
A Geometric Analysis of Transformer Representations via Optimal Transport¹

Yadnyesh Chakane
Independent Researcher

Sunishka Sharma
Adobe

Janhavi Khindkar
IIIT Hyderabad

Vishnu Vardhan Lanka
Independent Researcher

With
PIBBSS, Timaeus, & Apart Research

Abstract

Understanding the internal workings of transformers is a major challenge in deep learning; while these models achieve state-of-the-art performance, their multi-layer architectures operate as "black boxes," obscuring the principles that guide their information processing pathways. To address this, we used **Optimal Transport (OT)** to analyze the geometric transformations of representations between layers in both trained and untrained transformer models. By treating layer activations as empirical distributions, we computed layer-to-layer OT distances to quantify the extent of geometric rearrangement, complementing this with representation entropy measurements to track information content. Our results show that trained models exhibit a structured, three-phase information processing strategy (**encode-refine-decode**), characterized by an **information bottleneck**. This is evident from a U-shaped OT distance profile, where high initial costs give way to a low-cost "refinement" phase before a final, high-cost projection to the output layer. This structure is entirely absent in untrained models, which instead show chaotic, uniformly high-cost transformations. We conclude that OT provides a powerful tool for revealing the learned, efficient information pathways in neural networks, demonstrating that learning is not merely about fitting data, but about creating an organized, information-theoretically efficient pipeline to process representations.

¹ Research conducted at the Apart AI Safety X Physics Grand Challenge, 2025

Keywords: Optimal Transport, Interpretability, Transformers, AI Safety

1. Introduction

The proliferation of large-scale transformer models has marked a paradigm shift in artificial intelligence, enabling unprecedented capabilities in natural language processing and beyond. Yet, as these models grow in size and complexity, their internal decision-making processes become increasingly opaque. They are often treated as "black boxes," where the relationship between their architecture and their emergent reasoning capabilities remains poorly understood. This lack of interpretability is not merely an academic curiosity; it poses a significant challenge to deploying these systems safely and reliably. Without a principled understanding of *how* a model arrives at an answer, we cannot fully trust its outputs, diagnose its failures, or guard against unforeseen biases. The central objective of our research is to move beyond performance metrics and develop a methodology to illuminate the internal information processing pathways of these models.

This paper proposes a novel approach to this problem by analyzing the model's operations through a geometric lens, using the mathematical framework of **Optimal Transport (OT)**. We posit that the collection of activation vectors at any given layer of a transformer can be viewed as an empirical sample from a high-dimensional probability distribution—a snapshot of the information as it exists at that stage. Consequently, the function of each subsequent layer is to transform this distribution into a new one. Optimal Transport provides a powerful and sensitive "ruler" to measure the geometric cost of this transformation, quantifying the "work" required to reshape one layer's representation distribution into the next. This method offers a far more nuanced view than simpler metrics like Euclidean or Cosine distance, which fail to capture the collective, structural changes in the data.

Our core contribution is an empirical study that uses this OT-based methodology to contrast the behavior of a fully trained transformer with that of a randomly initialized, untrained model of the identical architecture. By feeding both models the same data and tracking the layer-wise OT distance and representation entropy, we can isolate and characterize the precise structural changes that emerge as a direct result of learning. This comparative analysis serves as a controlled experiment to answer a fundamental question: What is the geometric signature of a learned, effective representation strategy?

For this specific research work we have used **GPT-2 style transformer models** and the **GSMSK** dataset.

2. Motivations

We draw an analogy between the process of learning in a neural network and fundamental principles in physics, such as self-organization and the principle of least action. An untrained network can be viewed as a high-entropy, disordered system. We hypothesize that the process of learning is analogous to a physical system finding a low-energy, ordered state. The trained network, therefore, should follow a path of "least informational action" to transform inputs into outputs efficiently. Optimal Transport provides the exact mathematical tool to test this hypothesis. It allows us to quantify the "work" done by each layer by measuring the cost of reshaping the entire distribution of activations, offering a holistic, geometric perspective that is lost when analyzing individual neurons.

Without a clear understanding of a model's internal reasoning process, we cannot reliably predict its behavior in novel scenarios, diagnose its failures, or prevent it from generating harmful outputs. Our work aims to make these internal computations more transparent, which is a foundational step toward building more robust and auditable AI.

3. Proposed Research

3.1 Optimal Transport: A Geometric Ruler for Representations

To understand the layer-by-layer transformations within a neural network, we require a tool that can measure the "distance" between two sets of high-dimensional points. While standard metrics like Euclidean or Cosine distance can measure the change in individual vectors, they fail to capture the collective, structural change of the entire distribution of representations. For this, we turn to Optimal Transport (OT).

3.1.1 Theoretical Background

Optimal Transport originates from the problem of finding the most efficient way to transport a pile of earth, with a certain shape, into a hole of a different shape. In a probabilistic setting, it provides a principled way to measure the distance between two probability distributions, P and Q . The core idea is to find a "transport plan," denoted by γ , that specifies how to move the "mass" of distribution P to match distribution Q while minimizing a total transportation cost.

The modern formulation, known as the Kantorovich problem, defines the set of all valid transport plans, $\Pi(P, Q)$, as all joint probability distributions whose marginals are P and Q . The cost of moving a unit of mass from a point x to a point y is given by a cost function, $c(x, y)$. For our purposes, we use the squared Euclidean distance, $c(x, y) = \|x - y\|_2^2$, as it corresponds to the physical notion of "work." The squared 2-Wasserstein distance is then defined as the minimum possible cost over all valid transport plans:

$$W_2^2(P, Q) = \inf_{\gamma \in \Pi(P, Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_2^2 d\gamma(x, y)$$

The value W_2 gives a true metric distance between the distributions, capturing their geometric differences in a way that is sensitive to the underlying structure of the space.

3.1.2 The Need for Regularization: Sinkhorn Distance

While theoretically elegant, computing the exact Wasserstein distance is computationally expensive and suffers from statistical issues (the "curse of dimensionality") when applied to empirical samples in high dimensions, as is the case with neural network activations. To overcome this, we use **entropic regularization**. This approach adds a penalty term based on the entropy of the transport plan, $H(\gamma)$, to the optimization problem:

$$\text{OT}_\varepsilon(P, Q) = \min_{\gamma \in \Pi(P, Q)} \left\{ \sum_{i,j} \gamma_{ij} C_{ij} - \varepsilon H(\gamma) \right\}$$

Here, C is the cost matrix where $C_{jk} = \|x_j - y_k\|_2^2$, and $\varepsilon > 0$ is the regularization parameter. This regularization smooths the problem, making it strictly convex and solvable with a highly efficient, iterative algorithm known as the **Sinkhorn-Knopp algorithm**, or simply **Sinkhorn's algorithm**. We utilize the **Python Optimal Transport (POT)** library for this computation. The resulting value is often referred to as the **Sinkhorn distance**. This method is not only faster but also statistically more robust for high-dimensional data, making it the ideal choice for our analysis.

3.1.3 Calculating Representation Distance with OT

In our research, we apply this framework as follows:

1. **Form Empirical Distributions:** For any two layers, L_i and L_{i+1} , we pass a batch of data through the model and collect the sets of activation vectors, $\{x_1, \dots, x_n\}$ from L_i , $\{y_1, \dots, y_m\}$ from L_{i+1} . These form our two empirical distributions, P_n and Q_m .
2. **Compute the Cost Matrix:** We compute a cost matrix C of size $n \times m$, where each entry $C_{jk} = \|x_j - y_k\|_2^2$ is the squared Euclidean distance between an activation vector from the source layer and one from the target layer.
3. **Calculate Sinkhorn Distance:** We use the Sinkhorn algorithm with the cost matrix C and uniform probability vectors for our empirical distributions to compute the regularized OT cost. This final value is the "OT Distance" reported in our plots, representing the geometric work done by the network to transform the representation from layer L_i to L_{i+1} .
4. We repeat this process for all adjacent layer pairs.
5. We ran this analysis at different stages of training, on an untrained model and a fully trained one, to see how learnability affects the signal geometry.

3.2 Research Findings and Analysis

Our empirical investigation, centered on a comparative analysis of a trained versus an untrained transformer model, reveals a distinct and quantifiable "signature of

learning." By applying our Optimal Transport-based framework, we can move beyond simple accuracy metrics to observe the emergence of a sophisticated and efficient information processing strategy.

3.2.1 The Learned Strategy of the Trained Model

The trained model exhibits a highly structured and non-uniform geometric pathway across its layers. This is most clearly visualized in the plot of adjacent-layer OT distances, which reveals a distinct U-shaped profile. We identify three functional phases:

1. **Encoding Phase (Initial Layers):** The model performs its most significant geometric transformations in the initial layers. The OT distance is highest between initial layers, indicating a massive initial projection from the raw input embedding space into the model's internal representational geometry. This cost rapidly decreases, suggesting a swift consolidation of key features.
2. **Refinement Phase (Middle Layers):** Following the initial encoding, the model enters a long, stable phase characterized by remarkably low OT distances. In these middle layers, the model is not making large-scale changes but is instead performing subtle refinements on an already-formed representation. This suggests a period of iterative processing where abstract concepts are composed and manipulated with minimal geometric "work."
3. **Decoding Phase (Final Layers):** The pathway culminates in a final, dramatic spike in OT distance. The transformation into the final layer is the most geometrically significant of all, indicating that this layer has a specialized function: to project the final, abstract internal representation into the target output space.

This three-phase strategy is explained by the model's handling of information entropy. Our analysis shows that the trained model learns to create an **information bottleneck**. The representation entropy follows a corresponding U-shaped curve, dropping sharply to a minimum in the middle layers before rising again at the output. This low-entropy state aligns perfectly with the low-cost refinement phase, demonstrating that the model learns to compress information into an efficient, essential representation to perform its core reasoning.

3.2.2 The Random Walk of the Untrained Model

In stark contrast, the untrained model displays no such strategic organization. Its layer-wise OT distances are uniformly high and erratic, with no discernible phases. Each layer performs a large, costly, and seemingly random geometric transformation on its input. The concept of a low-cost refinement phase is entirely absent. Similarly, its entropy profile shows no evidence of a learned information bottleneck; after an initial random collapse, the entropy steadily increases, indicating an accumulation of noise rather than a structured processing of information. The geometric pathway of the untrained model is best described as a "random walk," lacking the purpose and efficiency that emerges through training.

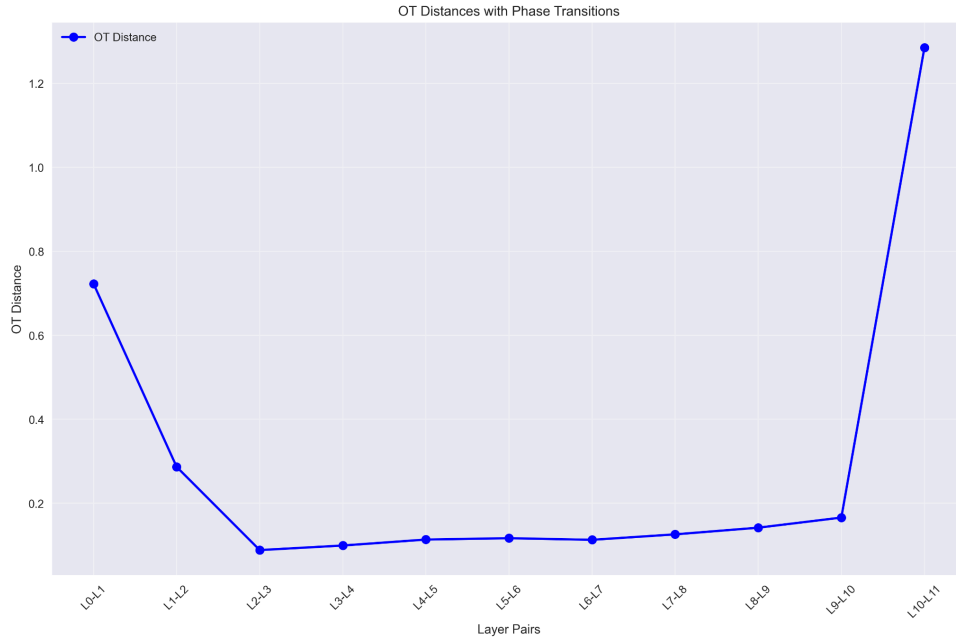


Figure 1 – OT Distances with Phase Transition (Learned Model)

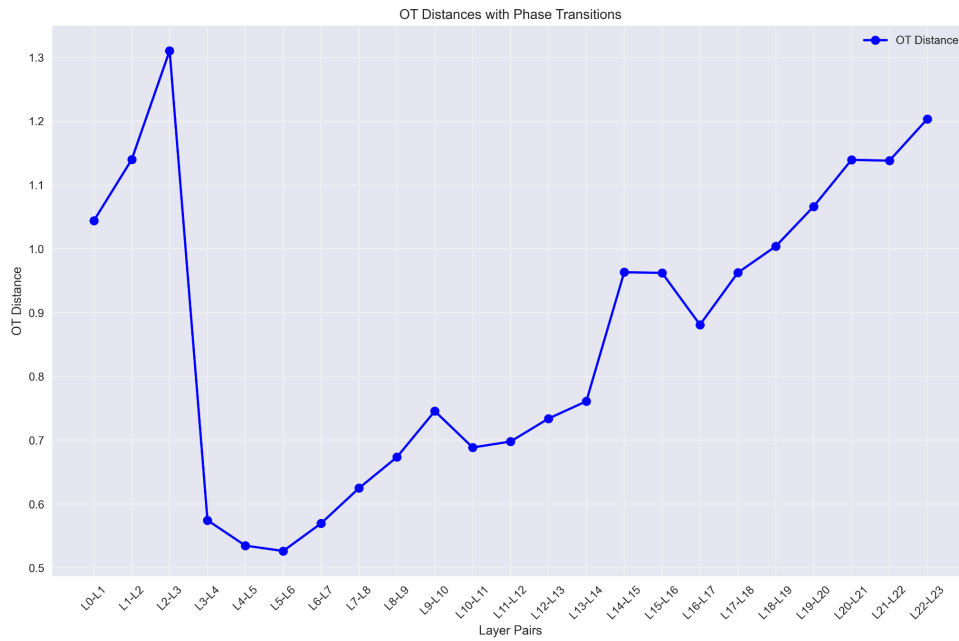


Figure 2 – OT Distances with Phase Transition (Unlearned Model)

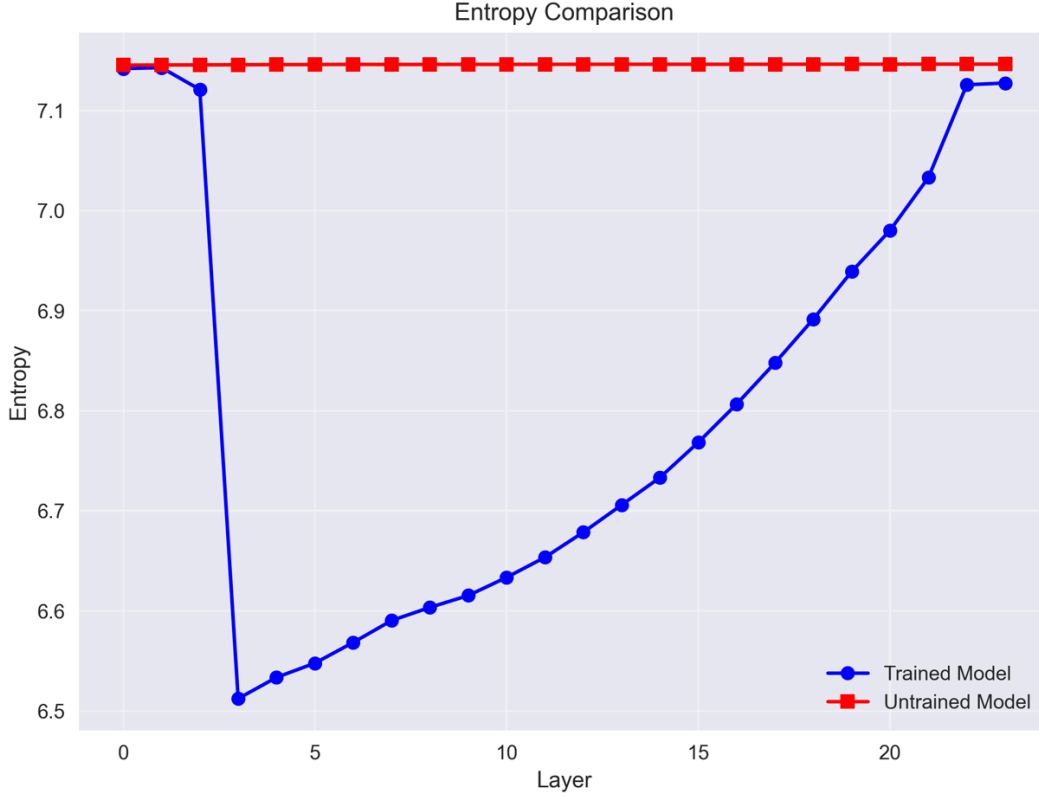


Figure 3 - The Emergence of the Information Bottleneck. A comparison of layer-wise representation entropy reveals that the trained model (blue) learns to compress information into a low-entropy state for refinement. In contrast, the untrained model (red) remains in a state of high disorder, demonstrating that the bottleneck is a learned, not inherent, property of the architecture.

4. Discussion and Conclusion

Our investigation into the internal geometry of transformer models has revealed a distinct and quantifiable "signature of learning." By employing Optimal Transport as a geometric ruler, we have shown that a trained transformer organizes its internal operations into a highly efficient, three-phase processing pipeline: a high-cost initial encoding, a long, low-cost refinement phase, and a final, high-cost decoding projection. The mechanism enabling this strategy is a learned **information bottleneck**, where the model compresses representations into a low-entropy state for efficient manipulation. This structured pathway stands in stark contrast to the behavior of an untrained model, which exhibits only a chaotic, high-cost "random walk" through its layers. This confirms that the efficient processing strategy is not an inherent property of the architecture but an emergent phenomenon of learning.

The central conclusion of this work is that learning in a deep neural network is synonymous with self-organization into an information-theoretically efficient system. The model does not merely learn to map inputs to correct outputs; it learns an economical pathway to do so, minimizing the "work" required to transform information. This physics-inspired perspective, viewing learning as a transition

from a high-entropy, disordered state to a low-energy, structured one, provides a powerful new lens for interpreting these complex systems.

This research opens several promising avenues for future inquiry. A key open question is the **universality** of this learned structure. Does the U-shaped OT profile appear across different architectures (e.g., Vision Transformers) and on different task modalities? Finally, correlating the specific properties of this geometric pathway - such as the depth of the information bottleneck - with final task performance could provide new insights into what makes certain models more effective than others, guiding the design of future architectures.

5. References

Chewi, S., Niles-Weed, J., & Rigollet, P. (2024). Statistical optimal transport. *arXiv preprint arXiv:2407.18163*, 3.

Valeriani, L., Doimo, D., Cuturello, F., Laio, A., Ansuini, A., & Cazzaniga, A. (2023). The geometry of hidden representations of large transformer models. *Advances in Neural Information Processing Systems*, 36, 51234-51252.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Balakrishnan, S., Manole, T., & Wasserman, L. (2025). Statistical Inference for Optimal Transport Maps: Recent Advances and Perspectives. *arXiv preprint arXiv:2506.19025*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., ... & Schulman, J. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

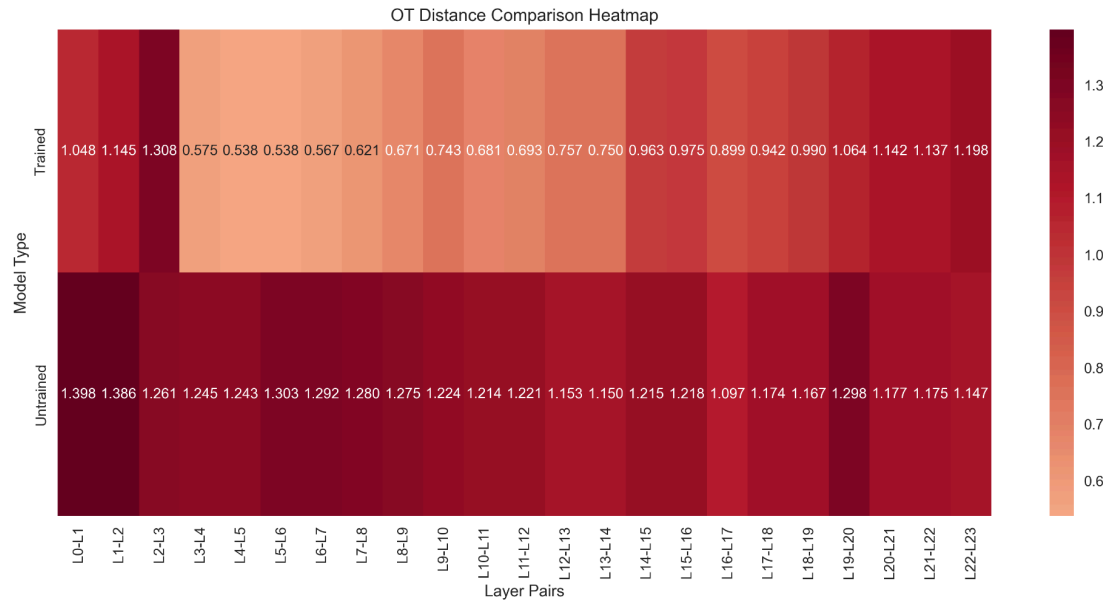
Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

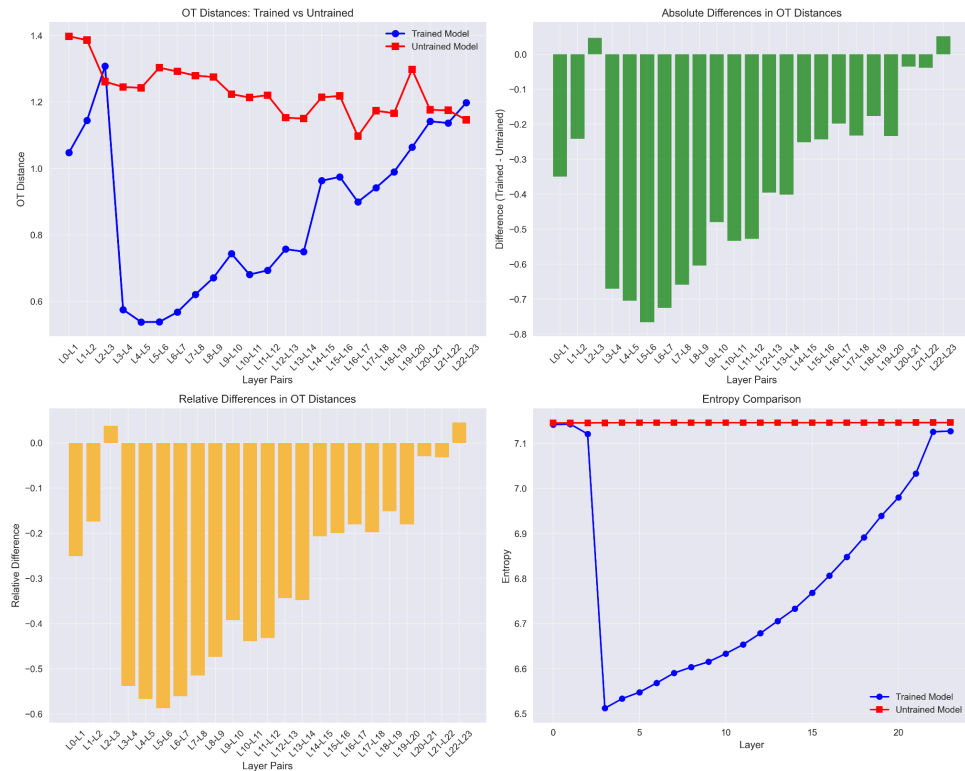
Olah, et al., "Zoom In: An Introduction to Circuits", Distill, 2020.

6. Appendix

A) OT Distance Comparison Heatmap



B) Trained Vs Untrained Model Statistics



C) Layer Statistics Comparison

