[1,2], Grigori Guitchounts[3]
[1]Network Science Institute, Northeastern University, [2] Institute for Experiential AI, Northeastern University, [3]Independent
fernando.je@northeastern.edu, g.guitchounts@alumni.harvard.edu

Abstract

As artificial intelligence models have exploded in scale and capability, understanding of their internal mechanisms remains a critical challenge. Inspired by the success of dynamical systems approaches in neuroscience, here we propose a novel framework for studying computations in deep learning systems. We focus on the residual stream (RS) in transformer models, conceptualizing it as a dynamical system evolving across layers. We find that activations of individual RS units exhibit strong continuity across layers, despite the RS being a non-privileged basis. Activations in the RS accelerate and grow denser over layers. In reduced-dimensional spaces, the RS follows a curved trajectory with attractor-like dynamics in the lower layers. These insights bridge dynamical systems theory and mechanistic interpretability, establishing a foundation for a "neuroscience of AI" that combines theoretical rigor with large-scale data analysis to advance our understanding of modern neural networks (Fernando & Guitchounts, 2025).

Transformer Residual Stream (RS) Activations Grow Dense and are Highly Correlated Over the Layers

In this work, we apply dynamical systems (DS) analysis to the residual stream of transformer models, which serve as the core information channel connecting successive layers. Our approach views the residual stream as a high-dimensional state evolving across layers. This perspective builds upon recent work on privileged bases in transformers (Elhage, Lasenby, & Olah, 2023; Gromov, Tirumala, Shapourian, Glorioso, & Roberts, 2025) and representational alignment across neural networks (Brown, Vyas, & Bansal, 2023; Khosla, Williams, McDermott, & Kanwisher, 2024).

Our initial investigation revealed that RS activations in Llama 3.1 (8B) at the last token position increase in magnitude over the layers (Fig. 1B). Most units showed low-magnitude activations in early layers, but progressively increased. Sorting the mean activations across $N = 1000$ data batches by the mean activation at the last layer, $\pi = \text{argsort}(\bar{\mathbf{h}}^{2L})$, revealed that activations not only grow dense as the layers progress, but that units tend to preserve their sign over the layers.

To quantify representation continuity between layers, we analyzed pairwise correlations between layers, examining within-layer transitions ($\mathbf{h}_l^{Attn} \rightarrow \mathbf{h}_l^{MLP}$) and cross-layer transitions ($\mathbf{h}_l^{MLP} \rightarrow \mathbf{h}_{l+1}^{Attn}$)(Fig. 1C). The within-layer correlations were consistently higher than the cross-layer correlations, suggesting different information processing regimes of attention and MLP operations. The correlation strength increased over layers for both types, starting high ($r > 0.8$) even in early layers. For each unit, the distribution of correlations over the 63 layer transitions was binned into intervals $[0, 1]$ to create a density plot (Fig. 1D), revealing that despite the RS being a nonprivileged basis, most units maintain strong correlations throughout the network.

We examined RS vector changes by measuring cosine similarity of layerwise vector pairs and computing representation velocity. Cosine similarity increased as a function of layer depth, with within-layer ($\mathbf{h}_l^{Attn} \rightarrow \mathbf{h}_l^{MLP}$) transitions more similar than cross-layer (Fig. 1E). The velocity profile showed a distinct acceleration pattern through the model, with relatively constant velocities in early layers followed by a slight increase in velocity, with the steepest acceleration occurring in the final third. (Fig. 1F). This progressive acceleration of representational change, combined with our observations of increasing activation magnitudes and correlations, suggests that the transformer RS systematically amplifies certain representational directions in later layers.

Analysis of mutual information (MI) for given RS units at successive layers revealed three key phenomena (Fig. 1G,H): (1) sharp MI decline in early layers, especially prominent at cross-layer transitions; (2) simultaneous increase in linear correlations between layers (Fig. 1C,D); and (3) coincidence with growing activation magnitudes through the layers (Fig. 1B).

The apparent paradox between decreasing MI and increasing correlations suggests the model redistributes information while favoring a simpler, linearly-aligned features in later layers.

RS Trajectories Exhibit Attractor-like Dynamics in Lower Layers

PCA demonstrated that early layers distribute variance across more dimensions than later layers, with later layers needing fewer principal components to explain the same amount of variance (Fig. 1I,J,K).

RS trajectories visualized in PCA space revealed systematic patterns in representations evolution, with consistent paths for individual trajectories and layer-wise means (Fig. 1A). This suggests a structured computation process, with slightly offset trajectories for the within-layer $\mathbf{h}_l^{Attn} \rightarrow \mathbf{h}_l^{MLP}$ and cross-layer $\mathbf{h}_l^{MLP} \rightarrow \mathbf{h}_{l+1}^{Attn}$ transitions.

To test trajectory stability, we performed perturbation analysis by "teleporting" RS states to various points in the PCA space at different layers (Fig. 1L). We hypothesized that RS progression through transformer layers might reveal attractor-like dynamics, such that moving the activations to various portions of this phase space would eventually bring them back to the mean trajectory.

Response to perturbations varied systematically with layer depth. Early layers (the first, i.e. layer 0, and layer 7) showed

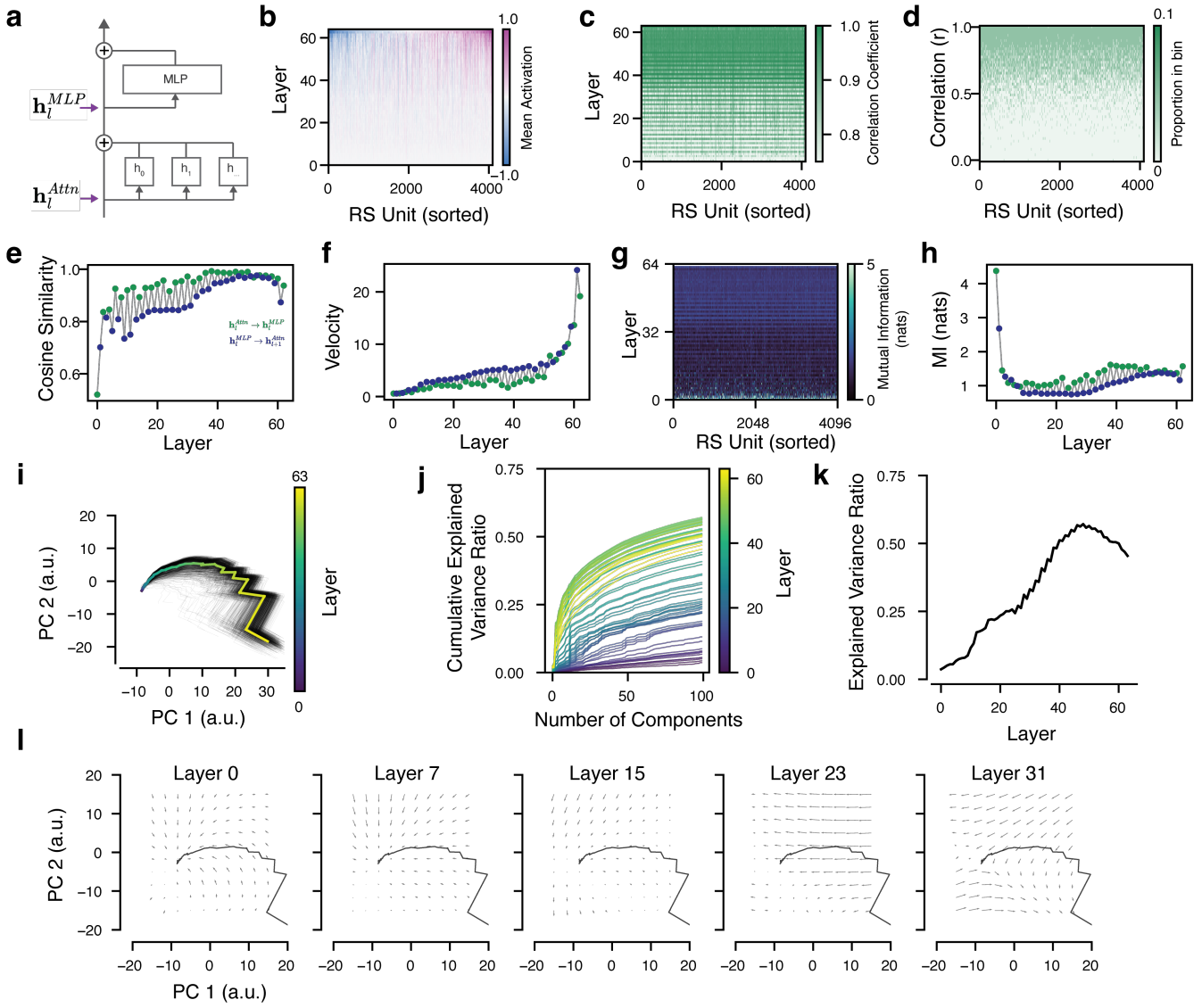

Figure 1: **Transformer residual stream (RS) activations grow dense over the layers, are highly correlated among successive layers, and exhibit nonstationary dynamics. A:** Activations of the transformer RS were captured before layernorm and the attention operation (pre-Attn) and before the MLP at each layer of Llama 3.1 8B, resulting in $\mathbf{64} \times \mathbf{4096}$. 'layers' by 'units'. Activations were analyzed at the last token position for data samples from the WIKITEXT-2-RAW-V1 dataset. **B:** Mean activations across $N = 1000$ samples. **C:** Correlations of activations for unit u between layer l and $l + 1$ over data samples. **D:** Histogram of correlations across layers for each unit. **E:** Cosine similarity among pairs of RS vectors $\mathbf{h}_l^{Attn} \rightarrow \mathbf{h}_l^{MLP}$ (green) and $\mathbf{h}_l^{MLP} \rightarrow \mathbf{h}_{l+1}^{Attn}$ (blue). **F:** Velocity V of the RS vectors. **G:** Mutual information (MI) among pairs of activations for unit u between layer l and $l + 1$ over data samples. **H:** MI over the layers, averaged across units in the RS. **Perturbation of RS trajectories Reveals Attractor-like Dynamics I:** Trajectories of $n = 1000$ individual (black) and mean (colored by layer) data samples in PCA space. **J:** Cumulative explained variance of the trajectories as a function of the number of components. **K:** Explained variance per layer using 100 PC components. **L:** Perturbation analysis in which trajectories were 'teleported' to various points, at various stages in the RS (indicated by layer number above each subplot). Gray line shows unperturbed control trajectory. Quiver arrows indicate direction and magnitude of teleported trajectories based on the successive 12 sublayers after teleportation.

a pull toward the initial states if perturbed. Perturbation at later layers showed systematic flows as well, but not toward the original positions. This preliminary result suggests that the transformer develops stable computational channels that maintain desired trajectories, possibly self-correcting through its dynamics. Further work is required to ascertain the computational consequences of such attractor-like dynamics and their architectural and/or training-related origins.

References

Brown, D., Vyas, N., & Bansal, Y. (2023). *On privileged and convergent bases in neural network representations.* Retrieved from https://arxiv.org/abs/2307.12941

Elhage, N., Lasenby, R., & Olah, C. (2023, March 16). Privileged bases in the transformer residual stream. *Transformer Circuits Thread*. Retrieved from https://transformer-circuits.pub/2023/privileged-basis/index.html (Anthropic)

Fernando, J., & Guitchounts, G. (2025). Transformer dynamics: A neuroscientific approach to interpretability of large language models. *arXiv preprint arXiv:2502.12131*.

Gromov, A., Tirumala, K., Shapourian, H., Glorioso, P., & Roberts, D. A. (2025). *The unreasonable ineffectiveness of the deeper layers.* Retrieved from https://arxiv.org/abs/2403.17887

Khosla, M., Williams, A. H., McDermott, J., & Kanwisher, N. (2024). Privileged representational axes in biological and artificial neural networks. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/early/2024/06/20/2024.06.20.599957 doi: 10.1101/2024.06.20.599957