

# Analyzing Spectral Information of Transformers

Vo Huu Anh Tuan  
Korea University  
South Korea

Anh Tong  
Korea University  
South Korea

## Abstract

We present a comprehensive framework for analyzing transformer models through spectral decomposition of their Jacobian matrices. Our method computes layer-wise Jacobians for both attention and MLP components, applying singular value decomposition to extract spectral properties that reveal the internal dynamics of transformer computation. Our preliminary analysis on GPT-2 model reveals interesting spectral radius dynamics - three-phase computational architecture: feature extraction, representation refinement, and output preparation. Our findings provide new mathematical insights into transformer information processing dynamics and offer a principled framework for mechanistic interpretability that complements existing attention-based analysis methods.

## Keywords

Transformer, LLM, Explainable AI, Spectral analysis

## 1 Introduction

Understanding the internal mechanisms of transformer architectures [15] remains one of the most pressing challenges in deep learning interpretability. While attention visualization and activation patching have provided valuable insights, these methods often lack the mathematical rigor needed to characterize the fundamental dynamics governing transformer computation. Recent advances in explainable artificial intelligence [6, 7, 9] have emphasized the need for principled mathematical frameworks that can reveal the underlying computational principles of these complex models.

In this work, we introduce a spectral analysis framework based on Jacobian decomposition that provides quantitative insights into transformer dynamics. Our approach computes layer-wise Jacobians for attention and MLP components, then applies singular value decomposition (SVD) to extract spectral properties including eigenvalue distributions, spectral radii. These metrics reveal fundamental characteristics of information processing, stability properties, and hierarchical organization within transformer models.

Our analysis of GPT-2’s spectral norms reveals a distinctive U-shaped pattern across layers: high values in the first and last layers with stable intermediate values. This pattern suggests a three-phase computational architecture where initial layers perform aggressive feature transformation, middle layers conduct stable incremental processing, and final layers concentrate information for output generation. The moderate spectral norms in intermediate layers contribute to training stability by avoiding gradient problems throughout the network depth. Additionally, we found that attention and feedforward blocks exhibit complementary spectral characteristics. Attention shows higher eigenvalue decay rates, indicating concentrated spectral energy in dominant modes that enable

focused, selective information routing. In contrast, feedforward blocks display more distributed spectral energy across multiple modes, facilitating broader representational refinement throughout the residual stream.

## 2 Related work

*Transformer Interpretability.* The field of transformer interpretability has evolved rapidly, with approaches ranging from attention visualization to mechanistic circuit discovery [10]. Recent work has emphasized the importance of causal interventions and systematic validation of interpretability claims [2]. However, most existing methods lack the mathematical foundation needed to characterize the fundamental dynamics of transformer computation.

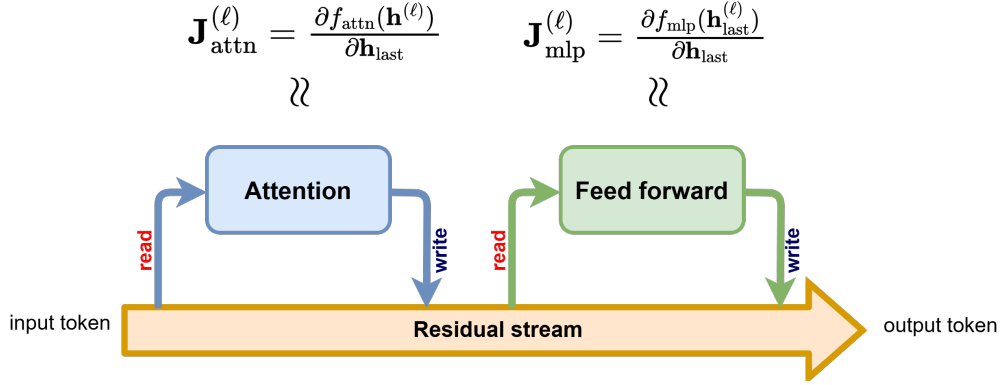
*Spectral Analysis in Neural Networks.* Spectral analysis has been applied to neural networks primarily in the context of training dynamics and generalization [4]. Recent work has explored eigenvalue distributions in attention mechanisms [12] and the role of spectral properties in model expressivity [11]. However, comprehensive Jacobian-based spectral analysis of transformers remains largely unexplored.

*Spectral Analysis in LLMs.* Staats et al. [13] uses Random Matrix Theory to analyze weight matrix spectra across BERT, Pythia, and LLaMA models. The research identifies systematic deviations from the Marchenko-Pastur law, indicating learned structures, and crucially demonstrates that small singular values significantly impact performance, contradicting traditional pruning assumptions. Recent gradient-based methods have evolved beyond simple attribution to sophisticated mechanistic analysis. Edge Attribution Patching (EAP) now outperforms automated circuit discovery methods by approximating activation patching with just two forward passes and one backward pass [3]. Recent work [8] showed that the integration of Layer-wise Relevance Propagation [1] with gradient-guided attention mechanisms has created more robust attribution methods specifically designed for transformer architectures.

*Neural ODE Perspectives.* The connection between transformers and dynamical systems has been explored through Neural Ordinary Differential Equation (ODE) frameworks [14]. Tong et al. [14] demonstrated the application of Lyapunov exponents to transformer analysis, while other research has examined attention mechanisms as particle systems [5]. These approaches provide theoretical motivation for our Jacobian-based analysis.

## 3 Background

The transformer architecture is built upon three fundamental building blocks: self-attention, multi-head attention, and feed-forward neural networks. We provide a concise overview of these core components below.



**Figure 1: Transformer block architecture showing attention and feedforward components with residual connections. The attention layer reads information from all token positions in the residual stream and writes updates back to the current token’s stream. The Jacobian captures how changes in input representations propagate through these transformations, providing a linear approximation of the local dynamics around the current operating point.**

*Self-attention mechanism.* The self-attention operation relies on three key elements: queries, keys, and values. Given input data  $X \in \mathbb{R}^{n \times d_x}$  and linear transformation matrices  $W_q, W_k, W_v \in \mathbb{R}^{d_{\text{attn}} \times d_x}$  for queries, keys, and values respectively, we compute  $Q = XW_q^T$ ,  $K = XW_k^T$ , and  $V = XW_v^T$ . The self-attention function is then defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_x}}\right)V \quad (1)$$

The self-attention mechanism can be understood as a process for identifying which positions in a sequence should receive focus. For a sequence  $X$  with elements  $x_1, x_2, \dots, x_n$ , each position  $t$  generates a query  $q_t = x_t W_q^T$  that is compared against all keys  $k_\tau = x_\tau W_k^T$  using the similarity function  $\kappa(q_t, k_\tau) = \frac{\exp(q_t k_\tau^T)}{\sum_s \exp(q_t k_s^T)}$ . The output at position  $t$  is computed as a weighted combination  $\sum_{\tau=1}^n \kappa(q_t, k_\tau) v_\tau$  where  $v_\tau = x_\tau W_v^T$ . Higher values of  $\kappa(q_t, k_\tau)$  indicate that the model should pay more attention to  $v_\tau$ .

*Multi-head attention.* This mechanism extends single-head attention by computing multiple attention functions in parallel and combining their outputs. It is formulated as:

$$\text{MultiHead} = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W_o \quad (2)$$

$$\text{where head}_i = \text{Attention}(Q_i, K_i, V_i) \quad (3)$$

where each head uses its own parameter matrices  $W_q^i, W_k^i, W_v^i$ . The matrix  $W_o$  serves as an output projection.

Multi-head attention enables the model to capture diverse representational patterns and encode more comprehensive information. This approach helps mitigate the sparsity issues that can arise from the softmax operation in standard attention.

*Feed-forward network.* The final transformer component is a position-wise feed-forward network consisting of two linear transformations with a non-linear activation function. It typically employs the Gaussian Error Linear Unit (GELU):

$$\text{FFN}(x) = \text{GELU}(xW_1^T + b_1)W_2^T + b_2 \quad (4)$$

where  $W_1$  and  $W_2$  are weight matrices and  $b_1, b_2$  are bias vectors.

Research has extensively studied the role of FFN layers, characterizing them as information storage mechanisms that contribute to the emergent capabilities observed in large-scale transformers [6].

## 4 Method

In this section, we describe our approach for analyzing transformer architectures through the spectral properties of their core components. We begin by motivating the importance of Jacobian analysis for attention and feedforward blocks, then detail our computational methods and key metrics.

### 4.1 Motivation: Why analyze Jacobian spectra of transformer components?

The residual stream in transformer architectures serves as a high-dimensional information highway where each layer reads from and writes to shared representational subspaces [10]. Understanding how attention and MLP transformations interact with this stream is crucial for mechanistic interpretability, as these operations fundamentally determine what information flows through the network and how it gets transformed.

The Jacobian matrices of attention and MLP blocks capture the local linear approximation of how these transformations modify the residual stream. Their spectral properties reveal several critical aspects:

- **Directional sensitivity and amplification:** The eigenvalue spectrum indicates which directions in the residual stream are amplified, preserved, or suppressed by each transformation. Large eigenvalues suggest directions of high sensitivity where small input changes lead to large output changes.
- **Subspace coordination:** Since layers can only interact by reading from and writing to overlapping subspaces of the residual stream, the eigenvectors reveal which representational directions each layer primarily operates on. This helps identify how different layers coordinate their computations.

- **Information routing:** The spectral decomposition shows how information gets routed between different subspaces. Eigenvalues near zero indicate directions where information is effectively deleted from the residual stream, while the corresponding eigenvectors show what types of features are being filtered out.
- **Representational efficiency:** The rank and condition number of the Jacobians reveal how efficiently each layer uses the available representational capacity of the residual stream, and whether transformations are well-conditioned or prone to numerical instabilities.

### 4.2 Jacobian computation

Our analysis centers on computing Jacobians for the two primary transformer components: attention blocks and MLP (feedforward) blocks (see Figure 1).

**Attention Jacobian:** For the attention function

$$f_{\text{attn}}(\mathbf{h}) = \text{Attention}(\text{LayerNorm}(\mathbf{h})),$$

we compute:

$$J_{\text{attn}}^{(\ell)} = \frac{\partial f_{\text{attn}}(\mathbf{h}^{(\ell)})}{\partial \mathbf{h}_{\text{last}}} \tag{5}$$

where  $\mathbf{h}_{\text{last}}$  represents the token embedding of the previous layer.

**MLP Jacobian:** For the feedforward function

$$f_{\text{mlp}}(\mathbf{h}) = \text{MLP}(\text{LayerNorm}(\mathbf{h})),$$

we compute:

$$J_{\text{mlp}}^{(\ell)} = \frac{\partial f_{\text{mlp}}(\mathbf{h}_{\text{last}}^{(\ell)})}{\partial \mathbf{h}_{\text{last}}} \tag{6}$$

In this paper, we focus on the last token provides computationally tractable analysis while capturing the model’s predictive dynamics.

*Remark.* The Jacobian matrices represent the linear component of the first-order Taylor expansion around the current activation state, revealing how small perturbations in the input representation propagate through each transformation. This linearization provides insight into the local sensitivity and directional preferences of attention and MLP operations within the high-dimensional residual stream.

### 4.3 Spectral metrics

There are several key metrics from the spectral decomposition that we consider in this paper:

**Spectral Radius:**

$$\rho(\mathbf{J}) = \max_i |\sigma_i| \tag{7}$$

indicating the maximum amplification factor and potential for instability.

**Eigenvalue Decay:**

$$\delta = \frac{|\sigma_2|}{|\sigma_1|} \tag{8}$$

characterizing the concentration of spectral energy in dominant modes.

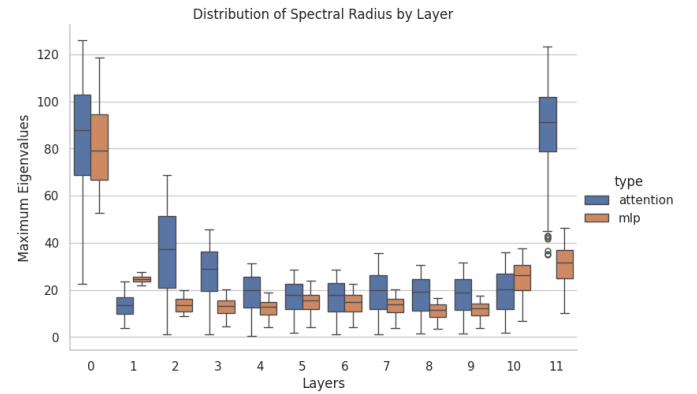
## 5 Experimental Results

We evaluate our spectral analysis framework on GPT-2 using the Tiny Shakespeare dataset. Our implementation uses the library and pre-trained models from Hugging Face.

### 5.1 Dataset

The Tiny Shakespeare corpus provides a controlled text domain for analyzing transformer behavior. We tokenize the corpus into individual sentences using NLTK’s sentence tokenizer and limit each sequence to 100 tokens to reduce computational cost of Jacobian computation.

### 5.2 Analysis on Spectral Radius



**Figure 2: Boxplot of spectral radius  $\rho(\mathbf{J})$  in individual layers in GPT-2 model after obtaining Jacobians from 600 samples.**

Figure 2 shows the spectral radius (spectral norm) of the Jacobian matrices with respect to the individual layers and the components of the transformer. The distribution of spectral norms across GPT-2’s 12 layers reveals interesting patterns that provide insights into the model’s information processing dynamics.

The high spectral norm in the first layer likely reflects the model’s need to project token embeddings into a rich representational space where subsequent layers can perform meaningful computations. The high values indicate strong directional preferences in how initial token representations are transformed.

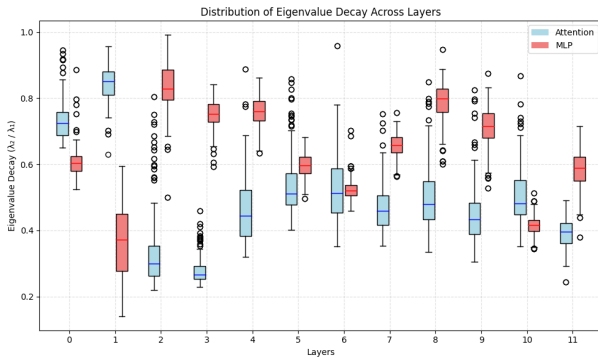
We also observe a high spectral norm in the last layer. This suggests significant amplification before the final prediction, possibly concentrating information relevant for next-token prediction into dominant spectral modes. This boundary effect aligns with the intuition that early layers perform feature extraction while final layers focus representations for output generation.

The intermediate layers (layer 2-10) show consistent spectral norms (typically 15-45), suggesting these intermediate layers perform more controlled, incremental transformations. This stability indicates that the middle layers avoid extreme amplifications that could lead to gradient instability or information loss.

Overall, the U-shaped pattern of spectral norms (high at boundaries, stable in middle) suggests a three-phase computation: aggressive initial transformation, stable intermediate processing, and

focused final amplification. The transformer architecture may naturally separate feature extraction, representation refinement, and output preparation into distinct computational regimes. The moderate spectral norms in middle layers likely contribute to training stability by avoiding the vanishing or exploding gradient problems that could arise from extreme eigenvalue distributions throughout the network depth.

### 5.3 Analysis on Eigenvalue Decay



**Figure 3: Boxplot of eigenvalue decay in individual layers in GPT-2 model after obtaining Jacobians from 300 samples.**

Figure 3 shows the eigenvalue decay rate measures how rapidly singular values decrease from the dominant mode, with higher values indicating more concentrated spectral energy. Attention blocks consistently exhibit higher decay rates compared to MLP blocks, indicating attention concentrates its transformations in fewer dominant directions. This spectral concentration difference suggests attention and MLP blocks serve complementary roles—attention provides focused, selective information routing while MLPs offer broader representational refinement across the residual stream dimensions.

## 6 Conclusion

This paper presented a preliminary approach for spectral analysis of transformer models through Jacobian decomposition. Our approach provides mathematically principled insights into transformer dynamics, revealing distinct spectral signatures for attention and feedforward components and systematic evolution patterns across layers.

*Future work.* The current approach requires expensive numerical differentiation for Jacobian computation. A promising direction is deriving analytical forms for the attention Jacobian directly from the attention matrices computed during the forward pass. Additionally, while our current work focuses on matrix-level spectral properties, future research should investigate how individual input vectors interact with these spectral structures. This could reveal how different types of semantic content (entities, relations, syntactic patterns) are processed differently by the spectral modes of each component.

## References

- [1] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS one* 10, 7 (2015), e0130140.
- [2] Hila Chefer, Shir Gur, and Lior Wolf. 2020. Transformer Interpretability Beyond Attention Visualization. *CoRR* abs/2012.09838 (2020). arXiv:2012.09838 <https://arxiv.org/abs/2012.09838>
- [3] Yingyi Chen, Qinghua Tao, Francesco Tonin, and Johan Suykens. 2023. Primal-attention: Self-attention through asymmetric kernel svd in primal representation. *Advances in Neural Information Processing Systems* 36 (2023), 65088–65101.
- [4] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. 2015. The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*. PMLR, 192–204.
- [5] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. 2025. A mathematical perspective on transformers. *Bull. Amer. Math. Soc.* 62, 3 (2025), 427–479.
- [6] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer Feed-Forward Layers Are Key-Value Memories. arXiv:2012.14913 [cs.CL]
- [7] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Science robotics* 4, 37 (2019), eaay7120.
- [8] Shahar Katz and Lior Wolf. 2024. Reversed Attention: On The Gradient Descent Of Attention Layers In GPT. *arXiv preprint arXiv:2412.17019* (2024).
- [9] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, et al. 2024. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion* 106 (2024), 102301.
- [10] Elhage Nelson, Nanda Neel, Olsson Catherine, Henighan Tom, Joseph Nicholas, Mann Ben, Askell Amanda, Bai Yuntao, Chen Anna, Conerly Tom, et al. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread* (2021).
- [11] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. 2019. On the spectral bias of neural networks. In *International conference on machine learning*. PMLR, 5301–5310.
- [12] Thiziri Nait Saada, Alireza Naderi, and Jared Tanner. 2024. Mind the Gap: a Spectral Analysis of Rank Collapse and Signal Propagation in Attention Layers. *arXiv preprint arXiv:2410.07799* (2024).
- [13] Max Staats, Matthias Thamm, and Bernd Rosenow. 2024. Small Singular Values Matter: A Random Matrix Analysis of Transformer Models. *arXiv preprint arXiv:2410.17770* (2024).
- [14] Anh Tong, Thanh Nguyen-Tang, Dongun Lee, Duc Nguyen, Toan Tran, David Leo Wright Hall, Cheongwoong Kang, and Jaesik Choi. 2025. Neural ODE Transformers: Analyzing Internal Dynamics and Adaptive Fine-tuning. In *The Thirtieth International Conference on Learning Representations*. <https://openreview.net/forum?id=XnDyddPcBT>
- [15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).